JMIR Medical Education

Impact Factor (2024): 3.2
Volume 10 (2024) ISSN 2369-3762 Editor-in-Chief: Blake J. Lesselroth, MD, MBI, FACP, FAMIA

Contents

Original Papers

XSL•FO
RenderX

XSL•FO
RenderX

XSL•FO
RenderX

XSL•FO
RenderX

XSL•FO
RenderX

XSL•FO
RenderX

Reviews

Viewpoints

XSL•FO
RenderX

XSL•FO
RenderX

Tutorials

Corrigenda and Addendas

Editorial

Short Paper

Research Letter

XSL•FO
RenderX

Original Paper

Challenges for Medical Students in Applying Ethical Principles to Allocate Life-Saving Medical Devices During the COVID-19 Pandemic: Content Analysis

Hsing-yen Hsieh[1], PhD; Chyi-her Lin[1,2], MD; Ruyi Huang[1,3,4,5], MPH, MD; Guan-chun Lin[1], PhD; Jhen-Yu Lin[3], MS; Clydie Aldana[1], BS

[1]School of Medicine for International Students, College of Medicine, I-Shou University, Kaohsiung, Taiwan

[2]Department of Pediatrics, E-Da Hospital, Kaohsiung, Taiwan

[3]Holistic Medicine, Department of Family and Community Medicine, E-Da Hospital, Kaohsiung, Taiwan

[4]Data Science Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan

[5]Division 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.

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e52711) doi:10.2196/52711

KEYWORDS

virtual patient; virtual patients; medical resources distribution; medical ethical education; COVID-19 pandemic; ethics; medical student; medical students; medical ethics; decision-making; ethical dilemna; 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

XSL•FO
RenderX

the ECMO instead? Please explain your decision and your reasons to support it."

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

Data Analysis

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

The qualitative analysis consisted of the following steps:

1. The coding items included the final decision of the students (for whom to use), which core medical ethical principles were applied with various degrees in their choices, and whether viewpoints other than ethics, such as medical guidelines or legislation, were mentioned.
2. The reasons for the students' final decisions were classified according to the patient they selected, either the 62-year-old younger patient or the 80-year-old patient with multiple comorbidities. Our analysis focused on encoding the ethical justifications provided by the medical students to support their final decisions. We omitted considerations related to their alternative choices during the decision-making process.
3. The classification of reasoning for those who made a decision was primarily based on the students' understanding and interpretations in their essays, which Persad et al [12] mentioned were equality, vulnerability, maximizing the quality of life, and contribution to society. The original resource allocation principles were designed for the distribution of medical supplies among a group of individuals. However, the present case pertains to the treatment decision for an individual patient, further complicated by the fact that one patient had already been put on a ventilator. By contextualizing the principles within the framework of the present case, we eliminated the applicability of 4 subprinciples: lottery, saving the most lives, reciprocity, and giving priority to the worst off (ie, sickest first).
4. If students displayed reluctance in making a choice, we also coded their explanations for the perception that ethical decision-making might not be suitable, categorizing these explanations as "undetermined" or "both unqualified."
5. The main reasons for the students' final decisions were classified into medical, legal, and ethical perspectives.
6. The coding process was independently judged by 2 researchers with expertise in qualitative research. Any inconsistencies in coding were resolved by reviewing the classification descriptions to refine the precision of category definitions and revisiting the context to ensure accurate coding.

Results

Student Demographics

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

XSL•FO
RenderX

Table 1. Basic information of the students.

Demographic	Medical students (n=67), n (%)
Sex	
Male	27 (40)
Female	40 (60)
Age (years)	
>25	61 (91)
≤25	6 (9)
Seniority year	
Third	33 (49)
Fourth	34 (51)
Country of origin	
The Kingdom of Eswatini	32 (48)
Saint Lucia	7 (10)
Belize	7 (10)
Kiribati	5 (7)
Honduras	3 (4)
The Marshall Islands	3 (4)
Saint Kitts and Nevis	3 (4)
Paraguay	2 (3)
Saint Vincent & The Grenadines	2 (3)
Palau	1 (1)
Haiti	1 (1)
Solomon Islands	1 (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 2. Choosing the most suitable patient for extracorporeal membrane oxygenation treatment.

Patient selected	Students (n=67), n (%)
80-year-old	39 (58)
62-year-old	22 (33)
Undetermined	5 (8)
Both unqualified	1 (1)

XSL•FO
RenderX

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.

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

[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.

XSL•FO
RenderX

Confidence in Ethical Decision-Making

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

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

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

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

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

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

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

Discussion

Principal Findings

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

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

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

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

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

Students were unfamiliar with philosophical and ethical reasoning and were under pressure to make ethical decisions about allocating life-saving medical modalities. They tended to

XSL•FO
RenderX

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.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Global distribution of international medical students.
[PNG File , 190 KB - mededu_v10i1e52711_app1.png]

References

1. Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, et al. Coronavirus pandemic (COVID-19). Our World in Data. URL: https://ourworldindata.org/coronavirus [accessed 2023-12-14]
2. Tsai DFC, Foo KF, Ku SC, Fang CC. The intensive care medical resource allocation and ethics in pandemic. Formos J Med 2020:605-616 [FREE Full text]
3. Cacchione PZ. Moral distress in the midst of the COVID-19 pandemic. Clin Nurs Res 2020 May;29(4):215-216. [doi: 10.1177/1054773820920385] [Medline: 32363981]
4. National Academy of Medicine. In: Gayle H, Foege W, Brown L, Kahn B, editors. Framework for Equitable Allocation of COVID-19 Vaccine. Washington, DC: The National Academies Press; 2020.
5. Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, et al. Fair allocation of scarce medical resources in the time of Covid-19. N Engl J Med 2020 May 21;382(21):2049-2055. [doi: 10.1056/NEJMsb2005114] [Medline: 32202722]
6. Epstein EG, Whitehead PB, Prompahakul C, Thacker LR, Hamric AB. Enhancing understanding of moral distress: the measure of moral distress for health care professionals. AJOB Empir Bioeth 2019;10(2):113-124. [doi: 10.1080/23294515.2019.1586008] [Medline: 31002584]
7. Schiffer AA, O'Dea CJ, Saucier DA. Moral decision-making and support for safety procedures amid the COVID-19 pandemic. Pers Individ Dif 2021 Jun;175:110714 [FREE Full text] [doi: 10.1016/j.paid.2021.110714] [Medline: 33551530]
8. O'Byrne L, Gavin B, McNicholas F. Medical students and COVID-19: the need for pandemic preparedness. J Med Ethics 2020 Sep;46(9):623-626 [FREE Full text] [doi: 10.1136/medethics-2020-106353] [Medline: 32493713]
9. Mueller D, Tollison R, Willett T. The utilitarian contract: a generalization of Rawls' theory of justice. Theor Decis 1974;4(3-4):345-367 [FREE Full text] [doi: 10.1007/bf00136654]
10. Gilabert P. The two principles of justice. In: Mandle J, Reidy D, editors. The Cambridge Rawls Lexicon. Cambridge, England: Cambridge University Press; 2014.
11. Barrow JM, Khandhar PB. Deontology. In: StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing; 2023.
12. Persad G, Wertheimer A, Emanuel EJ. Principles for allocation of scarce medical interventions. Lancet 2009 Jan 31;373(9661):423-431 [FREE Full text] [doi: 10.1016/S0140-6736(09)60137-9] [Medline: 19186274]
13. Asao S, Lewis B, Harrison JD, Glass M, Brock TP, Dandu M, et al. Ethics simulation in global health training (ESIGHT). MedEdPORTAL 2017 Jun 07;13:10590 [FREE Full text] [doi: 10.15766/mep_2374-8265.10590] [Medline: 30800792]
14. Modlin CE, C Vilorio A, Stoff B, L Comeau D, Gebremariam TH, Derbew M, et al. American medical trainee perspectives on ethical conflicts during a short-term global health rotation in Ethiopia: a qualitative analysis of 30 cases. Am J Trop Med Hyg 2021 Nov 01;106(2):398-411 [FREE Full text] [doi: 10.4269/ajtmh.21-0179] [Medline: 34724634]
15. The demography of medical schools: a discussion paper. British Medical Association. 2004. URL: https://puntsdevista.comb.cat/edicio9/Documents/2004%20UK%20Demography%20Schools%20Medicine.pdf [accessed 2023-12-14]
16. Alessandri F, Di Nardo M, Ramanathan K, Brodie D, MacLaren G. Extracorporeal membrane oxygenation for COVID-19-related acute respiratory distress syndrome: a narrative review. J Intensive Care 2023 Feb 08;11(1):5 [FREE Full text] [doi: 10.1186/s40560-023-00654-7] [Medline: 36755270]
17. Rabie A, Elhazmi A, Azzam MH, Abdelbary A, Labib A, Combes A, et al. Expert consensus statement on venovenous extracorporeal membrane oxygenation ECMO for COVID-19 severe ARDS: an international Delphi study. Ann Intensive Care 2023 May 02;13(1):36 [FREE Full text] [doi: 10.1186/s13613-023-01126-9] [Medline: 37129771]
18. Kuppler M, Kern C, Bach RL, Kreuter F. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. Front Sociol 2022;7:883999 [FREE Full text] [doi: 10.3389/fsoc.2022.883999] [Medline: 36299413]
19. Gesundheit N, Brutlag P, Youngblood P, Gunning WT, Zary N, Fors U. The use of virtual patients to assess the clinical skills and reasoning of medical students: initial insights on student acceptance. Med Teach 2009 Aug;31(8):739-742. [doi: 10.1080/01421590903126489] [Medline: 19811211]
20. Khin-Htun S, Kushairi A. Twelve tips for developing clinical reasoning skills in the pre-clinical and clinical stages of medical school. Med Teach 2019 Sep;41(9):1007-1011. [doi: 10.1080/0142159X.2018.1502418] [Medline: 30299197]
21. Posel N, Mcgee JB, Fleiszer DM. Twelve tips to support the development of clinical reasoning skills using virtual patient cases. Med Teach 2015;37(9):813-818. [doi: 10.3109/0142159X.2014.993951] [Medline: 25523009]
22. Pokhrel S, Chhetri R. A literature review on impact of COVID-19 pandemic on teaching and learning. High Educ Future 2021 Jan 19;8(1):133-141. [doi: 10.1177/2347631120983481]

XSL•FO
RenderX

23. Hancock KK, Minor CV. Pandemic hastens cleveland clinic's unified well-being strategy. Front Health Serv Manage 2021 Oct 01;38(1):4-13. [doi: 10.1097/HAP.0000000000000121] [Medline: 34431813]
24. Truog RD, Mitchell C, Daley GQ. The toughest triage - allocating ventilators in a pandemic. N Engl J Med 2020 May 21;382(21):1973-1975. [doi: 10.1056/NEJMp2005689] [Medline: 32202721]
25. Quill TE, Brody H. Physician recommendations and patient autonomy: finding a balance between physician power and patient choice. Ann Intern Med 1996 Nov 01;125(9):763-769. [doi: 10.7326/0003-4819-125-9-199611010-00010] [Medline: 8929011]
26. Vansteenkiste M, Simons J, Lens W, Sheldon KM, Deci EL. Motivating learning, performance, and persistence: the synergistic effects of intrinsic goal contents and autonomy-supportive contexts. J Pers Soc Psychol 2004 Aug;87(2):246-260. [doi: 10.1037/0022-3514.87.2.246] [Medline: 15301630]
27. Delaney J. Does unrealistic optimism undermine patient autonomy? Ethics Med Public Health 2023 Feb;26:100859 [FREE Full text] [doi: 10.1016/j.jemep.2022.100859]
28. Meyers C. Cruel choices: autonomy and critical care decision-making. Bioethics 2004 Apr;18(2):104-119. [doi: 10.1111/j.1467-8519.2004.00384.x] [Medline: 15146851]
29. Taylor RM. Ethical principles and concepts in medicine. Handb Clin Neurol 2013;118:1-9. [doi: 10.1016/B978-0-444-53501-6.00001-9] [Medline: 24182363]
30. Ogut E, Senol Y, Yildirim FB. Do learning styles affect study duration and academic success? Eur J Anat 2017;21(3):235-240 [FREE Full text]
31. Dehghani A. Factors affecting professional ethics development in students: a qualitative study. Nurs Ethics 2020 Mar;27(2):461-469. [doi: 10.1177/0969733019845135] [Medline: 31284820]
32. Levy N. Forced to be free? Increasing patient autonomy by constraining it. J Med Ethics 2014 May;40(5):293-300 [FREE Full text] [doi: 10.1136/medethics-2011-100207] [Medline: 22318413]
33. Berger JT. Moral distress in medical education and training. J Gen Intern Med 2014 Feb;29(2):395-398 [FREE Full text] [doi: 10.1007/s11606-013-2665-0] [Medline: 24146350]
34. Epstein EG, Haizlip J, Liaschenko J, Zhao D, Bennett R, Marshall MF. Moral distress, mattering, and secondary traumatic stress in provider burnout: a call for moral community. AACN Adv Crit Care 2020 Jun 15;31(2):146-157. [doi: 10.4037/aacnacc2020285] [Medline: 32525997]
35. Beauchamp TL, Childress JF. Principles of Biomedical Ethics 8th Edition. New York: Oxford University Press; 2013.
36. Heist BS, Kishida N, Deshpande G, Hamaguchi S, Kobayashi H. Virtual patients to explore and develop clinical case summary statement skills amongst Japanese resident physicians: a mixed methods study. BMC Med Educ 2016 Feb 01;16:39 [FREE Full text] [doi: 10.1186/s12909-016-0571-y] [Medline: 26830910]
37. Fincham JE. Response rates and responsiveness for surveys, standards, and the journal. Am J Pharm Educ 2008 Apr 15;72(2):43 [FREE Full text] [doi: 10.5688/aj720243] [Medline: 18483608]

Abbreviations

ECMO: extracorporeal membrane oxygenation

Edited by T de Azevedo Cardoso, AH Sapci, MD; submitted 13.09.23; peer-reviewed by E Ogut, I Mircheva; comments to author 28.10.23; revised version received 18.11.23; accepted 03.12.23; published 05.01.24.

Please cite as:
Hsieh HY, Lin CH, Huang R, Lin GC, Lin JY, Aldana C
Challenges for Medical Students in Applying Ethical Principles to Allocate Life-Saving Medical Devices During the COVID-19 Pandemic: Content Analysis
JMIR Med Educ 2024;10:e52711
URL: https://mededu.jmir.org/2024/1/e52711
doi:10.2196/52711
PMID:38050366

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

XSL•FO
RenderX

Original Paper

A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study

Friederike Holderried[1], MME, Dr med; Christian Stegemann–Philipps[1], Dr rer nat; Lea Herschbach[1], MSc; Julia-Astrid Moldt[1], MA; Andrew Nevins[2], Prof Dr; Jan Griewatz[1], MA; Martin Holderried[3], Dr med, Prof Dr; Anne Herrmann-Werner[1], MME, Prof Dr Med; Teresa Festl-Wietek[1], Dr rer nat; Moritz Mahling[1,4], Dr med, MHBA

[1]Tübingen Institute for Medical Education, Eberhard Karls University, Tübingen, Germany

[2]Division of Infectious Diseases, Stanford University School of Medicine, Stanford, CA, United States

[3]Department of Medical Development, Process and Quality Management, University Hospital Tübingen, Tübingen, Germany

[4]Department 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.

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e53961) doi:10.2196/53961

KEYWORDS

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 encounter [14], and courses frequently rely on simulated or real patients [15]. These formats are associated with high costs and a high organizational effort, however, which shortens the time available to acquire these skills. These restrictions often do not allow all students to interactively practice a skill or practice for more than 1 repetition [16]. Furthermore, learning in these settings often occurs supervised by the patient and peer group, thereby impacting performance and possibly inhibiting rather shy students from using the learning opportunity [17,18].

Chatbots offer a significant potential to overcome these restrictions, thereby enhancing the utility thereof in health care and medical education settings. Chatbots have thus become valuable tools in health care; their nonjudgmental and easily accessible nature makes them particularly well suited for responding to patient inquiries and concerns [19,20]. The use of chatbots in medical education offers equally promising opportunities. In particular, chatbots are of interest tool-wise in the area of virtual patients [21,22].

The advance of chatbots is significantly supported by the developments of LLMs, such as GPT, which progressed considerably in 2022 [23]. This progress in artificial intelligence (AI) technology opens up new horizons for innovative, cost-effective, and accessible learning methods [24,25]. GPT has performed surprisingly well regarding medical knowledge, including board exams [26-28]. The combination of excellent language skills and medical knowledge predispose GPT to perform as a chatbot. Moreover, LLMs allow for unsupervised and repeated learning, thereby enabling all students to learn for as long as it is needed. However, LLMs, such as GPT, are language models using a next-word prediction paradigm [29] and are thus prone to "hallucinations" (ie. producing nonsensical content) [30]. Moreover, LLMs are also known to occasionally escape prompts.

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.

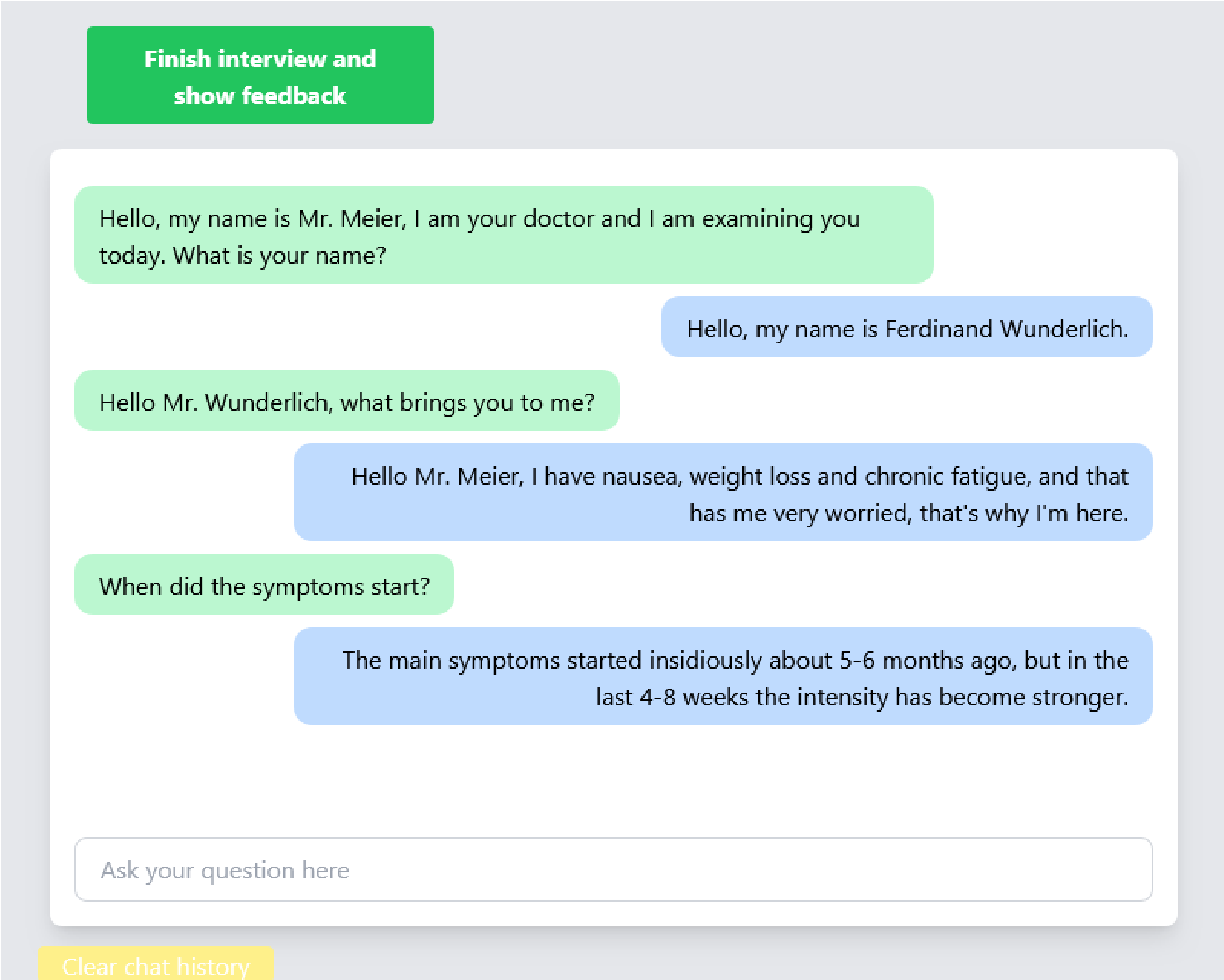
Chat Platform

To enable the interaction between students and GPT, we created a chat interface through which the students could post written

XSL•FO
RenderX

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.

XSL•FO
RenderX

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

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

Behavioral Prompt

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

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

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

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

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

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

XSL•FO
RenderX

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.

{

"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.

XSL•FO
RenderX

Table 2. Aspects, categories, and text examples used for rating.

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

XSL•FO
RenderX

Aspect and categories	Example(s)
−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[b] with empty answers due to technical reasons.*

[a]GPT: generative pretrained transformer.

[b]QAP: question-answer pair.

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.

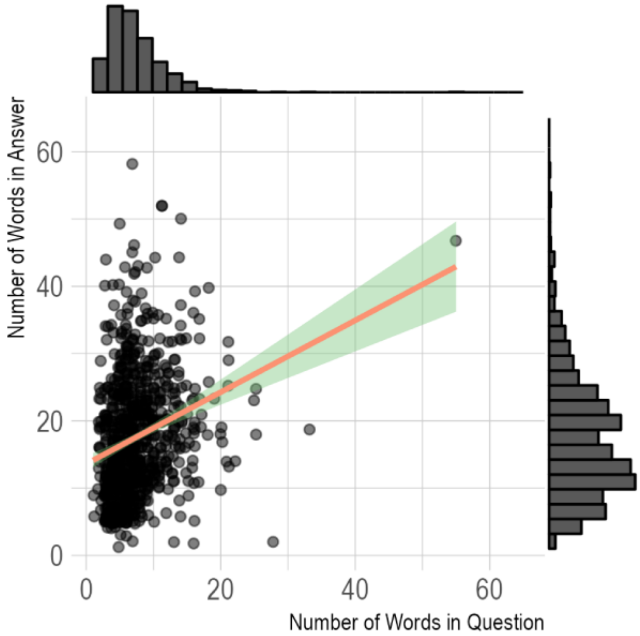
Conversation Length and Part of Conversation

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 $\rho=0.29$, 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%).

XSL•FO
RenderX

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.



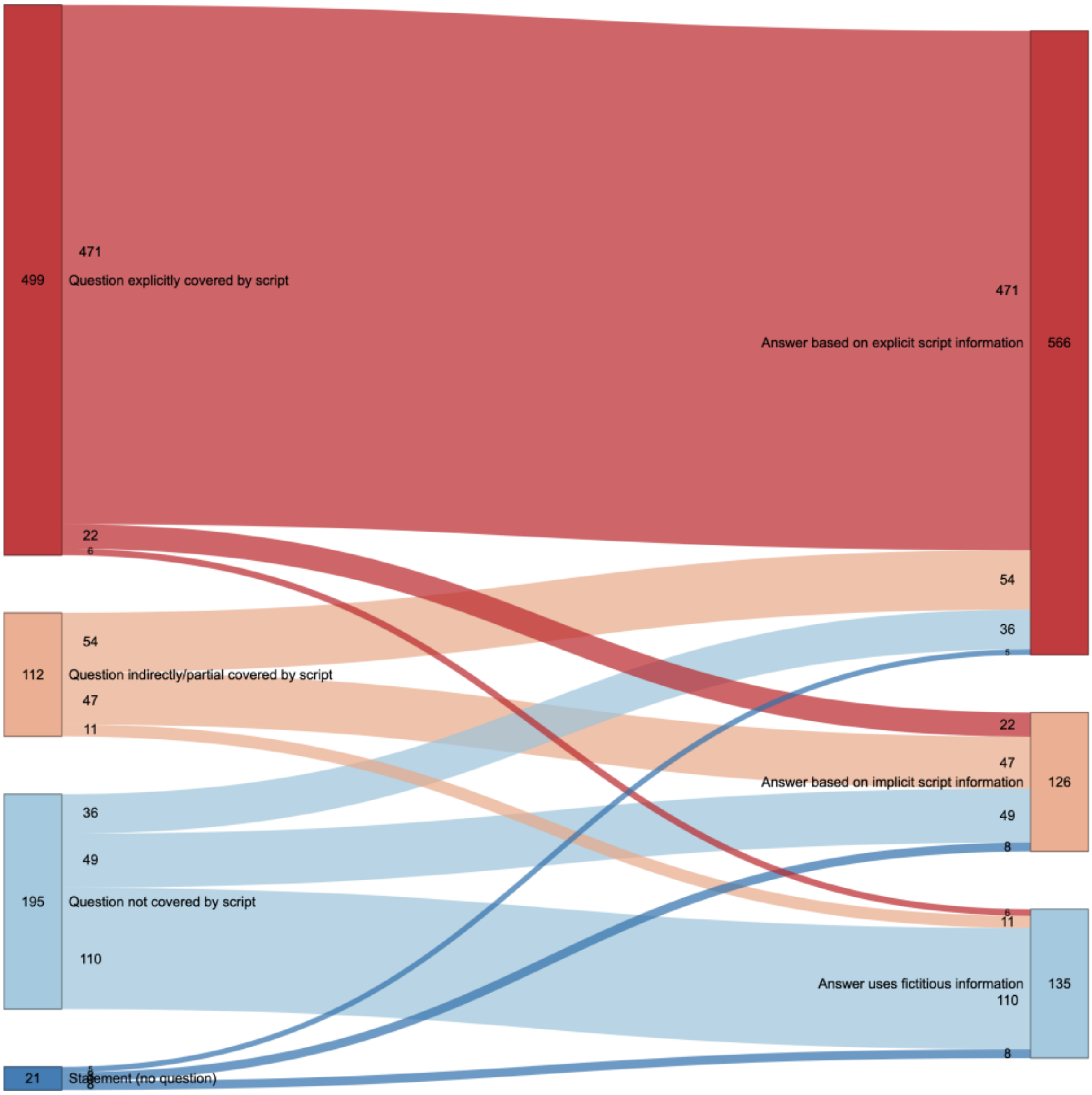
Content Analysis of Conversations

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).

XSL•FO
RenderX

Figure 3. Sankey plot for "Student's question covered by the illness script" and "GPT's answer uses the illness script" 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.



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

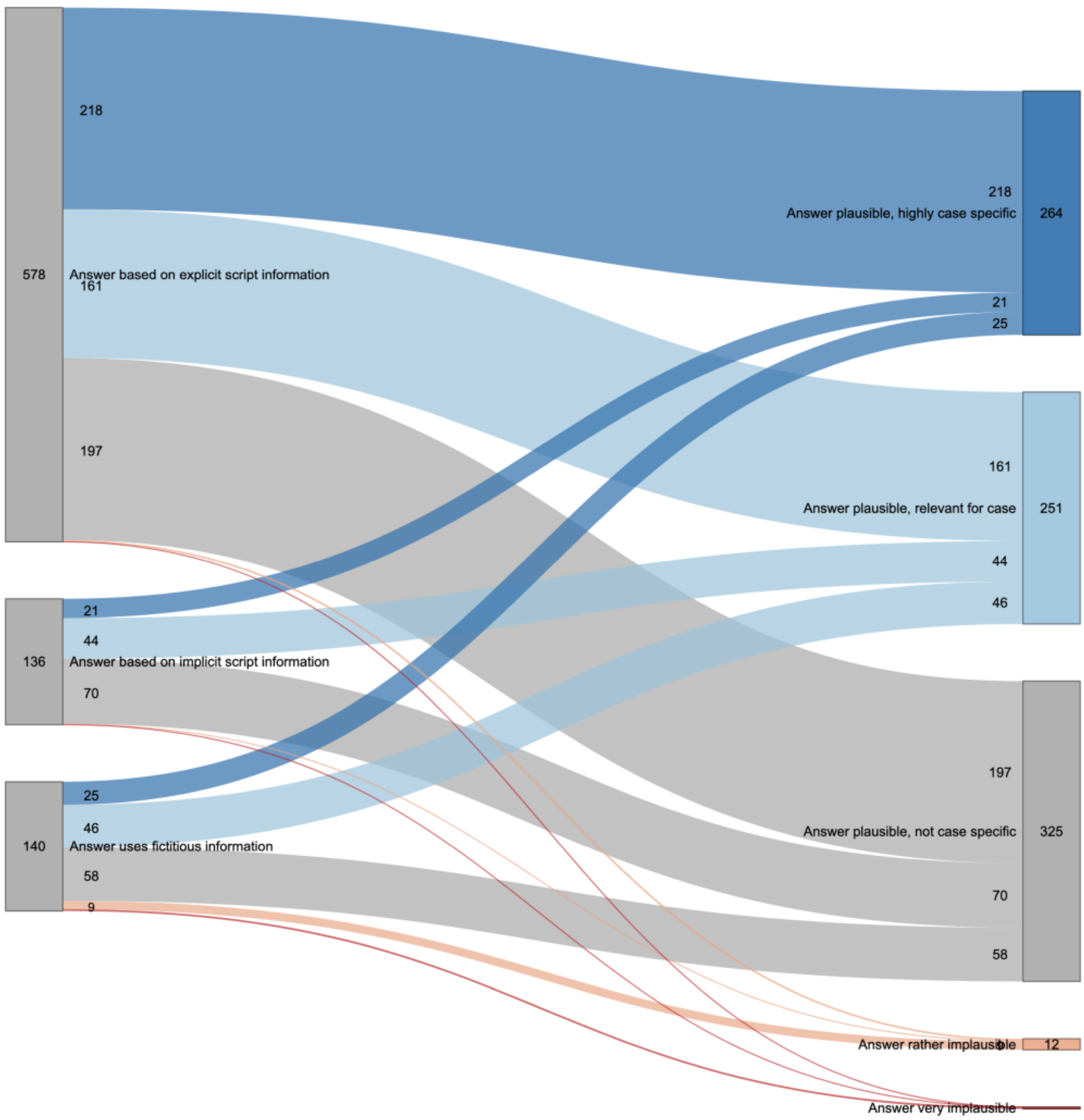
XSL•FO
RenderX

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.

XSL•FO
RenderX

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.").

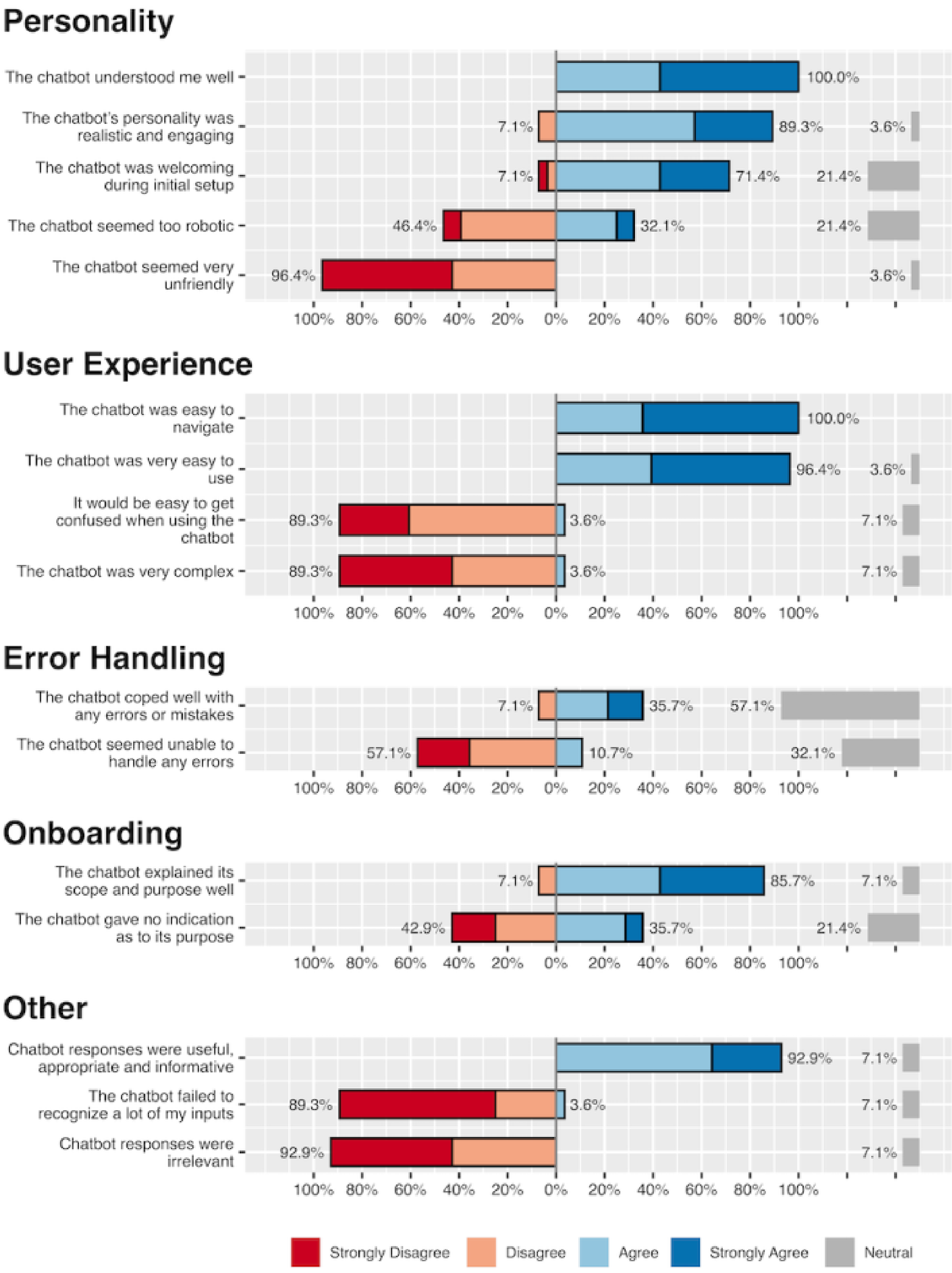
Among the 2 (0.2%) QAPs rated as very implausible, GPT escaped its role in 1 (50%) case. Herein, the participant asked about what can be seen in the physical exam, and the GPT-provided answer was, "Sorry, I am a language AI and do not have access to visual information. I can only provide information 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).

XSL•FO
RenderX

Figure 5. Results of the CUQ, grouped by category, as proposed by Holmes et al. 2023. Neutral responses are indicated on the right side of the figure. CUQ: Chatbot Usability Questionnaire.



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

XSL•FO
RenderX

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.

XSL•FO
RenderX

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

XSL•FO
RenderX

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

1. Brown J. How clinical communication has become a core part of medical education in the UK. Med Educ 2008 Mar;42(3):271-278. [doi: 10.1111/j.1365-2923.2007.02955.x] [Medline: 18275414]
2. Yaqoob Mohammed Al Jabri F, Kvist T, Azimirad M, Turunen H. A systematic review of healthcare professionals' core competency instruments. Nurs Health Sci 2021 Mar 26;23(1):87-102. [doi: 10.1111/nhs.12804] [Medline: 33386675]
3. de Haes H, Bensing J. Endpoints in medical communication research, proposing a framework of functions and outcomes. Patient Educ Couns 2009 Mar;74(3):287-294. [doi: 10.1016/j.pec.2008.12.006] [Medline: 19150197]
4. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009 Mar;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
5. Stewart MA. Effective physician-patient communication and health outcomes: a review. CMAJ 1995 May 01;152(9):1423-1433. [Medline: 7728691]
6. Goold SD, Lipkin M. The doctor-patient relationship: challenges, opportunities, and strategies. J Gen Intern Med 1999 Jan;14(Suppl 1):S26-S33. [doi: 10.1046/j.1525-1497.1999.00267.x] [Medline: 9933492]
7. Hausberg MC, Hergert A, Kröger C, Bullinger M, Rose M, Andreas S. Enhancing medical students' communication skills: development and evaluation of an undergraduate training program. BMC Med Educ 2012 Mar 24;12(1):16. [doi: 10.1186/1472-6920-12-16] [Medline: 22443807]
8. Deveugele M, Derese A, De Maesschalck S, Willems S, Van Driel M, De Maeseneer J. Teaching communication skills to medical students, a challenge in the curriculum? Patient Educ Couns 2005 Sep;58(3):265-270. [doi: 10.1016/j.pec.2005.06.004] [Medline: 16023822]

XSL•FO
RenderX

9. Noble LM, Scott-Smith W, O'Neill B, Salisbury H, UK Council of Clinical Communication in Undergraduate Medical Education. Consensus statement on an updated core communication curriculum for UK undergraduate medical education. Patient Educ Couns 2018 Sep;101(9):1712-1719. [doi: 10.1016/j.pec.2018.04.013] [Medline: 29706382]
10. Borowczyk M, Stalmach-Przygoda A, Doroszewska A, Libura M, Chojnacka-Kuraś M, Małecki ?, et al. Developing an effective and comprehensive communication curriculum for undergraduate medical education in Poland - the review and recommendations. BMC Med Educ 2023 Sep 07;23(1):645 [FREE Full text] [doi: 10.1186/s12909-023-04533-5] [Medline: 37679670]
11. Bachmann C, Pettit J, Rosenbaum M. Developing communication curricula in healthcare education: an evidence-based guide. Patient Educ Couns 2022 Jul;105(7):2320-2327. [doi: 10.1016/j.pec.2021.11.016] [Medline: 34887158]
12. von Fragstein M, Silverman J, Cushing A, Quilligan S, Salisbury H, Wiskin C, UK Council for Clinical Communication Skills Teaching in Undergraduate Medical Education. UK consensus statement on the content of communication curricula in undergraduate medical education. Med Educ 2008 Nov;42(11):1100-1107. [doi: 10.1111/j.1365-2923.2008.03137.x] [Medline: 18761615]
13. Palsson R, Kellett J, Lindgren S, Merino J, Semple C, Sereni D. Core competencies of the European internist: a discussion paper. Eur J Internal Med 2007 Mar;18(2):104-108. [doi: 10.1016/j.ejim.2006.10.002]
14. Keifenheim KE, Teufel M, Ip J, Speiser N, Leehr EJ, Zipfel S, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ 2015 Sep 28;15(1):159 [FREE Full text] [doi: 10.1186/s12909-015-0443-x] [Medline: 26415941]
15. Kaplonyi J, Bowles K, Nestel D, Kiegaldie D, Maloney S, Haines T, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ 2017 Dec 18;51(12):1209-1219. [doi: 10.1111/medu.13387] [Medline: 28833360]
16. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185 [FREE Full text] [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
17. Paradis E, Sutkin G. Beyond a good story: from Hawthorne effect to reactivity in health professions education research. Med Educ 2017 Jan 31;51(1):31-39. [doi: 10.1111/medu.13122] [Medline: 27580703]
18. Holmes S. Mitigating the Hawthorne effect using computer simulations. In: Serious Educational Game Assessment. Leiden, the Netherlands: Brill; 2011:175-187.
19. Jovanovic M, Baez M, Casati F. Chatbots as conversational healthcare services. IEEE Internet Comput 2021 May;25(3):44-51. [doi: 10.1109/mic.2020.3037151]
20. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019;5:2055207619871808 [FREE Full text] [doi: 10.1177/2055207619871808] [Medline: 31467682]
21. Frangoudes F, Hadjiaros M, Schiza E, Matsangidou M, Tsivitanidou O, Neokleous K. An overview of the use of chatbots in medical and healthcare education. In: Zaphiris P, Ioannou A, editors. Learning and Collaboration Technologies: Games and Virtual Environments for Learning. Cham: Springer International Publishing; 2021:170-184.
22. Moldt J, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659 [FREE Full text] [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
23. ChatGPT: optimizing language models for dialogue. Wayback Machine. 2022. URL: https://web.archive.org/web/20221130180912/https://openai.com/blog/chatgpt/ [accessed 2024-01-03]
24. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020 Sep 22;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]
25. Chung K, Park RC. Chatbot-based heathcare service with a knowledge base for cloud computing. Cluster Comput 2018 Mar 16;22(S1):1925-1937. [doi: 10.1007/s10586-018-2334-5]
26. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023 Jun 01;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
27. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
28. Ali R, Tang O, Connolly I, Zadnik SP, Shin J, Fridley J, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2022 May 17:2632. [doi: 10.1101/2023.03.25.23287743]
29. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv Preprint posted online 2023. [doi: 10.48550/arXiv.2303.12712]
30. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Mar 03;55(12):1-38. [doi: 10.1145/3571730]
31. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023 Jun 19;25:e43311 [FREE Full text] [doi: 10.2196/43311] [Medline: 37335593]

XSL•FO
RenderX

32. Maicher K, Danforth D, Price A, Zimmerman L, Wilcox B, Liston B, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Sim Healthc 2017;12(2):124-131. [doi: 10.1097/sih.0000000000000195]
33. ten Cate O, Custers EJFM, Durning SJ. Principles and Practice of Case-based Clinical Reasoning Education: A Method for Preclinical Students. Cham: Springer; 2018.
34. Welcome to the OpenAI developer platform. OpenAI. URL: https://platform.openai.com [accessed 2024-01-03]
35. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
36. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? 2019 Presented at: 31st European Conference on Cognitive Ergonomics; September 10-13, 2019; Belfast, UK. [doi: 10.1145/3335082.3335094]
37. Holmes S, Bond R, Moorhead A, Zheng J, Coates V, McTear M. Towards validating a chatbot usability scale. 2023 Presented at: DUXU 2023: 12th International Conference on Design, User Experience, and Usability; July 23–28, 2023; Copenhagen, Denmark p. 321. [doi: 10.1007/978-3-031-35708-4_24]
38. Holmes S, Bond R. Chatbot Usability Questionnaire (CUQ) Calculation Tool. Ulster University. URL: https://www.ulster.ac.uk/__data/assets/excel_doc/0010/478810/CUQ-Calculation-Tool.xlsx [accessed 2024-01-03]
39. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.
40. Huang J, Chang KCC. Towards reasoning in large language models: a survey. 2023 Presented at: ACL 2023: 61st Annual Meeting of the Association for Computational Linguistics; July 9-14, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.findings-acl.67]
41. Wason PC. Reasoning about a rule. Q J Exp Psychol 1968 Aug 01;20(3):273-281. [doi: 10.1080/14640746808400161] [Medline: 5683766]
42. Wason P, Johnson-Laird P. Psychology of Reasoning: Structure and Content. Cambridge, MA: Harvard University Press; 1972.
43. Galotti K. Approaches to studying formal and everyday reasoning. Psychol Bull 1989 May;105(3):331-351. [doi: 10.1037/0033-2909.105.3.331]
44. McHugh C, Way J. What is reasoning? Mind 2018 Jan 1;127(505):196. [doi: 10.1093/mind/fzw068]
45. Fagin R, Halpern J, Moses Y, Vardi M. Reasoning About Knowledge. Cambridge, MA: MIT Press; 2004.
46. Espejel JL, Ettifouri EH, Yahaya Alassan MS, Chouham EM, Dahhane W. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. Nat Lang Process J 2023 Dec;5:100032. [doi: 10.1016/j.nlp.2023.100032]
47. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv Preprint posted online 2023. [doi: 10.48550/arXiv.2302.04023]
48. Singh U, Aarabhi P. Can AI have a personality? 2023 Presented at: 2023 IEEE Conference on Artificial Intelligence (CAI); June 5-6, 2023; Santa Clara, CA p. 205-206 URL: https://ieeexplore.ieee.org/abstract/document/10194987/ [doi: 10.1109/cai54212.2023.00097]
49. Huang J, Wang W, Lam M, Li E, Jiao W, Lyu M. Revisiting the Reliability of Psychological Scales on Large Language Models. arXiv Preprint posted online 2023. [doi: 10.48550/arXiv.2305.19926]
50. Huang J, Lam M, Li E, Ren S, Wang W, Jiao W, et al. Emotionally numb or empathetic? Evaluating how LLMs feel using EmotionBench. arXiv Preprint posted online 2023. [doi: 10.48550/arXiv.2308.03656]
51. Serapio-García G, Safdari M, Crepy C, Sun L, Fitz S, Abdulhai M, et al. Personality traits in large language models. Research Square. 2023. URL: https://www.researchsquare.com/article/rs-3296728/latest [accessed 2024-01-03]
52. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ 2015 Oct 24;15:183 [FREE Full text] [doi: 10.1186/s12909-015-0468-1] [Medline: 26498479]
53. McNaughton N, Tiberius R, Hodges B. Effects of portraying psychologically and emotionally complex standardized patient roles. Teach Learn Med 1999 Jul;11(3):135-141. [doi: 10.1207/s15328015tl110303]
54. Liaw SY, Tan JZ, Lim S, Zhou W, Yap J, Ratan R, et al. Artificial intelligence in virtual reality simulation for interprofessional communication training: mixed method study. Nurse Educ Today 2023 Mar;122:105718 [FREE Full text] [doi: 10.1016/j.nedt.2023.105718] [Medline: 36669304]
55. Jani KH, Jones KA, Jones GW, Amiel J, Barron B, Elhadad N. Machine learning to extract communication and history-taking skills in OSCE transcripts. Med Educ 2020 Dec 10;54(12):1159-1170. [doi: 10.1111/medu.14347] [Medline: 32776345]
56. Tavarnesi G, Laus A, Mazza R, Ambrosini L, Catenazzi N, Vanini S, et al. Learning with virtual patients in medical education. CEUR-WS. URL: https://ceur-ws.org/Vol-2193/paper4.pdf [accessed 2024-01-03]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

API: application programming interface
CUQ: Chatbot Usability Questionnaire
GPT: generative pretrained transformer
LLM: large language model
QAP: question-answer pair

Edited by G Eysenbach; submitted 25.10.23; peer-reviewed by M Chatzimina, M Brown, Y Harada, A Khosla; comments to author 19.11.23; revised version received 09.12.23; accepted 14.12.23; published 16.01.24.

Please cite as:
Holderried F, Stegemann–Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, Holderried M, Herrmann-Werner A, Festl-Wietek T, Mahling M
A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study
JMIR Med Educ 2024;10:e53961
URL: https://mededu.jmir.org/2024/1/e53961
doi:10.2196/53961
PMID:38227363

©Friederike Holderried, Christian Stegemann–Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling. 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.

XSL•FO
RenderX

Original Paper

Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial

Akira Yamamoto[1], MD, PhD; Masahide Koda[2], MD, PhD; Hiroko Ogawa[3,4], MD, PhD; Tomoko Miyoshi[4,5], MD, PhD; Yoshinobu Maeda[1], MD, PhD; Fumio Otsuka[4], MD, PhD; Hideo Ino[5], MD, PhD

[1]Department of Hematology and Oncology, Okayama University Hospital, Okayama, Japan, Okayama, Japan

[2]Co-learning Community Healthcare Re-innovation Office, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan, Okayama, Japan

[3]Department of Primary Care and Medical Education, Dentistry and Pharmaceutical Sciences, Okayama University Graduate School of Medicine, Okayama, Japan, Okayama, Japan

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

[5]Center for Education in Medicine and Health Sciences, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan, Okayama, Japan

Corresponding Author:
Akira Yamamoto, MD, PhD
Department of Hematology and Oncology, Okayama University Hospital, Okayama, Japan
2-5-1 Shikata-cho, Kita-ku
Okayama, 700-8558
Japan
Phone: 81 86 235 7342
Fax: 81 86 235 7345
Email: ymtakira@gmail.com

Abstract

Background: Medical interviewing is a critical skill in clinical practice, yet opportunities for practical training are limited in Japanese medical schools, necessitating urgent measures. Given advancements in artificial intelligence (AI) technology, its application in the medical field is expanding. However, reports on its application in medical interviews in medical education are scarce.

Objective: This study aimed to investigate whether medical students' interview skills could be improved by engaging with AI-simulated patients using large language models, including the provision of feedback.

Methods: This nonrandomized controlled trial was conducted with fourth-year medical students in Japan. A simulation program using large language models was provided to 35 students in the intervention group in 2023, while 110 students from 2022 who did not participate in the intervention were selected as the control group. The primary outcome was the score on the Pre-Clinical Clerkship Objective Structured Clinical Examination (pre-CC OSCE), a national standardized clinical skills examination, in medical interviewing. Secondary outcomes included surveys such as the Simulation-Based Training Quality Assurance Tool (SBT-QA10), administered at the start and end of the study.

Results: The AI intervention group showed significantly higher scores on medical interviews than the control group (AI group vs control group: mean 28.1, SD 1.6 vs 27.1, SD 2.2; P=.01). There was a trend of inverse correlation between the SBT-QA10 and pre-CC OSCE scores (regression coefficient –2.0 to –2.1). No significant safety concerns were observed.

Conclusions: Education through medical interviews using AI-simulated patients has demonstrated safety and a certain level of educational effectiveness. However, at present, the educational effects of this platform on nonverbal communication skills are limited, suggesting that it should be used as a supplementary tool to traditional simulation education.

(JMIR Med Educ 2024;10:e58753) doi:10.2196/58753

KEYWORDS

medical interview; generative pretrained transformer; large language model; simulation-based learning; OSCE; artificial intelligence; medical education; simulated patients; nonrandomized controlled trial

XSL•FO
RenderX

Introduction

Medical interviews play a crucial role not only in the diagnostic process with patients but also in building trust and rapport [1]. Medical interviewing skills are necessary in medical practice and are categorized in the Japanese Model Core Curriculum for Medical Education under the categories of "Comprehensive Patient and Community Perspective" and "Clinical Competencies for Patient Care" [2]. In Japan, the Pre-Clinical Clerkship Objective Structured Clinical Examination (pre-CC OSCE), provided by the Public Interest Incorporated Association, Common Achievement Tests Organization, assesses fourth-year medical students for their competence and aptitude to participate in clinical clerkships [3]. This examination evaluates basic clinical skills, including medical interviewing. It is a nationwide standardized test with very limited flexibility in terms of feedback and the examination itself. Upon passing, medical students are expected to acquire the skills to conduct medical interviews through proper communication and gather necessary information before graduation through participatory clinical clerkships. The standard practice method involves learning medical interviewing in lectures, followed by practice sessions under the supervision of instructors and simulated patients [4].

However, opportunities for Japanese medical students to practice medical interviewing within the medical education curriculum are limited [5]. Japanese medical education has evolved by following the German model since the mid-19th century and the American model since the mid-20th century. As a unique development in Japan, standard curricula and nationwide common exams, including the pre-CC OSCE, have been introduced in medical schools across Japan, aiming to standardize medical education over the past quarter century. However, this has also restricted the autonomy of each university. The learning methods remain predominantly lecture-based and more flexible. In contrast, clinical-based learning methods such as problem-based learning and team-based learning have not yet been widely adopted in Western countries. Even after clinical clerkships, there are many restrictions on medical practice involving patients. This can be attributed to the fact that mandatory clinical training after graduation was implemented much later in Japan than in Western countries, and the integration between undergraduate medical education and postgraduate education is still underdeveloped. Furthermore, simulation education is effective across many fields for learners, not just medical interviewing, but the opportunities to use such education are limited in terms of both location and time [6]. Additionally, from educators' perspective, introducing medical interview education through simulation faces numerous barriers, including a lack of tutors, staff, simulated patients (including mannequins), and budget constraints [7].

Since the release of ChatGPT by OpenAI in the fall of 2022 [8], generative artificial intelligence (AI) technologies such as large language models (LLMs) have undergone rapid evolution and have been applied across various fields. In the medical domain, their integration is being considered in both clinical and research contexts [9]. One study demonstrated that LLMs can accurately answer questions of the United States Medical Licensing Examination (USMLE), demonstrating their use in medical education and assessment [10]. The COVID-19 pandemic accelerated the digital transformation from traditional bedside teaching to simulation education, including research into remote education models using chatbots [11,12]. However, research integrating LLMs into simulation education remains in its developmental phase [13].

In the field of medical interviewing, a survey of 3018 medical students revealed mixed feelings regarding the integration of LLMs. While some expressed concerns that it might deteriorate the patient-physician relationship, others were hopeful about the potential of AI technology in education, recognizing its dual value [14]. LLMs, which are distinct from previous deep learning–based algorithms, can predict the likelihood of a sequence of words based on the context of the preceding words. Natural and meaningful language sequences can be generated by learning from sufficient textual data. This capability led us to consider their application in practicing medical interviews.

In response to new advances in AI technology and the ongoing digital transformation and to alleviate the lack of educational resources for medical interview training, our team designed a simulation program to improve students' medical interview skills. This program uses GPT-4 Turbo to fulfill 2 roles: simulated patients and instructors providing feedback. To assess the educational impact of AI-assisted medical interview training on novice learners, specifically fourth-year medical students, we compared the scores from the clinical skills examination, pre-CC OSCE, between the control group, which practiced medical interviews only through traditional methods under the supervision of simulated patients and instructors, and the AI group, which received additional training through AI-simulated patient interviews. Since the medical students were preclinical clerkships, it was not possible to directly measure clinical competence. However, the pre-CC OSCE has shown a significant correlation with performance during clinical clerkships in Japanese medical student cohorts [15]. Notably, the scores of medical interviews have been identified as crucial predictors of performance during clinical clerkships. Therefore, in this study, the analysis was conducted using the scores from medical interviews.

Methods

Ethical Considerations

This educational research was approved by the institutional review board of Okayama University (2312-006). In this study, all data were anonymized and deidentified to ensure the privacy and confidentiality of the participants. No personally identifiable information was retained, and appropriate measures were taken to safeguard the participants' information. Furthermore, no compensation was provided to the participants for their involvement in the research.

Recruitment

As of November 2023, 35 fourth-year medical students at Okayama University, a national university in Japan, who consented to participate and had completed medical interview

XSL•FO
RenderX

practices at least once using our developed AI-simulated patient were designated as the intervention group (AI group, n=35). Fourth-year medical students from Okayama University as of November 2022 who had only a traditional educational program and did not participate in the intervention were selected as the control group (control group, n=110). The practice period was set to 1 month, and the students were provided with an educational environment that allowed them to practice at any time using their laptops or smartphones. After this 1-month training period, the students underwent the pre-CC OSCE, which served as the primary evaluation metric.

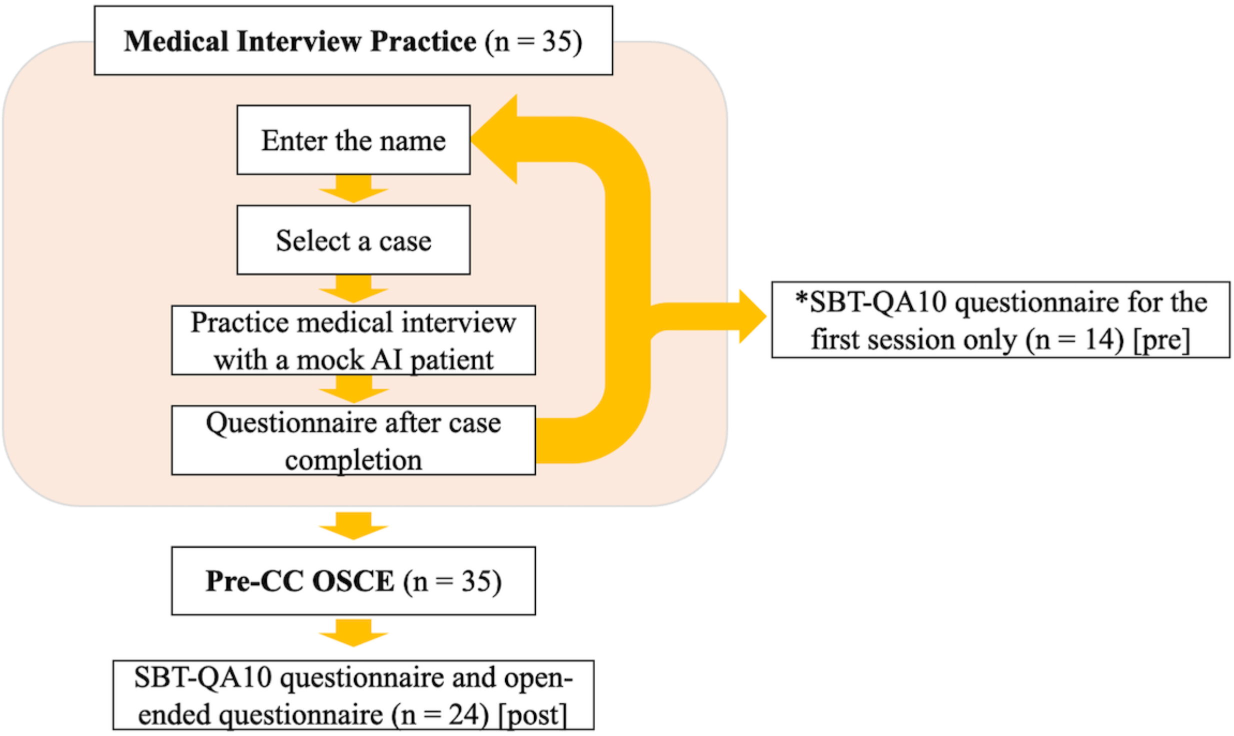
Educational Platforms

The responses of the AI-simulated patients were powered by GPT-4 Turbo, released in November 2023. We integrated it with the service "miibo" (miibo Corporation) through an application programming interface, which allows conversations with specified generative AI in a chat format. In this service, learners cannot see the prompts but can interact with fixed texts, such as case selections and questionnaires that do not involve AI, and choose from options and branch scenarios. While miibo is accessible via a web browser, it was also linked with LINE (LINE Corporation), which is widely used among students in Japan, for enhanced usability and to allow them to practice medical interviews via LINE as well. Learners could conduct interviews in chat format on either platform.

The GPT prompts were primarily composed of 3 elements: basic structure, case information, and feedback. The basic structure designated GPT-4 to act as the simulated patient and the learner as the physician practicing medical interviewing, with the emotional parameters fluctuating in response to the physician's statements. All outputs were in Japanese. The emotional parameters were set from 1 to 10 for 8 emotions—joy, sadness, anticipation, surprise, fear, disgust, trust, and anger—based on Ekman et al's [16] theory and Plutchik's [17] work. Initially, we loaded the case information into ChatGPT-4, ran a common prompt 3 times to estimate the initial emotional parameters, and set the average values. Case information included basic patient details, such as name, age, date of birth, and sex, along with relevant medical history. We prepared cases based on 8 primary symptoms, namely chest pain, abdominal pain, cough, heartburn, fatigue, fever, dizziness, and shortness of breath, which were developed and revised by multiple specialists. The feedback prompt was designed to provide feedback on general communication skills, elicitation of medically important information, and changes in patient emotions based on the conversation logs after the start of the medical interview. An example of a GPT prompt set on miibo is shown in Multimedia Appendix 1.

Consenting students could access the miibo platform page or a dedicated LINE account via a specified URL, where they could enter their name and select a case. After case selection, they were presented with a scenario starting with the patient entering the consultation room and initiating a greeting, marking the beginning of the medical interview. The conversations were primarily text-based, although voice input was also possible. After completing the medical interview practice, the session could be ended by clicking a button on the screen labeled "End medical interview" or by declaring it, followed by a transition to feedback within the miibo scenario. After the feedback, the conversation log was deleted, and the session proceeded to a questionnaire. After completing the questionnaire, participants were redirected to the case selection section, allowing them to repeat the practice of medical interviews as many times as they desired (Figure 1).

Figure 1. Research overview diagram. Pre-CC OSCE: Pre-Clinical Clerkship Objective Structured Clinical Examination.



XSL•FO
RenderX

Questionnaire

After completing the case, the questionnaire asked participants to rate the difficulty of the case on a 5-point scale and assess the realism of the AI-simulated patient, the sense of presence (interaction through emotions with the AI-simulated patient), and their levels of tension and anxiety on a 10-point scale. Participants were also asked to provide open-ended feedback on what they found good and bad about the experience. After completing the first practice session, students were asked to complete a questionnaire based on the Simulation-Based Training Quality Assurance Tool (SBT-QA10, prequestionnaire) [18] to evaluate the quality of the simulation training program. The SBT-QA10, a conventional evaluation tool for simulation training, was not used directly in this study but was partially modified to meet our specific needs. The item "I felt part of the team" was revised to reflect the sense of inclusion within a medical team comprising faculty members.

Additionally, while the medical interview practice solely involved conversations with AI, without direct visibility of faculty, all interaction logs were meticulously reviewed, and responses to questions were managed by faculty. Therefore, items related to support and interaction from faculties were retained. This questionnaire was administered again at the end of the study (postquestionnaire) by gathering open-ended feedback on the overall positive and negative experiences throughout the study.

Statistical Analysis

The primary outcome measure was the scores related to medical interviewing in the pre-CC OSCE. The pre-CC OSCE consisted of 2 evaluation formats: an overall performance evaluation (summary evaluation, scored from 1 to 6 points) and a score assessment based on individual skills according to a checklist (total score evaluation, scored from 1 to 31 points), both of which were targeted for assessment. As secondary outcome measures, we evaluated the SBT-QA10 and postcase practice questionnaires, specifically assessing the difficulty of the case, the realism of the simulated patient, interaction through emotions, and levels of anxiety and tension. The conversation logs from each practice session were also reviewed. At the start of practice, a unique ID was generated for each device and browser. This ID allowed for the accurate tracking of individual activity records when cross-referenced with the participant's initial name entry.

Statistical analysis was performed using Prism 9 for macOS (Version 9.5.1). The scores from the pre-CC OSCE were treated as interval data. In addition to the open-ended responses, the questionnaire used a Likert scale. The Mann-Whitney *U* test was used to compare 2 unrelated groups. Fisher exact test was applied to compare sex ratios, and Student *t* tests were used to compare backgrounds between groups based on grade point average (GPA; scored from 0.5 to 4.5). Multiple regression analysis was conducted with the pre-CC OSCE scores as the dependent variable and the questionnaire items as independent variables. The interpretation of correlation coefficients in this study follows the guidelines established by Hinkle et al [19]. According to their criteria, the strength of the correlation is categorized as follows: negligible (0.00-0.30), low (0.30-0.50), moderate (0.50-0.70), high (0.70-0.90), and very high (0.90-1.00). Missing values in the questionnaire items were excluded from the analysis. Additionally, only responses from participants who completed both the pre- and post-SBT-QA10 questionnaires were included in the analysis. The study was conducted with a feasible number of cases, and the effect size was evaluated by calculating Cohen *d* effect size using the pre-CC OSCE scores [19,20].

Results

Finally, the AI group that received LLM-based simulation education consisted of 35 of 87 students who had consented to participate in this study. In contrast, the control group comprised 110 students who had an opportunity to decline participation, but none chose to refuse. The effect size was calculated using the actual sample size and pre-CC OSCE scores, which revealed 0.48.

No significant differences were observed in the AI and control groups in the age, sex, or GPA of medically related subjects (Table 1). Regarding the medical interview practice, Multimedia Appendix 2 shows an abbreviated version of a representative conversation log and AI feedback, translated from Japanese to English.

Table 1. Background.

	AI[a] group	Control group	*P* value
Sex (female:male), n	15:20	34:76	.11[b]
Age (years), median (IQR)	22 (1)	23 (1)	.37[c]
GPA[d], mean (SD)	2.9 (0.5)	2.7 (0.6)	.10[e]

[a]AI: artificial intelligence.

[b]Fisher exact test was used for the sex ratio.

[c]The Mann-Whitney *U* test was used for the age.

[d]GPA: grade point average.

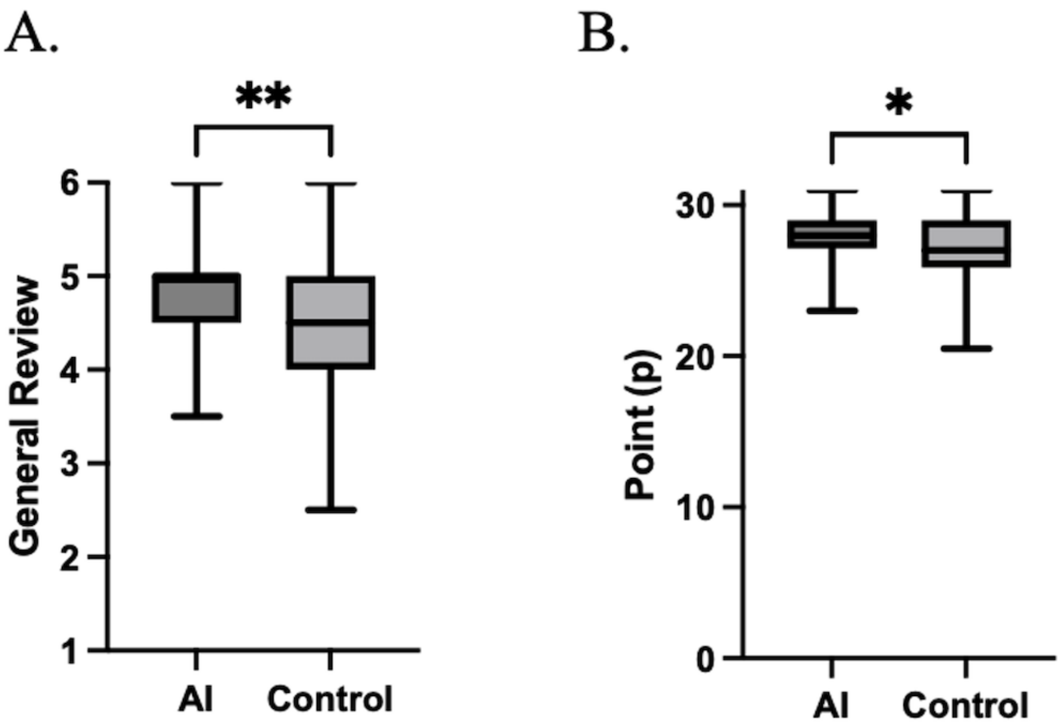
[e]The Student *t* test was used for the GPA (scored from 0.5 to 4.5) analysis.

Regarding the evaluation of educational effects, when comparing the scores for medical interviews in the pre-CC OSCE, the AI group scored significantly higher than the control group in both summary evaluations (AI vs control: 4.8, SD 0.7 vs 4.5, SD 0.7; 2-tailed; *P*=.007; maximum of 6 points, minimum of 1 point on a scale of 1-6) and total score evaluation (AI vs control: 28.1,

XSL•FO
RenderX

SD 1.6 vs 27.1, SD 2.2; 2-tailed; *P*=.01; maximum 31 points, minimum 0 points graded; Figure 2). Additionally, the passing score for the pre-CC OSCE has not been disclosed.

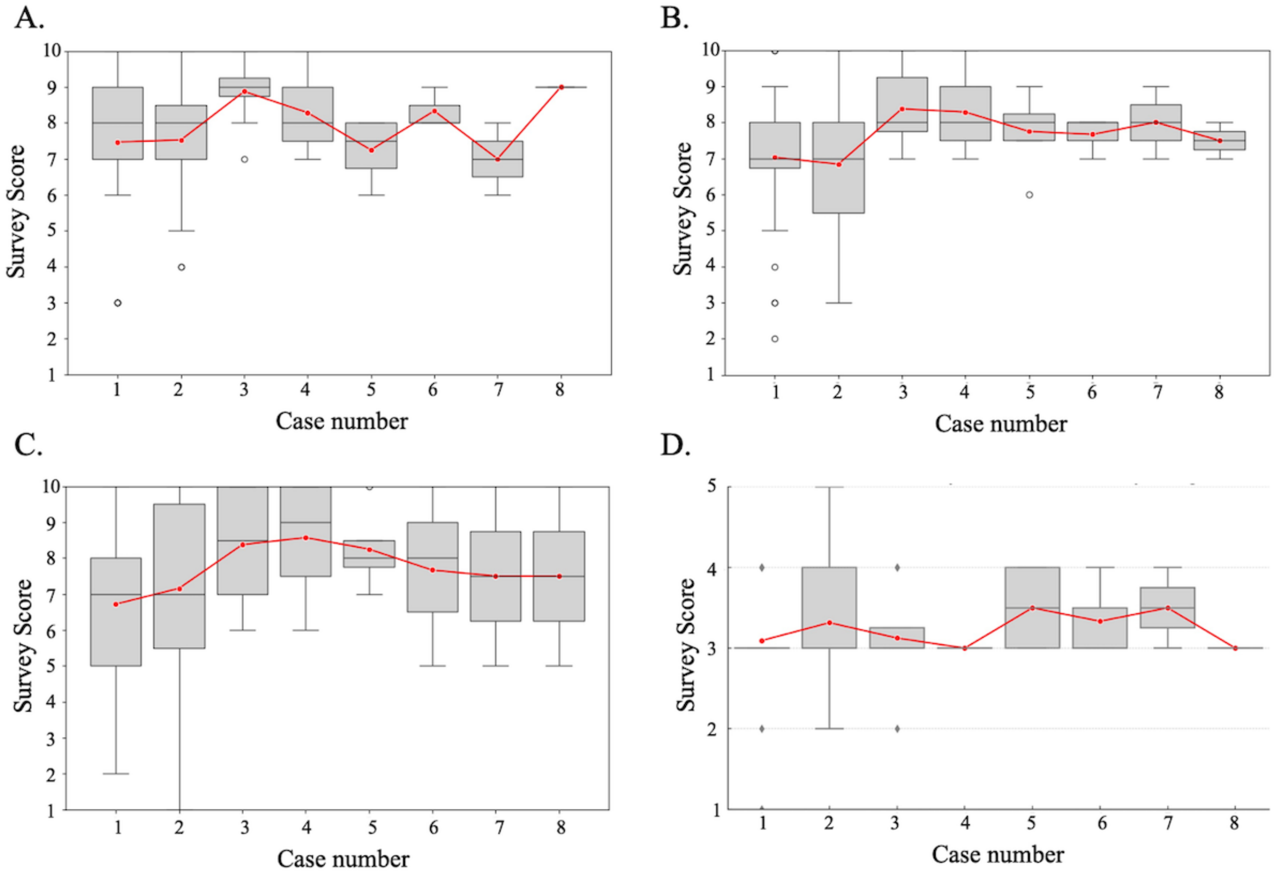
Figure 2. Pre-CC OSCE and LLMs-educational interventions. (A) Summary evaluation (maximum of 6 points, minimum of 1 point on a scale of 1-6). (B) Total score evaluation (maximum 31 points, minimum 0 points graded). Data were analyzed using the Mann-Whitney U test. LLM: large language model; pre-CC OSCE: Pre-Clinical Clerkship Objective Structured Clinical Examination. *P<.05, **P<.001.



The questionnaire results for each case regarding the realism of the AI-simulated patient, interaction through emotions, levels of anxiety and tension, and difficulty of the case are shown in Figure 3. The responses regarding the AI-simulated patients' reproducibility and interaction through emotions remained stable throughout, with median scores ranging from 7 to 9 for reproducibility and 7 to 8 for emotional interaction. Regarding the levels of anxiety and tension, it was observed that participants experienced them to some degree but without significant stress. Lastly, for the case difficulty, 75%(n=24) of the responses indicated it was "appropriate," 19%(n=6) found it "difficult," 3%(n=2) each considered it "easy" and "very easy," and 0% (n=0) found it "very difficult" in the first instance of the case (n=32). The response "appropriate" was the most common throughout the entire training period, ranging from 50% to 100%.

XSL•FO
RenderX

Figure 3. Questionnaire to be taken at the end of each case. (A) Artificial intelligence–simulated patient reproducibility is rated on a scale of 1–10, with 10 indicating "Very High Accuracy" and 1 indicating "no reproduction." (B) Interaction through emotions is rated on a scale from 1 to 10, where 10 signifies "Very Effective" and 1 signifies "Not Effective at All." (C) Anxiety and nervousness are rated on a scale from 1 to 10, with 10 indicating "Not Felt at All" and 1 meaning "Felt Very Strongly." (D) The difficulty of the case is rated on a scale from 1 to 5, where 1 represents "Very Easy" and 5 represents "Very Difficult.".



The scores for the SBT-QA10 in both the pre- and postquestionnaires were relatively high across all items, with the median scores ranging between 4 and 5 (Table 2). No significant changes were observed across all items from pre-questionnaire to postquestionnaire. Additionally, this analysis focused on the group (n=10) that responded to both the pre- and postsurveys. The results from separate analyses for the groups that only responded to the presurvey (n=14) and the postsurvey (n=24) are presented in Multimedia Appendix 3. No significant changes in trends were observed among these groups.

XSL•FO
RenderX

Table 2. Evaluation of the simulation program by SBT-QA10[a].

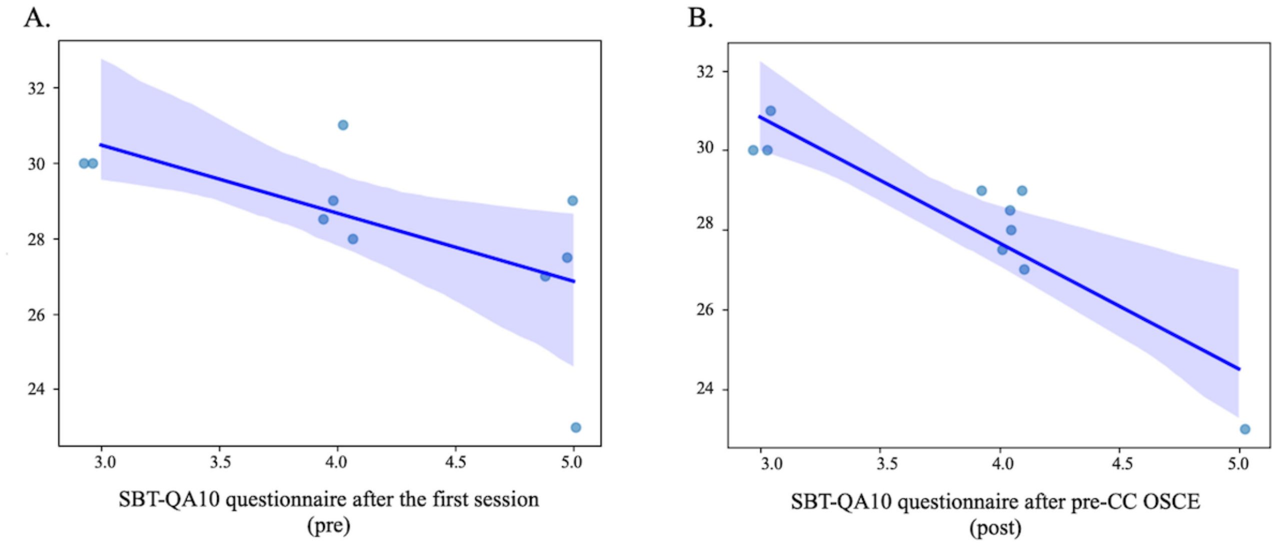
	SBT-QA10 questionnaire after the first session (pre), median (IQR)	SBT-QA10 questionnaire after pre-CC OSCE[b] (post), median (IQR)	*P* value (Wilcoxon test)
I felt part of the team (medical stuff team, including faculty)	4.0 (1.3)	4.0 (1.0)	.13
The faculty member(s) interacted well with me	4.5 (1.0)	4.0 (1.0)	.50
Being observed did not intimidate me	4.0 (2.3)	5.0 (2.0)	.50
I felt I was able to act as independently as I wanted to	4.0 (1.0)	4.0 (1.0)	>.99
I felt adequately supported by the faculty member(s)	4.0 (1.0)	4.0 (0.3)	.63
I felt that the scenario was realistic	5.0 (1.0)	4.5 (1.0)	>.99
I understood the purpose of the scenario	4.0 (1.3)	4.5 (1.0)	.38
It did not require a lot of mental effort to play my role in the scenario	4.0 (2.3)	4.0 (2.3)	>.99
I was not distracted by non-relevant objects and events during the scenario	4.0 (1.3)	4.0 (1.5)	>.99
I was focused on being involved in the scenario	4.5 (1.0)	4.5 (1.0)	>.99

[a]SBT-QA10: Simulation-Based Training Quality Assurance Tool. The results of the SBT-QA10 administered after the first session (pre) and pre-CC OSCE (post) for a sample size of 10 are presented for each item. Before-and-after comparisons were analyzed using the Wilcoxon test.

[b]Pre-CC OSCE: Pre-Clinical Clerkship Objective Structured Clinical Examination.

Next, we evaluated the group that received AI education to determine which subgroup achieved higher scores on the pre-CC OSCE. Given the high correlation coefficient of 0.75 between the total score evaluation and summary evaluation of the pre-CC OSCE and considering multicollinearity, we focused solely on the total score evaluation for further analysis, incorporating various questionnaire items, GPA and age in a multiple regression analysis. Among these, a consistent trend was observed with the SBT-QA10, where many items showed a negative correlation with the pre-CC OSCE scores. Specifically, the item "I felt part of the team" showed this trend statistically significant in both pre- (coefficients –1.8, SE 0.77; P=.047; R^2=0.41) and postevaluations (coefficients –3.2, SE 0.54; P<.001; R^2=0.81; Figure 4). When analyzing the total scores of each item in relation to the pre-CC OSCE scores to illustrate the overall trend, a negative correlation was observed; however, none were statistically significant. The analysis results of the combined pre- and post-SBT-QA10 scores are presented in Table 2, including the items of "I felt part of the team" (Table 3). In addition, the results from separate analyses for the groups that only responded to the presurvey (n=14) and the postsurvey (n=24) are presented in Multimedia Appendix 4. The multiple regression analysis revealed consistent negative trends across both the excluded groups and the entire dataset. No significant differences in all the items were observed, including the item 'I felt part of the team' in the pre-and postsurveys.

Figure 4. Multiple regression analysis of pre-CC OSCE and SBT-QA10. Multiple regression analysis of the "I felt part of the team (medical stuff team, including faculty)" item in the SBT-QA10 questionnaire. (A) Pre. (B) Post. Pre-CC OSCE: Pre-Clinical Clerkship Objective Structured Clinical Examination; SBT-QA10: Simulation-Based Training Quality Assurance Tool.



XSL•FO
RenderX

Table 3. Multiple regression analysis for analyzing pre-CC OSCE[a] scores and SBT-QA10[b].

Item	Const			Independent variable			R^2	Adjusted R^2
	β (SE)	*t* test (*df*)	*P* value	β (SE)	*t* test (*df*)	*P* value		
Pre-1: "I felt like part of the team"	35.9 (3.3)	11.0	<.001	–1.8 (0.8)	–2.4	.047	0.41	0.33
Post-1: "I felt like part of the team"	40.3 (2.1)	19.3	<.001	–3.2 (0.5)	–5.8	<.001	0.81	0.79
Total scores of pre-SBT-QA10	36.7 (6.9)	5.3	<.001	–2.0 (1.6)	–1.2	.25	0.16	0.05
Total scores of post-SBT-QA10	37.1 (7.1)	5.3	<.001	–2.1 (1.7)	–1.3	.25	0.16	0.06

[a]pre-CC OSCE: Pre-Clinical Clerkship Objective Structured Clinical Examination.

[b]SBT-QA10: Simulation-Based Training Quality Assurance Tool. Scores showed an overall inverse correlation with pre-CC OSCE scores, whereas the other items did not show a consistent trend. To illustrate the overall trend, the combined total scores of the SBT-QA10 from both pre and post are presented, including the item of "I felt part of the team" (pre 1 and post 1), in terms of which a significant difference was observed.

The AI group provided detailed feedback on both the advantages and disadvantages of the simulation system, summarized as follows:

- Positive aspects Practicality of training with AI: Participants could practice realistically with an emotional AI, akin to interacting with an actual patient. Convenience and accessibility: Training was available on an easy-to-use platform such as LINE, allowing participants to practice alone without a supervisor and offering flexibility in time and frequency of practice. Increased confidence through practice: Participants gained an understanding of the flow of a medical interview and learned essential questions relevant to clinical settings. Educational value and skill improvement: The training provided practical experience in medical interviewing and valuable feedback that helped improve skills, teaching participants how to inquire in various clinical situations.
- Negative aspects Dialogue and communication with AI: Participants encountered unnatural responses, such as repetitive expressions like "I'm worried" and instability in feedback. Technical and functional aspects: Issues included typing and response time delays, system errors, and operational inconveniences like incorrectly sent messages. Participants suggested that incorporating voice input might improve the experience. Comparison of medical interviews with pre-CC OSCE: Differences were noted in the information provided to simulated patients compared to pre-CC OSCE settings. Some participants appreciated the AI's superior conversational abilities, whereas others found the AI's casual speaking manner distracting.

Discussion

This study is the first to quantitatively verify the effectiveness of entrusting all aspects of medical interview education to AI, from acting as simulated patients to providing feedback as evaluators. It was found that the AI group, for which medical interview practice by LLM-based simulated patients was added to traditional medical interview education when practicing with simulated human patients, scored higher in the pre-CC OSCE medical interviews compared to the control group that only practiced with simulated human patients [4].

As previously reported, medical students do not resist the use of AI in medical education [21], which was evident in this study. The educational style of this study, which allowed students to practice using their smartphones and PCs, enabled them to practice repeatedly at their convenience, as mentioned in the open-ended feedback. This measure not only improved medical interview skills but also reduced anxiety due to a lack of practice and enhanced self-efficacy, suggesting a positive impact on the examination results. Although the 2 groups were from different academic years and might have confounding background factors, basic information such as GPA remained consistent between the groups. This educational method supported by LLMs has the potential to reduce financial and time costs for instructors and simulated patients. This study demonstrates that incorporating this method can effectively supplement the existing shortcomings in medical interview education, thus proving beneficial. However, there are limitations as outlined below. While improvements and applications are anticipated in the future, currently, platforms like this LLM-based medical interview practice should be cautiously used as supplementary tools to traditional simulation education.

Evaluating Clinical Significance

To verify the statistical significance of this study, the effect size was examined and found to be a moderate effect size [20,22]. Additionally, the minimal clinically significant difference (MCID), an indicator that represents the slightest change of clinical relevance to patients and health care providers, was used to evaluate the meaningfulness of the pre-CC OSCE scores. Unfortunately, there are limited references available that provide specific scores for setting MCID based on pre-CC OSCE scores. We considered the average score minus one standard deviation of the Match group from the study by Horita et al [23] as a reference value for MCID, which was calculated to be 26.9. Initially presented in percentage form in the source study, this value was converted to a point scale to align with the metrics used in our research. In this study, while the Control group's average score was approximately equal to the MCID, the

XSL•FO
RenderX

Intervention group significantly exceeded this benchmark. This suggests that the intervention could have led to clinically meaningful improvements in pre-CC OSCE scores. However, it is important to note that there are various methods for setting an MCID, and given the limited studies, this should be regarded as only one reference point.

Association Between Pre-CC OSCE Scores and AI Educational Interventions

When exploring which subgroups within the AI group tended to score higher on the pre-CC OSCE, there was an inverse correlation with the SBT-QA10 scores. Educators used the SBT-QA10 to understand the various perceptions experienced by learners during simulation education. High SBT-QA10 scores are generally thought to reflect positive experiences during simulations, leading to subsequent learning. The overall trend of high scores in this study suggests that the training had a positive impact on learners. However, subgroup analysis revealed results that contradict this implication. Unlike traditional simulation education with human-simulated patients, simulations conducted on one's smartphone or laptop allow for learning in a mentally safe state, potentially resulting in effortless learning within the comfort zone of students, thereby diminishing its effectiveness [24]. Conversely, for students who felt challenged, this may have created a learning zone that enhanced the learning effect.

Additionally, the SBT-QA10 is based on research in Western cultures, and this study, targeting learners in a Japanese cultural context, may require a different interpretation. People from Asian cultures have been reported to be stricter in self-evaluations. This cultural difference may have influenced the results significantly [25,26]. It is, therefore, considered important to adjust the learning environment, such as the difficulty level of cases, while constantly checking feedback from learners and educational outcomes because a good learning environment can vary among learners. However, there is a possibility that some extreme values are influencing the overall trend, as shown in Figure 4. Furthermore, as demonstrated in Multimedia Appendix 4, changing the comparison group eliminates the statistical significance previously observed, although a consistent negative trend is still evident. This suggests that the reliability of the data may be weak. Therefore, the interpretation of this trend should be approached with caution.

Fabrications by LLMs

Although concerns about fabrication by LLMs have been raised in various contexts [27], their occurrence in this study was limited, and no expressions deviating from the case settings were observed. During the alpha-testing phase with GPT-3.5 Turbo, fabrications were somewhat common, especially in instances where the AI began playing the role of the doctor instead of the simulated patient early in the conversation. Although modifications to the prompts somewhat mitigated this issue with GPT-3.5 Turbo, the change to GPT-4 and GPT-4 Turbo significantly reduced fabrications to a practical level of improvement [28].

The behavioral anomalies of AI in this study can be summarized as follows: The first concern is violations related to public order and morals based on OpenAI's guidelines. Upon analyzing the conversation logs, it was evident that the students' inputs did not contain any issues, indicating that the observed discrepancies were due to inaccuracies in the AI output. As the students were preinformed about the possibility of such errors, they could continue with their medical interviews by starting another consultation, preventing it from becoming a significant issue. The second point is related to fabrication in the feedback. For instance, despite confirming the patient's date of birth and name, there were a few cases in which the feedback suggested that these were not confirmed. This issue was thought to be caused by the prompts treating "confirming the patient's date of birth and name" as a continuous stream of information, and it was resolved by breaking down the information into separate elements. While prompt adjustments could improve some aspects, the specifications of GPT, which only allow reference to a certain amount of context window and have a limit on the amount of conversation that can be stored, are also considered to be contributing factors [29].

Safety

No excessive tension or anxiety associated with learning was observed during the simulations. Furthermore, responses from the GPT throughout the study period did not contain any statements that could harm learners' safety, and no students reported such concerns.

Limitations

This study was conducted with voluntary participation in educational research without using more desirable intervention methods, such as randomized controlled trials. The emphasis was on equality of educational opportunities, keeping the opportunities of traditional practice with simulated patients. Although consent was obtained from many students, only some of them actually participated in the medical interview practice sessions. This phenomenon can be attributed to unique cultural factors in Japan. Specifically, Japanese medical students often feel a strong inclination to meet others' expectations when explaining the research, leading them to provide consent [30]. However, this consent might not always reflect their genuine willingness, resulting in a lower actual participation rate. Consequently, the sample size was limited.

In addition, this study employed LLM-simulated patients' interventions and evaluated their effectiveness through a simulation-based assessment such as the pre-CC OSCE. However, reports suggest that qualitative improvements in simulators do not directly cause clinical skill enhancement, underscoring the importance of conducting clinical skill assessments in real-world settings as much as possible [31]. As this study focused on pre-clinical clerkship medical students, the assessment was limited to an indirect and short-term evaluation of clinical skills using medical interview scores from the pre-CC OSCE [15]. Therefore, we plan to conduct long-term evaluations of this program for clinical clerkship students and early-career physicians in actual clinical settings in future studies. Moreover, since this platform is text-based, its capacity to handle non-verbal communication is restricted. For instance,

XSL•FO
RenderX

similar to how GPT-4 can partially recognize visual and voice information, further advancements in LLM technologies that could better recognize and process human emotions and sensory inputs may help overcome this limitation. Currently, LLM-based medical interview simulation training should serve as a supplemental tool to existing medical interview education and is not yet capable of fully replacing traditional methods. Nonverbal communication skills, which are crucial, are still best developed through instructor-led training involving human-simulated patients. This study was conducted as a pilot project for the future application of LLMs in medical interview training.

Plan

This study suggests the potential for a significant reduction in the workload of instructors and simulated patients in medical interview practice while maintaining educational effects for medical students. Furthermore, the introduction of LLM-simulated patients to clinical skill examinations such as the pre-CC OSCE is conceivable. It holds promise not only for educating young doctors but also for the lifelong education of doctors, including simulations for handling complex cases in clinical settings. However, when introducing LLM simulations into medical education, caution is necessary regarding ethical considerations and accuracy, as previously pointed out. Completely replacing traditional instructor-led training with AI carries risks, and further studies thereon are required [13,21].

Improvements in prompts and the evolution of AI technology suggest that more realistic and accurate simulation education can be expected in the future. The integration of AI into medical education is inevitable; however, it has the potential to disrupt traditional medical education practices. Educators must remain vigilant regarding the potential positive and negative impacts of this integration [32]. Concurrently, it is essential to continue research on AI-mediated medical education to explore its applicability and limitations.

Conclusions

Education on medical interviewing using LLM-simulated patients demonstrated superior educational effectiveness while maintaining safety. This platform holds promise for multifaceted applications in the field of medical education in the future. It should be noted that this study only assessed short-term impacts and did not directly evaluate clinical skills. Additionally, due to the extremely limited educational effects on nonverbal communication skills, it is currently advisable to use this platform as a supplementary tool in medical interview training. Given the occurrence of fabrications and the opaque nature of LLM technology across various companies, caution and intense monitoring by tutors are essential when incorporating LLM-based educational platforms into medical education.

Acknowledgments

We extend our sincere thanks to Noriyuki Yamashita of the Center for Education in Medicine and Health Sciences, Okayama University Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama, Japan, for expertly managing our budget and meetings and ensuring the smooth progression of our project. We also thank Kana Nishikawa of the Instruction and Student Section, School of Medicine, Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama University, for her pivotal role in compiling performance data and liaising with the Public Interest Incorporated Association, Common Achievement Tests Organization, which was crucial for our funding and data integrity. Their contributions were invaluable to our work. This research was supported by Nishikawa Medical Foundation, Japan Medical Education Foundation, and donations to the Department of Primary Care and Medical Education, Dentistry and Pharmaceutical Sciences, Okayama University Graduate School of Medicine, Okayama, Japan.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt for medical interview practice.

[DOCX File , 25 KB - mededu_v10i1e58753_app1.docx]

Multimedia Appendix 2

Representative communication log and feedback in medical interview practice.

[DOCX File , 18 KB - mededu_v10i1e58753_app2.docx]

Multimedia Appendix 3

The simulation program was evaluated using the Simulation-Based Training Quality Assurance Tool, which was conducted separately in the groups that responded to the pre-survey (n=14) and the post-survey (n=24).

[DOCX File , 17 KB - mededu_v10i1e58753_app3.docx]

Multimedia Appendix 4

XSL•FO
RenderX

Multiple regression analysis for analyzing Pre-Clinical Clerkship Objective Structured Clinical Examination scores and Simulation-Based Training Quality Assurance Tool was conducted separately in the groups that responded to the presurvey (n=14) and the postsurvey (n=24).
[DOCX File , 17 KB - mededu_v10i1e58753_app4.docx]

References

1. Dang BN, Westbrook RA, Njue SM, Giordano TP. Building trust and rapport early in the new doctor-patient relationship: a longitudinal qualitative study. BMC Med Educ 2017 Feb 02;17(1):32 [FREE Full text] [doi: 10.1186/s12909-017-0868-5] [Medline: 28148254]
2. Japanese Ministry of Health Law. Medical Education Model Core Curriculum. 2023. URL: https://www.mhlw.go.jp/content/10900000/001026762.pdf [accessed 2024-02-06]
3. Learning and assessment items related to the skills and attitudes required of students participating in clinical participatory clinical practice (Version 4.2). Common Achievement Tests Organization. 2022. URL: https://www.cato.or.jp/pdf/osce_42.pdf [accessed 2024-02-06]
4. Cleland JA, Abe K, Rethans J. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach 2009 Jun;31(6):477-486. [doi: 10.1080/01421590903002821] [Medline: 19811162]
5. Abe K, Suzuki T, Fujisaki K, Ban N. Demographic characteristics of standardized patients (SPs) and their satisfaction and burdensome in Japan: the first report of a nationwide survey. Med Educ (Japan) 2007;38(5):301-307.
6. Bokken L, Rethans JJ, Scherpbier AJJA, van der Vleuten CPM. Strengths and weaknesses of simulated and real patients in the teaching of skills to medical students: a review. Simul Healthc 2008;3(3):161-169. [doi: 10.1097/SIH.0b013e318182fc56] [Medline: 19088660]
7. Nara N, Beppu M, Tohda S, Suzuki T. The introduction and effectiveness of simulation-based learning in medical education. Intern Med 2009;48(17):1515-1519 [FREE Full text] [doi: 10.2169/internalmedicine.48.2373] [Medline: 19721295]
8. Optimizing language models for dialogue. OpenAI. 2022. URL: https://openai.com/blog/chatgpt/ [accessed 2024-02-06]
9. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
11. Ajab S, Pearson E, Dumont S, Mitchell A, Kastelik J, Balaji P, et al. An alternative to traditional bedside teaching during COVID-19: high-fidelity simulation-based study. JMIR Med Educ 2022 May 09;8(2):e33565 [FREE Full text] [doi: 10.2196/33565] [Medline: 35404828]
12. Kaur A, Singh S, Chandan JS, Robbins T, Patel V. Qualitative exploration of digital chatbot use in medical education: a pilot study. Digit Health 2021;7:20552076211038151 [FREE Full text] [doi: 10.1177/20552076211038151] [Medline: 34513002]
13. Komasawa N, Yokohira M. Simulation-based education in the artificial intelligence era. Cureus 2023 Jun;15(6):e40940 [FREE Full text] [doi: 10.7759/cureus.40940] [Medline: 37496549]
14. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 09;22(1):772 [FREE Full text] [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
15. Komasawa N, Terasaki F, Nakano T, Kawata R. Relationships between objective structured clinical examination, computer-based testing, and clinical clerkship performance in Japanese medical students. PLoS One 2020;15(3):e0230792 [FREE Full text] [doi: 10.1371/journal.pone.0230792] [Medline: 32214357]
16. Ekman P, Sorenson ER, Friesen WV. Pan-cultural elements in facial displays of emotion. Science 1969;164(3875):86-88. [doi: 10.1126/science.164.3875.86] [Medline: 5773719]
17. Plutchik R. Emotion, a Psychoevolutionary Synthesis. New York City: Harper & Row; 1980.
18. Ekelund K, O'Regan S, Dieckmann P, Østergaard D, Watterson L. Evaluation of the simulation based training quality assurance tool (SBT-QA10) as a measure of learners' perceptions during the action phase of simulation. BMC Med Educ 2023 May 01;23(1):290 [FREE Full text] [doi: 10.1186/s12909-023-04273-6] [Medline: 37127593]
19. Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin; 2003.
20. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. England, UK: Routledge; 1988.
21. Shimizu I, Kasai H, Shikino K, Araki N, Takahashi Z, Onodera M, et al. Developing Medical Education Curriculum Reform Strategies to Address the Impact of Generative AI: Qualitative Study. JMIR Med Educ 2023 Nov 30;9:e53466 [FREE Full text] [doi: 10.2196/53466] [Medline: 38032695]
22. Hattie J. Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. 1st Edition. London: Routledge; 2008.
23. Horita S, Park YS, Son D, Eto M. Computer-based test (CBT) and OSCE scores predict residency matching and National Board assessment results in Japan. BMC Med Educ 2021 Feb 02;21(1):85 [FREE Full text] [doi: 10.1186/s12909-021-02520-2] [Medline: 33531010]

XSL•FO
RenderX

24. Brown M. Comfort zone: model or metaphor? J Outdoor Environ Education 2008 Apr 1;12(1):3-12. [doi: 10.1007/bf03401019]
25. Krendl AC, Pescosolido BA. Countries and cultural differences in the stigma of mental illness: the East–West divide. J Cross-Cultural Psychol 2020 Feb 21;51(2):149-167. [doi: 10.1177/0022022119901297]
26. Hsin A, Xie Y. Explaining Asian Americans' academic advantage over whites. Proc Natl Acad Sci U S A 2014 Jun 10;111(23):8416-8421 [FREE Full text] [doi: 10.1073/pnas.1406402111] [Medline: 24799702]
27. 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]
28. Hallucination leaderboard: comparing LLM performance at producing hallucinations when summarizing short documents. GitHub. URL: https://github.com/vectara/hallucination-leaderboard [accessed 2024-02-06]
29. GPT-4 Turbo in the OpenAI API. OpenAI. 2023. URL: https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api [accessed 2024-02-06]
30. Masaki S, Ishimoto H, Asai A. Contemporary issues concerning informed consent in Japan based on a review of court decisions and characteristics of Japanese culture. BMC Med Ethics 2014 Feb 04;15:8 [FREE Full text] [doi: 10.1186/1472-6939-15-8] [Medline: 24495473]
31. Bard A, Forsberg L, Wickström H, Emanuelson U, Reyher K, Svensson C. Clinician motivational interviewing skills in 'simulated' and 'real-life' consultations differ and show predictive validity for 'real life' client change talk under differing integrity thresholds. PeerJ 2023;11:e14634 [FREE Full text] [doi: 10.7717/peerj.14634] [Medline: 37810783]
32. Qadir J. Engineering education in the Era of ChatGPT: promise and pitfalls of generative AI for education. 2023 Presented at: IEEE Global Engineering Education Conference (EDUCON); May 1-4, 2023; Kuwait, Kuwait.

Abbreviations

AI: artificial intelligence
GPA: grade point average
LLM: large language model
MCID: minimal clinically important difference
Pre-CC OSCE: Pre-Clinical Clerkship Objective Structured Clinical Examination
SBT-QA10: Simulation-Based Training Quality Assurance Tool
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 25.03.24; peer-reviewed by H Berg, N Komasawa, VW Hui, DG Leung, CN Lo; comments to author 15.05.24; revised version received 04.06.24; accepted 15.08.24; published 23.09.24.

Please cite as:
Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, Ino H
Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial
JMIR Med Educ 2024;10:e58753
URL: https://mededu.jmir.org/2024/1/e58753
doi:10.2196/58753
PMID:39312284

©Akira Yamamoto, Masahide Koda, Hiroko Ogawa, Tomoko Miyoshi, Yoshinobu Maeda, Fumio Otsuka, Hideo Ino. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.09.2024. 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.

XSL•FO
RenderX

Using ChatGPT in Nursing: Scoping Review of Current Opinions

You Zhou[*], MSN; Si-Jia Li[*], MSN; Xing-Yi Tang, PhD; Yi-Chen He, MSN; Hao-Ming Ma, PhD; Ao-Qi Wang, MSN; Run-Yuan Pei, BSN; Mei-Hua Piao, PhD

School of Nursing, Chinese Academy of Medical Sciences, Peking Union Medical College, No. 33 Badachu Road, Shijingshan District, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Mei-Hua Piao, PhD

Abstract

Background: Since the release of ChatGPT in November 2022, this emerging technology has garnered a lot of attention in various fields, and nursing is no exception. However, to date, no study has comprehensively summarized the status and opinions of using ChatGPT across different nursing fields.

Objective: We aim to synthesize the status and opinions of using ChatGPT according to different nursing fields, as well as assess ChatGPT's strengths, weaknesses, and the potential impacts it may cause.

Methods: This scoping review was conducted following the framework of Arksey and O'Malley and guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews). A comprehensive literature research was conducted in 4 web-based databases (PubMed, Embase, Web of Science, and CINHAL) to identify studies reporting the opinions of using ChatGPT in nursing fields from 2022 to September 3, 2023. The references of the included studies were screened manually to further identify relevant studies. Two authors conducted studies screening, eligibility assessments, and data extraction independently.

Results: A total of 30 studies were included. The United States (7 studies), Canada (5 studies), and China (4 studies) were countries with the most publications. In terms of fields of concern, studies mainly focused on "ChatGPT and nursing education" (20 studies), "ChatGPT and nursing practice" (10 studies), and "ChatGPT and nursing research, writing, and examination" (6 studies). Six studies addressed the use of ChatGPT in multiple nursing fields.

Conclusions: As an emerging artificial intelligence technology, ChatGPT has great potential to revolutionize nursing education, nursing practice, and nursing research. However, researchers, institutions, and administrations still need to critically examine its accuracy, safety, and privacy, as well as academic misconduct and potential ethical issues that it may lead to before applying ChatGPT to practice.

(JMIR Med Educ 2024;10:e54297) doi:10.2196/54297

KEYWORDS

ChatGPT; large language model; nursing; artificial intelligence; scoping review; generative AI; nursing education

Introduction

Artificial intelligence (AI) was defined as a machine system that can make predictions, recommendations, and decisions influencing real or virtual environments based on a human-defined objective [1]. In recent years, with the rapid development of computer science, AI technology represented by machine learning, deep learning, and natural language processing has made amazing progress and achievements in the field of health care and been widely used in clinical practice, and has demonstrated a diagnostic performance that is not inferior to, or even better than human beings in some cases [2,3]. In the fields of nursing, AI is also playing an important role, including optimizing nursing processes [4], providing more personalized care [5], making health care more accessible [6], etc.

ChatGPT is an AI chatbot developed by OpenAI based on the third generation of the generative pretrained transformer architecture [7]. Since its release in November 2022, ChatGPT has attracted widespread attention and interest across the academic and scientific communities. Based on deep learning algorithms and natural language processing techniques, and trained with massive amounts of data from the internet, books, and articles, ChatGPT can automatically identify users' inputs and generate appropriate responses to simulate the interactive dialogue and feedback process between humans [8]. In the field of clinical medicine, ChatGPT has exhibited its ability to assist in disease diagnosis, and it was reported the correct diagnosis rate of ChatGPT-3 was about 93.3% in 10 differential diagnoses [9]. At the same time, ChatGPT has also shown great potential in assisting nursing. For example, ChatGPT could help nurses to improve documentation by standardizing the terms and concepts, thus reducing nurses' workload [10].

XSL•FO
RenderX

However, there are also widespread concerns about using ChatGPT in health care.

First, since ChatGPT's training data came from the internet and lacked transparency, researchers have expressed concerns about its accuracy, usability, and safety in clinical practice [11]. Second, during clinical application, considering the potential inconsistency between the training data and the clinical application scenarios, ChatGPT may endure implicit bias and data-shift problems, as well as artificial hallucinations caused by them, which may lead to insecurity issues and care inequity [12,13]. Overreliance on ChatGPT can also weaken nurses' judgment and lead to workforce deskilling. Third, in the academic publishing world, ChatGPT has caused broader discussions about academic integrity due to the difficulty of reviewers and available technologies in distinguishing content written by AI and a human [14]. In addition, especially in the field of education, although ChatGPT can help simplify administrative work, more educators expressed concerns that overdependence and complete trust in ChatGPT may cause and reinforce automation bias, and prevent students from developing abilities of critical thinking [15].

There have been extensive discussions about the application of ChatGPT in nursing. However, to date, no study has comprehensively summarized the perceptions on using ChatGPT in different nursing domains. Therefore, the aim of this study was to synthesize the opinions and acceptance of using ChatGPT from different application scenarios in nursing, as well as the strengths and weaknesses of ChatGPT and its possible impacts, to provide a reference for the future development of a large language model (LLM) that is more appropriate for nursing education and practice.

Methods

Study Design

This scoping review was conducted according to the 5-step methodological framework proposed by Arksey and O'Malley [16] (identifying the research question, identifying relevant studies, study selection, charting the data, and collating, summarizing, and reporting the results). The reporting of the review was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [17].

Identifying the Research Questions

1. How is ChatGPT used in different nursing fields, and what are the opinions and acceptance of this technology?
2. What are the strengths, weaknesses, ethical considerations, and potential impacts of the application of ChatGPT in nursing?

Identifying Relevant Studies

A comprehensive literature search was conducted in 4 web-based databases (PubMed, Embase, Web of Science, and CINHAL) from 2022 to September 3, 2023, to identify studies reporting the opinions and acceptance of using ChatGPT in nursing fields. Two reviewers (YCH and XYT) screened the references of the included articles to further identify relevant studies.

To include as many studies as possible, the search terms were not limited strictly. The search terms in PubMed included two key topic areas: ("ChatGPT" OR "Chatbot*" OR "Large language model" OR "LLM" OR "LLMs") AND ("Nursing" OR "Nurse*"). The search, using a combination of keywords and Boolean operators, was designed to comprehensively cover the intersection of ChatGPT and nursing.

Study Selection

The inclusion criteria were as follows: (1) articles associated with the application or opinions of ChatGPT in nursing fields, such as nursing education, nursing practice, nursing academic writing, etc; (2) any types of articles including original articles, review articles, preprints, protocols, editorials, letters to editor, correspondence, and case reports; and (3) English publications. We excluded studies without available full-text and nonhuman studies.

All identified articles were first imported into the EndNote X9 (Clarivate Analytics) software to manually remove duplicates. Then, two reviewers (YZ and SJL) independently screened the titles and abstracts through the Rayyan application according to the inclusion and exclusion criteria to include studies for further full-text assessment. Any disagreements were resolved through consensus by consulting another reviewer (MHP).

Charting the Data

According to the research question, two reviewers (XYT and YCH) independently extracted and synthesized pertinent information using an Excel sheet, including authors, year of publication, country, study design, objective of study, study results (opinions or findings of using ChatGPT in nursing), fields of concern, and suggestions or recommendations for future studies. Any disagreements were resolved through consulting another reviewer (MHP).

Collating, Summarizing, and Reporting the Results

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showed the process of study selection. Two researchers (YZ and SJL) independently used an inductive approach to analyze and thematically summarize the contents of the included studies to identify the opinions and acceptance of similarities and differences about using ChatGPT in nursing. On this basis, the opinions extracted from studies were further synthesized and categorized according to different nursing fields in which ChatGPT was applied (such as nursing education, nursing practice, nursing research, nursing writing, etc). A table of supplement material in Multimedia Appendix 1 were also created to demonstrate the status and opinions of using ChatGPT in nursing.

Results

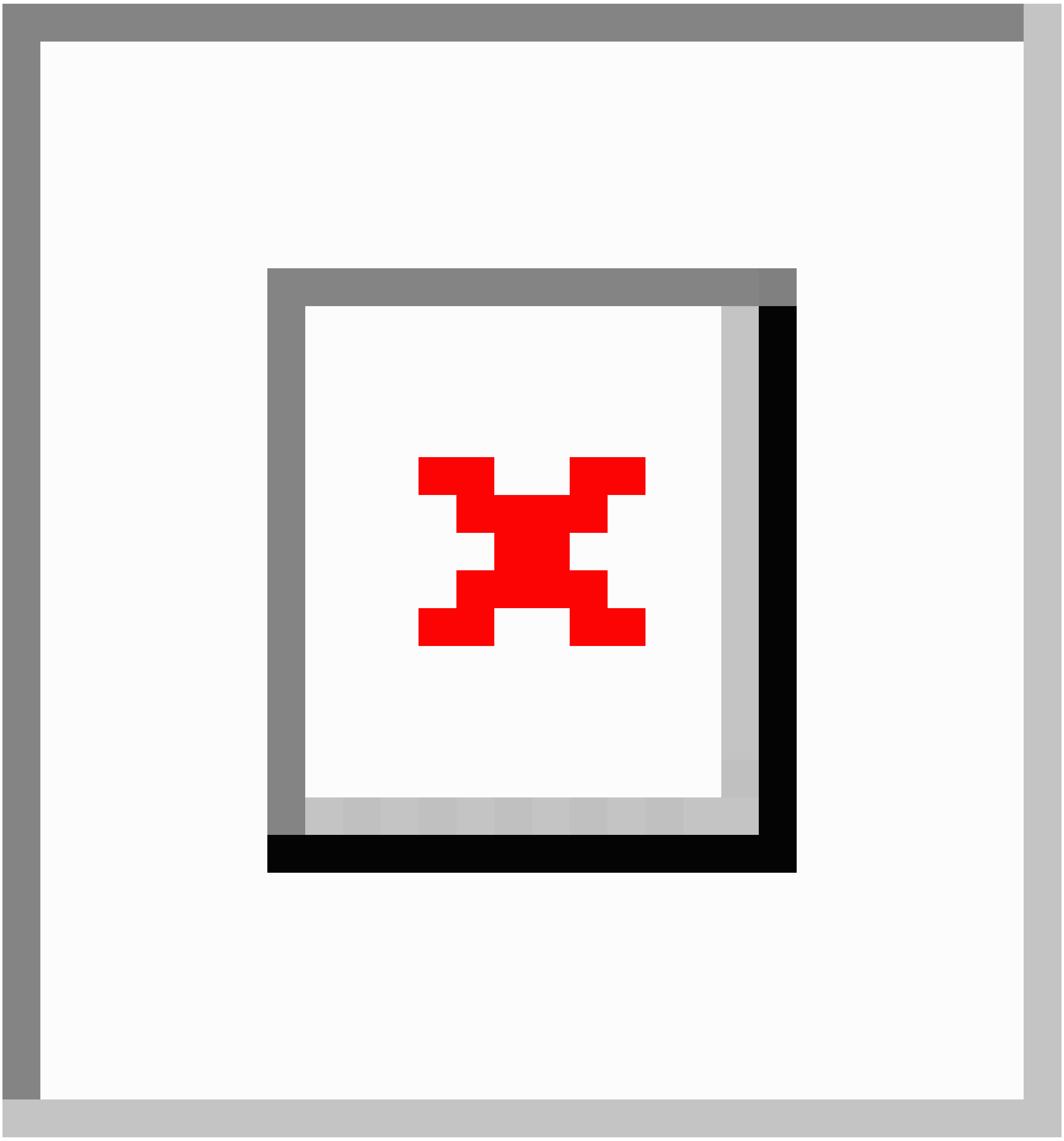
Search Results

Figure 1 showed the process of literature selection. A total of 320 studies were identified from the initial literature search. After removing the duplicates (n=135), 185 studies were

XSL•FO
RenderX

identified for titles and abstracts screening, of which 47 studies meeting the inclusion criteria were allowed for full-text evaluation. Finally, 17 studies were excluded, and 30 studies were included in this review.

Figure 1. PRISMA flow diagram of study selection. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



Study Characteristics

Table 1 summarized the characteristics of the included studies. All 30 studies were published in 2023. The United States (7/30), Canada (5/30), and China (4/30) were countries with the most publications, accounting for more than 50% of all publications. In terms of study design, more than half of the studies were editorials (12/30) as well as letters to the editor (6/30), only 3 were original articles, and this study's design was unclear in 7 studies. Table 2 presented the fields of concern of the included studies. Most studies focused on the application of ChatGPT in nursing education (n=20). Other fields of concern included using ChatGPT in nursing practice (n=10), nursing research (n=2), nursing academic writing (n=2), nursing examination (n=2), and nursing future (n=1). Six studies addressed the use of ChatGPT in multiple fields of nursing [15,18-22].

XSL•FO
RenderX

Table . Characteristics of included studies.

Characteristics		Studies, n (%)
Year of publication		
	2022	0 (0)
	2023	30 (100)
Country		
	United States	7 (23.33)
	Canada	5 (16.67)
	China	4 (13.33)
	Japan	2 (6.67)
	United Kingdom	2 (6.67)
	Australia	1 (3.33)
	Belgium	1 (3.33)
	Brazil	1 (3.33)
	Cambodia	1 (3.33)
	Indonesia	1 (3.33)
	Iraq	1 (3.33)
	Malta	1 (3.33)
	Netherlands	1 (3.33)
	Singapore	1 (3.33)
	Turkey	1 (3.33)
Study design		
	Editorial	12 (40)
	Not specific	7 (23.33)
	Letter to editor	6 (20)
	Article	3 (10)
	Debate essay	1 (3.33)
	Comment	1 (3.33)
Main fields of concern		
	Nursing education	20 (66.67)
	Nursing practice	10 (33.33)
	Nursing research	2 (6.67)
	Nursing academic writing	2 (6.67)
	Nursing examination	2 (6.67)
	Future nursing	1 (3.33)
	Multi-fields of nursing[a]	6 (20)

[a]Six studies addressed the use of ChatGPT in multiple fields of nursing [15,18-22].

XSL•FO
RenderX

Table . Fields of concern of the included studies.

Author	Year	Country	Study design	Nursing education	Nursing practice	Nursing research	Nursing academic writing	Nursing examination	Nursing future
Abdulai and Hung [18]	2023	Canada	Comment	✓	✓	✓	—[a]	—	—
Ahmed [23]	2023	Iraq	Letter	—	✓	—	—	—	—
Allen and Woodnutt [24]	2023	United Kingdom	Editorial	✓	—	—	—	—	—
Archibald and Clark [25]	2023	Canada	Editorial	✓	—	—	—	—	—
Berşe et al [19]	2023	Turkey	Letter	✓	✓	—	—	—	—
Castonguay et al [26]	2023	Canada	Not specific	✓	—	—	—	—	—
Chan et al [27]	2023	Hong Kong, China	Not specific	✓	—	—	—	—	—
Choi et al [28]	2023	Hong Kong, China	Not specific	✓	—	—	—	—	—
da Silva [29]	2023	Japan	Editorial	—	—	—	✓	—	—
Draganic [30]	2023	United States	Editorial	✓	—	—	—	—	—
Lim [31]	2023	United States	Editorial	✓	—	—	—	—	—
Frith [32]	2023	United States	Not specific	✓	—	—	—	—	—
Gunawan [33]	2023	Indonesia	Editorial	—	—	—	—	—	✓
Heerschap [20]	2023	Canada	Not specific	✓	✓	—	—	—	—
Irwin et al [21]	2023	Australia	Editorial	✓	✓	—	—	—	—
Kleebayoon and Wiwanikit [34]	2023	Cambodia	Letter	✓	—	—	—	—	—
Koo [35]	2023	Taiwan, China	Letter	✓	—	—	—	—	—
Moons and Van Bulck [22]	2023	Belgium	Editorial	—	✓	✓	—	—	—
O'Connor [36]	2023	United States	Editorial	✓	—	—	—	—	—
Odom-Forren [37]	2023	United States	Editorial	—	✓	—	—	—	—
Scerri and Morin [38]	2023	Malta	Editorial	—	✓	—	—	—	—
Shay [15]	2023	United States	Not specific	✓	✓	—	—	—	—
Siegerink et al [39]	2023	Netherlands	Editorial	—	—	—	✓	—	—
Sun and Hoelscher [40]	2023	United States	Article	✓	—	—	—	—	—

XSL•FO
RenderX

Author	Year	Country	Study design	Nursing education	Nursing practice	Nursing research	Nursing academic writing	Nursing examination	Nursing future
Taira et al [41]	2023	Japan	Article	—	—	—	—	✓	—
Tam et al [42]	2023	Singapore	Not specific	✓	—	—	—	—	—
Thakur et al [43]	2023	Canada	Letter	✓	—	—	—	—	—
Vitorino and Júnior [44]	2023	Brazil	Letter	✓	—	—	—	—	—
Woodnutt et al [45]	2023	United Kingdom	Debate essay	—	✓	—	—	—	—
Zong et al [46]	2023	China	Article	—	—	—	—	✓	—

[a]A blank space indicates that the content is not covered in the corresponding article.

ChatGPT and Nursing Education

Existing research has shown that ChatGPT has great potential in the field of nursing education. For educators, ChatGPT can be used for curriculum development, drafting course materials, and generating practice tests, which can simplify teachers' course preparation and assessment tasks [15,42,43]. Teachers can use ChatGPT to simulate patient encounters, providing students with an interactive learning experience to practice skills such as communication and assessment to enhance education [21,36,42]. For students, since ChatGPT has the function of instant feedback, it can be used as a tool to quickly acquire knowledge and skills, helping to improve learning efficiency and time management [19,27,40]. Students can also create individualized learning plans and obtain personalized feedback from ChatGPT, and use it to develop their writing skills, which will help motivate students to carry out independent learning and improve the efficiency and accuracy of the writing process [21,35,36,42-44]. In addition, ChatGPT has been believed to improve students' digital literacy [26,42].

However, opposition exists at the same time. The researchers argue that using ChatGPT in nursing education may lead to plagiarism in assignments and academic dishonesty, given its superior ability to generate textual content [21,28,31,36]. It is also for this reason that, ChatGPT may undermine the nursing education assessment system that is now based on essays and assignments [24,36]. Students' excessive use of ChatGPT may lead to reduced course participation [15]. Moreover, due to the nature of passive acceptance, over-reliance on ChatGPT will be detrimental to students' ability to transform information into knowledge, as well as critical thinking, literature retrieval, and evidence synthesis [15,20,28,31,32,42].

ChatGPT and Nursing Practice

The current view is that nurses can provide an unprecedented personalized care to patients based on ChatGPT; at the same time, patients can use it for health consultations, information about the status of their diseases and symptoms, and about their treatments [23]. In addition, due to the advantages of rapid assistance and rapid resource accessibility, ChatGPT can be used as a tool for nurses to quickly access information, helping nurses to keep up to date with information about patients' illnesses, treatments, and medications, which is conducive to optimizing time management and providing high-quality care for patients [37,38,40].

However, despite the promising applications, there are still some problems and limitations in applying ChatGPT to nursing practice. First, ChatGPT cannot guarantee the security and confidentiality of the information uploaded to the servers. Therefore, inputting detailed and private information of patients to it may lead to a leakage of patients' privacy [18-20,23,38]. Second, unlike search engines, ChatGPT does not search the internet to find the best answer to a question, but rather analyzes a large amount of data and then predicts the next most likely word in the answer, and therefore may output incorrect or biased information [19,20,37,38,45]. What's more, nursing is a human-centered discipline, and a major disadvantage of chatbots is that they do not have the unique emotions and empathy of humans. Communication based on ChatGPT may make communication between nurses and patients impersonal and lacking in empathy, which may have a negative impact on the nurse-patient relationship [18,19,23,27,37,38].

ChatGPT and Nursing Research, Writing, and Examination

There are also widespread concerns about using ChatGPT in academia and publishing. As ChatGPT is not an individual nor can it be held responsible for the content it generates, scholars argued that the decision to list ChatGPT as a coauthor was wrong and undesirable [39]. In addition, researchers had attempted to complete the nursing examinations using ChatGPT. Taira et al [41] found that ChatGPT demonstrated a stable, very close passing level in the 2019 - 2023 Japanese National Nurse Examinations, however, ChatGPT showed some limitations in dealing with questions in complex situations. Zong et al [46] tested ChatGPT's performance on the 2017 - 2021 Chinese National Nurse Licensing Examination. The results showed that ChatGPT did not pass the examination in any of the years but scored equally close to the passing score [46].

XSL•FO
RenderX

Discussion

Principal Findings

This scoping review aimed to summarize the opinions and acceptance of published studies on the use of ChatGPT in nursing fields. The results of our study indicated that, nursing research on ChatGPT is still in its infancy and few original research has been conducted. ChatGPT has the potential to provide nursing students with personalized study guides, provide patients with high-level personalized care plans, and greatly facilitate research and academic writing efforts, but at the same time, it can also lead to automation bias, nurse-patient mistrust, and potential ethical issues caused by misinformation, and academic misconduct issues. Discussion about using ChatGPT in nursing education, nursing practice, and nursing research and academic writing remains heated and the researchers have not yet reached a unanimous opinion.

Considering the global nursing shortage, the cultivation of exceptional nurses has become an important issue in the field of nursing education. Therefore, when new technologies are available, what role they can play in nursing education is of particular interest. First, ChatGPT can assist teaching. For example, ChatGPT's superior generative and analytical capabilities can help teachers reduce their workload by converting complex learning materials into easy-to-understand classroom content and assisting in grading students' work [47]. Second, ChatGPT can facilitate changes in learning methods. ChatGPT can generate outlines to assist with literature reviews; create realistic clinical cases and scenarios to help medical students improve their diagnostic skills; and act as a personal tutor to create personalized learning plans and materials based on students' abilities and learning feedback to improve learning efficiency [47,48]. In addition, ChatGPT was found to improve information skills in nursing students. In a study by Rahman and Watanobe [49], ChatGPT was found to assist students in generating code, checking code errors, and debugging and optimizing code. This is very important. With the advent of the digital age, programming will likely become a required course for nursing education and an essential skill for nurses in the future. ChatGPT's significant help in programming learning is very meaningful to the learning of nursing informatics and cultivation of digital literacy for nursing students.

Although ChatGPT has demonstrated potential benefits in nursing education, opposition emerges from researchers. Academic writing is crucial for students' success, yet crafting a research paper is a daunting task, even for experienced writers. ChatGPT plays a vital role in assisting with the writing process, but also raises issues about academic dishonesty, particularly when students become overly dependent on it [50]. In addition, students can also exploit ChatGPT for cheating during examinations, thus undermining the integrity of these assessments [51,52]. Furthermore, the use of ChatGPT in nursing education also brings ethical considerations such as data privacy and security. Students may share personal thoughts, feelings, and experiences while using ChatGPT, posing potential risks associated with the collection of this sensitive information [53].

Therefore, when integrating nursing education and the emerging technology, educators should comprehensively consider the strengths and limitations of ChatGPT. Educators and educational institutions should embrace this technology with an open mind and avoid simply banning its use. In practice, educators should teach students to critically evaluate and properly use ChatGPT to avoid overreliance; and use diverse teaching methods to encourage them to acquire skills of critical and independent thinking, and clinical reasoning. It is also critical to address and resolve ethical concerns, such as finding a balance between data privacy and correctly using ChatGPT. Moreover, educational institutions or educational administrations ought to establish guidelines and consensus or systems regarding the proper use of ChatGPT in nursing education.

In addition to nursing education, researchers also showed great interest in how ChatGPT can be applied to and improve nursing practice. ChatGPT empowers patients with health consultations and can help nurses to give personalized patient care by acting as an information tool. In a study by Kuroiwa et al [54], patients achieved accurate self-diagnosis of carpal tunnel syndrome and lumbar spinal stenosis by ChatGPT. ChatGPT seems to have the potential to become a patient self-management and condition monitoring tool outside the hospital. Therefore, future research could attempt to develop a ChatGPT-based chatbot and integrate it into existing mobile health (mHealth) intervention programs and platforms, exploring the role of mHealth interventions integrated with a LLM on symptom control and lifestyle change in patients with chronic diseases.

However, ethical concerns (ie, security and confidentiality, accuracy and bias in information output, and the lack of human empathy) also exist, and some issues are inevitable due to the nature of AI. For instance, the disclosure of patients' privacy and provision of incorrect information may damage the trusting relationship between patients and nurses. Additionally, compassion emerges from interpersonal relationships and social interactions with persons, thus chatbots were considered to lack the capacity for compassion [55]. However, some consumer informatics studies found that chatbots seemed to be better at projecting the impression of empathy. In the study by Chen et al [56], a chatbot provided high-quality, empathetic, and easy-to-read answers to cancer-related questions on social media that were comparable to those provided by doctors. While the issue of empathy seems to be resolved, it is worth pondering whether chatbots will still be able to balance empathy and ethics to provide reliable answers to patients' questions in the face of complex and varied real-life clinical environments and problems.

Given these concerns, implementing risk management strategies to control these risks is crucial. First, data confidentiality is essential when applying ChatGPT in nursing practice, and patients should be provided with informed consent and told not to disclose private personal information. Second, information provided by ChatGPT may be inaccurate and biased, thus professionals' interventions such as reviewing the information developed by ChatGPT, and addressing bias in decision-making processes are necessary. Third, although ChatGPT can greatly improve nurses' efficiency, it still cannot replace the important role of nurses. Future nurses should emphasize the human touch and ethical considerations in nursing processes and conduct

XSL•FO
RenderX

more research to determine the support resources needed to effectively use this technology [19].

The concerns regarding using ChatGPT in other nursing fields also exist. As far as research and academic writing is concerned, several studies have now listed ChatGPT as a coauthor [36,57,58]. However, Palagani et al [59] found that although ChatGPT can generate article content as well as references as requested by the author, most of the references were incorrect or nonexistent. As a supportive tool for academic writing, ChatGPT can assist researchers in conducting a literature review and correcting grammatical errors to improve writing quality [60]. However, the abuse of ChatGPT may carry a great risk of leading to academic misconduct. In a study by Gao et al [14], reviewers indicated that it was difficult to distinguish between content generated by AI and human. Although recognition tools such as GPTZero and GPT-2 Output Detector (OpenAI) are already available, accurately identifying AI-generated content in submitted manuscripts will still be a daunting task as chatbot algorithms are iterated and optimized. Therefore, future research should focus on the development of recognition tools for AI-generated content and try to optimize the language style of different languages to improve the detection performance.

Scholars also explored ChatGPT's capability to pass nursing licensing examinations and found that although it approached the passing threshold, it failed to meet the required passing standards. Considering that ChatGPT was developed primarily based on English-language data, and that there are differences in health care policies, regulations, languages, and cultures in various countries, this may partly explain why ChatGPT could not pass the examinations. This emphasizes an important ethical concern about the applicability and fairness of using AI in different health care settings. To address this issue, incorporating a wider range of languages and cultural contexts may be the future aim of AI technologies' development.

Future Directions

First, from the perspective of nursing education, educators should instruct students on the proper use of ChatGPT. Teachers should inform students to consciously consider LLMs such as ChatGPT as information search engines and learning assistants to avoid overreliance. Further, the most important thing is to cultivate students' critical thinking and information discernment skills so that they can recognize artificial hallucination and extract useful information provided by ChatGPT while discarding untrue and false contents. Additionally, educational institutions could establish guidelines and consensus about the proper use of ChatGPT in nursing education to standardize the current state of using LLMs in the educational profession. Second, in the context of nursing practice, given the potential of applying ChatGPT into symptom management and lifestyle change in patients with chronic diseases, a ChatGPT-based chatbot could be developed and integrated into mHealth intervention programs, and patients' private data can be secured by setting access rights and encrypting private data. In addition, more research and multiple efforts are required to identify the support resources needed to apply ChatGPT into nursing practice. Specifically, laws and regulations, and ethical standards for using LLMs in clinical practice are still to be introduced by the government and health care management agency; in terms of health care organizations, use guidelines and training curricula should be developed according to local application scenarios, patients' needs, and nurses' qualifications in the future; for researchers and developers, there is still a need for further diagnostic accuracy evaluation and usability testing to enhance the reliability of ChatGPT in complex clinical environments. Third, regarding nursing research, future research should concentrate on developing advanced tools to identify AI-generated content. To enhance the applicability and fairness of using ChatGPT, incorporating a broader spectrum of languages and cultural contexts may be the future aim of AI technologies' advancement.

Limitations

This study also had some limitations. First, this study only included publications in English, which may lead to a certain publication bias. Second, the search deadline for this study was September 3, 2023, considering the rapidly growing publication volume of studies on the application of ChatGPT in nursing, further reviews are still needed in the future to include more studies to enrich our findings. In addition, given the small number of original studies available about ChatGPT and nursing, this review included a wide range of types and quality of studies, and some of the low-quality studies may compromise the generalizability of the results of this study.

Conclusions

As an emerging AI technology, ChatGPT has received a lot of attention and generated intense discussion in various nursing fields. Although at present, there is still a lack of original studies about its practical application in nursing, ChatGPT has showed great potential to revolutionize nursing education, nursing practice, and nursing research. However, before it can be applied to practice, researchers, institutions, and administrations still need to critically examine the privacy, safety, and accuracy as well as academic misconduct and potential ethical issues it may lead to.

Acknowledgments

The author thanked all authors of this study. This study was funded by the nonprofit central research institute fund of Chinese Academy of Medical Sciences (grant 2023-RC320-01).

Data Availability

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

XSL•FO
RenderX

Authors' Contributions

YZ and MHP designed this study. YZ and SJL conducted literature searching. YZ and SJL screened and reviewed the articles. XYT and YCH extracted the data from included studies. YZ drafted the manuscript. MHP provided guidance and approved the final draft. All authors contributed to the development of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on the study content of the included studies.
[PDF File, 190 KB - mededu_v10i1e54297_app1.pdf]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File, 97 KB - mededu_v10i1e54297_app2.pdf]

References

1. Phillips SP, Spithoff S, Simpson A. Artificial intelligence and predictive algorithms in medicine: promise and problems. Can Fam Physician 2022 Aug;68(8):570-572. [doi: 10.46747/cfp.6808570] [Medline: 35961724]
2. Rauschecker AM, Rudie JD, Xie L, et al. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. Radiology 2020 Jun;295(3):626-637. [doi: 10.1148/radiol.2020190283] [Medline: 32255417]
3. Krishnan G, Singh S, Pathania M, et al. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. Front Artif Intell 2023;6:1227091. [doi: 10.3389/frai.2023.1227091] [Medline: 37705603]
4. Pailaha AD. The impact and issues of artificial intelligence in nursing science and healthcare settings. SAGE Open Nurs 2023;9:23779608231196847. [doi: 10.1177/23779608231196847] [Medline: 37691725]
5. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
6. Al Kuwaiti A, Nazer K, Al-Reedy A, et al. A review of the role of artificial intelligence in healthcare. J Pers Med 2023 Jun 5;13(6):951. [doi: 10.3390/jpm13060951] [Medline: 37373940]
7. OpenAI. ChatGPT: Optimizing Language Models for Dialogue: OpenAI; 2022.
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Dig Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
10. Gosak L, Pruinelli L, Topaz M, Štiglic G. The ChatGPT effect and transforming nursing education with generative AI: discussion paper. Nurse Educ Pract 2024 Feb;75:103888. [doi: 10.1016/j.nepr.2024.103888] [Medline: 38219503]
11. Shorey S, Mattar C, Pereira TLB, Choolani M. A scoping review of ChatGPT's role in healthcare education and research. Nurse Educ Today 2024 Apr;135:106121. [doi: 10.1016/j.nedt.2024.106121] [Medline: 38340639]
12. Liu Z, Zhang L, Wu Z, et al. Surviving ChatGPT in healthcare. Front Radiol 2023;3:1224682. [doi: 10.3389/fradi.2023.1224682] [Medline: 38464946]
13. Pendergrast T, Chalmers Z. Authors' reply: a use case for generative AI in medical education. JMIR Med Educ 2024 Jun 7;10:e58370. [doi: 10.2196/58370] [Medline: 38860619]
14. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75. [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
15. Shay A. ChatGPT: implications for faculty, students, and patients: May 19, 2023. Clin Nurse Spec 2023;37(5):245-246. [doi: 10.1097/NUR.0000000000000770] [Medline: 37595200]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Tricco AC, Lillie E, Zarin W. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Abdulai AF, Hung L. Will ChatGPT undermine ethical values in nursing education, research, and practice? Nurs Inq 2023 Jul;30(3):37101311. [doi: 10.1111/nin.12556] [Medline: 37101311]
19. Berşe S, Akça K, Dirgar E, Kaplan Serin E. The role and potential contributions of the artificial intelligence language model ChatGPT. Ann Biomed Eng 2024 Feb;52(2):130-133. [doi: 10.1007/s10439-023-03296-w]
20. Heerschap C. Use of artificial intelligence in wound care education. Wounds Int 2023;14(2):12-15. [Medline: 164494659]

XSL•FO
RenderX

21. Irwin P, Jones D, Fealy S. What is ChatGPT and what do we do with it? Implications of the age of AI for nursing and midwifery practice and education: an editorial. Nurse Educ Today 2023 Aug;127:105835. [doi: 10.1016/j.nedt.2023.105835] [Medline: 37267643]
22. Moons P, Van Bulck L. ChatGPT: can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. Eur J Cardiovasc Nurs 2023 Oct 19;22(7):e55-e59. [doi: 10.1093/eurjcn/zvad022] [Medline: 36752788]
23. Ahmed SK. The impact of ChatGPT on the nursing profession: revolutionizing patient care and education. Ann Biomed Eng 2023 Nov;51(11):2351-2352. [doi: 10.1007/s10439-023-03262-6] [Medline: 37266721]
24. Allen C, Woodnutt S. Can ChatGPT pass a nursing exam? Int J Nurs Stud 2023 Sep;145:104522. [doi: 10.1016/j.ijnurstu.2023.104522] [Medline: 37354792]
25. Archibald MM, Clark AM. ChatGTP: what is it and how can nursing and health science education use it? J Adv Nurs 2023 Oct;79(10):3648-3651. [doi: 10.1111/jan.15643] [Medline: 36942780]
26. Castonguay A, Farthing P, Davies S, et al. Revolutionizing nursing education through AI integration: a reflection on the disruptive impact of ChatGPT. Nurse Educ Today 2023 Oct;129:105916. [doi: 10.1016/j.nedt.2023.105916] [Medline: 37515957]
27. Chan MMK, Wong ISF, Yau SY, Lam VSF. Critical reflection on using ChatGPT in student learning: benefits or potential risks? Nurse Educ 2023;48(6):E200-E201. [doi: 10.1097/NNE.0000000000001476] [Medline: 37348135]
28. Choi EPH, Lee JJ, Ho MH, Kwok JYY, Lok KYW. Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. Nurse Educ Today 2023 Jun;125:105796. [doi: 10.1016/j.nedt.2023.105796] [Medline: 36934624]
29. da Silva JAT. Is ChatGPT a valid author? Nurse Educ Pract 2023 Mar. [doi: 10.1016/j.nepr.2023.103600]
30. Draganic K. Artificial intelligence: opportunities and challenges in NP education. Nurse Pract 2023 Apr 1;48(4):6. [doi: 10.1097/01.NPR.0000000000000023] [Medline: 36975742]
31. Lim F. Machine-generated writing and chatbots: nursing education’s fear of the unknown. Nurs Educ Perspect 2023;44(4):203-204. [doi: 10.1097/01.NEP.0000000000001147] [Medline: 37417856]
32. Frith KH. ChatGPT: disruptive educational technology. Nurs Educ Perspect 2023;44(3):198-199. [doi: 10.1097/01.NEP.0000000000001129] [Medline: 37093697]
33. Gunawan J. Exploring the future of nursing: insights from the ChatGPT model. Belitung Nurs J 2023;9(1):1-5. [doi: 10.33546/bnj.2551] [Medline: 37469634]
34. Kleebayoon A, Wiwanitkit V. ChatGPT and the teaching of contemporary nursing: comment. J Clin Nurs 2023 Oct;32(19-20):37194403. [doi: 10.1111/jocn.16762] [Medline: 37194403]
35. Koo M. Harnessing the potential of chatbots in education: the need for guidelines to their ethical use. Nurse Educ Pract 2023 Mar;68:103590. [doi: 10.1016/j.nepr.2023.103590] [Medline: 36870226]
36. O’Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? Nurse Educ Pract 2023 Jan;66:103537. [doi: 10.1016/j.nepr.2022.103537] [Medline: 36549229]
37. Odom-Forren J. The role of ChatGPT in perianesthesia nursing. J Perianesth Nurs 2023 Apr;38(2):176-177. [doi: 10.1016/j.jopan.2023.02.006] [Medline: 36965923]
38. Scerri A, Morin KH. Using chatbots like ChatGPT to support nursing practice. J Clin Nurs 2023 Aug;32(15-16):4211-4213. [doi: 10.1111/jocn.16677] [Medline: 36880216]
39. Siegerink B, Pet LA, Rosendaal FR, Schoones JW. ChatGPT as an author of academic papers is wrong and highlights the concepts of accountability and contributorship. Nurse Educ Pract 2023 Mar;68:103599. [doi: 10.1016/j.nepr.2023.103599] [Medline: 36898252]
40. Sun GH, Hoelscher SH. The ChatGPT storm and what faculty can do. Nurse Educ 2023;48(3):119-124. [doi: 10.1097/NNE.0000000000001390] [Medline: 37043716]
41. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation study. JMIR Nurs 2023 Jun 27;6:e47305. [doi: 10.2196/47305] [Medline: 37368470]
42. Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. Nursing education in the age of artificial intelligence powered chatbots (AI-chatbots): are we ready yet? Nurse Educ Today 2023 Oct;129:105917. [doi: 10.1016/j.nedt.2023.105917] [Medline: 37506622]
43. Thakur A, Parikh D, Thakur A. ChatGPT in nursing education: is there a role for curriculum development? Teach Learn Nurs 2023 Jul;18(3):450-451. [doi: 10.1016/j.teln.2023.03.011]
44. Vitorino LM, Júnior GHY. ChatGPT and the teaching of contemporary nursing: and now professor? J Clin Nurs 2023 Nov;32(21-22):7921-7922. [doi: 10.1111/jocn.16706] [Medline: 37004198]
45. Woodnutt S, Allen C, Snowden J, et al. Could artificial intelligence write mental health nursing care plans? J Psychiatr Ment Health Nurs 2024 Feb;31(1):79-86. [doi: 10.1111/jpm.12965] [Medline: 37538021]
46. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. Med Educ. Medical Education. Preprint posted online on 2023. [doi: 10.1101/2023.07.09.23292415]

XSL•FO
RenderX

47. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607. [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
48. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. BMC Med Educ 2024 May 15;24(1):536. [doi: 10.1186/s12909-024-05534-8] [Medline: 38750546]
49. Rahman MM, Watanobe Y. ChatGPT for education and research: opportunities, threats, and strategies. Appl Sci (Basel) 2023;13(9):5783. [doi: 10.3390/app13095783]
50. Honavar SG. Eye of the AI storm: exploring the impact of AI tools in ophthalmology. Indian J Ophthalmol 2023;71(6):2328-2340. [doi: 10.4103/IJO.IJO_1478_23] [Medline: 37322638]
51. Currie GM. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? Semin Nucl Med 2023 Sep;53(5):719-730. [doi: 10.1053/j.semnuclmed.2023.04.008] [Medline: 37225599]
52. Susnjak T. ChatGPT: the end of online exam integrity? arXiv. Preprint posted online on Dec 19, 2022. [doi: 10.48550/arXiv.2212.09292]
53. Vaccino-Salvadore S. Exploring the ethical dimensions of using ChatGPT in language learning and beyond. Languages 2023;8(3):191. [doi: 10.3390/languages8030191]
54. Kuroiwa T, Sarcon A, Ibara T, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023 Sep 15;25:e47621. [doi: 10.2196/47621] [Medline: 37713254]
55. White D, Katsuno H. Cultural anthropology for social emotion modeling: principles of application toward diversified social signal processing. 2019 Presented at: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW); Cambridge, United Kingdom.
56. Chen D, Parsa R, Hope A, et al. Physician and artificial intelligence chatbot responses to cancer questions from social media. JAMA Oncol 2024 Jul 1;10(7):956-960. [doi: 10.1001/jamaoncol.2024.0836] [Medline: 38753317]
57. Benichou L, ChatGPT. The role of using ChatGPT AI in writing medical scientific articles. J Stomatol Oral Maxillofac Surg 2023 Oct;124(5):101456. [doi: 10.1016/j.jormas.2023.101456] [Medline: 36966950]
58. Curtis N, ChatGPT. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. Pediatr Infect Dis J 2023 Apr 1;42(4):275. [doi: 10.1097/INF.0000000000003852] [Medline: 36757192]
59. Palagani D, Counter P, James T. ChatGPT-generated literature review: quod erat demonstrandum or ends justifying the means? Clin Otolaryngol 2023 Nov;48(6):929-930. [doi: 10.1111/coa.14097] [Medline: 37673422]
60. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. J Multidiscip Healthc 2023;16:1513-1520. [doi: 10.2147/JMDH.S413470] [Medline: 37274428]

Abbreviations

AI: artificial intelligence
LLM: large language model
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by B Lesselroth; submitted 04.11.23; peer-reviewed by A Nes, QW Wu, SA Rasbi, S Ahmed, X Wu; revised version received 25.07.24; accepted 19.08.24; published 19.11.24.

Please cite as:
Zhou Y, Li SJ, Tang XY, He YC, Ma HM, Wang AQ, Pei RY, Piao MH
Using ChatGPT in Nursing: Scoping Review of Current Opinions
JMIR Med Educ 2024;10:e54297
URL: https://mededu.jmir.org/2024/1/e54297
doi:10.2196/54297

© You Zhou, Si-jia Li, Xing-Yi Tang, Yi-Chen He, Hao-Ming Ma, Ao-Qi Wang, Run-Yuan Pei, Mei-Hua Piao. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.11.2024. 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.

XSL•FO
RenderX

The Scope of Virtual Reality Simulators in Radiology Education: Systematic Literature Review

Shishir Shetty[1], PhD; Supriya Bhat[2], MDS; Saad Al Bayatti[1], MSc; Sausan Al Kawas[1], PhD; Wael Talaat[1], PhD; Mohamed El-Kishawi[3], PhD; Natheer Al Rawi[1], PhD; Sangeetha Narasimhan[1], PhD; Hiba Al-Daghestani[1], MSc; Medhini Madi[4], MDS; Raghavendra Shetty[5], PhD

1

2

3

4

5

Corresponding Author:
Supriya Bhat, MDS

Abstract

Background: In recent years, virtual reality (VR) has gained significant importance in medical education. Radiology education also has seen the induction of VR technology. However, there is no comprehensive review in this specific area. This review aims to fill this knowledge gap.

Objective: This systematic literature review aims to explore the scope of VR use in radiology education.

Methods: A literature search was carried out using PubMed, Scopus, ScienceDirect, and Google Scholar for articles relating to the use of VR in radiology education, published from database inception to September 1, 2023. The identified articles were then subjected to a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–defined study selection process.

Results: The database search identified 2503 nonduplicate articles. After PRISMA screening, 17 were included in the review for analysis, of which 3 (18%) were randomized controlled trials, 7 (41%) were randomized experimental trials, and 7 (41%) were cross-sectional studies. Of the 10 randomized trials, 3 (30%) had a low risk of bias, 5 (50%) showed some concerns, and 2 (20%) had a high risk of bias. Among the 7 cross-sectional studies, 2 (29%) scored "good" in the overall quality and the remaining 5 (71%) scored "fair." VR was found to be significantly more effective than traditional methods of teaching in improving the radiographic and radiologic skills of students. The use of VR systems was found to improve the students' skills in overall proficiency, patient positioning, equipment knowledge, equipment handling, and radiographic techniques. Student feedback was also reported in the included studies. The students generally provided positive feedback about the utility, ease of use, and satisfaction of VR systems, as well as their perceived positive impact on skill and knowledge acquisition.

Conclusions: The evidence from this review shows that the use of VR had significant benefit for students in various aspects of radiology education. However, the variable nature of the studies included in the review reduces the scope for a comprehensive recommendation of VR use in radiology education.

(JMIR Med Educ 2024;10:e52953) doi:10.2196/52953

KEYWORDS

virtual reality; simulators; radiology education; medical imaging; radiology; education; systematic review; literature review; imaging; meta analysis; student; students; VR; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

The use of technology in education helps students achieve improved acquisition of professional knowledge and practical skills [1-3]. Virtual reality (VR) is a modern technology that simulates experience by producing 3D interactive situations and presenting objects in a virtual world with spatial dimensions [4,5]. VR technology can be classified as nonimmersive or immersive [6]. In a nonimmersive VR, the simulated 3D environment is experienced through a computer monitor [6]. On the other hand, an immersive VR provides a sense of presence in a computer-generated environment, created by producing realistic sights, sounds, and other sensations that replicate a user's physical presence in a virtual environment [6,7]. Using VR technology, a person can look about the artificial world, navigate around in it, and interact with simulated objects or items [5,8]. Due to the broad nature of VR technology,

XSL•FO
RenderX

it has many applications, some of which are in the field of medicine [9,10].

The use of VR in medicine started in the 1990s when medical researchers were trying to create 3D models of patients' internal organs [11-13]. Since then, VR use in the field of medicine and general health care has increased substantially to cover many areas including medical education. Radiology education has also come to see the use of VR technology in the recent past [14]. The use of VR in radiology education enables students to practice radiography in a virtual environment, which is radiation free [15]. Additionally, the use of VR enables effective and repeatable training. This allows trainees to recognize and correct errors as they occur [16,17]. The aim of this review is to explore the scope of VR in radiology education.

Methods

This systematic review has been performed using the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) guidelines [18] [Checklist 1]).

Information Sources and Study Selection

The bibliographic databases used were PubMed, Scopus, ScienceDirect, and Google Scholar. A systematic literature search was conducted for articles published from database inception to September 1, 2023. Topic keywords were used to generate search strings. The search strings that were used are provided in Table 1. Only the first 10 pages of Google Scholar results were exported. The identified studies were then subjected to a study selection process. The search string for ScienceDirect was shorter because the database only allows a maximum of 8 Boolean operators, hence the sting had to be shortened. The search in PubMed was limited to the title and abstract. The searches in Scopus and ScienceDirect were limited to title, abstract, and keywords.

Table . Search strings used in the systematic review.

Database	Search string
PubMed and Scopus	("virtual reality" OR "immersive reality" OR "simulated reality" OR simulator OR simulate) AND (radiology OR radiography OR imaging OR radiologist) AND (education OR teaching)
ScienceDirect and Google Scholar	("virtual reality" OR "immersive reality" OR "simulated reality" OR simulator) AND (radiology OR radiography OR imaging) AND (education OR teaching)

Inclusion and Exclusion Criteria

Original research articles written in the English language were included in the review. Studies conducted on medical, dental, and allied health sciences students (undergraduate and postgraduate) from any part of the world were included in the review. Studies exploring the use of VR learning in radiology education were included.

Narrative reviews, scoping reviews, systematic reviews, meta-analyses, editorials, and commentaries were excluded. Studies that did not align with the required study objective were excluded.

Method of Quality Assessment

Randomized controlled trials (RCTs) and randomized experimental studies were appraised using the RoB 2 tool from the Cochrane Collaboration [19]. A visualization of the risk-of-bias assessment was done using the web-based *robvis* tool [20]. Cross-sectional studies were appraised using the appraisal checklist for analytical cross-sectional studies from the Joanna Briggs Institute [21].

Data Extraction

Each article included in the review was summarized in a table, including basic study characteristics. The extracted attributes were study author(s), publication year, study design, type and number of participants, type of radiology education under study, and the outcome being assessed. The extracted data are provided in Table 2.

XSL•FO
RenderX

Table . Data extraction table of the studies included in the systematic review.

Study	Study design	Participants	Aspect of radiology	Study outcome
Ahlqvist et al [22]	RCT[a]	31 first-year radiologic technology student	Diagnostic radiology	Assessing radiographic image quality
Bridge et al [23]	Randomized experimental trial	48 medical imaging students	General radiology	Student satisfaction and technical skills (ie, patient positioning, equipment positioning, and mean proficiency)
Gunn et al [24]	Randomized experimental trial	45 medical imaging student	Diagnostic radiology	Technical radiographic skills
Gunn et al [25]	Cross-sectional study	28 medical imaging students and 38 radiation therapy students	Interventional radiology	Students' perceived confidence in performing diagnostic and planning CT[b] scans
Jensen et al [26]	Cross-sectional study	10 radiography students	General radiology	Self-perceived clinical readiness of radiography students regarding the acquisition of wrist radiographs
Kato et al [27]	Randomized experimental trial	30 first-year radiologic technology student	General radiology	Radiographic skills proficiency
Nilsson et al [28]	Randomized experimental trial	57 dental students	Oral radiology	Interpretation of spatial relations in radiographs using parallax
Nilsson et al [29]	Randomized experimental trial	45 dental students	Oral radiology	Interpretation of spatial relations in radiographs using parallax
O'Connor and Rainford [30]	Randomized experimental trial	191 radiography students	General radiology	Patient preparation, room preparation, patient care, radiographic technique, and image appraisal
O'Connor et al [15]	Cross-sectional study	105 first-year radiography students	General radiology	Reporting student experience
Rainford et al [31]	Cross-sectional study	35 radiography students and 100 medical students	Interventional radiology	Reporting student experience
Rowe et al [32]	Randomized experimental trial	188 radiography students	General radiology	Technical skills (ie, duration of the exam, frequency of machinery movement, frequency of incorrect machinery movement, frequency of radiographic exposure errors, and frequency of patient positioning errors)
Sapkaroski et al [33]	Cross-sectional study	92 medical radiation science students	General radiology	Reporting student experience
Sapkaroski et al [34]	RCT	76 first-year radiography students	Radiation technology	Patient positioning
Sapkaroski et al [35]	RCT	76 radiography students	General radiology	Students' perception about developing radiographic hand positioning skills.
Shanahan [36]	Cross-sectional study	86 first-year radiography students	General radiology	Reporting student perception
Wu et al [37]	Cross-sectional study	18 medical students	General radiology	Reporting student perception

[a]RCT: randomized controlled trial.

[b]CT: computed tomography.

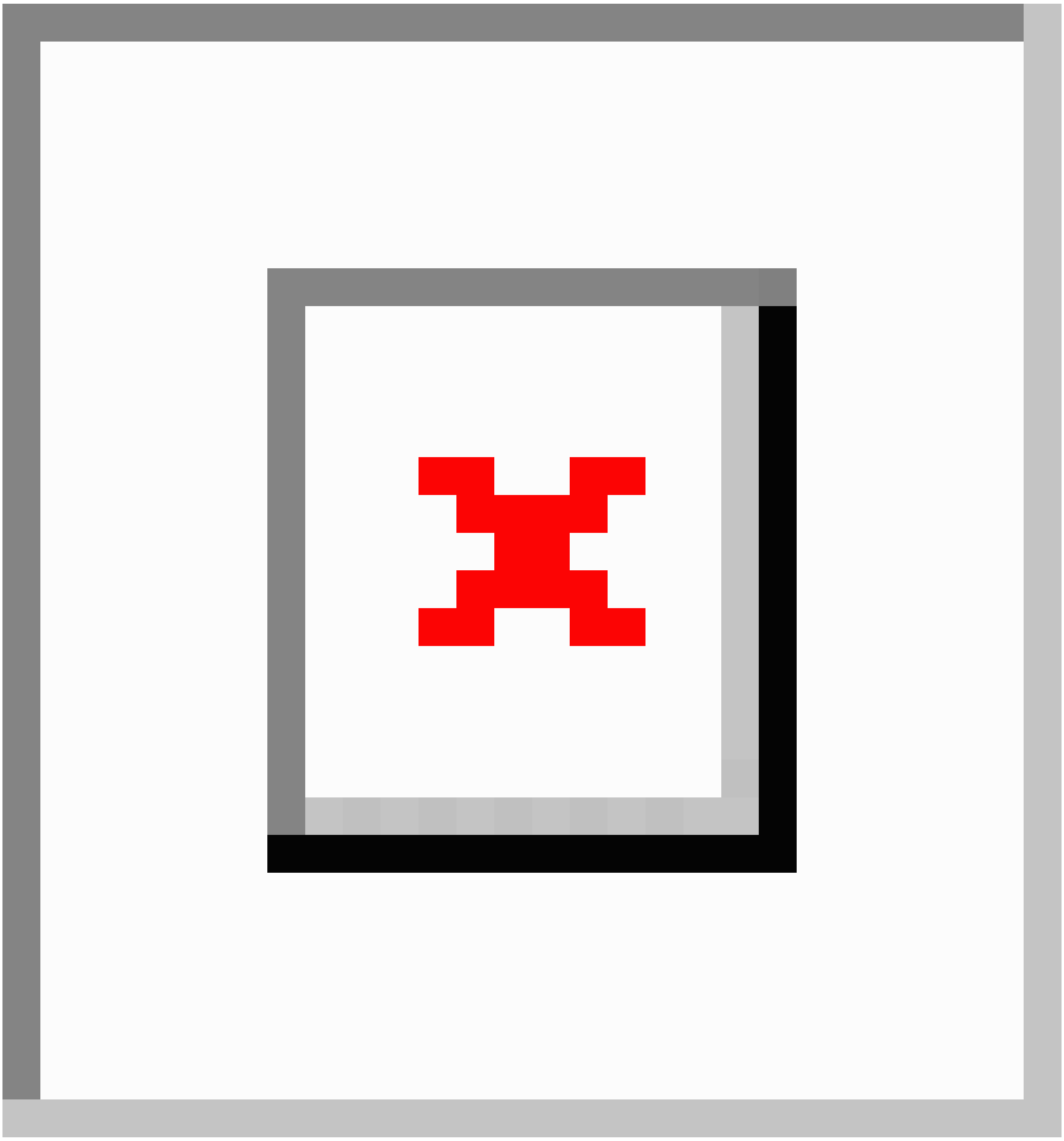
XSL•FO
RenderX

Results

Search Results

The database search identified a total of 2877 studies; 374 (13%) studies were from PubMed, 2169 (75.4%) were from Scopus, 234 (8.1%) were from ScienceDirect, and 100 (3.5%) were from Google Scholar. Before the screening procedure, 37 duplicates were removed. During title and abstract screening, 2808 articles were excluded since they did not align with the eligibility criteria. The remaining 32 articles were then subjected to a full-text review, and 15 were excluded for reasons provided in Figure 1, which shows the study selection process [38]. At the end of the process, 17 studies were found eligible for inclusion in the review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the study selection process.



Characteristics of Included Studies

Among the 17 studies, 3 (18%) RCTs, 7 (41%) randomized experimental trials, and 7 (41%) cross-sectional studies were included. The studies encompassed various aspects of radiology education, including dental radiology [28,29], diagnostic radiology [22,24], and interventional radiology [25,31].

Results of Quality Assessment

Among the 7 cross-sectional studies, 2 (29%) scored "good" in overall quality and the remaining 5 (71%) scored "fair." The

XSL•FO
RenderX

results for the quality appraisal of cross-sectional studies are shown in Table 3. Studies were appraised using the checklist for analytical cross-sectional studies from the Joanna Briggs Institute [21].

Among the 10 randomized trials, 3 (30%) had a low risk of bias, 5 (50%) showed some concerns, and 2 (20%) had a high risk of bias. These results are shown in Table 4. RCTs were appraised using the RoB 2 tool from the Cochrane Collaboration [19]. A risk-of-bias graph (Figure 2) and a risk-of-bias summary (Figure 3) are also provided.

Table . Appraisal for cross-sectional studies included in the systematic review.

Study	Item 1[a]	Item 2[b]	Item 3[c]	Item 4[d]	Item 5[e]	Item 6[f]	Item 7[g]	Item 8[h]	Overall quality
Gunn et al [25]	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Good
Jensen et al [26]	Yes	Yes	Yes	No	No	N/A[i]	Yes	Yes	Fair
O'Connor et al [15]	Yes	Yes	Yes	Yes	No	N/A	Yes	No	Fair
Rainford et al [31]	Yes	Yes	Yes	Yes	No	N/A	Yes	Yes	Fair
Sapkaroski et al [33]	No	Yes	Yes	Unclear	No	N/A	Yes	Unclear	Fair
Shanahan [36]	No	Yes	Unclear	Yes	Yes	Unclear	Yes	Yes	Good
Wu et al [37]	Yes	Yes	Yes	Yes	No	N/A	Yes	Yes	Fair

[a]Item 1: were the criteria for inclusion in the sample clearly defined?

[b]Item 2: were the study subjects and the setting described in detail?

[c]Item 3: was the exposure measured in a valid and reliable way?

[d]Item 4: were objective, standard criteria used for measurement of the condition?

[e]Item 5: were confounding factors identified?

[f]Item 6: were strategies to deal with confounding factors stated?

[g]Item 7: were the outcomes measured in a valid and reliable way?

[h]Item 8: was appropriate statistical analysis used?

[i]N/A: not assessable.

XSL•FO
RenderX

Table . Risk-of-bias assessment for randomized trials included in the systematic review.

Study	D1[a]	D2[b]	D3[c]	D4[d]	D5[e]	Overall
Ahlqvist et al [22]	Low	Low	High	Low	Some concerns	Some concerns
Bridge et al [23]	Low	Low	Some concerns	Low	Low	Some concerns
Gunn et al [24]	Low	Low	Low	Low	Low	Low
Kato et al [27]	Some concerns	Low	Low	Low	Low	Some concerns
Nilsson et al [28]	Some concerns	Low	Low	Low	Low	Low
Nilsson et al [29]	Low	Low	High	Low	Some concerns	High
O'Connor and Rainford [30]	High	Low	Low	Low	Some concerns	Some concerns
Rowe et al [32]	Low	High	Some concerns	Low	Low	Some concerns
Sapkaroski et al [34]	Low	Some concerns	Some concerns	High	Low	High
Sapkaroski et al [35]	Low	Low	Low	Low	Low	Low

[a]D1: risk of bias arising from the randomization process.

[b]D2: risk of bias due to deviations from the intended interventions (effect of assignment to intervention).

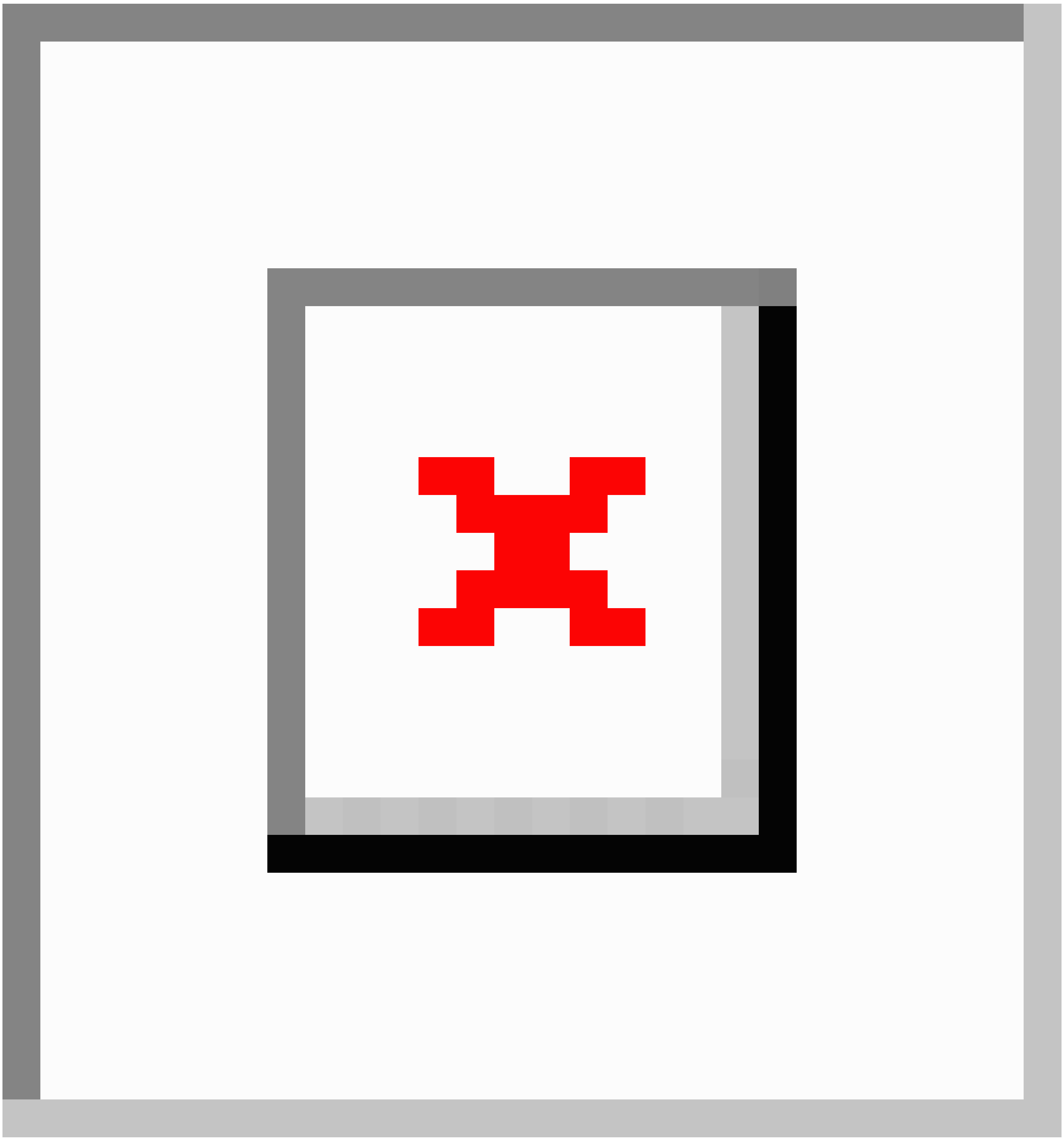
[c]D3: risk of bias due to missing outcome data.

[d]D4: risk of bias in measurement of the outcome.

[e]D5: risk of bias in selection of the reported result.

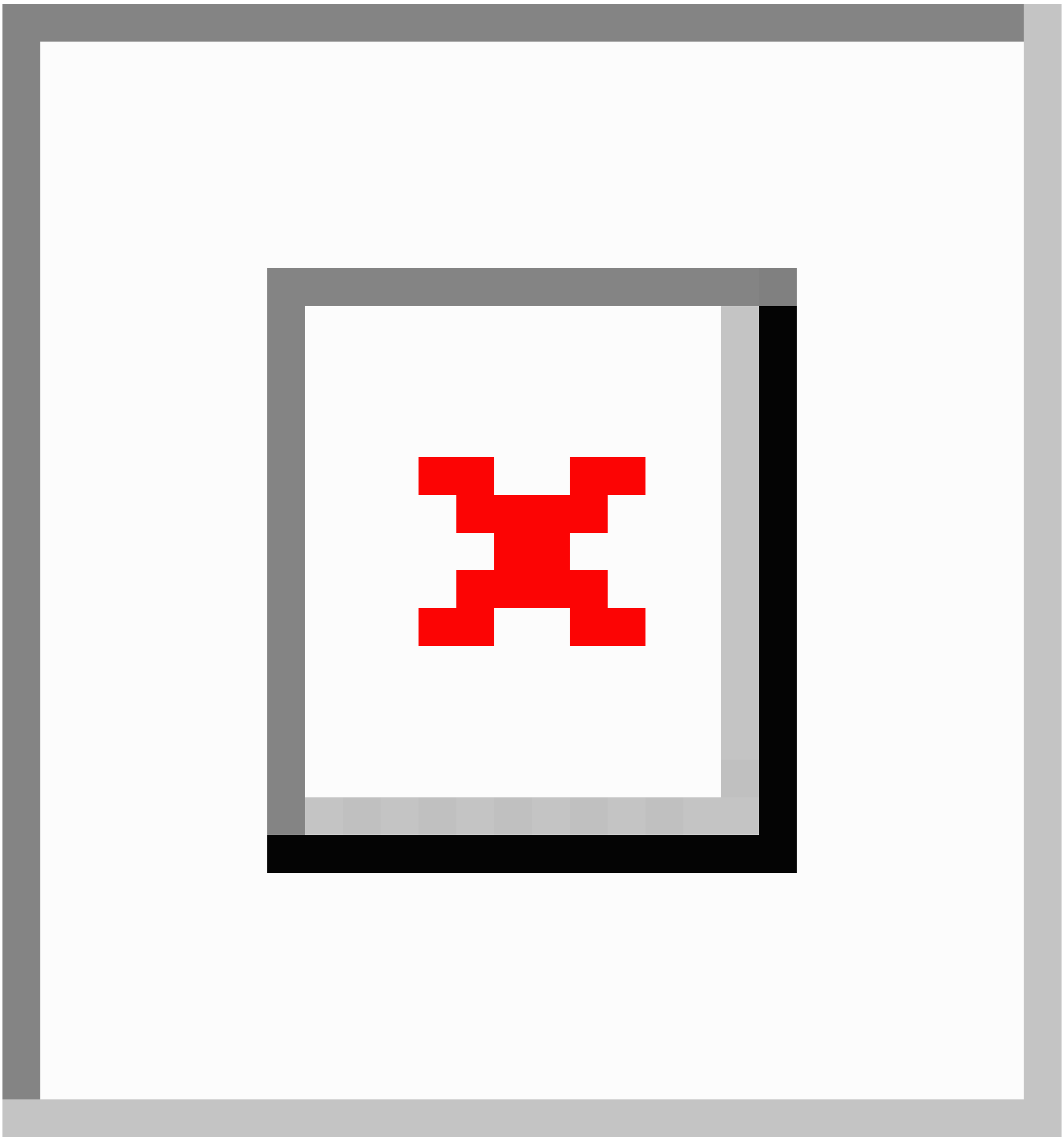
XSL•FO
RenderX

Figure 2. Risk-of-bias graph using a traffic light plot for different domains (D1 to D5) [22-24,27-30,32,34,35].



XSL•FO
RenderX

Figure 3. Weighted bar plots displaying the distribution of risk-of-bias judgments within each bias domain.



Type of VR Hardware and Software Used in the Studies

The studies used a wide range of VR software and hardware. Some of the studies used 3D simulation software packages displayed on 2D desktop computers [22,24,25,36], whereas others used headsets for an immersive VR environment [15,23,26,35,37]. The most used VR teaching software were the CETSOL VR Clinic software [33,35], Virtual Medical Coaching VR software [15,30,32], Projection VR (Shaderware) software [36], SieVRt VR system (Luxsonic Technologies) [37], medical imaging training immersive environment software [23], VR CT Sim software [25], VitaSim ApS software [26], VR X-Ray (Skilitics and Virtual Medical Coaching) software [27], and radiation dosimetry VR software (Virtual Medical Coaching Ltd) [31].

Effect of VR Teaching on Skill Acquisition

Ahlqvist et al [22] looked at how virtual simulation can be used as an effective tool to teach quality assessment of radiographic images. They also compared how it faired in comparison to traditional teaching. The study reported a statistically significant improvement in proficiency from before training to after training. Additionally, the study reported that the proficiency score improvement for the VR-trained students was higher than that for the students trained using conventional method.

In the study conducted by Sapkaroski et al [34], students in the VR group demonstrated significantly better patient positioning

XSL•FO
RenderX

skills compared to those in the conventional role-play group. The positioning parameters that were assessed were digit separation and palm flatness (the VR group scored 11% better), central ray positioning onto the third metacarpophalangeal joint (the VR group scored 23% better), and a control position projection of an oblique hand. The results for the control position projection indicated no significant difference in positioning between the 2 groups [34].

Bridge et al [23] also performed a performance comparison between students trained by VR and traditional methods. They assessed skills about patient positioning, equipment positioning, and time taken to complete a performative role-play. Students in the VR group performed better than those in the control group, with 91% of them receiving an overall score of above average (>3). The difference in mean group performance was statistically significant (P=.0366). Similarly, Gunn et al [24] reported improved and higher role-play skill scores for students trained using VR software simulation compared to those trained on traditional laboratory simulation. The mean role-play score for the VR group was 30.67 and that for the control group was 28.8 [24].

Another study reported that students trained using VR performed significantly better (ranked as "very good" or "excellent") than the control group (conventional learning) in skills such as patient positioning, selecting exposure factors, centering and collimating the x-ray beam, placing the anatomical marker, appraisal of image quality, equipment positioning, and procedure explanation to the patient [30]. Another recently conducted study found that the VR-taught group achieved better test duration and fewer errors in moving equipment and positioning a patient. There was no significant difference in the frequency of errors in the radiographic exposure setting such as source-to-image distance between the VR and the physical simulation groups [32].

Nilsson et al [28] developed a test to evaluate the student's ability to interpret 3D information in radiographs using parallax. This test was applied to students before and after training. There was a significantly larger (P<.01) pre-post intervention mean score for the VR group (3.11 to 4.18) compared to the control group (3.24 to 3.72). A subgroup analysis was also performed, and students with low visuospatial ability in the VR group had a significantly higher improvement in the proficiency test compared to those in the control group. The same authors conducted another follow-up study to test skill retention [29]. Net skill improvement was calculated as the difference in test scores after 8 months. The results from the proficiency test showed that the ability to interpret spatial relations in radiographs 8 months after the completion of VR training was significantly better than before VR training. The students who trained conventionally showed almost the same positive trend in improvement. The group difference was smaller and not statistically significant. This meant that, 8 months after training, the VR group and the traditionally trained group had the same skill level [29].

Among the included studies, only 1 reported that the VR group had lower performance in proficiency tests and radiographic skill tests, compared to a conventionally trained group. The study, conducted in 2022, showed that the proficiency of the VR group was significantly lower than that of the conventional technique group in performing lateral elbow and posterior-anterior chest radiography [27]. An itemized rubric evaluation used in the study revealed that the VR group also had lower performance in most of the radiographic skills, such as locating and centering of the x-ray beam, side marker placement, positioning the x-ray image detector, patient interaction, and process control and safety [27]. The study concluded that VR simulation can be less effective than real-world training in radiographic techniques, which requires palpation and patient interaction. These results may be different from those of other studies due to different outcome evaluation methods and since they used head-mounted display VR coaching, whereas the other studies, except O'Connor et al [15], used VR on a PC monitor.

All of the studies except Kato et al [27] agreed that VR use was more effective for students in developing radiographic and radiologic skills. Despite this general agreement, there were slight in-study variations in learning outcomes, which made some of the studies look at factors that may influence skill and knowledge acquisition during VR use. In studies such as Bridge et al [23], it was noted that the arrangement of equipment had the greatest influence on the overall score. After performing a multivariable analysis, Gunn et al [24] reported that there was no effect of age, gender, and gaming skills or activity on the outcome of VR learning. In the study by Shanahan [36], a few students (19/84, 23%) had previously used VR simulation software. This had no bearing on the learning outcomes. Another observation in the same study was that student age was found to significantly affected the student's confidence about skill acquisition after VR training [36].

Students' Perception of VR Uses for Learning

The findings from the study by Gunn et al [25] revealed that 68% of students agreed or strongly agreed that VR simulation was significantly helpful in learning about computed tomography (CT) scanning. In another study by Jensen et al [26], 90% of the students strongly agreed that VR simulators could contribute to learning radiography, with 90% reporting that the x-ray equipment in the VR simulation was realistic. In the study by Wu et al [37], most of the students (55.6%) agreed or somewhat agreed that VR use was useful in radiology education. Similarly, 83% of the students in Shanahan's [36] study regarded VR learning with an ease of use. In the same study, students also reported that one of the major benefits of VR learning include using the simulation to repeat activities until being satisfied with the results (95% of respondents). Students also stated that VR enabled them to quickly see images and understand if changes needed to be made (94%) [36]. In the study by Gunn et al [25], 75% of medical imaging students agreed on the ease of use and software enjoyment in VR simulated learning. In the same study, 57% of the students reported a positive perceived usefulness of VR. Most respondents (80%) in the study by Rainford et al [31] favored the in-person VR experience over web-based VR. Similarly, 58% of the respondents in the study conducted by O'Connor et al [15] reported enjoying learning using VR simulation. In the study by Wu et al [37], 83.3% of students agreed or strongly agreed that they enjoyed using VR for learning. Similarly, the

XSL•FO
RenderX

studies by Rainford et al [31] and O'Connor et al [15] reported student recommendation of 87% and 94%, respectively, for VR as a learning tool.

Students' Perceived Skill and Knowledge Acquisition

In the study by Bridge et al [23], students who trained using VR reported an increase in perceived skill acquisition and high levels of satisfaction. The study authors attributed this feedback to the availability of "gold standards" that showed correct positioning techniques, as well as instant feedback provided by the VR simulators. Gunn et al [25] examined students' confidence in performing a CT scan in a real clinical environment after using VR simulations as a learning tool. The study reported an increase (from before to after training) in the students' perceived confidence in performing diagnostic CT scans. Similarly, the study by Jensen et al [26] reported that the use of VR had influenced students' self-perceived readiness to perform wrist x-ray radiographs. The study, however, found no significant difference in pre- and posttraining (perceived preparedness) scores. The pre- and posttraining scores were 75 (95% CI 54-96) and 77 (95% CI 59-95), respectively. The study by O'Connor et al [15] looked at the effect of VR on perceived skill adoption. Most of the students in the study reported high levels of perceived knowledge acquisition in the areas of beam collimation, anatomical marker placement, centering of the x-ray tube, image evaluation, anatomical knowledge, patient positioning, and exposure parameter selection to their VR practice. However, most students felt that VR did not contribute to their knowledge of patient dose tracking and radiation safety [15]. In the study by Rainford et al [31], 73% of radiography and medical students felt that VR learning increased their confidence across all relevant learning outcomes. The biggest increase in confidence level was regarding their understanding of radiation safety matters [31]. Sapkaroski et al [33] performed a self-perception test to see how students viewed their clinical and technical skills after using VR for learning. In their study, students reported a perceived improvement in their hand and patient positioning skills. Their study also compared 2 software, CETSOL VR Clinic and Shaderware. The cohort who used CETSOL VR Clinic had higher scores on perceived improvement [33]. Sapkaroski et al [35] compared the student's perception scores on the educational enhancement of their radiographic hand positioning skills, after VR or clinical role-play scenario training. Although the VR group scored higher, there was no significant difference between the scores for the 2 groups [35]. In the study by Shanahan [36], when the perception of skill development was evaluated, most of the students reported that the simulation positively developed their technical (78%), radiographic image evaluation (85%), problem-solving (85%), and self-evaluation (88%) abilities. However, in the study by Kato et al [27], there was no difference in the perceived acquisition of knowledge among students using traditional teaching and VR-based teaching.

Discussion

Principal Findings

The results presented in this review reveal strong evidence for the effectiveness of VR teaching in radiology education, particularly in the context of skill acquisition and development [22,24,27,30,32,34].

In this review, quality appraisal of the cross-sectional studies revealed that the strategies for deal with confounding factors was one of the factors directly affecting the reliability of the results. Similarly, the appraisal of the randomized trials revealed that the bias arising due to missing outcome data was one of the factors directly affecting the reliability of the results.

All the studies found that VR-based teaching had a positive impact on various areas of radiographic and radiologic skill development. In comparison to the traditional way of teaching, only 1 study by Kato et al [27] reported VR teaching as inferior to traditional teaching. The studies consistently reported better improvements in proficiency, patient positioning outcomes, equipment handling, and radiographic techniques among students trained using VR. According to Nilsson et al [29], O'Connor et al [15], and Wu et al [37], the improvements were due to the immersive and interactive nature of VR simulations, which allowed learners to engage with radiological scenarios in a dynamic and hands-on manner. The studies also revealed that VR learning has the ability to easily and effectively introduce students to new skills. It was also found that existing skills could be improved, mainly through simulation feedback that happens in real time during training [22,24,28,30,36].

The improvement of skills after VR training have been noted in different domains, including patient positioning, equipment positioning, equipment knowledge, assessment of radiographic image quality, and patient interaction. Improvement was also observed in other skills such as as central ray positioning, source-to-image distance, image receptor placement, and side marker placement [22,24,30,32,34]. Two studies, Nilsson et al [28] and Nilsson et al [29], looked at how VR affected the students' ability to interpret 3D information in radiographs using parallax. They both reported a positive effect. Nilsson et al [29] also gave insights into the long-term benefits of VR training in radiology. Eight months after training, the control (traditionally taught) group in Nilsson et al [29] showed a slight increase in skills, but the VR-trained group still maintained a significantly higher skill level. This finding shows the enduring impact of VR-based education on skill acquisition in radiology. Although most studies supported the effectiveness of VR in radiology education, 1 study reported contrasting results [27]. VR-trained students were found to perform worse than traditionally trained students in conducting lateral elbow and posterior-anterior chest radiography in Kato et al [27]. This difference in results was, according to the authors, attributed to the use of a different rubric evaluation method and the use of a head-mounted display–based immersive VR system, which was not used in other studies. These 2 reasons may be the reason for the variation in study findings.

A wide range of VR software with different functions were used in the studies. In addition to acquiring radiographic images, the CETSOL VR Clinic software facilitated students to interact with their learning environment [33,35]. Students using the Virtual Medical Coaching VR software performed imaging exercise on a virtual patient with VR headsets and hand controllers [15,30,32]. The SieVRt VR system displayed Digital

XSL•FO
RenderX

Imaging and Communications in Medicine format images in a virtual environment, thus facilitating teaching [37]. The medical imaging training immersive environment simulation software provided automated feedback to the learners including a rerun of procedures, thus highlighting procedural errors [23]. The VR CT Sim software allowed the student virtually to perform the complete CT workflow [25]. Students could manipulate patient positioning and get feedback from the VitaSim ApS software [26]. The VR X-Ray software allowed students to manipulate radiographic equipment and patient's position with a high level of immersive experience [27]. Radiation dosimetry VR software facilitated virtual movement of the staff and equipment to radiation-free areas, thus optimizing radiation protection [31].

The included studies also looked at factors that could influence skill acquisition when VR is used in radiology education. Bridge et al [23], Gunn et al [24], Kato et al [27], and Shanahan [36] investigated factors such as age, gender, prior gaming experience, and familiarity with VR technology. However, these factors were shown to have no significant effect on VR learning outcomes. This shows that VR education can equally accommodate a wide range of learners, regardless of experience or existing attributes.

Across several studies, positive feedback emerged regarding the utility, ease of use, enjoyment, and perceived impact on skill and knowledge acquisition. The included studies consistently reported positive perceptions of VR use among students [25,26,37]. Gunn et al [25] reported that a significant proportion of medical imaging and radiation therapy students found the use of VR simulation to be significantly helpful in learning about CT scanning. Similarly, Jensen et al [26] and Wu et al [37] reported that a majority of students agreed on the usefulness of VR in radiology education. Another aspect that received positive feedback was the ease of use. Students liked the ability to repeat tasks until they were satisfied with the results and the ability to quickly visualize radiographs to determine the need for revisions [36]. Rainford et al [31] and O'Connor and Rainford [30] found that most students would recommend VR as a learning tool to other students.

Several studies investigated student's perceptions of skill and knowledge acquisition when using VR for radiology education. Bridge et al [15] and O'Connor et al [23] discovered an increase in students' perceived acquisition of radiographic skills. Gunn et al [25] reported an increase in students' perceived confidence to perform CT scans after learning using VR simulations. According to Rainford et al [31], a large percentage of radiography and medical students felt that VR learning boosted their confidence across all relevant learning outcomes, with the highest levels of confidence recorded in radiation safety. Sapkaroski et al [33] discovered that after using VR for learning, students experienced an improvement in their hand and patient placement skills. In summary, the positive feedback from the students shows that VR use in radiology education is a useful, engaging, and effective teaching tool. This perceived acquisition of skills is backed by the results from the proficiency tests.

The VR modalities used in some of the studies allowed remote assistance from an external agent (teacher), as the VR training is conducted in front of a screen while being part of a team, with the teacher making constant corrections and indications [22,24,27]. However, researchers are looking into VR systems with artificial intelligence–supported tutoring, which includes the assessment of learners, generation of learning content, and automated feedback [39].

Conclusion

Findings from the included studies show that VR-based teaching offers substantial benefits in various aspects of radiographic and radiologic skill development. The studies consistently reported that students educated using VR systems improved significantly in overall proficiency, patient positioning, equipment knowledge, equipment handling, and radiographic techniques. However, the variable nature of the studies included in the review reduces the scope for a comprehensive recommendation of VR use in radiology education. A key contributing factor to relatively better learning outcomes was the immersive and interactive nature of VR systems, which provided real-time feedback and dynamic learning experiences to students. Factors such as age, gender, gaming experience, and familiarity with VR systems did not significantly influence learning outcomes. This shows that VR can be used for diverse groups of students when teaching radiology. Students generally provided positive feedback about the utility, ease of use, and satisfaction of VR, as well as its perceived impact on skill and knowledge acquisition. These students' reports show the value of VR as an important, interesting, and effective tool in radiology education.

Conflicts of Interest

None declared.

Checklist 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.

[PDF File, 465 KB - mededu_v10i1e52953_app1.pdf]

References

1. Ding X, Li Z. A review of the application of virtual reality technology in higher education based on Web of Science literature data as an example. Front Educ 2022 Nov 17;7:1048816. [doi: 10.3389/feduc.2022.1048816]

XSL•FO
RenderX

2. Oyelere SS, Bouali N, Kaliisa R, Obaido G, Yunusa AA, Jimoh ER. Exploring the trends of educational virtual reality games: a systematic review of empirical studies. Smart Learning Environments 2020 Oct 19;7:31. [doi: 10.1186/s40561-020-00142-7]
3. Zamyatina NV, Ushakova IA, Mandrikov VB. Medical education in digital transformation. In: Proceedings of the International Scientific Conference “Digitalization of Education: History, Trends and Prospects” (DETP 2020): Atlantis Press; 2020:324-327. [doi: 10.2991/assehr.k.200509.059]
4. Kardong-Edgren S, Farra SL, Alinier G, Young HM. A call to unify definitions of virtual reality. Clin Simul Nurs 2019 Jun;31:28-34. [doi: 10.1016/j.ecns.2019.02.006]
5. Sousa Santos B, Dias P, Pimentel A, et al. Head-mounted display versus desktop for 3D navigation in virtual reality: a user study. Multimed Tools Appl 2009 Jan;41(1):161-181. [doi: 10.1007/s11042-008-0223-2]
6. Hamad A, Jia B. How virtual reality technology has changed our lives: an overview of the current and potential applications and limitations. Int J Environ Res Public Health 2022 Sep 8;19(18):11278. [doi: 10.3390/ijerph191811278] [Medline: 36141551]
7. Kaplan-Rakowski R, Meseberg K. Immersive media and their future. In: Branch RM, Lee H, Tseng S, editors. Educational Media and Technology Yearbook: Springer; 2019, Vol. 42:143-153. [doi: 10.1007/978-3-030-27986-8_13]
8. Kuliga SF, Thrash T, Dalton RC, Hölscher C. Virtual reality as an empirical research tool — exploring user experience in a real building and a corresponding virtual model. Comput Environ Urban Syst 2015 Nov;54:363-375. [doi: 10.1016/j.compenvurbsys.2015.09.006]
9. Guan H, Xu Y, Zhao D. Application of virtual reality technology in clinical practice, teaching, and research in complementary and alternative medicine. Evid Based Complement Alternat Med 2022 Aug 11;2022:1373170. [doi: 10.1155/2022/1373170] [Medline: 35990836]
10. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct;6(3):181-185. [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
11. Satava RM, Jones SB. Current and future applications of virtual reality for medicine. Proc IEEE 1998 Mar;86(3):484-489. [doi: 10.1109/5.662873]
12. Satava RM. Medical virtual reality. the current status of the future. Stud Health Technol Iform 1996;29:100-106. [Medline: 10163742]
13. Rosenberg LB, Stredney D. A haptic interface for virtual simulation of endoscopic surgery. Stud Health Technol Inform 1996;29:371-387. [Medline: 10172846]
14. Chytas D, Salmas M, Demesticha T, et al. A review of the use of virtual reality for teaching radiology in conjunction with anatomy. Cureus 2021 Dec 5;13(12):e20174. [doi: 10.7759/cureus.20174] [Medline: 35004000]
15. O’Connor M, Stowe J, Potocnik J, Giannotti N, Murphy S, Rainford L. 3D virtual reality simulation in radiography education: the students' experience. Radiography (Lond) 2021 Feb;27(1):208-214. [doi: 10.1016/j.radi.2020.07.017] [Medline: 32800641]
16. Hart R, Karthigasu K. The benefits of virtual reality simulator training for laparoscopic surgery. Curr Opin Obstet Gynecol 2007 Aug;19(4):297-302. [doi: 10.1097/GCO.0b013e328216f5b7] [Medline: 17625408]
17. Thomas DJ, Singh D. Letter to the editor: virtual reality in surgical training. Int J Surg 2021 May;89:105935. [doi: 10.1016/j.ijsu.2021.105935] [Medline: 33819684]
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372(71):n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
19. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
20. McGuinness LA, Higgins JPT. Risk-of-bias visualization (robvis): an R package and shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
21. Critical appraisal tools. JBI. URL: https://jbi.global/critical-appraisal-tools [accessed 2024-04-25]
22. Ahlqvist JB, Nilsson TA, Hedman LR, et al. A randomized controlled trial on 2 simulation-based training methods in radiology: effects on radiologic technology student skill in assessing image quality. Simul Healthc 2013 Dec;8(6):382-387. [doi: 10.1097/SIH.0b013e3182a60a48] [Medline: 24096919]
23. Bridge P, Gunn T, Kastanis L, et al. The development and evaluation of a medical imaging training immersive environment. J Med Radiat Sci 2014 Sep;61(3):159-165. [doi: 10.1002/jmrs.60] [Medline: 26229652]
24. Gunn T, Jones L, Bridge P, Rowntree P, Nissen L. The use of virtual reality simulation to improve technical skill in the undergraduate medical imaging student. Interactive Learning Environments 2018;26(5):613-620. [doi: 10.1080/10494820.2017.1374981]
25. Gunn T, Rowntree P, Starkey D, Nissen L. The use of virtual reality computed tomography simulation within a medical imaging and a radiation therapy undergraduate programme. J Med Radiat Sci 2021 Mar;68(1):28-36. [doi: 10.1002/jmrs.436] [Medline: 33000561]
26. Jensen J, Graumann O, Jensen RO, et al. Using virtual reality simulation for training practical skills in musculoskeletal wrist x-ray - a pilot study. J Clin Imaging Sci 2023 Jul 11;13:20. [doi: 10.25259/JCIS_45_2023] [Medline: 37559875]

XSL•FO
RenderX

27. Kato K, Kon D, Ito T, Ichikawa S, Ueda K, Kuroda Y. Radiography education with VR using head mounted display: proficiency evaluation by rubric method. BMC Med Educ 2022 Jul 28;22(1):579. [doi: 10.1186/s12909-022-03645-8] [Medline: 35902953]
28. Nilsson TA, Hedman LR, Ahlqvist JB. A randomized trial of simulation-based versus conventional training of dental student skill at interpreting spatial information in radiographs. Simul Healthc 2007;2(3):164-169. [doi: 10.1097/SIH.0b013e31811ec254] [Medline: 19088619]
29. Nilsson TA, Hedman LR, Ahlqvist JB. Dental student skill retention eight months after simulator-supported training in oral radiology. J Dent Educ 2011 May;75(5):679-684. [Medline: 21546602]
30. O'Connor M, Rainford L. The impact of 3D virtual reality radiography practice on student performance in clinical practice. Radiography (Lond) 2023 Jan;29(1):159-164. [doi: 10.1016/j.radi.2022.10.033] [Medline: 36379142]
31. Rainford L, Tcacenco A, Potocnik J, et al. Student perceptions of the use of three-dimensional (3-D) virtual reality (VR) simulation in the delivery of radiation protection training for radiography and medical students. Radiography (Lond) 2023 Jul;29(4):777-785. [doi: 10.1016/j.radi.2023.05.009] [Medline: 37244141]
32. Rowe D, Garcia A, Rossi B. Comparison of virtual reality and physical simulation training in first-year radiography students in South America. J Med Radiat Sci 2023 Jun;70(2):120-126. [doi: 10.1002/jmrs.639] [Medline: 36502536]
33. Sapkaroski D, Baird M, McInerney J, Dimmock MR. The implementation of a haptic feedback virtual reality simulation clinic with dynamic patient interaction and communication for medical imaging students. J Med Radiat Sci 2018 Sep;65(3):218-225. [doi: 10.1002/jmrs.288] [Medline: 30006966]
34. Sapkaroski D, Baird M, Mundy M, Dimmock MR. Quantification of student radiographic patient positioning using an immersive virtual reality simulation. Simul Healthc 2019 Aug;14(4):258-263. [doi: 10.1097/SIH.0000000000000380] [Medline: 31274828]
35. Sapkaroski D, Mundy M, Dimmock MR. Virtual reality versus conventional clinical role-play for radiographic positioning training: a students' perception study. Radiography (Lond) 2020 Feb;26(1):57-62. [doi: 10.1016/j.radi.2019.08.001] [Medline: 31902456]
36. Shanahan M. Student perspective on using a virtual radiography simulation. Radiography (Lond) 2016 Aug;22(3):217-222. [doi: 10.1016/j.radi.2016.02.004]
37. Wu Y, Mondal P, Stewart M, Ngo R, Burbridge B. Bringing radiology education to a new reality: a pilot study of using virtual reality as a remote educational tool. Can Assoc Radiol J 2023 May;74(2):251-263. [doi: 10.1177/08465371221142515] [Medline: 36471627]
38. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022 Mar 27;18(2):e1230. [doi: 10.1002/cl2.1230] [Medline: 36911350]
39. King S, Boyer J, Bell T, Estapa A. An automated virtual reality training system for teacher-student interaction: a randomized controlled trial. JMIR Serious Games 2022 Dec 8;10(4):e41097. [doi: 10.2196/41097] [Medline: 36480248]

Abbreviations

CT: computed tomography
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
VR: virtual reality

Edited by A Hasan Sapci, TDA Cardoso; submitted 23.09.23; peer-reviewed by F Jorge, S Kassutto; revised version received 01.02.24; accepted 31.03.24; published 08.05.24.

Please cite as:
Shetty S, Bhat S, Al Bayatti S, Al Kawas S, Talaat W, El-Kishawi M, Al Rawi N, Narasimhan S, Al-Daghestani H, Madi M, Shetty R
The Scope of Virtual Reality Simulators in Radiology Education: Systematic Literature Review
JMIR Med Educ 2024;10:e52953
URL: https://mededu.jmir.org/2024/1/e52953
doi:10.2196/52953

© Shishir Shetty, Supriya Bhat, Saad Al Bayatti, Sausan Al Kawas, Wael Talaat, Mohamed El-Kishawi, Natheer Al Rawi, Sangeetha Narasimhan, Hiba Al-Daghestani, Medhini Madi, Raghavendra Shetty. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.5.2024. 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

XSL•FO RenderX

information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Inverted Classroom Teaching of Physiology in Basic Medical Education: Bibliometric Visual Analysis

Zonglin He[1,2,*], MBBS, MPhil; Botao Zhou[1,*], MBBS; Haixiao Feng[3,*], BA, MS; Jian Bai[1,4], BA; Yuechun Wang[1], PhD

1

2

3

4

[*]these authors contributed equally

Corresponding Author:
Yuechun Wang, PhD

Abstract

Background: Over the last decade, there has been growing interest in inverted classroom teaching (ICT) and its various forms within the education sector. Physiology is a core course that bridges basic and clinical medicine, and ICT in physiology has been sporadically practiced to different extents globally. However, students' and teachers' responses and feedback to ICT in physiology are diverse, and the effectiveness of a modified ICT model integrated into regular teaching practice in physiology courses is difficult to assess objectively and quantitatively.

Objective: This study aimed to explore the current status and development direction of ICT in physiology in basic medical education using bibliometric visual analysis of the related literature.

Methods: A bibliometric analysis of the ICT-related literature in physiology published between 2000 and 2023 was performed using CiteSpace, a bibliometric visualization tool, based on the Web of Science database. Moreover, an in-depth review was performed to summarize the application of ICT in physiology courses worldwide, along with identification of research hot spots and development trends.

Results: A total of 42 studies were included for this bibliometric analysis, with the year 2013 marking the commencement of the field. University staff and doctors working at affiliated hospitals represent the core authors of this field, with several research teams forming cooperative relationships and developing research networks. The development of ICT in physiology could be divided into several stages: the introduction stage (2013 - 2014), extensive practice stage (2015 - 2019), and modification and growth stage (2020 - 2023). Gopalan C is the author with the highest citation count of 5 cited publications and has published 14 relevant papers since 2016, with a significant surge from 2019 to 2022. Author collaboration is generally limited in this field, and most academic work has been conducted in independent teams, with minimal cross-team communication. Authors from the United States published the highest number of papers related to ICT in physiology (18 in total, accounting for over 43% of the total papers), and their intermediary centrality was 0.24, indicating strong connections both within the country and internationally. Chinese authors ranked second, publishing 8 papers in the field, although their intermediary centrality was only 0.02, suggesting limited international influence and lower overall research quality. The topics of ICT in physiology research have been multifaceted, covering active learning, autonomous learning, student performance, teaching effect, blended teaching, and others.

Conclusions: This bibliometric analysis and literature review provides a comprehensive overview of the history, development process, and future direction of the field of ICT in physiology. These findings can help to strengthen academic exchange and cooperation internationally, while promoting the diversification and effectiveness of ICT in physiology through building academic communities to jointly train emerging medical talents.

(JMIR Med Educ 2024;10:e52224) doi:10.2196/52224

KEYWORDS

flipped classroom; flipped classroom teaching; physiology; scientific knowledge map; hot topics; frontier progress; evolution trend; classroom-based; bibliometric visual analysis; bibliometric; visual analysis; medical education; teaching method; bibliometric analysis; visualization tool; academic; academic community; inverted classroom

XSL•FO
RenderX

Introduction

In recent decades, student-centered active learning strategies have been implemented in numerous educational institutions worldwide as an alternative to traditional passive learning strategies such as didactic lecturing [1]. As a novel teaching mode, inverted classroom teaching (ICT), first proposed by Lage et al [2] in 2020, is now widely used to enhance the engagement of students in the active learning process. ICT, also known as "flipped classroom teaching," promotes student participation, engagement, and identification of necessary resources and needs to meet learning objectives by repurposing classroom time for student-centered learning activities [3,4]. The teaching materials are made available for self-study outside of the classroom, while ICT also emphasizes active learning by assigning preclass tasks to students with clear learning objectives. ICT represents a significant advancement in modern classroom design, and its potential for promoting student-centered learning is particularly noteworthy.

Medical institutions were among the first to shift away from traditional didactic methods toward student-centered learning, which has been shown to motivate and empower students to be life-long learners, foster self-growth, and encourage receiving and applying up-to-date information and techniques in various medical fields [5,6]. Since it was first proposed as a teaching model [2], ICT has been used in almost all fields of education, especially in basic medicine and clinical medicine, and has become a focus of educational research. A recent bibliometric analysis on ICT revealed its ability to reallocate the teaching content taught in traditional classrooms outside the classroom for students to study on their own before the class. The resulting saved classroom time is then used for various student-centered learning activities such as problem-based and inquiry-based learning [4,7,8]. With the COVID-19 pandemic wreaking havoc around the globe, ICT has been increasingly incorporated into online teaching and is regarded as a promising and flexible approach for securing high-quality teaching via different forms of teaching media [9]. Despite the overwhelming benefits and compelling cases, researchers have also reported negative examples and disadvantages of using active-learning strategies, such as students lacking learning motivation [10,11], increased workload for both faculty and students [12], longer preparation time [12], and reluctance to discuss the teaching content with peers [13]. Moreover, a systematic theoretical and practical system of ICT in medical education has not yet been established.

Physiology is a bridging course between basic and clinical medicine, which is a core course for students in medicine and related subjects. Physiology is typically scheduled in the first semester of the second year of medical school. This course is often considered challenging for students in the early stages of their medical education owing to its highly conceptual nature, the significant cognitive effort required to acquire academic information, and the combined laboratory experiments associated with theoretical knowledge [9,14,15]. To a certain extent, the history and development of inverted teaching in physiology may serve as a window to probe into the general picture of the use of ICT in basic medical education. However, there is still a vast knowledge gap in the development and application of ICT in physiology courses; for example, it remains unclear how ICT in physiology evolved from the information era to the digital and artificial intelligence era. With the development of CiteSpace, a powerful visualization and analysis software, it has now become feasible to depict and visualize science knowledge graphs [16], including the outline and timeline of ICT in physiology, which can help to address these knowledge gaps in a more quantitative manner than possible with traditional qualitative methods such as a scoping review.

Therefore, in this study, we performed a visual analysis of the ICT in physiology literature from the Web of Science (WoS) database with CiteSpace. The aim was to explore the temporal evolution context and spatial distribution networks of ICT in physiology; investigate the cooperation network among authors, institutions, and countries publishing research in this field using co-occurrence analysis; and uncover hot research topics and development trends through cocitation analysis of references, authors, and journals, along with keyword co-occurrence and clustering analyses.

Methods

Search Strategy

We selected the WoS Core Collection as the data source for this study. To capture a broad range of potentially eligible articles, we used the following search terms with Boolean operators: ("flipped classroom" OR "flipped classroom teaching" OR "flipped study" OR "flipped learning" OR "flipped teaching" OR "flipped instruction" OR "inverted teaching" OR "inverted learning" OR "inverted study" or "inverted classroom" OR "inverted instruction") in all fields AND ("Physiology") in all fields. The time span was set from January 2000 to April 2023, and the data were collected on December 11, 2023. Only journal articles indexed in the WoS Core Collection were used to gather data. This database was selected because it is the longest-established citation tracking database, which includes quality indices such as Journal Citation Reports [17], provides a well-recognized subject classification for research journals, and permits the easy download of a considerable number of stored references [18].

Study Selection Criteria

The search was performed in English to obtain the largest number of documents in the WoS data set on the use of ICT in physiology education. The following inclusion criteria were applied: (1) document type=articles, (2) language=English, (3) years of publication=2000-2023 (November). The exclusion criteria were (1) studies in a field not related to medicine or pedagogy; (2) not published in English; (3) categorized as books, chapters, theses, protocols, study outlines, government publications, posters, editorial materials, duplicates, or nonpeer-reviewed articles; and (4) published outside of the time frame of 2000-2023.

Upon applying the above search strategy, 632 indices were retrieved in the WoS data set and 295 records were screened after removing 237 studies using automation tools from the database. Before further screening and retrieval of the full texts

XSL•FO
RenderX

of the references, all 294 indices with detailed citation records and bibliometric information were exported in both record and reference formats, saved as plain-text files, and stored in the .txt format. The stored records were then input into the CiteSpace software for visualization, as indicated by the user manual [19], which generated clustered plots of bibliometric references and differentiated various topics. The relevant articles pertaining to inverted classroom pedagogy were identified by examining the visualized clusters and topics, and irrelevant literature was excluded by adhering to the guidelines in the CiteSpace manual. In brief, in the cluster plots, irrelevant topics are presented in isolated clusters without citation networks; hence, these dots, representing the irrelevant literature, were removed from the eligible references after reviewing the titles and abstracts.

The full text of the included articles was downloaded and reviewed by two authors independently (YW and ZH), and a consensus was reached through discussion between the two reviewers in the case of any disagreements. In total, 253 studies were excluded after title and abstract screening and a total of 42 articles were included for the final analysis. The flowchart of study selection is provided in Multimedia Appendix 1 and the details of the excluded studies with reasons for exclusion are provided in Multimedia Appendix 2.

Data Analysis Process

CiteSpace 6.1.R6 software was used to visually analyze the literature related to ICT in physiology published up to November 2023. CiteSpace is a knowledge visualization software developed by Chaomei Chen at Drexel University and is now a widely used knowledge mapping tool in various fields of education and teaching [20]. CiteSpace can measure and visualize literature collections in broad fields of the natural and social sciences using cocitations of references, authors, and journals; the co-occurrence of authors, keywords, institutes, and countries; and cluster analysis to create a scientific knowledge network map, explore the critical path of the evolution of the discipline, and analyze the hot spot research topics and frontier trends clearly and scientifically.

In this study, we analyzed the overall national and regional distributions and cooperation of the authors of ICT in physiology research papers through the constructed network cooperation map, and then determined the knowledge base and the core authors of ICT in physiology research through analysis of the literature and author cocitation networks. We further identified the “star” journals publishing research in this field through a cocitation analysis of the source journals. Finally, the hot spot keywords were determined through keyword co-occurrence and clustering analysis based on the frequency and centrality of the keywords, which were used to further explore the hot topics of worldwide research on ICT in physiology. Overall, the methodology used in this study involved cooperative network analysis and cocitation analysis.

Cooperative network analysis was used to identify core authors, leading research institutions, and national/regional cooperation in ICT in physiology research. The nodes in the graph are represented by circles, with larger circles indicating a greater number of items represented, such as papers, authors, institutions, references, and countries. In CiteSpace, intermediary or between centrality is used as a critical indicator of node importance, which is characterized by the shortest number of paths passing through a node. Nodes with a centrality value above 0.1 are considered to be important. In this study, the circle size represents the cited frequency of an article, with purple circles indicating high centrality; thus, larger and deeper-purple circles suggest greater importance of the study in ICT in physiology research.

Cocitation analysis was used to identify relationships between cited articles, authors, and journals in the field of ICT in physiology research. For example, if two articles (or authors or journals) A and B are cited simultaneously by a third article, then a cocitation relationship exists between them. Frequent citation of articles (or authors or journals) together suggests that their research topics, including concepts, theories, or methods, are likely related. Cocitation analysis ranks key papers according to their citation frequency and explains the correlation between their contents and directions through the centrality value. This analysis can also infer literature clusters from various papers that are published during the same period, indicating hot spots in the field. The frequency and relevance of citations represent hot spots in scientific research over time, and these core documents form the knowledge base for the hot spots. In turn, the knowledge base clarifies the cutting-edge nature of the research, as frequently cited papers constitute the corresponding knowledge base [21].

Results

Publication Trends in ICT in Physiology

The year 2013 marks the commencement of the field, in which Tune et al [22] were the first to publish a research paper related to ICT in physiology. The research volume then increased yearly, reaching its peak in 2022. According to the number of publications, different stages of ICT in physiology development can be defined. Before 2017, there were only a small number of papers related to ICT in physiology, marking 2013 - 2017 as the gradual upward stage. In 2018, there was a slight decrease in the number of published papers on the topic, which may be due to the conflicts between conventional teaching and incorporating ICT into physiology teaching, indicating the need for more modification and reflection in practice. Hence, 2018 - 2019 can be considered as the adaptation period. The second gradual upward period appeared during the COVID-19 outbreak in 2021 and then peaked in 2022, indicating a boom period for this field of study.

Authors’ Cooperative Network

An author’s contribution to the area of ICT in physiology can be identified by their significant publications and cooperative connections with other authors, which facilitates understanding the progress in ICT in physiology [23]. Author collaboration appears to be generally limited, and most academic work in this field is conducted in independent teams with minimal cross-team communication.

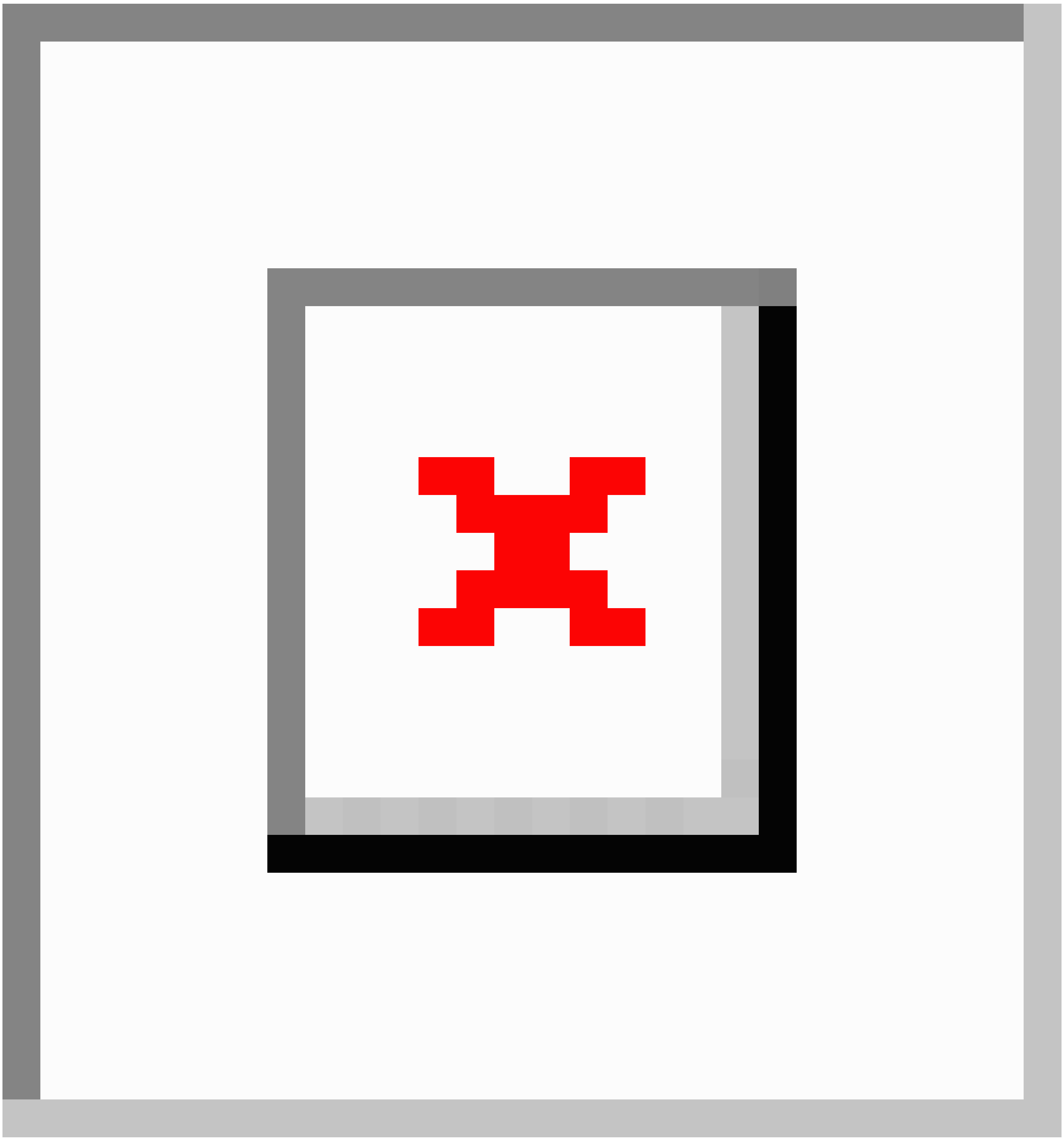
As shown in Figure 1A, the research author cooperation map highlights various research partnership teams, particularly those

XSL•FO
RenderX

surrounding the authors Gopalan C, Gillam-Krakauer M, and multiple researchers with cooperative connections. Gopalan C has the highest citation count with 5 publications, followed by authors Carbajal MM, Falck AJ, Johnston LC, Feng D, Luo Z, French H, Dadiz R, Vasquez MM, and Gray MM who collaborated on three records with a citation count of 3 each, as depicted in Multimedia Appendix 3.

Since 2016, Gopalan C has published 14 relevant papers, with a significant surge from 2019 to 2022, as illustrated in Figure 1A. Gopalan C, Bingen H, Tveit B, Steindal S, and Krumsvik R have jointly published three papers centered on nursing education [24-26], indicating a stable partnership among these authors who conducted a series of studies on ICT in nursing education. Additionally, some other authors, including Feng D and Luo Z from Central South University in China, have coauthored two papers [27,28].

Figure 1. Network analysis map. The collaboration networks for (**A**) authors and national/regional collaboration (N=145, E=400) and (**B**) institutions (N=85, E=216) in the field of inverted teaching in physiology. Node size (N) corresponds to the frequency of inverted teaching in physiology publications from each author/institution. The connecting lines (E) represent collaborative connections between authors/institutions, with thicker lines indicating more frequent collaboration.



XSL•FO
RenderX

National/Regional and Institutional Cooperative Networks

Overall, the extent of collaboration between nations and research institutions is relatively weak, with very low centrality, and the research power of countries is uneven. As seen in Figure 1 and Multimedia Appendix 4, US-based authors published the highest number of inverted teaching in phyisology–related papers (18 in total, accounting for over 43% of the total papers). Moreover, their intermediary centrality is 0.24, indicating that they have strong connections and are highly engaged in international cooperation. Chinese authors ranked second, publishing 8 papers; however, their intermediary centrality is only 0.02, suggesting that their papers have limited international influence and lower overall quality, providing little influential power in the field of ICT in physiology. Australia ranked third, with a centrality of 0.01 covered by 3 papers on ICT in physiology.

As shown in Figure 1B and Multimedia Appendix 5, universities and affiliated hospitals are the primary institutions that have published ICT in physiology–related papers. Southern Illinois University Edwardsville and Duke University in the United States have published the most papers in this field since 2018 and have significantly contributed to ICT in physiology research. Other institutions, including the University of Washington, University of Texas, Vanderbilt University, Central South University, University of Rochester, University of Pennsylvania, Yale University, University of Washington, and Baylor College of Medicine, contributed 3 research papers each. As seen from the year color bar on the left bottom corner of Figure 1B, most nodes are labeled in orange, indicating that most institutions published these articles in 2021 and 2022. Specifically, most of the studies performed in the United States are labeled in green and yellow, corresponding to earlier years, indicating the pioneering role of universities in the United States for ICT in physiology research; in particular, authors from Southern Illinois University published an ICT in physiology paper in 2017, which is earlier than most institutions contributing to this field.

Cocitation Analysis of References

The highly cited literature on ICT in physiology is summarized in Table 1, which shows the top 15 most influential articles in this field of research ranked by citation frequency and mediation centrality published between 2013 and 2020. The top-ranked item by citation counts is by Chen et al [21], which was published in 2017 with a citation count of 8 and a centrality of 0.26, followed by the paper published by McLaughlin et al [29] in 2014, also with a citation count of 8.

Table . Top 12 influential papers on inverted teaching in physiology published in the last decade (2013 - 2023).

Cited reference	Citation count	Centrality	Publication year
Chen et al [21]	8	0.26	2017
McLaughlin et al [29]	8	0	2014
Tune et al [22]	5	0	2013
Gilboy et al [30]	4	0.12	2015
Pierce and Fox [31]	4	0.15	2012
Betihavas et al [1]	4	0.5	2016
Xiao et al [32]	4	0.17	2018
Hew and Lo [33]	4	0.2	2018
Day [34]	3	0.07	2018
French et al [35]	3	0.04	2018
Blair et al [36]	3	0.06	2020
Freeman et al [37]	3	0.08	2014
Akçayır and Akçayır [38]	3	0.3	2018
Foldnes [39]	3	0.03	2016
Gross et al [40]	3	0.03	2015

Cocitation Analysis

As shown in Multimedia Appendix 6 and Multimedia Appendix 7, Gopalan C was found to be the most cited author with a count of 5 and a centrality of 0.02.

Multimedia Appendix 8 summarizes the top 10 journals that published ICT in physiology papers. *Advances in Physiology Education* was the first journal to publish ICT in physiology research papers and has maintained the highest frequency of citations from 2013 to 2022 (also see Multimedia Appendix 6). Additionally, journals such as *Computers & Education* and *The Internet and Higher Education* have also provided considerable attention to this topic, implying that modern educational technologies such as information science and the internet play a crucial role in facilitating the inverted classroom mode.

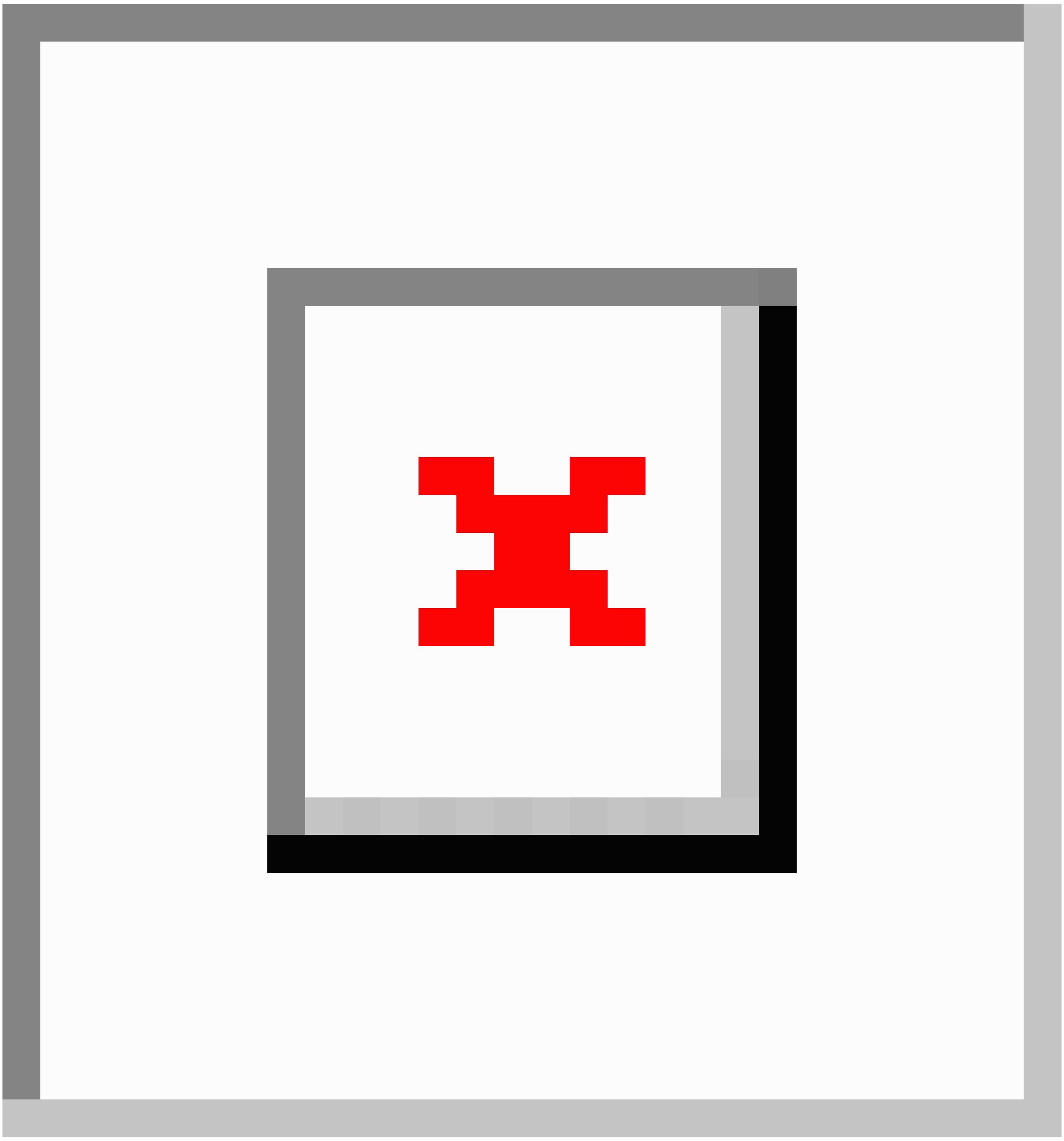
Research Hot Spots Suggested by Keyword Co-Occurrence Analysis

Figure 2 presents the coexistence diagram of ICT in physiology keywords, with each node representing a keyword and the font size indicating the node's size; that is, a larger font indicates that the keyword appears more frequently. The cluster labels

XSL•FO
RenderX

obtained from the keyword cluster analysis can indirectly reflect the leading research topics, while the timeline map of the keyword clusters can demonstrate the leading research topics by time. Table 2 lists the top keyword clusters in ICT in physiology research according to the number of occurrences and centrality of each keyword, demonstrating that the top keywords are "flipped classroom," "active learning," "student performance," "performance," and "medical education." Figure 2A shows that from 2013 to 2022, research on ICT in physiology focused on medical education, performance, engagement, active learning, online teaching, and other aspects. According to the intermediary centrality, "flipped classroom" (0.69) is the most influential keyword, followed by "medical education" (0.2) and "education" (0.14).

Figure 2. Co-occurrence map and appearance history of keywords in literature related to inverted teaching in physiology. (**A**) The map of keyword clusters and the timeline map (N=131, E=459). (**B**) The co-occurrence map of keywords (N=233, E=795). The node size, N, corresponds to the frequency of publications from each journal. The connecting lines, E, represent collaborative connections between journals, with thicker lines indicating more frequent collaboration.



XSL•FO
RenderX

Table . Main keywords in research related to inverted teaching in physiology.

Keywords	Co-occurrence number	Mediator centrality	First year of appearance
flipped classroom	22	0.69	2013
medical education	8	0.2	2016
education	7	0.14	2020
performance	6	0.14	2014
engagement	6	0.26	2015
flipped teaching	5	0.19	2018
student performance	4	0.04	2013
medical students	4	0.19	2016
active learning	4	0.05	2016
online teaching	3	0.06	2021
instruction	3	0.11	2015
classroom	3	0.09	2015
modified team-based learning	2	0.02	2017
dental education	2	0	2017
classroom model	2	0.05	2017
science	2	0.03	2017
faculty	2	0	2021
covid-19 pandemic	2	0	2022
physiology education	2	0.04	2016
bioscience	2	0.01	2019
perceptions	2	0.04	2022
higher education	2	0.01	2019
medical student	2	0.01	2018
students	2	0.01	2015
efficacy	2	0.09	2020
physiology	2	0.02	2015
learning preference	2	0.01	2021
student perceptions	2	0.02	2016
too	2	0	2022
blended learning	2	0.03	2020
online flipped classroom	2	0	2022
intrinsic motivation	2	0.03	2014
self-determination theory	2	0	2021
learning style	2	0.15	2016

The keyword co-occurrence analysis showed that in addition to the retrieved topic term “flipped classroom,” “medical education” ranked the highest in terms of word frequency and ranked the third highest according to mediator centrality, reflecting that active learning is a hot topic in ICT in physiology research. The keywords ”education,” “performance,” and “engagement” followed closely behind, with the centrality being 0.14, 0.14, and 0.26, respectively (Table 2). This indicates that researchers in the field of ICT in physiology have been paying relatively more attention to performance aspects, which could reflect the effectiveness and satisfaction of ICT in physiology. The keywords “engagement” and “perceptions” also had high co-occurrence numbers and mediator centrality.

Research Hot Spots and Frontier Topics Suggested by Keyword Cluster Analysis

Based on other keywords in the same cluster and the popular words obtained by the latent semantic analysis/indexing algorithm, it was found that many popular words in each cluster reflected the current hot spots of ICT in physiology and had

XSL•FO
RenderX

good consistency with the hot spot topics obtained by the co-occurrence analysis of keywords (see Multimedia Appendix 9), such as active learning, self-directed learning, student characteristics, learning preferences, learning styles, modified team-based learning, learning environment, flipped design, student engagement, and undergraduate medical education, among others.

Discussion

Principal Findings

In this study, we used CiteSpace software to visually analyze the literature related to the use of ICT in physiology published between 2000 and 2023 retrieved from the WoS database. The results of this bibliometric analysis showed that the core authors publishing in the field of ICT in physiology include staff from universities and affiliated hospitals. Some research teams have also formed cooperative relationships. Research in ICT in physiology mainly focuses on active learning, autonomous learning, student performance, teaching effectiveness, blended teaching, personalized flipped teaching, and other related topics.

Overall, studies on the ICT model in the context of physiology remain scarce, with limited collaboration among authors and a consequent lack of a cohesive research network. Regional growth in this field is uneven and international disparities are evident. Despite the many established benefits of ICT, it is not widely used in various nations and regions. This may be attributable to the fact that the development of the ICT model is still in its infancy, and a mature theoretical structure is needed and must be tested over a wide range of professional specialties. In this sense, relevant researchers must increase interaction and collaboration, investigate systematic teaching techniques appropriate for various disciplines, and perform practical testing and assessment of the model. In the future, research power can be integrated to form a cohesive unit through cooperation among research institutions to promote further breakthroughs in ICT research in the context of physiology.

Development of ICT in Physiology

The ICT model has undergone three stages of development, including the introduction stage (2013 - 2014), extensive practice stage (2015 - 2019), and modification and growth stage (2020 - 2022).

Several studies have confirmed that an active-learning strategy is associated with improved student performance, a reduced failure rate, and better learning achievements in basic and clinical medical education [37,41]. Shaffer [42] reported that anatomy course objectives were achieved at a much higher rate after incorporating an active teaching style compared to the achievement rate following traditional teaching. Furthermore, in the clinical discipline, Qutub [43] reported the considerable effectiveness of ICT as an active learning style in a hematology course, enabling students to obtain desirable knowledge and improve their academic performance; moreover, students recognized that ICT as an active learning style was more beneficial than the traditional teaching approach. In 2016, Betihavas et al [1] performed a systematic review of 9 studies on the use of ICT in nursing education and reported that nursing students achieved similar or higher academic outcomes with ICT than with a conventional teaching strategy; however, the students indicated a mixed sense of satisfaction.

Other researchers in medical education and health science programs have reported similar results. For example, in an analysis of 274 papers, Barranquero-Herbosa et al [44] found that ICT in nursing education improves performance and is well-received by both students and instructors. O'Connor et al [45] concluded that reversing the flow of classroom teaching improves academic performance, develops self-directed learning skills, and consolidates acquired knowledge through active learning strategies. Sultan [46] found that flipping the classroom gives students more time for active learning, peer collaboration, and applying and analyzing theoretical knowledge. Moreover, McLean et al [47] showed that ICT could improve students' preparation, attendance, and participation in the course Medical Sciences 4200, an elective nonthesis-based course that covers content related to physiology, biochemistry, and immunology.

With COVID-19 wreaking havoc worldwide in early 2020, the strict and rapid public health measures put forward led to the suspension of face-to-face education and the transfer of the classroom to online meetings, which also corresponds to the application of blended learning as a pedagogical approach based on a combination of online and face-to-face education processes [48]. This necessary shift during the pandemic greatly facilitated the implementation of ICT in various subjects and expanded the use of other types of education tools. For instance, Bawazeer et al [49] reported the use of polls in virtual sessions on physiology, pharmacology, and pathology courses to assess students' engagement, understanding, performance, and attendance, and found improvements in understanding and permormance. Feng et al [28] reported that incorporating the inverted classroom and a team-based learning strategy in the online setting can enhance the learning outcomes for students in a clinical immunology laboratory course. Although the pandemic and the availability of novel technologies have made blended learning a "new normal" in medical education, the successful adaptation of blended learning requires sufficient teacher training as well as the adoption of appropriate technologies by educational institutions [50].

The Role of ICT in Medical Education

In 2018, Chung et al [51] performed a systematic review on the use of ICT in nursing education, which showed that the basic flipped classroom mode has been frequently used in nursing education; nevertheless, the effects of ICT on learning behavior in physiology courses were not clearly investigated and only a few studies included in the review reported the use of after-class activities to engage students in facilitating the applications of the knowledge learned. Moreover, Lin and Hwang [52] reviewed studies on ICT papers published up to 2017 based on the technology-enhanced learning model, and noted that little attention was paid to the assessment of learners' higher-order thinking skills and their degree of preparation or cognitive load. Similar findings have also been reported in relation to the application of ICT in subjects other than medicine, including mathematics [53].

XSL•FO
RenderX

Nevertheless, there is no doubt that ICT can efficiently engage students in learning sessions, even during the pandemic [54]. Research investigating students' perceptions and performance revealed that students have high levels of acceptance for a virtual flipped teaching approach, which was already evident prior to the COVID-19 pandemic [9,55-57].

Lack of a Cohesive Research Network in ICT in Physiology Research

Acknowledging the importance of international cooperation and the role of different countries contributing to research on ICT in physiology may facilitate communication and collaboration among countries. With the highest number of published papers, authors from the United States have been the primary contributors to research on the applications of ICT in physiology courses since 2013.

The positive effects of ICT largely depend on an effective classroom design [58]. Designing an effective inverted classroom, guiding students to engage in inverted classroom learning, and personalizing the ICT to enhance teaching effectiveness and student learning outcomes have increasingly become the main topics of ICT research. These are common challenges encountered by teachers and students in ICT. Since a layered teaching approach adapted to the learning, teaching, and classroom conditions can maximize the expected benefits, various ICT approaches have been developed to date, such as partially inverted classrooms [59], Small Private Online Courses–based inverted classrooms [60], and lecture-based inverted classrooms [61].

Current Hot Spots of ICT in Physiology Research

There are currently three main topics generally discussed in the field of ICT: preparation before class, classroom activities, and consolidation after class [23]. The current hot spots of research in ICT for physiology worldwide focus on active learning, inverted classroom design, student perception and engagement, teaching effectiveness, and teaching evaluation, among others, while the scope of the research includes students, teachers, school teaching management, and national educational guidelines and policies. Moreover, our results are consistent with previous bibliometric studies related to the research on ICT in other fields [62]. For instance, a recent review by Cheng et al [62] on the top 100 highly cited ICT papers similarly showed that researchers in this field have largely focused on students' learning achievements and learning behaviors rather than directly comparing the benefits of inverted and traditional learning. Similarly, Meral et al [63] reported that motivation, perception, and academic achievement/performance were the most common topics in the ICT studies published between 2010 and 2019.

Regarding the research hot spots suggested by the analysis of keywords, we identified the following main areas of focus of research on ICT in physiology at present: (1) ICT theories, including active learning and independent learning; (2) ICT strategies, including inverted design, student characteristics, learning style, learning preference, learning environment, educational technology, and student participation; and (3) ICT evaluations, including academic performance, student performance, and student satisfaction. Specific to disciplines and programs, the field of research on ICT in physiology covers clinical medicine, stomatology, nursing, pharmacy, and veterinary medicine, among others. With respect to the courses, ICT approaches can be applied to general physiology, gastrointestinal and renal physiology, exercise physiology, physiology lab courses, and introductory biology. The applicable levels of education include graduate, undergraduate, professional training, and adult continuing education.

Study Strengths and Limitations

This study has both strengths and limitations. To our knowledge, this is the first study to map the current ICT studies in physiology specifically rather than considering the whole field of ICT. Moreover, the visualization of the quantitative results provides a convenient and comprehensible understanding of the current publication status of studies, research hot spots, and development trends in the field of ICT for physiology.

Although all attempts were made to include relevant nouns and terms in the literature retrieval process, some relevant papers may have nevertheless been missed. Additionally, the search only incorporated "physiology" as the keyword for the teaching subject, which may have led to evidence selection bias in which research that covers all medical courses rather than physiology alone may have been missed and could not be incorporated into the study for analysis. In addition, the search was limited to the WoS database, which may have excluded some important non-English publications. Moreover, each subject has unique characteristics in the application of an inverted teaching model, and the results and conclusions reached based on the analysis of this study may not necessarily be generalized to other subjects; thus, these results should be interpreted with caution.

Conclusion

This study analyzes literature on ICT in physiology, identifying core authors, research topics, and development stages. To date, research in this field has focused on active and autonomous learning, student performance, the teaching effect, blended teaching, and personalized flipping teaching models. The development of ICT is linked to modern information technology, the COVID-19 pandemic, educational teaching concepts, and related teaching reform policies. Based on these findings, further academic exchanges and cooperation in applications of ICT in physiology are encouraged, which can highlight the potential of this teaching model to train the next generation of excellent medical talents.

Acknowledgments

This study has been supported by the educational science programs for research projects: Higher Education Special Project in Guangdong Province (2021GXJK259), Medical Teaching and Education Management Reform Research Project in Jinan University

XSL•FO
RenderX

(2021YXJG005), Annual Experimental Teaching Curriculum Reform Special Project (SYJG202202), and "Four New" Experimental Teaching Curriculum Reform Project at Jinan University (SYJG202301).

Authors' Contributions

YW and JB conceptualized the study. ZH, BZ, and HF contributed to the methodology. ZH and BZ contributed to data visualization. JB and YW supervised the study. ZH, BT, HF, JB, and YW wrote the initial draft of the manuscript. ZH, BZ, HF, YW, and JB contributed to manuscript review and editing. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flowchart of literature selection.
[PNG File, 65 KB - mededu_v10i1e52224_app1.png]

Multimedia Appendix 2
Excluded studies with reasons.
[DOCX File, 89 KB - mededu_v10i1e52224_app2.docx]

Multimedia Appendix 3
The top 12 authors who published relevant papers on inverted teaching in physiology.
[DOCX File, 16 KB - mededu_v10i1e52224_app3.docx]

Multimedia Appendix 4
Distribution of countries publishing papers related to inverted teaching in physiology.
[DOCX File, 15 KB - mededu_v10i1e52224_app4.docx]

Multimedia Appendix 5
The top 12 institutions publishing papers related to inverted teaching in physiology.
[DOCX File, 16 KB - mededu_v10i1e52224_app5.docx]

Multimedia Appendix 6
(A) The cited reference analysis map of inverted teaching in physiology: N=235, E=684. Node size (N) corresponds to the frequency of publications from each reference. The connecting lines (E) represent collaborative connections between authors, with thicker lines indicating more frequent collaboration. (B) Analysis of cocited journals (N=236, E=996). Node size (N) corresponds to the frequency of publications from each journal. The connecting lines (E) represent citation connections between references, with thicker lines indicating more frequent cocitations.
[PNG File, 570 KB - mededu_v10i1e52224_app6.png]

Multimedia Appendix 7
The most influential authors of inverted teaching in physiology research.
[DOCX File, 15 KB - mededu_v10i1e52224_app7.docx]

Multimedia Appendix 8
Primary journals that publish research papers in the field of inverted classroom teaching in physiology.
[DOCX File, 18 KB - mededu_v10i1e52224_app8.docx]

Multimedia Appendix 9
Information of the main clusters of keywords in research related to inverted teaching in physiology.
[DOCX File, 17 KB - mededu_v10i1e52224_app9.docx]

References

1. Betihavas V, Bridgman H, Kornhaber R, Cross M. The evidence for ‘flipping out’: a systematic review of the flipped classroom in nursing education. Nurse Educ Today 2016 Mar;38:15-21. [doi: 10.1016/j.nedt.2015.12.010] [Medline: 26804940]

XSL•FO
RenderX

2. Lage MJ, Platt GJ, Treglia M. Inverting the classroom: a gateway to creating an inclusive learning environment. J Econ Educ 2000 Jan;31(1):30-43. [doi: 10.1080/00220480009596759]
3. Persky AM, McLaughlin JE. The flipped classroom - from theory to practice in health professional education. Am J Pharm Educ 2017 Aug;81(6):118. [doi: 10.5688/ajpe816118] [Medline: 28970619]
4. Goodman BE, Martin DS, Williams JL. Teaching human cardiovascular and respiratory physiology with the station method. Adv Physiol Educ 2002 Dec;26(1-4):50-56. [doi: 10.1152/advan.00034.2001] [Medline: 11850328]
5. Taylor DCM, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE guide no. 83. Med Teach 2013 Nov;35(11):e1561-e1572. [doi: 10.3109/0142159X.2013.828153] [Medline: 24004029]
6. Ramnanan CJ, Pound LD. Advances in medical education and practice: student perceptions of the flipped classroom. Adv Med Educ Pract 2017 Jan;8:63-73. [doi: 10.2147/AMEP.S109037] [Medline: 28144171]
7. del Arco I, Mercadé-Melé P, Ramos-Pla A, Flores-Alarcia Ò. Bibliometric analysis of the flipped classroom pedagogical model: trends and strategic lines of study. Front Educ 2022 Sep;7:1022295. [doi: 10.3389/feduc.2022.1022295]
8. Julia J, Afrianti N, Soomro KA, et al. Flipped classroom educational model (2010-2019): a bibliometric study. Eur J Ed Res 2020 Oct 15;9(4):1377-1392. [doi: 10.12973/eu-jer.9.4.1377]
9. Beason-Abmayr B, Caprette DR, Gopalan C. Flipped teaching eased the transition from face-to-face teaching to online instruction during the COVID-19 pandemic. Adv Physiol Educ 2021 Jun 1;45(2):384-389. [doi: 10.1152/advan.00248.2020] [Medline: 33961513]
10. Pence PL, Franzen SR, Kim MJ. Flipping to motivate: perceptions among prelicensure nursing students. Nurse Educ 2021 Jan;46(1):43-48. [doi: 10.1097/NNE.0000000000000814] [Medline: 32175953]
11. Birgili B, Demir Ö. An explanatory sequential mixed-method research on the full-scale implementation of flipped learning in the first years of the world’s first fully flipped university: departmental differences. Comput Educ 2022 Jan;176:104352. [doi: 10.1016/j.compedu.2021.104352]
12. Foster G, Stagl S. Design, implementation, and evaluation of an inverted (flipped) classroom model economics for sustainable education course. J Clean Prod 2018 May;183:1323-1336. [doi: 10.1016/j.jclepro.2018.02.177]
13. Critz CM, Knight D. Using the flipped classroom in graduate nursing education. Nurse Educ 2013 Sep;38(5):210-213. [doi: 10.1097/NNE.0b013e3182a0e56a] [Medline: 23969751]
14. Slominski T, Grindberg S, Momsen J. Physiology is hard: a replication study of students' perceived learning difficulties. Adv Physiol Educ 2019 Jun 1;43(2):121-127. [doi: 10.1152/advan.00040.2018] [Medline: 30835145]
15. Colthorpe KL, Abe H, Ainscough L. How do students deal with difficult physiological knowledge? Adv Physiol Educ 2018 Dec 1;42(4):555-564. [doi: 10.1152/advan.00102.2018] [Medline: 30192189]
16. Wong PC, Chen C, Gorg C, Shneiderman B, Stasko J, Thomas J. Graph analytics-lessons learned and challenges ahead. IEEE Comput Grap Appl 2011 Sep;31(5):18-29. [doi: 10.1109/MCG.2011.72]
17. Journal citation reports. Clarivate. URL: https://clarivate.com/products/scientific-and-academic-research/research-analytics-evaluation-and-management-solutions/journal-citation-reports/ [accessed 2024-06-07]
18. Rojas-Sánchez MA, Palos-Sánchez PR, Folgado-Fernández JA. Systematic literature review and bibliometric analysis on virtual reality and education. Educ Inf Technol 2023;28(1):155-192. [doi: 10.1007/s10639-022-11167-5] [Medline: 35789766]
19. Chen C. The CiteSpace Manual. 2014. URL: http://cluster.ischool.drexel.edu/~cchen/citespace/CiteSpaceManual.pdf [accessed 2023-12-15]
20. Chen C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci 2006 Feb;57(3):359-377. [doi: 10.1002/asi.20317]
21. Chen F, Lui AM, Martinelli SM. A systematic review of the effectiveness of flipped classrooms in medical education. Med Educ 2017 Jun;51(6):585-597. [doi: 10.1111/medu.13272] [Medline: 28488303]
22. Tune JD, Sturek M, Basile DP. Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. Adv Physiol Educ 2013 Dec;37(4):316-320. [doi: 10.1152/advan.00091.2013] [Medline: 24292907]
23. Sun L, Yang L, Wang X, Zhu J, Zhang X. Hot topics and frontier evolution in college flipped classrooms based on mapping knowledge domains. Front Public Health 2022 Aug;10:950106. [doi: 10.3389/fpubh.2022.950106] [Medline: 36091514]
24. Bingen HM, Steindal SA, Krumsvik R, Tveit B. Nursing students studying physiology within a flipped classroom, self-regulation and off-campus activities. Nurse Educ Pract 2019 Feb;35:55-62. [doi: 10.1016/j.nepr.2019.01.004] [Medline: 30690317]
25. Bingen HM, Tveit B, Krumsvik RJ, Steindal SA. Nursing students’ experiences with the use of a student response system when learning physiology. Nordic J Digit Literacy 2019 Nov 6;14(1-2):37-53. [doi: 10.18261/issn.1891-943x-2019-01-02-04]
26. Bingen HM, Steindal SA, Krumsvik RJ, Tveit B. Studying physiology within a flipped classroom: the importance of on-campus activities for nursing students' experiences of mastery. J Clin Nurs 2020 Aug;29(15-16):2907-2917. [doi: 10.1111/jocn.15308] [Medline: 32353915]
27. Xu Y, Chen C, Feng D, Luo Z. A survey of college students on the preference for online teaching videos of variable durations in online flipped classroom. Front Public Health 2022 Mar;10:838106. [doi: 10.3389/fpubh.2022.838106] [Medline: 35356026]

XSL•FO
RenderX

28. Feng Y, Zhao B, Zheng J, Fu Y, Jiang Y. Online flipped classroom with team-based learning promoted learning activity in a clinical laboratory immunology class: response to the COVID-19 pandemic. BMC Med Educ 2022 Dec 3;22(1):836. [doi: 10.1186/s12909-022-03917-3] [Medline: 36463210]
29. McLaughlin JE, Roth MT, Glatt DM, et al. The flipped classroom: a course redesign to foster learning and engagement in a health professions school. Acad Med 2014 Feb;89(2):236-243. [doi: 10.1097/ACM.0000000000000086] [Medline: 24270916]
30. Gilboy MB, Heinerichs S, Pazzaglia G. Enhancing student engagement using the flipped classroom. J Nutr Educ Behav 2015;47(1):109-114. [doi: 10.1016/j.jneb.2014.08.008] [Medline: 25262529]
31. Pierce R, Fox J. Vodcasts and active-learning exercises in a "flipped classroom" model of a renal pharmacotherapy module. Am J Pharm Educ 2012 Dec 12;76(10):196. [doi: 10.5688/ajpe7610196] [Medline: 23275661]
32. Xiao N, Thor D, Zheng M, Baek J, Kim G. Flipped classroom narrows the performance gap between low- and high-performing dental students in physiology. Adv Physiol Educ 2018 Dec 1;42(4):586-592. [doi: 10.1152/advan.00104.2018] [Medline: 30251890]
33. Hew KF, Lo CK. Flipped classroom improves student learning in health professions education: a meta-analysis. BMC Med Educ 2018 Mar 15;18(1):38. [doi: 10.1186/s12909-018-1144-z] [Medline: 29544495]
34. Day LJ. A gross anatomy flipped classroom effects performance, retention, and higher - level thinking in lower performing students. Anat Sci Educ 2018 Nov;11(6):565-574. [doi: 10.1002/ase.1772] [Medline: 29356452]
35. French H, Gray M, Gillam-Krakauer M, et al. Flipping the classroom: a national pilot curriculum for physiology in neonatal–perinatal medicine. J Perinatol 2018 Oct;38(10):1420-1427. [doi: 10.1038/s41372-018-0185-9] [Medline: 30087455]
36. Blair RA, Caton JB, Hamnvik OP. A flipped classroom in graduate medical education. Clin Teach 2020 Apr;17(2):195-199. [doi: 10.1111/tct.13091] [Medline: 31512400]
37. Freeman S, Eddy SL, McDonough M, et al. Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci U S A 2014 Jun 10;111(23):8410-8415. [doi: 10.1073/pnas.1319030111] [Medline: 24821756]
38. Akçayır G, Akçayır M. The flipped classroom: a review of its advantages and challenges. Comput Educ 2018 Nov;126:334-345. [doi: 10.1016/j.compedu.2018.07.021]
39. Foldnes N. The flipped classroom and cooperative learning: evidence from a randomised experiment. Act Learn High Educ 2016 Mar;17(1):39-49. [doi: 10.1177/1469787415616726]
40. Gross D, Pietri ES, Anderson G, Moyano-Camihort K, Graham MJ. Increased preclass preparation underlies student outcome improvement in the flipped classroom. CBE Life Sci Educ 2015;14(4):ar36. [doi: 10.1187/cbe.15-02-0040] [Medline: 26396151]
41. Kyriakoulis K, Patelarou A, Laliotis A, et al. Educational strategies for teaching evidence-based practice to undergraduate health students: systematic review. J Educ Eval Health Prof 2016 Sep;13:34. [doi: 10.3352/jeehp.2016.13.34] [Medline: 27649902]
42. Shaffer JF. Student performance in and perceptions of a high structure undergraduate human anatomy course. Anat Sci Educ 2016 Nov;9(6):516-528. [doi: 10.1002/ase.1608] [Medline: 26990231]
43. Qutob H. Effect of flipped classroom approach in the teaching of a hematology course. PLoS One 2022 Apr;17(4):e0267096. [doi: 10.1371/journal.pone.0267096] [Medline: 35446895]
44. Barranquero-Herbosa M, Abajas-Bustillo R, Ortego-Maté C. Effectiveness of flipped classroom in nursing education: a systematic review of systematic and integrative reviews. Int J Nurs Stud 2022 Nov;135:104327. [doi: 10.1016/j.ijnurstu.2022.104327] [Medline: 35944288]
45. O'Connor EE, Fried J, McNulty N, et al. Flipping radiology education right side up. Acad Radiol 2016 Jul;23(7):810-822. [doi: 10.1016/j.acra.2016.02.011] [Medline: 27066755]
46. Sultan AS. The flipped classroom: an active teaching and learning strategy for making the sessions more interactive and challenging. J Pak Med Assoc 2018 Apr;68(4):630-632. [Medline: 29808055]
47. McLean S, Attardi SM, Faden L, Goldszmidt M. Flipped classrooms and student learning: not just surface gains. Adv Physiol Educ 2016 Mar;40(1):47-55. [doi: 10.1152/advan.00098.2015] [Medline: 26847257]
48. Tonbuloğlu B, Tonbuloğlu İ. Trends and patterns in blended learning research (1965-2022). Educ Inf Technol 2023 Apr 3:1-32. [doi: 10.1007/s10639-023-11754-0] [Medline: 37361774]
49. Bawazeer MA, Aamir S, Othman F, Alkahtani R. Students engagement using polls in virtual sessions of physiology, pathology, and pharmacology at King Saud bin Abdulaziz University for Health Sciences during COVID-19 pandemic: a cross-sectional study. BMC Med Educ 2023 Apr 21;23(1):276. [doi: 10.1186/s12909-023-04253-w] [Medline: 37085845]
50. Bozkurt A. A retro perspective on blended/hybrid learning: systematic review, mapping and visualization of the scholarly landscape. J Interact Media Educ 2022 Jul 21;2022(1):2. [doi: 10.5334/jime.751]
51. Chung CJ, Lai CL, Hwang GJ. Roles and research trends of flipped classrooms in nursing education: a review of academic publications from 2010 to 2017. Interact Learn Environ 2021 Aug 18;29(6):883-904. [doi: 10.1080/10494820.2019.1619589]
52. Lin HC, Hwang GJ. Research trends of flipped classroom studies for medical courses: a review of journal publications from 2008 to 2017 based on the technology-enhanced learning model. Interact Learn Environ 2019 Nov 17;27(8):1011-1027. [doi: 10.1080/10494820.2018.1467462]

XSL•FO
RenderX

53. Yang QF, Lin CJ, Hwang GJ. Research focuses and findings of flipping mathematics classes: a review of journal publications based on the technology-enhanced learning model. Interact Learn Environ 2021 Aug 18;29(6):905-938. [doi: 10.1080/10494820.2019.1637351]
54. Gopalan C, Daughrity S, Hackmann E. The past, the present, and the future of flipped teaching. Adv Physiol Educ 2022 Jun 1;46(2):331-334. [doi: 10.1152/advan.00016.2022] [Medline: 35357955]
55. Gopalan C, Butts-Wilmsmeyer C, Moran V. Virtual flipped teaching during the COVID-19 pandemic. Adv Physiol Educ 2021 Dec 1;45(4):670-678. [doi: 10.1152/advan.00061.2021] [Medline: 34498940]
56. Bryson JR, Andres L. Covid-19 and rapid adoption and improvisation of online teaching: curating resources for extensive versus intensive online learning experiences. J Geogr High Educ 2020 Oct 1;44(4):608-623. [doi: 10.1080/03098265.2020.1807478]
57. Fogg KC, Maki SJ. A remote flipped classroom approach to teaching introductory biomedical engineering during COVID-19. Biomed Eng Educ 2021;1(1):3-9. [doi: 10.1007/s43683-020-00001-4] [Medline: 35146488]
58. Paralikar S, Shah CJ, Joshi A, Kathrotia R. Acquisition of higher-order cognitive skills (HOCS) using the flipped classroom model: a quasi-experimental study. Cureus 2022 Apr;14(4):e24249. [doi: 10.7759/cureus.24249] [Medline: 35602838]
59. Lax N, Morris J, Kolber BJ. A partial flip classroom exercise in a large introductory general biology course increases performance at multiple levels. J Biol Educ 2017 Oct 2;51(4):412-426. [doi: 10.1080/00219266.2016.1257503]
60. Zhang XM, Yu JY, Yang Y, Feng CP, Lyu J, Xu SL. A flipped classroom method based on a small private online course in physiology. Adv Physiol Educ 2019 Sep 1;43(3):345-349. [doi: 10.1152/advan.00143.2018] [Medline: 31305152]
61. Ulrich LM, Palacios S, Kirkby SE. Flipped classroom vs. engaging lecture style for pulmonary physiology. Am J Respir Crit Care Med 2022 May;205:A3957. [doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A3957]
62. Cheng SC, Hwang GJ, Lai CL. Critical research advancements of flipped learning: a review of the top 100 highly cited papers. Interact Learn Environ 2022 Oct 3;30(9):1751-1767. [doi: 10.1080/10494820.2020.1765395]
63. Meral E, Teke D, Güler M, Namli ZB. General trends of studies on flipped classroom model: bibliometric mapping and content analysis. Int Online J Educ Teach 2020;8(2):564-587 [FREE Full text]

Abbreviations

ICT: inverted classroom teaching
WoS: Web of Science

Edited by P Kanzow, TDA Cardoso; submitted 02.09.23; peer-reviewed by I Maniu, Q Chen, V Katavic; revised version received 02.04.24; accepted 02.04.24; published 25.06.24.

Please cite as:
He Z, Zhou B, Feng H, Bai J, Wang Y
Inverted Classroom Teaching of Physiology in Basic Medical Education: Bibliometric Visual Analysis
JMIR Med Educ 2024;10:e52224
URL: https://mededu.jmir.org/2024/1/e52224
doi:10.2196/52224

© Zonglin He, Botao Zhou, Haixiao Feng, Jian Bai, Yuechun Wang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 25.6.2024. 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.

XSL•FO
RenderX

Evolution of Chatbots in Nursing Education: Narrative Review

Fang Zhang[1,*], BA; Xiaoliu Liu[2,*], MMed; Wenyan Wu[2,*], MMed; Shiben Zhu[3], MSc

1

2

3

*these authors contributed equally

Corresponding Author:
Shiben Zhu, MSc

Abstract

Background: The integration of chatbots in nursing education is a rapidly evolving area with potential transformative impacts. This narrative review aims to synthesize and analyze the existing literature on chatbots in nursing education.

Objective: This study aims to comprehensively examine the temporal trends, international distribution, study designs, and implications of chatbots in nursing education.

Methods: A comprehensive search was conducted across 3 databases (PubMed, Web of Science, and Embase) following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Results: A total of 40 articles met the eligibility criteria, with a notable increase of publications in 2023 (n=28, 70%). Temporal analysis revealed a notable surge in publications from 2021 to 2023, emphasizing the growing scholarly interest. Geographically, Taiwan province made substantial contributions (n=8, 20%), followed by the United States (n=6, 15%) and South Korea (n=4, 10%). Study designs varied, with reviews (n=8, 20%) and editorials (n=7, 18%) being predominant, showcasing the richness of research in this domain.

Conclusions: Integrating chatbots into nursing education presents a promising yet relatively unexplored avenue. This review highlights the urgent need for original research, emphasizing the importance of ethical considerations.

(JMIR Med Educ 2024;10:e54987) doi:10.2196/54987

KEYWORDS

nursing education; chatbots; artificial intelligence; narrative review; ChatGPT

Introduction

Nursing education, crucial for positive patient-professional relationships [1,2] and continuous professional development [3], holds a pivotal position in global health care systems [4], driving progress [5] and integrating technological advancements to enhance patient-centered care [6,7]. A study on oncology nursing provided compelling evidence for nurses, addressing challenges and advocating for specialized education and safety measures in the escalating global cancer burden [8]. A recent meta-analysis of 12 studies with 821 participants evaluated the role of virtual reality in nursing education, which revealed substantial enhancements in knowledge but identified no distinguishable disparities in skills, satisfaction, confidence, and performance time, underscoring the imperative for additional investigations in these domains [9]. Another study explored the usability and feasibility of extended reality smart glasses in core nursing skill training for undergraduate students, uncovering positive effects on engagement, learning satisfaction, and competency improvement and highlighting the potential of smart glasses as an impactful educational strategy in nursing training [10]. However, nursing education encounters obstacles such as a worldwide scarcity of nursing expertise [11], uneven distribution of resources [12], potential disparities between theoretical and practical aspects [9], restricted interdisciplinary collaboration [13], insufficient opportunities for professional development [14], and the ramifications of the global COVID-19 pandemic [15].

In the swiftly evolving landscape of artificial intelligence (AI) and smartphone proliferation, the integration of large language models such as ChatGPT into chatbots is emerging as a trend, with chatbots progressively showcasing the potential to revolutionize mental health [16], behavior [17], and knowledge [18] within the dynamic and advancing field of deep learning. Recent studies on education have accentuated the use of chatbots to deliver personalized learning experiences [19,20] by tailoring content delivery to the unique needs of individual students, thereby augmenting comprehension and retention. Concurrently, chatbots provide an easily accessible platform for continuous learning [21], affording students the opportunity to retrieve information at their convenience and cultivating a culture of self-directed learning. Moreover, the interactive attributes of chatbots facilitate real-time feedback, permitting the prompt rectification of misconceptions and fostering a more profound

XSL•FO
RenderX

grasp of intricate health care concepts [22]. The adaptability of chatbots caters to diverse learning styles, ensuring inclusivity in education [23]. Despite these advantages, few studies investigate the integration, development, and feasibility of chatbots within nursing education.

Our aim is to meticulously investigate and amalgamate the existing literature pertaining to the integration of chatbots in nursing education by reviewing selected articles. By scrutinizing studies sourced from 3 prominent databases (PubMed, Embase, and Web of Science), we highlight insightful perspectives on the evolving role of chatbots in nursing education. Approaching this investigation with the perspective of a reviewer, we seek to contribute a nuanced and well-supported analysis of the existing literature on this topic.

Methods

Search Strategy

We devised pertinent search queries concerning nursing education and chatbots, with the designated search terms detailed in Section 1 in Multimedia Appendix 1. A thorough investigation encompassing 3 databases—PubMed, Embase, and Web of Science—was carried out from their individual inception dates to November 16, 2023.

Eligibility Criteria for Study Inclusion

The eligibility criteria were devised in accordance with the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) framework [24]. The study inclusion criteria were meticulously outlined to ensure the accuracy and relevance of the selected research. The specified population comprised nurses or nursing students, including managers and clinical nurses, with a deliberate exclusion of doctors and other professional personnel. The intervention criteria encompassed any chatbot intervention, including chatbot apps, messaging, and web-based interventions, while excluding interventions not specifically focused on chatbots or lacking communication with them. The comparator conditions involved conventional education methods, such as face-to-face or drug interventions, excluding the integration of chatbot interventions. The exclusion criteria also considered comparators that included chatbot interventions at comparable rates but with differing frequencies. The outcomes of interest included results relevant to nursing education, covering levels of medical knowledge, nurses' engagement with chatbots, and the improvement of practical skills. The study design inclusion criteria accepted any design. Detailed eligibility criteria are shown in Section 2 in Multimedia Appendix 1.

Selection Process and Outcomes of Interest

The search findings were imported into Covidence (Veritas Health Innovation) while adhering to established protocols. The screening process involved 2 stages. Initially, titles and abstracts were screened, followed by a thorough review of full-text articles. Duplicated papers were removed using Covidence prior to the screening stages to ensure the integrity of the selection process. Three authors (SZ, XL, and WW) independently and in duplicate executed all screening stages and data extraction, resolving any discrepancies through consultation with the senior author (FZ). To ensure precision and uniformity in data, we formulated a comprehensive data extraction form (SZ and WW) that underwent subsequent refinement (SZ and FZ), in alignment with guidelines from the *Cochrane Handbook for Systematic Reviews of Interventions* [25]. Before full extraction, the form underwent a pilot test on a subset of included studies. Extracted details from all included studies (SZ, XL, and WW) included elements such as publication details (study ID, title, and year), author particulars (lead author contact information), study specifics (country, study design, and objectives), and conclusions.

Study Design and Statistical Analysis

This was a narrative review. After the screening process, we successfully gathered comprehensive data, encompassing publication details (study ID, title, and year), author particulars (contact information for the lead author), study specifics (country, study design, and objectives), and conclusions. Subsequently, we categorized this data based on the respective year, country, and study design. To provide a visual representation of the trends observed, we conducted percentage calculations for each category. These percentages were then used to illustrate the trend over time and to convey the distribution of studies across various categories.

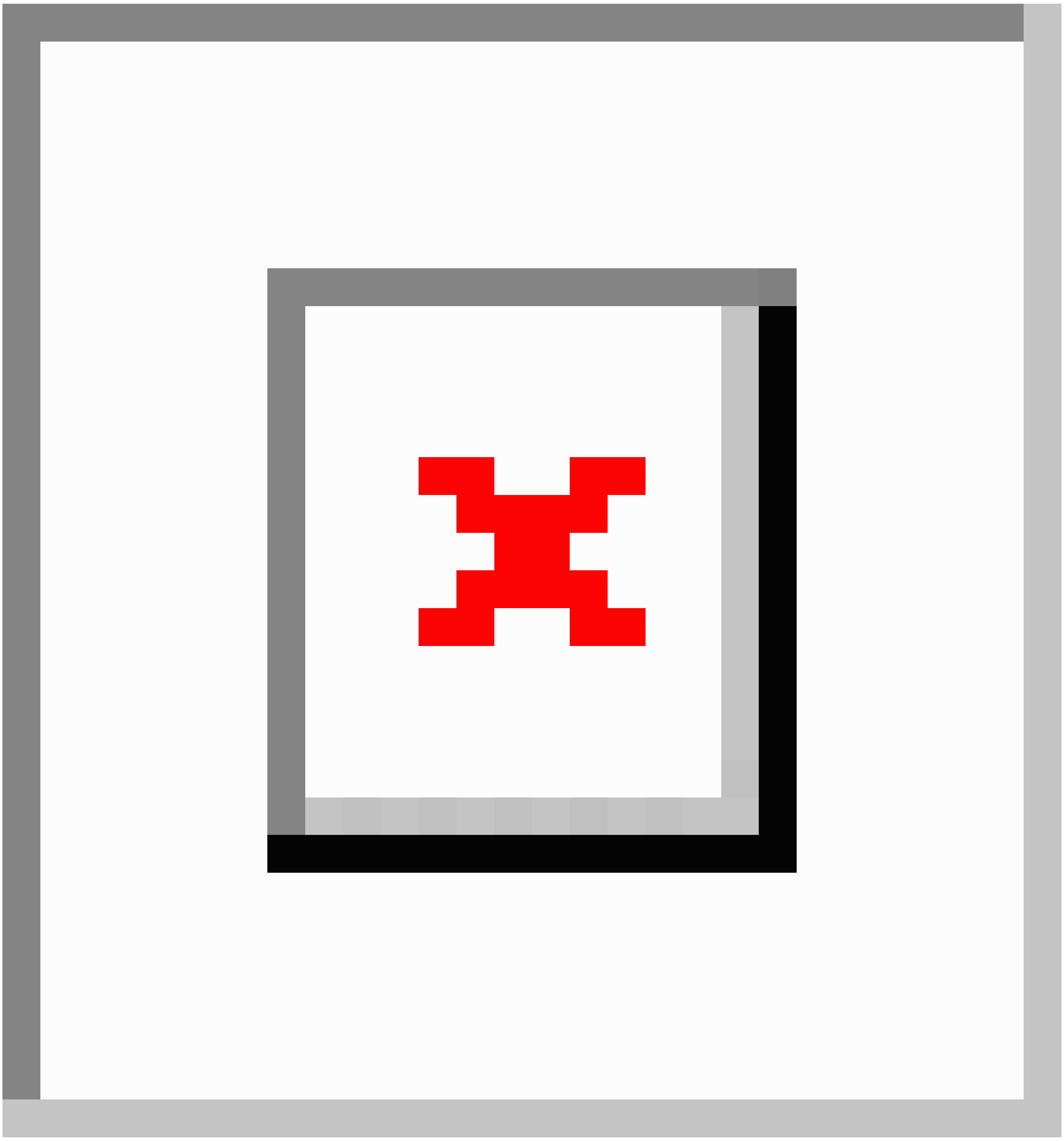
Results

In total, 38,412 distinct records were identified. Subsequently, an eligibility assessment was conducted on 77 full-text articles, with 3 articles not retrieved, as depicted in Figure 1. Out of these, 37 were subsequently excluded, resulting in the inclusion of 40 articles that met the eligibility criteria for synthesis [26-65].

Between 2010 and 2020, on average, 1 article was published every 3-4 years, culminating in a total of 3 articles, contributing to 8% of the 40 publications. However, a noticeable upswing occurred in 2021, with the publication of 3 (8%) articles. In 2022, the count increased to 6 (15%) articles. The most notable surge transpired in 2023, with the publication of 28 articles, accounting for a substantial 70% of the total publications (Figure 2).

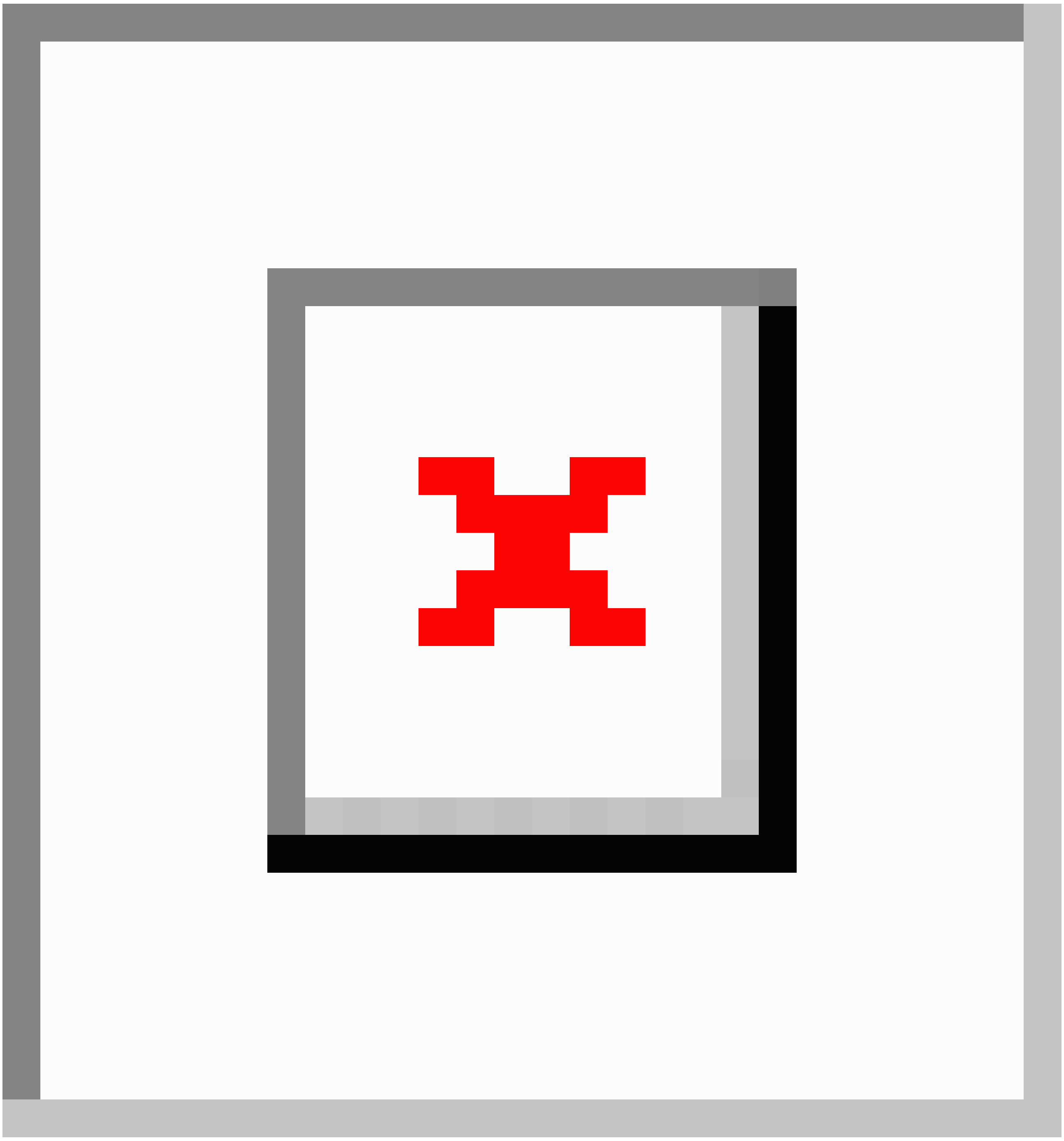
XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram showing the study selection process.



XSL•FO
RenderX

Figure 2. Temporal fluctuations in articles published from 2010 to 2023.



Taiwan province contributed 20% (8/40) of the total articles [31,32,34,35,37,42-44]. Following closely was the United States, contributing 15% (6/40) of the articles [39,40,46,52,55,59]. South Korea secured the third position, representing 10% (4/40) of the articles [41,47,48,63]. Canada [26,28,30], Mainland China [33,50,64], and Singapore [57,58,60] each contributed 8% (3/40) of the articles. Turkey [29,62] contributed 5% (2/40) of the articles. Other countries such as Australia [45], France [38], Germany [49], Hong Kong [36], India [56], Iraq [27], Italy [51], Japan [61], Malta [54], the United Kingdom [53], and Ukraine [65] each contributed 3% (1/40) of the articles.

In our review, the predominant study design was reviews, constituting 20% (8/40) of the total articles [36,46,49,50,56,59,60]. Reviews exemplify a meticulous synthesis of existing literature, providing comprehensive insights and analyses on specific topics. Editorials, comprising 18% (7/40) of the articles [28,39,45,47,52-54], serve as platforms for commentary, opinions, or perspectives on current issues and developments in the field. Commentaries constituted 10% (4/40) of the articles [26,30,35,64], offering critical reflections, analyses, or perspectives on specific subjects. Letters to the editor, making up 8% (3/40) of the articles [27,29,61], provide readers with a platform to express opinions, raise concerns, or offer feedback on published content. Quasi-experimental studies comprised 8% (3/40) of the articles [41,48,58], employing experimental methods without random assignment. Constituting 5% (2/40) of the articles, teaching tips offer valuable insights into effective educational strategies [34,55]. Randomized

XSL•FO
RenderX

controlled trials (RCTs), considered the gold standard in experimental design, constituted 5% (2/40) of the articles [42,62]. Experimental design, symbolizing systematic investigation, was embodied in 3% (1/40) of the articles [31]. Empirical articles, grounded in observations and experiences, constituted 3% (1/40) of the articles [32]. Phenomenological studies, delving into lived experiences and perceptions, comprised 3% (1/40) of the articles [33]. Proof-of-concept studies, showcasing the feasibility of an idea or approach, constituted 3% (1/40) of the articles [38]. Mini reviews, furnishing concise overviews, comprised 3% (1/40) of the articles [65]. Descriptive qualitative studies, concentrating on detailed exploration, accounted for 3% (1/40) of the articles [40]. Experimental studies, engaging in controlled testing, made up 3% (1/40) of the articles [43]. Systematic reviews, characterized by methodical literature synthesis, represented 3% (1/40) of the articles [44]. Articles centering on experimentation methodology represented 3% (1/40) of the articles [51]. Development studies, exploring the creation of new methodologies or tools, constituted 3% (1/40) of the articles [57]. Lastly, articles classified as communications, conveying crucial information or updates, represented 3% (1/40) of the articles (Table 1) [63].

XSL•FO
RenderX

Table . Overview of the extracted studies.

Study	Title	Country	Study design	Year
Abdulai and Hung [26]	Will ChatGPT Undermine Ethical Values in Nursing Education, Research, and Practice?	Canada	Commentary	2023
Ahmed [27]	The Impact of ChatGPT on the Nursing Profession: Revolutionizing Patient Care and Education	Iraq	Letter to editor	2023
Archibald and Clark [28]	ChatGTP: What Is It and How Can Nursing and Health Science Education Use It?	Canada	Editorial	2023
Berşe et al [29]	The Role and Potential Contributions of the Artificial Intelligence Language Model ChatGPT	Turkey	Letter to editor	2023
Castonguay et al [30]	Revolutionizing Nursing Education Through AI Integration: A Reflection on the Disruptive Impact of ChatGPT	Canada	Commentary	2023
Chang et al [32]	Promoting Students' Learning Achievement and Self-Efficacy: A Mobile Chatbot Approach for Nursing Training	Taiwan	Empirical article	2022
Chang et al [31]	Chatbot-Facilitated Nursing Education: Incorporating a Knowledge-Based Chatbot System Into a Nursing Training Program	Taiwan	Experimental design	2022
Chan et al [64]	Critical Reflection on Using ChatGPT in Student Learning: Benefits or Potential Risks?	China	Commentary	2023
Chen and Kuo [34]	Applying the Smartphone-Based Chatbot in Clinical Nursing Education	Taiwan	Teaching tip	2022
Chen et al [33]	Need Assessment for History-Taking Instruction Program Using Chatbot for Nursing Students: A Qualitative Study Using Focus Group Interviews	China	Phenomenological study	2023
Cheng [35]	Transformation in Nursing Education: Development and Implementation of Diverse Innovative Teaching	Taiwan	Commentary	2021
Choi et al [36]	Chatting or Cheating? The Impacts of ChatGPT and Other Artificial Intelligence Language Models on Nurse Education	Hong Kong	Review	2023
Chuang et al [37]	The Design and Application of a Chatbot in Clinical Nursing Education	Taiwan	Review	2021

XSL•FO
RenderX

Study	Title	Country	Study design	Year
Daniel et al [38]	Answering Hospital Caregivers' Questions at Any Time: Proof-of-Concept Study of an Artificial Intelligence–Based Chatbot in a French Hospital	France	Proof-of-concept study	2022
Teixeira da Silva [61]	Is ChatGPT a Valid Author?	Japan	Letter to editor	2023
Dave et al [65]	ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations	Ukraine	Mini review	2023
de Gagne [39]	The State of Artificial Intelligence in Nursing Education: Past, Present, and Future Directions	United States	Editorial	2023
Friedman and Goldschmidt [40]	Let Me Introduce You to Your First Virtual Patient	United States	Descriptive qualitative study	2014
Han et al [41]	Analysis of the Effect of an Artificial Intelligence Chatbot Educational Program on Non-Face-to-Face Classes: A Quasi-Experimental Study	South Korea	Quasi-experimental study	2022
Hsu and Chen [43]	Personalized Medical Terminology Learning Game: Guess the Term	Taiwan	Experimental study	2023
Hsu [42]	Mastering Medical Terminology With ChatGPT and Termbot	Taiwan	RCT[a]	2023
Hwang et al [44]	How Artificial Intelligence (AI) Supports Nursing Education: Profiling the Roles, Applications, and Trends of AI in Nursing Education Research (1993‐2020)	Taiwan	Systematic review	2022
Irwin et al [45]	What is ChatGPT and What Do We Do with It? Implications of the Age of AI for Nursing and Midwifery Practice and Education: An Editorial	Australia	Editorial	2023
Johnson et al [46]	When To Err Is Inhuman: An Examination of the Influence of Artificial Intelligence-Driven Nursing Care on Patient Safety	United States	Review	2023
Jung [47]	Challenges for Future Directions for Artificial Intelligence Integrated Nursing Simulation Education	South Korea	Editorial	2023
Kang et al [48]	Awareness of Using Chatbots and Factors Influencing Usage Intention Among Nursing Students in South Korea: A Descriptive Study	South Korea	Quasi-experimental study	2023
Krüger et al [49]	ChatGPT: Curse or Blessing in Nursing Care?	Germany	Review	2023

XSL•FO
RenderX

Study	Title	Country	Study design	Year
Liu et al [50]	The Application of Chat Generative Pre-trained Transformer in Nursing Education	China	Review	2023
Mascitti et al [51]	COACH BOT - Modular e-Course With Virtual Coach Tool Support	Italy	Experimentation methodology	2010
Miao and Ahn [52]	Impact of ChatGPT on Interdisciplinary Nursing Education and Research	United States	Editorial	2023
O'Connor [53]	Open Artificial Intelligence Platforms in Nursing Education: Tools for Academic Progress or Abuse?	United Kingdom	Editorial	2023
Scerri and Morin [54]	Using Chatbots Like ChatGPT to Support Nursing Practice	Malta	Editorial	2023
Seney et al [55]	Using ChatGPT to Teach Enhanced Clinical Judgment in Nursing Education	United States	Teaching tip	2023
Sharma and Sharma [56]	A Holistic Approach to Remote Patient Monitoring, Fueled by ChatGPT and Metaverse Technology: The Future of Nursing Education	India	Review	2023
Shorey et al [57]	A Virtual Counseling Application Using Artificial Intelligence for Communication Skills Training in Nursing Education: Development Study	Singapore	Development Study	2019
Shorey et al [58]	Evaluation of a Theory-Based Virtual Counseling Application in Nursing Education	Singapore	Quasi-experimental study	2023
Sun and Hoelscher [59]	The ChatGPT Storm and What Faculty Can Do	United States	Review	2023
Tam et al [60]	Nursing Education in the Age of Artificial Intelligence Powered Chatbots (AI-Chatbots): Are We Ready Yet?	Singapore	Review	2023
Uslu and van Giersbergen [62]	The Effects of Manikin-Based and Standardized-Patient Simulation on Clinical Outcomes: A Randomized Prospective Study	Turkey	RCT	2023
Ye et al [63]	Development of a Chatbot Program for Follow-Up Management of Workers' General Health Examinations in Korea: A Pilot Study	South Korea	Communication	2021

[a]RCT: randomized controlled trial.

XSL•FO
RenderX

Discussion

Principal Findings

In this paper, we comprehensively examined the temporal trends, international distribution, study designs, and implications of chatbots in nursing education to map the challenges and issues to address in the future. Our analysis highlights significant findings, including a marked increase in research publications in 2023, reflecting growing interest in this field. Contributions from Taiwan province, the United States, and South Korea illustrate the global scope of chatbot integration in nursing education. The diverse study designs reviewed, ranging from reviews and editorials to quasi-experimental studies, indicate the extensive research exploring chatbots' role in enhancing personalized instruction, patient-care simulations, and critical thinking. Despite these advancements, challenges such as the need for rigorous original research, funding, ethical considerations, and resource distribution disparities remain. Furthermore, addressing these issues through international collaboration and targeted research will be crucial for fully realizing the potential of chatbots in nursing education.

AI language models such as chatbots have caused a revolution in nursing education through the provision of personalized and interactive learning activities. Chatbots are implemented in nursing education for personalized instruction, patients-care simulation, and critical thinking enhancement. Chatbots in health care are used for teleconsultation to improve communication skills, support clinical judgment, and enable remote patient monitoring. Chatbots are a key component in addressing the global shortages of knowledge and resources in nursing training. They bridge theoretical and practical aspects, thereby illustrating the potential of this technology to revolutionize learning processes and change the face of health care services and education.

This study aims to shed light on the evolution of chatbots in nursing education through data analysis of temporal trends. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram facilitates a systematic search procedure, which guarantees a transparent and strict methodology. Indeed, articles published in 2023 accounted for 70% (28/40) of the included articles, which might be due to either increased scholarly interest or intensified effort. This study tries to delve into the technological education aspect of health care, which is a rapidly expanding area. Consequently, it will provide a comprehensive reflection of the dynamic and developing educational sector.

This study provides a new approach about how AI and mobile communication can be applied in and influence nursing education. Chatbots and AI integration can be seen as a technical invention with thrilling effects on mental health, behavior, and knowledge in relation to the field of deep learning. The analysis stresses the sole benefits of chatbots in education, that is, chatbots provide the capacity for individualized learning [27,31,32,39,43,44,47,48,50,51,53,56,60]. The studies focus on problems in nursing education that involve the shortage of global knowledge, condition differences, and lack of relationship between theory and practice [29,35,45,49,58] and illustrate the ways chatbots can cope with these issues.

A detailed study of the worldwide distribution and categorization of chatbot research on nursing education is carried out with reference to international contexts, highlighting major contributions. The participation of United States and South Korea is notable, and Taiwan province has the largest share, accounting for 20% (8/40) of all articles. This regional perspective highlights the universal nature of adding chatbots to nursing education. As the research methodology analysis reveals, reviews cover 20% (8/40) of the articles, providing exhaustive summaries of the present literature. A diverse range of designs that includes commentaries, quasi-experimental studies, teaching tips, and RCTs explains the extensive and varied research on chatbots in nursing education.

In spite of the huge benefits, there are some barriers that nursing education will face as they try to incorporate chatbots. Original research such as RCTs or cohort studies is the most important part of confirming the efficiency of conversational bots. Funding research about advanced techniques and the application of rigorous process need high levels both of staff and finance. The integrity and the security problems of chatbots that provide wrong advice are highlighted, demonstrating the need for correcting the technical problems in order to ensure ethical and secure operations. Funding should be set aside to close resource distribution disparities [39,40,47,57-59], so that students from disadvantaged backgrounds can also have an opportunity to have access to technologically advanced educational resources. Collaboration among those in the academic, technical, and health care disciplines is indispensable as an effort to develop supportive surroundings for the application of chatbots to nursing education globally.

This study demonstrates the substantial changes that chatbots bring into nursing education to make nursing practice more enjoyable. This integration aims at resolving several issues, including the lack of competitiveness from a global perspective and economic disparity, in essence to establish an integrated and dynamic learning environment. Analyzing the small components of chatbots and conducting research on the feasibility, pros, and cons are necessary aims for the future of education [44]. The lack of original research forces us to rely more on the already existing qualitative studies such as commentaries and editorials. Above all, great attention should be given to privacy and ethics when integrating current technologies into the health care education system.

There are some limitations. First, the study only provides a description of the changes over time in articles related to chatbots in nursing education, as well as the distribution of regions and types of articles. Due to the lack of original studies, it does not show the characteristics of papers included in the final analysis. Second, there is uncertainty about whether the specific research topics related to chatbots in nursing education are consistent between countries. Third, there is a lack of in-depth quantitative exploration and discussion regarding the specific application directions of chatbots in nursing education, preventing the formulation of more constructive recommendations.

XSL•FO
RenderX

Conclusion

Integrating chatbots into nursing education presents a promising yet relatively unexplored avenue. This review highlights the urgent need for original research, emphasizing the importance of ethical considerations. This exploration contributes to the evolving landscape of technology in health care education, bridging gaps and fostering a learner-centric approach aligned with contemporary health care demands.

Authors' Contributions

SZ contributed to conceptualization, methodology, data curation, formal analysis, writing—original draft preparation, and writing—review and editing. XL contributed to methodology, data curation, and writing—original draft preparation. WW contributed to methodology, data curation, and writing—original draft preparation. FZ contributed to conceptualization, methodology, project administration, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategies and eligibility criteria for study inclusion.

[DOCX File, 30 KB - mededu_v10i1e54987_app1.docx]

References

1. Langendyk V, Hegazi I, Cowin L, Johnson M, Wilson I. Imagining alternative professional identities: reconfiguring professional boundaries between nursing students and medical students. Acad Med 2015 Jun;90(6):732-737. [doi: 10.1097/ACM.0000000000000714] [Medline: 25901875]
2. Suikkala A, Koskinen S, Leino-Kilpi H. Patients' involvement in nursing students' clinical education: a scoping review. Int J Nurs Stud 2018 Aug;84:40-51. [doi: 10.1016/j.ijnurstu.2018.04.010] [Medline: 29763831]
3. King R, Taylor B, Talpur A, et al. Factors that optimise the impact of continuing professional development in nursing: a rapid evidence review. Nurse Educ Today 2021 Mar;98:104652. [doi: 10.1016/j.nedt.2020.104652] [Medline: 33190952]
4. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010 Dec 4;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5] [Medline: 21112623]
5. Bhutta ZA, Chen L, Cohen J, et al. Education of health professionals for the 21st century: a global independent commission. Lancet 2010 Apr 3;375(9721):1137-1138. [doi: 10.1016/S0140-6736(10)60450-3] [Medline: 20362799]
6. Pepito JA, Locsin R. Can nurses remain relevant in a technologically advanced future? Int J Nurs Sci 2019 Oct 4;6(1):106-110. [doi: 10.1016/j.ijnss.2018.09.013] [Medline: 31406875]
7. Alhalal E, Alrashidi LM, Alanazi AN. Predictors of patient-centered care provision among nurses in acute care setting. J Nurs Manag 2020 Sep;28(6):1400-1409. [doi: 10.1111/jonm.13100] [Medline: 32667691]
8. Challinor JM, Alqudimat MR, Teixeira TOA, Oldenmenger WH. Oncology nursing workforce: challenges, solutions, and future strategies. Lancet Oncol 2020 Dec;21(12):e564-e574. [doi: 10.1016/S1470-2045(20)30605-7] [Medline: 33212044]
9. Chen FQ, Leng YF, Ge JF, et al. Effectiveness of virtual reality in nursing education: meta-analysis. J Med Internet Res 2020 Sep 15;22(9):e18290. [doi: 10.2196/18290] [Medline: 32930664]
10. Kim SK, Lee Y, Yoon H, Choi J. Adaptation of extended reality smart glasses for core nursing skill training among undergraduate nursing students: usability and feasibility study. J Med Internet Res 2021 Mar 2;23(3):e24313. [doi: 10.2196/24313] [Medline: 33650975]
11. Marć M, Bartosiewicz A, Burzyńska J, Chmiel Z, Januszewicz P. A nursing shortage - a prospect of global and local policies. Int Nurs Rev 2019 Mar;66(1):9-16. [doi: 10.1111/inr.12473] [Medline: 30039849]
12. Hashish EAA. Nurses' perception of organizational justice and its relationship to their workplace deviance. Nurs Ethics 2020 Feb;27(1):273-288. [doi: 10.1177/0969733019834978] [Medline: 30982425]
13. Zhou Y, Li Z, Li Y. Interdisciplinary collaboration between nursing and engineering in health care: a scoping review. Int J Nurs Stud 2021 May;117:103900. [doi: 10.1016/j.ijnurstu.2021.103900] [Medline: 33677250]
14. Mlambo M, Silén C, McGrath C. Lifelong learning and nurses' continuing professional development, a metasynthesis of the literature. BMC Nurs 2021 Apr 14;20(1):62. [doi: 10.1186/s12912-021-00579-2] [Medline: 33853599]
15. Leaver CA, Stanley JM, Goodwin Veenema T. Impact of the COVID-19 pandemic on the future of nursing education. Acad Med 2022 Mar 1;97(3S):S82-S89. [doi: 10.1097/ACM.0000000000004528] [Medline: 34789661]
16. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
17. Singh B, Olds T, Brinsley J, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023 Jun 23;6(1):118. [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]

XSL•FO
RenderX

18. Yang HS, Wang F, Greenblatt MB, Huang SX, Zhang Y. AI chatbots in clinical laboratory medicine: foundations and trends. Clin Chem 2023 Nov 2;69(11):1238-1246. [doi: 10.1093/clinchem/hvad106] [Medline: 37664912]
19. Kuhail MA, Alturki N, Alramlawi S, Alhejori K. Interacting with educational chatbots: a systematic review. Educ Inf Technol 2023 Jan;28(1):973-1018. [doi: 10.1007/s10639-022-11177-3]
20. Smutny P, Schreiberova P. Chatbots for learning: a review of educational chatbots for the Facebook Messenger. Computers & Education 2020 Jul;151:103862. [doi: 10.1016/j.compedu.2020.103862]
21. Suhaili SM, Salim N, Jambli MN. Service chatbots: a systematic review. Expert Syst Appl 2021 Dec;184:115461. [doi: 10.1016/j.eswa.2021.115461]
22. Garcia Valencia OA, Thongprayoon C, Jadlowiec CC, Mao SA, Miao J, Cheungpasitporn W. Enhancing kidney transplant care through the integration of chatbot. Healthcare (Basel) 2023 Sep 12;11(18):2518. [doi: 10.3390/healthcare11182518] [Medline: 37761715]
23. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. Int J Educ Technol High Educ 2023 Oct 31;20(1):56. [doi: 10.1186/s41239-023-00426-1]
24. Amir-Behghadami M, Janati A. Population, intervention, comparison, outcomes and study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. Emerg Med J 2020 Jun;37(6):387. [doi: 10.1136/emermed-2020-209567] [Medline: 32253195]
25. Higgins JPT, Thomas J, Chandler J, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.4: Cochrane; 2023. URL: www.training.cochrane.org/handbook [accessed 2024-06-04]
26. Abdulai AF, Hung L. Will ChatGPT undermine ethical values in nursing education, research, and practice? Nurs Inq 2023 Jul;30(3):e12556. [doi: 10.1111/nin.12556] [Medline: 37101311]
27. Ahmed SK. The impact of ChatGPT on the nursing profession: revolutionizing patient care and education. Ann Biomed Eng 2023 Nov;51(11):2351-2352. [doi: 10.1007/s10439-023-03262-6] [Medline: 37266721]
28. Archibald MM, Clark AM. ChatGTP: what is it and how can nursing and health science education use it? J Adv Nurs 2023 Oct;79(10):3648-3651. [doi: 10.1111/jan.15643] [Medline: 36942780]
29. Berşe S, Akça K, Dirgar E, Kaplan Serin E. The role and potential contributions of the artificial intelligence language model ChatGPT. Ann Biomed Eng 2024 Feb;52(2):130-133. [doi: 10.1007/s10439-023-03296-w] [Medline: 37378876]
30. Castonguay A, Farthing P, Davies S, et al. Revolutionizing nursing education through AI integration: a reflection on the disruptive impact of ChatGPT. Nurse Educ Today 2023 Oct;129:105916. [doi: 10.1016/j.nedt.2023.105916] [Medline: 37515957]
31. Chang CY, Kuo SY, Hwang GH. Chatbot-facilitated nursing education: incorporating a knowledge-based chatbot system into a nursing training program. Educ Tech Soc 2022 Jan;25(1):15-27 [FREE Full text]
32. Chang CY, Hwang GJ, Gau ML. Promoting students' learning achievement and self‐efficacy: a mobile chatbot approach for nursing training. Br J Educ Technol 2022 Jan;53(1):171-188. [doi: 10.1111/bjet.13158]
33. Chen Y, Lin Q, Chen X, et al. Need assessment for history-taking instruction program using chatbot for nursing students: a qualitative study using focus group interviews. Digit Health 2023 Jun;9:20552076231185435. [doi: 10.1177/20552076231185435] [Medline: 37426591]
34. Chen YT, Kuo CL. Applying the smartphone-based chatbot in clinical nursing education. Nurse Educ 2022;47(2):E29. [doi: 10.1097/NNE.0000000000001131] [Medline: 34711755]
35. Cheng SF. Transformation in nursing education: development and implementation of diverse innovative teaching [Article in Chinese]. Hu Li Za Zhi 2021 Dec;68(6):4-5. [doi: 10.6224/JN.202112_68(6).01] [Medline: 34839484]
36. Choi EPH, Lee JJ, Ho MH, Kwok JYY, Lok KYW. Chatting or cheating? the impacts of ChatGPT and other artificial intelligence language models on nurse education. Nurse Educ Today 2023 Jun;125:105796. [doi: 10.1016/j.nedt.2023.105796] [Medline: 36934624]
37. Chuang YH, Chen YT, Kuo CL. The design and application of a chatbot in clinical nursing education [Article in Chinese]. Hu Li Za Zhi 2021 Dec;68(6):19-24. [doi: 10.6224/JN.202112_68(6).04] [Medline: 34839487]
38. Daniel T, de Chevigny A, Champrigaud A, et al. Answering hospital caregivers' questions at any time: proof-of-concept study of an artificial intelligence-based chatbot in a French hospital. JMIR Hum Factors 2022 Oct 11;9(4):e39102. [doi: 10.2196/39102] [Medline: 35930555]
39. de Gagne JC. The state of artificial intelligence in nursing education: past, present, and future directions. Int J Environ Res Public Health 2023 Mar 10;20(6):4884. [doi: 10.3390/ijerph20064884] [Medline: 36981790]
40. Friedman SA, Goldschmidt K. Let me introduce you to your first virtual patient. J Pediatr Nurs 2014;29(3):281-283. [doi: 10.1016/j.pedn.2014.03.021] [Medline: 24704180]
41. Han JW, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ 2022 Dec 1;22(1):830. [doi: 10.1186/s12909-022-03898-3] [Medline: 36457086]
42. Hsu MH. Mastering medical terminology with ChatGPT and Termbot. Health Edu J 2023 Jun;83(4):352-358. [doi: 10.1177/00178969231197371]
43. Hsu MH, Chen YH. Personalized medical terminology learning game: guess the term. Games Health J 2023 Apr;13(2):84-92. [doi: 10.1089/g4h.2023.0054] [Medline: 37699207]

XSL•FO
RenderX

44. Hwang GJ, Tang KY, Tu YF. How artificial intelligence (AI) supports nursing education: profiling the roles, applications, and trends of AI in nursing education research (1993–2020). Interact Learn Environ 2022 Jun 26;32(1):1-20. [doi: 10.1080/10494820.2022.2086579]
45. Irwin P, Jones D, Fealy S. What is ChatGPT and what do we do with it? implications of the age of AI for nursing and midwifery practice and education: an editorial. Nurse Educ Today 2023 Aug;127:105835. [doi: 10.1016/j.nedt.2023.105835] [Medline: 37267643]
46. Johnson EA, Dudding KM, Carrington JM. When to err is inhuman: an examination of the influence of artificial intelligence-driven nursing care on patient safety. Nurs Inq 2024 Jan;31(1):e12583. [doi: 10.1111/nin.12583] [Medline: 37459179]
47. Jung S. Challenges for future directions for artificial intelligence integrated nursing simulation education. Korean J Women Health Nurs 2023 Sep;29(3):239-242. [doi: 10.4069/kjwhn.2023.09.06.1] [Medline: 37813667]
48. Kang SR, Kim SJ, Kang KA. Awareness of using chatbots and factors influencing usage intention among nursing students in South Korea: a descriptive study. Child Health Nurs Res 2023 Oct;29(4):290-299. [doi: 10.4094/chnr.2023.29.4.290] [Medline: 37939675]
49. Krüger L, Krotsetis S, OpenAI's Generative Pretrained Transformer 3 (GPT-3) Model, Nydahl P. ChatGPT: curse or blessing in nursing care? [Article in German]. Med Klin Intensivmed Notfmed 2023 Oct;118(7):534-539. [doi: 10.1007/s00063-023-01038-3] [Medline: 37401955]
50. Liu J, Liu F, Fang J, Liu S. The application of Chat Generative Pre-trained Transformer in nursing education. Nurs Outlook 2023;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
51. Mascitti I, Feituri M, Funghi F, Correnti S, Galassi L. COACH BOT - modular e-course with virtual coach tool support. In: Filipe J, Fred A, Sharp B, editors. Proceedings of the 2nd International Conference on Agents and Artificial Intelligence - Volume 2: ICAART: SciTePress; 2010:115-120. [doi: 10.5220/0002589901150120]
52. Miao H, Ahn H. Impact of ChatGPT on interdisciplinary nursing education and research. Asian Pac Isl Nurs J 2023 Apr 24;7:e48136. [doi: 10.2196/48136] [Medline: 37093625]
53. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? Nurse Educ Pract 2023 Jan;66:103537. [doi: 10.1016/j.nepr.2022.103537] [Medline: 36549229]
54. Scerri A, Morin KH. Using chatbots like ChatGPT to support nursing practice. J Clin Nurs 2023 Aug;32(15-16):4211-4213. [doi: 10.1111/jocn.16677] [Medline: 36880216]
55. Seney V, Desroches ML, Schuler MS. Using ChatGPT to teach enhanced clinical judgment in nursing education. Nurse Educ 2023;48(3):124. [doi: 10.1097/NNE.0000000000001383] [Medline: 36857593]
56. Sharma M, Sharma S. A holistic approach to remote patient monitoring, fueled by ChatGPT and Metaverse technology: the future of nursing education. Nurse Educ Today 2023 Dec;131:105972. [doi: 10.1016/j.nedt.2023.105972] [Medline: 37757713]
57. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A virtual counseling application using artificial intelligence for communication skills training in nursing education: development study. J Med Internet Res 2019 Oct 29;21(10):e14658. [doi: 10.2196/14658] [Medline: 31663857]
58. Shorey S, Ang ENK, Ng ED, et al. Evaluation of a theory-based virtual counseling application in nursing education. Comput Inform Nurs 2023 Jun 1;41(6):385-393. [doi: 10.1097/CIN.0000000000000999] [Medline: 36728150]
59. Sun GH, Hoelscher SH. The ChatGPT storm and what faculty can do. Nurse Educ 2023;48(3):119-124. [doi: 10.1097/NNE.0000000000001390] [Medline: 37043716]
60. Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. Nursing education in the age of artificial intelligence powered chatbots (AI-chatbots): are we ready yet? Nurse Educ Today 2023 Oct;129:105917. [doi: 10.1016/j.nedt.2023.105917] [Medline: 37506622]
61. Teixeira da Silva JA. Is ChatGPT a valid author? Nurse Educ Today 2023 Mar;68:103600. [doi: 10.1016/j.nepr.2023.103600] [Medline: 36906947]
62. Uslu Y, van Giersbergen MY. The effects of manikin-based and standardized-patient simulation on clinical outcomes: a randomized prospective study. Cyprus J Med Sci 2023 Aug 9;8(4):271-275. [doi: 10.4274/cjms.2022.2022-12]
63. Ye BJ, Kim JY, Suh C, et al. Development of a chatbot program for follow-up management of workers' general health examinations in Korea: a pilot study. Int J Environ Res Public Health 2021 Feb 23;18(4):2170. [doi: 10.3390/ijerph18042170] [Medline: 33672158]
64. Chan MMK, Wong ISF, Yau SY, Lam VSF. Critical reflection on using ChatGPT in student learning: benefits or potential risks? Nurse Educ 2023;48(6):E200-E201. [doi: 10.1097/NNE.0000000000001476] [Medline: 37348135]
65. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

PICOS: Population, Intervention, Comparison, Outcome, and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by TDA Cardoso; submitted 29.11.23; peer-reviewed by H Feng, J Shen, S Bisht, T Hebda; revised version received 16.05.24; accepted 22.05.24; published 13.06.24.

Please cite as:
Zhang F, Liu X, Wu W, Zhu S
Evolution of Chatbots in Nursing Education: Narrative Review
JMIR Med Educ 2024;10:e54987
URL: https://mededu.jmir.org/2024/1/e54987
doi:10.2196/54987

© Fang Zhang, Xiaoliu Liu, Wenyan Wu, Shiben Zhu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.6.2024. 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.

XSL•FO
RenderX

Review

Measuring the Digital Competence of Health Professionals: Scoping Review

Anne Mainz[1], MSc; Julia Nitsche[2], MSc; Vera Weirauch[1,3], MA; Sven Meister[1,3], Prof Dr

[1]Health Informatics, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[2]Department of Didactics and Educational Research in Health Science, Faculty of Health, School of Medicine, Witten/Herdecke University, Witten, Germany

[3]Department Healthcare, Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Corresponding Author:
Anne Mainz, MSc
Health Informatics
Faculty of Health, School of Medicine
Witten/Herdecke University
Pferdebachstraße 11
Witten, 58448
Germany
Phone: 49 2302 926 78627
Email: anne.mainz@uni-wh.de

Abstract

Background: Digital competence is listed as one of the key competences for lifelong learning and is increasing in importance not only in private life but also in professional life. There is consensus within the health care sector that digital competence (or digital literacy) is needed in various professional fields. However, it is still unclear what exactly the digital competence of health professionals should include and how it can be measured.

Objective: This scoping review aims to provide an overview of the common definitions of digital literacy in scientific literature in the field of health care and the existing measurement instruments.

Methods: Peer-reviewed scientific papers from the last 10 years (2013-2023) in English or German that deal with the digital competence of health care workers in both outpatient and inpatient care were included. The databases ScienceDirect, Scopus, PubMed, EBSCOhost, MEDLINE, OpenAIRE, ERIC, OAIster, Cochrane Library, CAMbase, APA PsycNet, and Psyndex were searched for literature. The review follows the JBI methodology for scoping reviews, and the description of the results is based on the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.

Results: The initial search identified 1682 papers, of which 46 (2.73%) were included in the synthesis. The review results show that there is a strong focus on technical skills and knowledge with regard to both the definitions of *digital competence* and the measurement tools. A wide range of competences were identified within the analyzed works and integrated into a validated competence model in the areas of technical, methodological, social, and personal competences. The measurement instruments mainly used self-assessment of skills and knowledge as an indicator of competence and differed greatly in their statistical quality.

Conclusions: The identified multitude of subcompetences illustrates the complexity of digital competence in health care, and existing measuring instruments are not yet able to reflect this complexity.

(JMIR Med Educ 2024;10:e55737) doi:10.2196/55737

KEYWORDS

digital competence; digital literacy; digital health; health care; health care professional; health care professionals; scoping review

XSL•FO
RenderX

Introduction

Background

The 2006 European Parliament recommendation on key competences for lifelong learning lists digital competences as 1 of the 8 key competences for every citizen to enable personal fulfillment, active citizenship, social cohesion, and employability in our modern society [1]. Therefore, it is no surprise that the digital transformation within the health care sector, involving new processes and technologies [2], has completely changed the demands on people working in health care professions. Digital competence in health care is needed [3,4]. According to Vitello et al [5], competence is “the ability to integrate and apply contextually-appropriate knowledge, skills and psychosocial factors (e.g., beliefs, attitudes, values and motivations) to consistently perform successfully within a specified domain.” Salman et al [6] divide competence into 2 aspects: hard and soft. The hard aspects of competence include knowledge, skill, and behavior, whereas the soft aspects include character traits, motives, attitudes, values, and self-image. Together, all these aspects determine the performance or output—both visible and invisible—of an individual in a particular job. *Competence*, in contrast to *competency*, is attached to the person rather than to a task or activity [5], which fits better within this work because we are focusing not on specific digital activities but on how professionals deal with digital technologies when working in the health care domain. This is why we concentrate on competence in this work.

The updated version of the digital competence framework for citizens (DigComp 2.2) [7] divides digital competences for private individuals into 5 main dimensions: information and data literacy, communication and collaboration, digital content and creation, safety, and problem-solving. Specific knowledge, skills, and attitudes are assigned to each of these dimensions. Along with the requirements for digital competence in private life, there are certain requirements to be met before one can be considered digitally competent in professional life in the health care sector.

Unfortunately, to date, there is no standard definition for the construct *digital competence* within the health care domain. Although the topic of interest is *digital competence*, the term *digital literacy* was also considered because this term is more common in English-speaking countries, and both concepts are often used synonymously [8]. Currently, for both terms, different understandings exist [9]. In this review, the semantic meaning of the terms is important, that is, *the skills and characteristics required to navigate the (professional) digital world*.

The lack of a uniform definition also leads to problems in determining digital competence for health professionals: authors criticize the lack of validated and up-to-date instruments to measure digital literacy or digital competence in this field [10,11]. With existing measurement tools, the focus is solely on technical skills; the related aspects that also affect the use of digital technologies are neglected [10].

Therefore, the objective of this research was to create an overview of how digital competence is defined and measured among health care professionals and thus to provide a holistic picture.

Research Questions

Primarily, the following questions will be answered with the help of the literature review:

- What definitions exist of the digital competence of health care professionals?
  - What are the similarities and differences among the various definitions?
  - On which basic models are the different definitions based?
- What possibilities exist for measuring the digital competence of health care professionals?
  - Which dimensions of digital competence are measured?
  - How are the dimensions measured (self-assessment, performance tasks, etc)?
  - Have the assessment tools been validated? What quality criteria have been applied?

Methods

Overview

To provide a systematic overview of existing research literature on digital literacy in health professions, we conducted a scoping review [12]. The review follows the JBI methodology for scoping reviews [13] (based on the works of Arksey and O’Malley [14] and Levac et al [15]), which follows these steps: (1) defining and aligning the objectives and questions; (2) developing and aligning the inclusion criteria with the objectives and questions; (3) describing the planned approach to evidence searching, selection, data extraction, and presentation of the evidence; (4) searching for the evidence; (5) selecting the evidence; (6) analysis of the evidence; (7) presentation of the results; and (8) summarizing the evidence in relation to the purpose of the review, making conclusions, and noting any implications of the findings.

The review was planned beforehand by AM and SM, including choosing the review method, formulating the research questions, selecting the databases, phrasing the search terms, and determining the eligibility criteria. AM screened the search results, during which process there was regular professional exchange with another author, VW. The results were reviewed by SM, VW, and JN. AM, SM, VW, and JN all have experience in conducting scoping reviews.

To ensure the high quality and informative value of the results report, the description of the results is based on the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [12,16] (Multimedia Appendix 1). In addition, an evaluation protocol was prepared in advance of the review and made publicly available on OSF [17].

Search Strategy

The literature search took place in April 2023 and used the databases ScienceDirect, Scopus, PubMed, EBSCOhost (which provides results from MEDLINE, OpenAIRE, ERIC, and

XSL•FO
RenderX

OAIster), Cochrane Library, CAMbase, APA PsycNet, and Psyndex. The search term used was as follows: (“digital competence” OR “digital literacy”) AND (“medical professional” OR “healthcare professional” OR “healthcare worker” OR “physician assistant” OR “health professional”).

Fixed combinations of terms (such as digital literacy) are placed in quotation marks. Parentheses are used to force the right evaluation order of the expression. No adjacent terms were added so as not to make assumptions about the nature of the terms of interest. These were combined with various health worker designations. Neutral terms were chosen for the designation of nonmedical personnel to achieve a neutral and comprehensive understanding for different health professions. The keywords were linked with the Boolean operator “OR” to show results with at least one of the given terms. The operator “AND” ensures that all search results contain both “digital competence” or “digital literacy” and a health worker designation. The search term was developed through several trial cycles of a combination of terms. These were entered into the different databases and, based on the search results, terms were added or removed. The results are shown in Table 1.

Table 1. Results of the database search. The search term (“digital competence” OR “digital literacy”) AND (“medical professional” OR “healthcare professional” OR “healthcare worker” OR “physician assistant” OR “health professional”) was used for each database (N=1682).

Database	Results, n (%)
ScienceDirect	594 (35.31)
Scopus	361 (21.46)
PubMed	15 (0.89)
EBSCOhost (MEDLINE, OpenAIRE, ERIC, and OAIster)	706 (41.97)
Cochrane Library	6 (0.36)
CAMbase	0 (0)
APA PsycNet	0 (0)
Psyndex	0 (0)

Eligibility Criteria

This scoping review considered peer-reviewed publications that were research articles, book chapters, review articles, or conference papers published within the last 10 years (2013-2023). Papers in either English or German were included.

The articles address the digital competence of health care workers in both outpatient and inpatient care. They come from medical, technical, or educational research fields. Papers from the patient’s perspective or those that address eHealth literacy or digital health literacy, defined as the “skills, knowledge and resources to search for, find, understand, evaluate and apply health information [from the internet]” [18], were excluded because the concept of interest is more concerned with the understanding of information rather than with the professional use of digital technologies. The overall eligibility criteria for this scoping review are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria for the scoping review.

Inclusion criteria

- Peer-reviewed publications
- Research articles, book chapters, review articles, or conference papers
- Research field: medical, technical, or educational
- Subject: articles addressing digital competence or digital literacy
- Population: health care workers in both outpatient and inpatient care and students and graduates of health care professions
- Period: articles published from 2013 to 2023
- Language: English or German

Exclusion criteria

- Not peer-reviewed publications
- Research field: any research field other than medical, technical, or educational
- Subject: articles addressing eHealth literacy or digital health literacy
- Population: patients
- Period: articles published before 2013
- Language: other than English or German

XSL•FO
RenderX

Article Screening and Data Extraction

According to the recommendations of Moher et al [19], these steps are followed in the study selection process: first, duplicates are removed from the initial search results, after which the remaining publications are evaluated based on their titles, keywords, and abstracts and, subsequently, checked for suitability based on the full texts. The eligible papers are included in the review [19]. We followed the recommended process and, from the eligible papers, extracted and listed the following data in a Microsoft Excel sheet that was developed a priori but refined iteratively: authors, year of publication, country of origin, type of survey, and target group.

Synthesis of Results

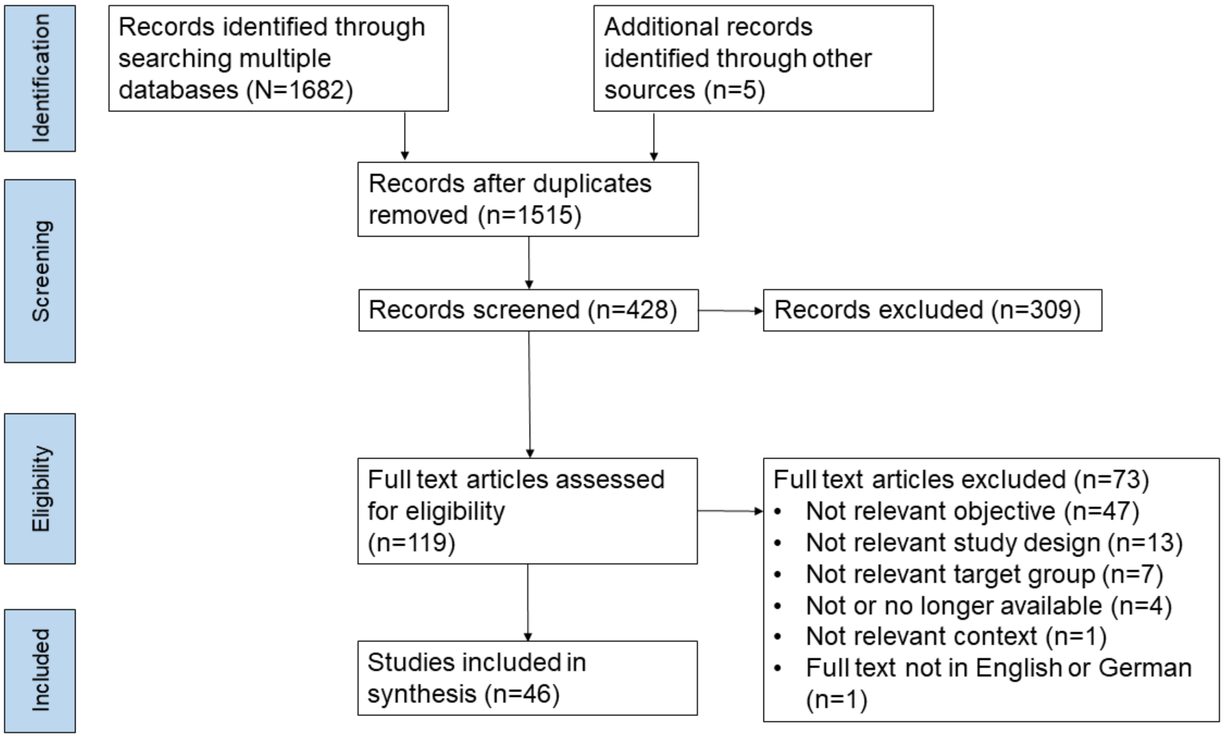
We present the characteristics of the selected studies, with a comparison of the drafted definitions of digital competence. In addition, we report the fundamental frameworks, models, and research papers that originally specified these definitions. We have collected and clustered all competences mentioned in the eligible papers. The structuring of the competences identified in the works was based on the competence categories according to the competence model developed by Hecklau et al [20], who cluster competences into technical, methodological, social, and personal competences to achieve clarity and transparency of the competence model. This clustering was adopted within our work to organize the determined competences. Finally, we explicitly examine the papers in which digital literacy assessment tools are used, with a consideration of the origin of the questionnaires, the form of measurement, and an assessment of their statistical quality.

Results

Selection of Sources of Evidence

The initial search identified 1682 papers (Table 1), of which 1510 (89.77%) remained after duplicates were removed. After applying the inclusion criteria (time period, type, and language) and screening the titles, of the 1510 papers, 428 (28.34%) were available for preselection, which, after the screening of the abstracts, reduced to 119 (27.8%) titles. Finally, after consideration of the full texts, of the 1682 papers identified through the initial search, 46 (2.73%) were included in this scoping review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the number of articles identified, screened, assessed for eligibility, and included in the final analysis.



Characteristics of Sources of Evidence

The selected papers were largely published from 2020 onward (35/46, 76%), indicating an increase in the perceived relevance of digital literacy among health professionals in the scientific world. In the years prior (2013-2019), only 11 (24%) of the 46 papers were published, with a slightly perceptible increase from 1 (9%) paper in 2014 to 4 (36%) in 2019. Of the 46 papers, the maximum number was published in 2020 (n=15, 33%); in subsequent years, the number of papers decreased to 8 (17%) in 2021 and 6 (13%) in 2022, and in 2023, a total of 6 (13%) papers had been published until May of that year. Table 2 shows the key data of the included papers.

XSL•FO
RenderX

Table 2. Key data of the included papers.

Authors	Year	Country	Type of study	Target group
Awami [21]	2020	Libya	Quantitative study	Health care professionals
Barbosa et al [22]	2023	Austria, Belgium, Croatia, Denmark, Finland, France, Italy, Malta, Netherlands, Norway, Poland, Portugal, and United Kingdom	Quantitative study	Radiotherapists
Brice and Almond [23]	2020	Australia	Scoping review	Health care professionals
Brown et al [24]	2020	Australia	Quantitative study	Nurses
Burzynska et al [25]	2023	Poland	Quantitative study	Physicians
Butler-Henderson et al [26]	2020	Australia	Meta-analysis	Health care professionals
Cabero-Almenara et al [27]	2021	Spain	Quantitative study	Health science lecturers
Cham et al [28]	2022	Australia	Quantitative study	Students of health professions
Coldwell-Neilson et al [9]	2019	Australia	Framework development	Optometry students
Evangelinos and Holley [29]	2014	United Kingdom	Qualitative interview	Health care professionals
Faihs et al [30]	2022	Germany	Quantitative study	Medical students
Golz et al [31]	2021	Switzerland	Quantitative study	Health care professionals
Hallit et al [32]	2020	Lebanon	Quantitative study	Pharmacists
Hilty et al [33]	2021	United States	Scoping review	Health care professionals
Holt et al [34]	2020	Denmark	Quantitative study	Nursing students
Jarva et al [35]	2022	Finland	Qualitative interview	Health care professionals
Jarva et al [36]	2023	Finland	Questionnaire development	Health care professionals
Jimenez et al [37]	2020	Singapore	Scoping review	Health care professionals
Jose et al [38]	2023	Chile	Scoping review	Health care professionals
Kaihlanen et al [39]	2021	Finland	Quantitative study	Nurses
Kayser et al [40]	2022	Denmark	Quantitative study	Health care professionals
Kim and Jeon [41]	2020	South Korea	Quantitative study	Nursing students
Konttila et al [42]	2019	Finland	Systematic review	Health care professionals
Kuek and Hakkennes [11]	2020	Australia	Quantitative study	Health care professionals
Longhini et al [10]	2022	Italy	Systematic review	Health care professionals
MacLure and Steward [43]	2018	United Kingdom	Qualitative interview	Pharmacists
MacLure and Steward [44]	2016	United Kingdom	Systematic review	Pharmacists
Matthews [45]	2021	United Kingdom	Systematic review	Health graduates
McGregor et al [46]	2017	Australia	Qualitative interview	Health graduates
Montebello et al [47]	2016	Malta	SWOT[a] analysis	Students of health professions
Nazeha et al [4]	2020	Singapore	Scoping review	Health care professionals
O'Connor and LaRue [48]	2021	United Kingdom	Framework development	Nurses
Oo et al [49]	2021	Myanmar	Quantitative study	Health care professionals
Poncette et al [50]	2020	Germany	Mixed methods study	Medical students
Pontefract and Wilson [51]	2019	United Kingdom	Qualitative interview	Health care professionals
Rachmani et al [52]	2020	Indonesia	Quantitative study	Health care professionals
Reixach et al [53]	2022	Spain	Quantitative study	Health care professionals
Shiferaw et al [54]	2020	Ethiopia	Quantitative study	Health care professionals

XSL•FO
RenderX

Authors	Year	Country	Type of study	Target group
Skiba et al [55]	2017	United States	Historical development report	Health care professionals
Tegegne et al [56]	2023	Ethiopia	Quantitative study	Health care professionals
Värri et al [57]	2020	Finland	Framework development	Students of health professions
Vehko et al [58]	2019	Finland	Quantitative study	Nurses
Virtanen et al [59]	2021	Finland	Systematic review	Health care professionals
Vissers et al [60]	2018	International	Quantitative study	Physiotherapy students
Whittaker et al [61]	2020	New Zealand	Quantitative study	Health care professionals
Wubante et al [62]	2023	Ethiopia	Quantitative study	Health care professionals

[a]SWOT: strengths, weaknesses, opportunities, and threats.

The majority of the articles were published in Australia (7/46, 15%), Finland (7/46, 15%), and the United Kingdom (6/46, 13%). The remaining papers were distributed worldwide: Ethiopia (3/46, 7%); Denmark, Germany, Singapore, Spain, and the United States (2/46, 4% each); and Chile, Indonesia, Italy, Lebanon, Libya, Malta, Myanmar, New Zealand, Poland, South Korea, Switzerland, 13 countries in Europe (Austria, Belgium, Croatia, Denmark, Finland, France, Italy, Malta, Netherlands, Norway, Poland, Portugal, and the United Kingdom), and the rest of the world (1/46, 2% each).

The types of papers were mainly distributed between quantitative studies (23/46, 50%) and reviews (scoping reviews, systematic reviews, and meta-analyses; 11/46, 24%). Less represented were qualitative interviews (5/46, 11%) and framework development (3/46, 7%), as well as questionnaire development; mixed methods study; strengths, weaknesses, opportunities, and threats analysis; and historical development report (1/46, 2% each).

The papers' target group was largely unspecific, with most of them addressing *health care professionals* (25/46, 54%). Other papers addressed specifically *nurses* (4/46, 9%), *pharmacists* (3/46, 7%), *health graduates* (2/46, 4%), *health science lecturers* (1/46, 2%), *physicians* (1/46, 2%), and *radiotherapists* (1/46, 2%). Some of the papers were aimed at students: students of health professions in general (3/46, 7%), medical students and nursing students (2/46, 4% each), and optometry students and physiotherapy students (1/46, 2% each).

Definition of Data Literacy

The main difficulty concerning the literature analysis was that some of the papers used the term *digital literacy* but actually referred to a different concept (especially *eHealth literacy*). When selecting the papers for review, articles that dealt, in terms of semantics, with concepts other than *data literacy* were sorted out.

Most of the papers provided definitions in which digital competence is composed of various dimensions of competence. There was a strong focus on skills in the formulated definitions of digital competence [9,21,22,25-33,35,36,39,40,42,43,45,47,48,50-55,59,60,62]. Many papers (27/46, 59%) also stated in their definitions that certain kinds of knowledge are necessary for competence [4,10,22,23,25-28,30-33,36,39,40,42,47,49-55,59,60,62]. Some of the papers (17/46, 37%) proposed that the attitude toward technical issues should be considered a component of competence [4,10,11,21,24,27,28,30-33,36,42,47,49,54,59]. Other papers (6/46, 13%) added that former experiences with digital topics play a crucial role in forming competence [28,31,40,42,43,46]. According to Konttila et al [42], experiences are the base for the emergence of attitudes. Other works mentioned motivation (7/46, 15%) [31,35,36,40,42,57,59], practices (2/46, 4%) [9,31], consciousness (2/46, 4%) [9,54], fears (2/46, 4%) [11,43], goals (1/46, 2%) [25], identity (1/46, 2%) [9], self-awareness (1/46, 2%) [28], and strategies (1/46, 2%) [54] as part of competence. These competence dimensions provide a framework for the required competence areas, which are described in the *Identified Competence Areas and Competences* subsection.

The definitions used are either the results of scoping reviews or frameworks where many individual results have been merged (15/46, 33%) [4,10,23,26,28,33,37,42,44,46,48,51,52,59,61]. Alternatively, they are based on other, explicitly named works, such as DigComp 2.2 [7] (4/46, 9%) [22,29,54,56]; the European framework for the digital competence of educators [63] (1/46, 2%) [27]; the technology acceptance model [64] and the unified theory of acceptance and use of technology [65] (1/46, 2%) [11]; the accreditation of competence in information and communication technologies by the government of Catalonia [66] (1/46, 2%) [53]; the Educause Center for Analysis and Research [67] (1/46, 2%) [60]; the General Confidence with Computer Use Scale [68] (1/46, 2%) [32]; the eHealth literacy questionnaire [69] (1/46, 2%) [40]; the eHealth literacy assessment toolkit [70] (1/46, 2%) [34]; the Self-Assessment of Nursing Informatics Competencies Scale [71] (1/46, 2%) [24]; a scale assessing the informatics competencies for nurses [72] (1/46, 2%) [39]; a scale assessing digital literacy with regard to information and communication technology [73] (1/46, 2%) [41]; the definition by Konttila et al [42] (1/46, 2%) [31]; the definition by Ferrari [74] (1/46, 2%) [21]; the definition by Bawden [75] (1/46, 2%) [25]; the definition by Sharpe and Beetham [76] (1/46, 2%) [9]; the definition by Hecklau et al [20] (1/46, 2%) [38]; the definition by Gretton and Honeymen [77] (2/46, 4%) [43,44]; the Health Education England definition [78] (1/46, 2%) [45]; the Jisc 7 elements of digital literacies (1/46, 2%) [47]; the World Health Organization's *Electronic Health Records: A Manual For Developing Countries* [79]

XSL•FO
RenderX

(1/46, 2%) [49]; and the definition by Skiba et al [80] (1/46, 2%) [57]. No information was provided in 4 (9%) of the 46 studies [30,50,58,62] about the basis of the definition used. Montebello et al [47] refers to the Jisc 7 elements of digital literacies as basis for their digital literacy definition but the original source is not available anymore.

Identified Competence Areas and Competences

Overview

Within the included papers, competences in the 4 main competence areas according to the model developed by Hecklau et al [20] were identified: multiple competences could be grouped into technical, methodological, social, and personal competences. All these competences, classified into 4 competence areas, are described in the following paragraphs and depicted in Textbox 2.

Textbox 2. The identified competences grouped into different competence areas.

Competence areas and competences

- Technical competences
- Basic computer competence [4,9,11,21-25,27-29,32,33,35-39,41,43-45,47-49,51-54,56,57,62]
- Basic competence to use wireless devices [21,23-25,37,49]
- Applied digital health skills [4,10,22,24,26,29,30,33,35,37,39,40,42,43,46,48,50-53,55,57,58,61,62]
- Anticipation of advanced and future digital competences [30,37,38,41,48,50,57]
- Administration of technology [4,23,45]
- Ethical aspects of digitalization [4,36,37,48,50,57,58]
- Legal aspects of digitalization [4,37,48,50,52]
- Methodological competences
- Data and information processing competence [4,9,21,22,24-26,29-31,35,37,38,40,41,44,45,47,48,50-57,62]
- Continuous learning [4,9,23,25,28-30,32,38,41,45-47,49,54,55,57,62]
- Project management [4,57,61]
- Research competence [4,37,45,47,57]
- Problem-solving [22,35,38,41,54,56,62]
- Social competences
- Working in teams [9,23,29,35,38,41,42,45,47,50,51,53-55,62]
- Communication competence [4,9,22,29-31,35,36,38,42,43,45,47,49-51,54-57,59,62]
- Networking skills [38,47,50]
- Teaching [27,45]
- Focus on patients [4,10,35-37,48,50,55,57]
- Personal competences
- Innovative behavior [23,38,45,50]
- Self-reflection [35,53,54]
- Critical thinking [22,25,54]
- Creativity [38,54]
- Professionalism [23]

Technical Competences

Multiple subcompetences of technical competences were identified: the ones mentioned most often were *basic computer competence*, meaning knowledge of different computer components and basic computer concepts [21,32,43]; and skills in using hardware (eg, switching equipment on and off and operating input and output devices) [49,62]. Internet use, consisting of navigating the internet, knowledge of various internet sources, and finding and downloading articles, is part of basic computer competence [24,25,28,37,43,52,62]. The users should be able to use and install software [24,28,32,33,37,49,52,62] and especially be able to use information and communication technology, including understanding the basic concepts and components of information and communication technology and designing, creating, integrating, publishing, and revising content [4,9,22,23,27,35-38,41,43-45,47-49,53,54,56,57,62]. Another part of basic computer competence is file management and

XSL•FO
RenderX

comprehensive knowledge of file formats, the creation of documents and folder structure [37,49], and IT security (eg, using passwords and antivirus tools) [22,29,37,38,45,52,54,56,62].

Another subcompetence mentioned was *basic competence to use wireless devices*, consisting of operating hardware [49], using the internet [21,37], managing files [21,37], and using applications [21,37].

Existing competences can be transferred to eHealth contexts to achieve the foundation for *applied digital health skills* [46]. Here, one of the largest areas is the use of health applications, meaning the use of various digital health solutions for treatment planning, diagnostics, treatment, processing imaging data, and so on [22,24,33,35,40,42,48,57,58]. This includes the management of electronic patient records [22,24,37,43,49,51,57,58,62], the use of wearables and mobile health apps [30,57], the administration of electronic documentation [4,37], and the use of health information systems [37,52,55,57]. In addition, health professionals need skills and knowledge about specific data protection and security requirements of their profession [4,30,48,53]. Furthermore, digitally competent health care workers need to be able to establish new technologies in their work environments and participate in the design, implementation, and evaluation of systems, as well as seek available resources, formulate ethical decisions technical wise, and promote the use of IT in health environments [4,24,42,48,50,57].

A further subcompetence is the *anticipation of advanced and future digital competences*, where users stay informed about the current state of the art of digital technologies and the competences that are necessary to use these [38,41], as well as how certain technologies will develop in the future, which play a role in the future of health care (eg, big data, artificial intelligence, robotics, and genomics) [30,37,48,50].

One crucial aspect of technical competence is the *administration of technology*, which encompasses planning, implementation, optimization, and operation or management, as well as the control of technological products or tools, processes, and services [4,23,45].

Knowledge about *ethical aspects* [4,36,37,48,50,57,58], such as freedom of choice, privacy, autonomy, and fairness [36], as well as the *legal aspects of digitalization* [4,37,48,50,52,62], in particular regarding the regulation of medical practice and medical devices [50] and the protection of patient data as well as confidentiality when processing data [52], is equally important when handling new technologies to enable data protection and data security.

Methodological Competences

The competence to *process data and information* consists of finding [4,23,24,26,37,44,47,52,53,62], evaluating [21,23,25,37,43,47,50-53,57,62], creating [23,24,44,49,51], managing [4,23,24,26,29,30,47-49,52,53,57], sharing or communicating [4,23,26,31,44,47,53,57], analyzing [4,26,37,50,53], visualizing [4], and interpreting [24,26,47,49] information or data; deriving actions or decisions [50]; being well versed in data protection and security [50,51]; and knowing the difference among data, information, and knowledge [48].

In addition, the ability to *continuously learn* is a fundamental component of digital competence. Learning is described as using educational methods such as teaching, training, storytelling, discussion, and targeted research to acquire knowledge, skills, values, beliefs, and habits [23]. It includes the anticipation of service and training needs and, for future digital literacy skills [57], learning how to use new technologies [29,49,62] and acquiring new concepts, methods, and tools [23], especially by using digital teaching and learning resources [4,29,41,47].

Digitally competent health professionals should also be proficient in *project management* to be able to introduce new operating models and lead IT-based change in their field [4,57,61].

They should be able to use IT for research support and innovations [4] as well as for assessment and continuous improvement of their own skills, their work community skills development, and the development of electronic services [57] through *research competence*.

Problem-solving competence can be interpreted as both dealing with digital problems [22,35,38,54,56] and solving problems through digital means [41,54,56,62].

Social Competences

To engage digitally in the social work environment, digitally competent health professionals must be able to *work in teams*, meaning they should be able to work cooperatively or collaboratively [9,23,38,41,45,47,50,53,62]; take a leadership role [38]; deal with diverse teams consisting of members with different demographics, from different professions, and with different personality traits [38,51]; be willing to compromise for the sake of group harmony [38]; and establish collegial support to create positive digital experiences [35,42].

Another basic requirement to work in (digital) teams is *communication competence* using a wide range of communication methods [50], including digital communication [4,9,30,38,57,62] (eg, web-based meetings and consultations and the use of social media [57] within the team [36,57] and with patients [35,36]). Digitally competent health professionals need to know the correct vocabulary [57] and, with this knowledge, the ability to share knowledge [38].

Networking skills are evident in the use of knowledge networks, where health professionals participate in digital networks for learning and research and develop an open-access mentality [38,47,50].

Health professionals should not only be able to gain knowledge but also to pass it on: *teaching* is an important part of digital literacy. Health professionals could impart their knowledge using digital resources and provide these resources to learners, assess their learning success, and increase not only their own but also the learners' digital literacy [27,45].

Another important part of digital literacy is keeping the *focus on patients* by considering the patients' digital needs and evaluating their digital skills, as well as considering their

XSL•FO
RenderX

willingness to use digital services to provide services that they feel safe to use and capable of using [35,57]. In addition, health professionals should promote the use of IT among patients through support and empowerment for self-management, IT guidance (eg, guides and web-based materials), and support in finding information [4,57].

Personal Competences

To be digitally competent, health professionals need *innovative behavior* as a personality trait, meaning they should have the spirit of invention and lifelong determination [23,38,45,50]. The initiative to conceive, consider, try out, or apply new ideas, products, processes, and procedures to their individual work role or their work unit without fear of change [23] is essential to drive the transformation process of health care forward [50].

Another relevant ability for health professionals is *self-reflection* with regard to their own digital competence [35,53,54] and the identification of personal and professional needs to apply technical solutions [53].

Other personal traits mentioned as relevant for digital competence are *critical thinking* [22,25,54] and *creativity* [38,54]. Critical thinking is mentioned in connection with information evaluation [25] or gaining new information within a professional context [22,54]. Creativity is of use when knowledge is built up [54] or a task has to be approached with an innovative mindset [38].

Professionalism is defined as the behavior, demeanor, and attitude of a person in a work environment and is considered a useful quality rather than a requirement of a role [23], but it is a characteristic that is beneficial to health professionals wishing to be digitally competent.

Measurement Instruments

Of the 46 included papers, 25 (54%) used different questionnaires to evaluate the digital literacy of health professionals. The majority of the questionnaires used (15/25, 60%) [21,22,25,28,30-32,36,49,50,52,53,58,61,62] were developed originally for these papers. Others used existing questionnaires or frameworks (Textbox 3) such as the Self-Assessment of Nursing Informatics Competencies Scale [71] in the study by Brown et al [24]; a scale assessing the informatics competencies for nurses [72] in the study by Kaihlanen et al [39]; the eHealth literacy assessment toolkit [70] in the study by Holt et al [34]; the eHealth literacy questionnaire [69] in the study by Kayser et al [40]; the General Confidence with Computer Use Scale [68] in the study by Hallit et al [32]; the attitudes and digital literacy toward information and communication technology scale [73] in the study by Kim and Yeon [41]; the Educause Center for Analysis and Research [67] in the study by Vissers et al [60]; the technology acceptance model [64] and the unified theory of acceptance and use of technology [65] in the study by Kuek and Hakkennes [11]; DigComp 2.2 [7] in the studies by Barbosa et al [22], Shiferaw et al [54], and Tegegne et al [56]; the European framework for the digital competence of educators [63] in the study by Cabero-Almenara et al [27]; and the accreditation of competence in information and communication technologies by the government of Catalonia [66] in the study by Reixach et al [53].

XSL•FO
RenderX

Textbox 3. Underlying work for the questionnaires used in the studies.

Underlying work and corresponding studies

- Technology acceptance model [64] and unified theory of acceptance and use of technology [65]
- Kuek and Hakkennes [11]
- Updated version of the digital competence framework for citizens [7]
- Barbosa et al [22], Shiferaw et al [54], and Tegegne et al [56]
- Self-Assessment of Nursing Informatics Competencies Scale [71]
- Brown et al [24]
- Informatics competencies scale for nurses [72]
- Kaihlanen et al [39]
- eHealth literacy assessment toolkit [70]
- Holt et al [34]
- eHealth literacy questionnaire [69]
- Kayser et al [40]
- General Confidence with Computer Use Scale [68]
- Hallit et al [32]
- Attitudes and digital literacy toward information and communication technology scale [73]
- Kim and Yeon [41]
- Educause Center for Analysis and Research [67]
- Vissers et al [60]
- European framework for the digital competence of educators [63]
- Cabero-Almenara et al [27]
- Accreditation of competence in information and communication technologies by the government of Catalonia [66]
- Reixach et al [53]

Digital literacy was measured in various forms, and some questionnaires used different combinations of measurement forms (Textbox 4). The specific items of the questionnaires considered in the review are categorized thematically herein. In many surveys, participants provided a self-assessment of specific skills and knowledge. Often, they had to assign certain abilities or confidence levels to themselves [11,22,24,25,27,28,30-32,34,36,39-41,49,52-54,56,58,61,62]. Other questionnaires collected participants' attitudes toward technical topics [11,21,24,30,31,36,40,41,50,62]. Some items dealt with the experiences or needs of participants with regard to (further) training in digital topics [21,25,30,49,50,53,56,62]. Another way of measuring digital literacy involved requesting access to different technologies, such as smartphones, laptop computers, or tablet devices, for private or professional use [28,32,49,60,62] or the frequency of use of these technologies [11,25,28,40,60]. Other items addressed user behavior: what the devices were used for [24,49,60], and which applications were used [21,24].

The questionnaires differed greatly in their statistical quality. Some have not been validated in any statistical form [21,25,28,39,50,58,60-62], whereas others were only tested on internal consistency [41,49,53], and several were verified with different reliability and validity tests [11,22,24,27,30-32,34,36,40,52,54,56].

XSL•FO
RenderX

Textbox 4. Different measurement forms of digital literacy with item examples.

Measurement form and item examples

- Self-assessment [11,22,24,25,27,28,30-32,34,36,39-41,49,52-54,56,58,61,62]
- "I can use the most common computer programs and services (e.g. email, intranet) in my work." [36]
- "How well do you feel you master the following skills required to use information systems?" [58]
- Attitudes [11,21,24,30,31,36,40,41,50,62]
- "I believe that new digital technologies will fundamentally change medicine in the next few years." [30]
- "The transfer to digital services is a positive change." [36]
- Experiences, needs of education, or training [21,25,30,49,50,53,56,62]
- "I would benefit from additional trainings/courses in the field of shaping digital competences." [25]
- "On a personal level, would you like to have specific training in any of the following areas? eg. Digital culture, participation and citizenship using digital tools." [53]
- Access to technology [28,32,49,60,62]
- "Do you think you have internet access in your office?" [62]
- "Owning a computer." [32]
- Frequency of use [11,25,28,40,60]
- "Please state how often you use the following in your work and in your personal life: computers, Microsoft Office applications, smartphones, tablets, email, the internet, and social media (i.e. Facebook, Twitter and Instagram)." [11]
- "How often do you use the internet?" [60]
- User behavior [21,24,49,60]
- "I use MS Excel for work." [21]
- "What is the purpose of [sic] you use a computer?: work, education, communication, entertainment, and playing games" [49]

Discussion

Principal Findings

The selected literature sources show the increasing scientific interest in digital literacy in health care and the worldwide spread of this development. There is a focus on quantitative research, although, because the available survey instruments were considered insufficient to determine digital literacy, researchers often developed their own. The underlying definitions are based on a variety of approaches and sources, which highlights the need for a structured overview. Most of the definitions focused on skills and knowledge as indicators of competence. *Soft aspects*, as described by Salman et al [6], were also mentioned by authors but less frequently and in many different forms. Attitude, experience, and motivation were mentioned most often. Behavior, which is a *hard aspect* according to Salman et al [6], was not addressed explicitly in the definitions provided in the included papers.

The identified competences have been categorized according to the competence categories formulated by Hecklau et al [20]. The determined technical competences include basic computer competence, basic competence to use wireless devices, applied digital health skills, anticipation of advanced and future digital competences, administration of technology, ethical aspects of digitalization, and legal aspects of digitalization. Data and information processing competence, continuous learning, project management, research competence, and problem-solving were mentioned in the literature as methodological competences. The following were classified as social competences: working in teams, communication competence, networking skills, teaching, and focus on patients. Personal competences include innovative behavior, self-reflection, critical thinking, creativity, and professionalism.

The results confirm that existing measurement tools focus solely on technical areas [10], and other related aspects, such as the identified competences from the methodological, social, and personal areas in other nonquantitative works, have not been taken into account. Unlike what Longhini et al [10] and Kuek and Hakkennes [11] stated, many of the questionnaires used had high statistical quality and were verified with different reliability and validity tests. The questionnaires largely measure digital literacy via self-assessment. Some also use items relating to attitudes, experiences, access to technology, frequency of use, and use behavior.

The allocation of competences to the categories was sometimes not trivial and not clearly distinguishable; for example, *teaching* could be categorized as both a social and a methodological competence. How the partial competence areas are connected also remains unanswered in these works. Hurst [81] describes 3 possible dependency relationships: a general factor model where basic competence is composed of equally important subaspects, an additive model where the individual subaspects have a juxtaposed relationship, or a hierarchical model where basic subcompetences and higher-level competences exist that build on each other [81]. A more complex consideration of the relationships among the individual competences, for example,

XSL•FO
RenderX

through a factor analysis, would also be conceivable and should be investigated in subsequent research work. Some of the skills identified are specifically linked to digital topics, but others are more general and *analog* in nature, especially in the social and personal categories. Therefore, mutual influences among the competences are not only conceivable but also probable.

Limitations

One limitation of this literature review is that, because of the very nature of scoping reviews, the quality of the included works was not considered in the review process, and all papers were included in the synthesis, irrespective of quality [14]. This may have led to inferior works being included in the results and being placed on an equal footing with high-quality works. When constructing the search term, no wildcards were used, which limited the search of potential fitting literature, which must be specified as a further limitation. In addition, more variants of the job title *medical professional* could have been used to maximize the search results. Another limitation could be the practical implementation of the selection of papers and their evaluation by just 1 author. Although the procedure was planned as a team, and the results were discussed extensively, the process was carried out by only 1 person.

Future Directions

This literature review focuses solely on the terms *digital competence* and *digital literacy* and provides an overview of the use of these closely related terms. A larger literature review that includes other adjacent topics, such as *informatics competences*, or refers to specific digital activities in the health care sector, such as *telemedicine competences*, would heighten the credibility in terms of an overall semantic understanding of the concept of competence when dealing with all sorts of digital technologies. Within this work, which aimed at an understanding of the specifically named term *digital competence*, the addition of related concepts would not be possible without the development of an initial understanding of this concept, which the authors have developed in the course of this work.

A further enrichment of an in-depth analysis would be the addition of specific medical specialties. The aim of this work was the nonspecific and generalizable consideration of required digital skills in health care, but, of course, every profession has its individual (digital) requirements that are worth considering.

Conclusions

The review shows that the interest in digital literacy as a research topic in health care is currently on the rise but that the understanding of this rather abstract term is widely divergent. A uniform definition and use of terms is needed. The existence of hard and soft aspects of competence, as described by Salman et al [6], was confirmed by many of the used definitions, but which of the identified aspects contribute to what extent needs to be investigated further. Furthermore, the multitude of subcompetences illustrates the complexity of digital competence that needs to be taken into account when developing a measurement instrument. Well-validated questionnaires exist, these focus solely on technical aspects. The competency model identified in this work can be used as a starting point for factor analysis of the identified competences or questionnaire development.

Acknowledgments

This research was funded by the Bundesministerium für Gesundheit (BMG; German Federal Ministry of Health; ZMI5-2523FEP30B).

Data Availability

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

Authors' Contributions

All authors contributed to the conceptualization, formal analysis, visualization, validation, and writing of the original and revised drafts. AM and SM developed the methodology design. AM conducted the literature screening and data curation. SM contributed supervision and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.

[PDF File (Adobe PDF File), 515 KB - mededu_v10i1e55737_app1.pdf]

References

1. Recommendation of the European parliament and of the council of 18 December 2006 on key competences for lifelong learning. Official Journal of the European Union. 2006 Dec 30. URL: https://tinyurl.com/3r4ne9bk [accessed 2024-03-09]
2. Kraus S, Schiavone F, Pluzhnikova A, Invernizzi AC. Digital transformation in healthcare: analyzing the current state-of-research. J Bus Res 2021 Feb;123:557-567. [doi: 10.1016/j.jbusres.2020.10.030]

XSL•FO
RenderX

3. Davies AC, Davies A, Abdulhussein H, Hooley F, Eleftheriou I, Hassan L, et al. Educating the healthcare workforce to support digital transformation. Stud Health Technol Inform 2022 Jun 06;290:934-936. [doi: 10.3233/SHTI220217] [Medline: 35673156]
4. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 05;22(11):e22706 [FREE Full text] [doi: 10.2196/22706] [Medline: 33151152]
5. Vitello S, Greatorex J, Shaw S. What is competence? A shared interpretation of competence to support teaching, learning and assessment. Cambridge University Press & Assessment. 2021. URL: https://tinyurl.com/4ke9yr7c [accessed 2024-03-09]
6. Salman M, Ganie SA, Saleem I. The concept of competence: a thematic review and discussion. Eur J Train Dev 2020 May 25;44(6/7):717-742. [doi: 10.1108/ejtd-10-2019-0171]
7. European Commission, Joint Research Centre, Vuorikari R, Kluzer S, Punie Y. DigComp 2.2, The Digital Competence framework for citizens – With new examples of knowledge, skills and attitudes. Publications Office of the European Union. 2022. URL: https://op.europa.eu/en/publication-detail/-/publication/50c53c01-abeb-11ec-83e1-01aa75ed71a1/language-en [accessed 2024-03-09]
8. Spante M, Hashemi SS, Lundin M, Algers A. Digital competence and digital literacy in higher education research: systematic review of concept use. Cogent Educ 2018 Oct 23;5(1):1519143. [doi: 10.1080/2331186x.2018.1519143]
9. Coldwell-Neilson J, A Armitage J, J Wood-Bradley R, Kelly B, Gentle A. Implications of updating digital literacy – a case study in an optometric curriculum. IISIT 2019;16:033-049. [doi: 10.28945/4285]
10. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]
11. Kuek A, Hakkennes S. Healthcare staff digital literacy levels and their attitudes towards information systems. Health Informatics J 2020 Mar 15;26(1):592-612 [FREE Full text] [doi: 10.1177/1460458219839613] [Medline: 30983476]
12. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
13. Peters MD, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: scoping reviews (2020 version). In: Aromataris E, Munn Z, editors. Joanna Briggs Institute Reviewer's Manual. Adelaide, Australia: Joanna Briggs Institute; 2020.
14. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
15. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021 Mar 29;10(1):89 [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
17. Mainz A. Digital literacy of health professionals: a scoping review. Open Science Framework. 2023 Feb. URL: https://osf.io/9whxu [accessed 2024-03-14]
18. Busse TS, Nitsche J, Kernebeck S, Jux C, Weitz J, Ehlers JP, et al. Approaches to improvement of digital health literacy (eHL) in the context of person-centered care. Int J Environ Res Public Health 2022 Jul 07;19(14):8309 [FREE Full text] [doi: 10.3390/ijerph19148309] [Medline: 35886158]
19. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-9, W64 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
20. Hecklau F, Galeitzke M, Flachs S, Kohl H. Holistic approach for human resource management in industry 4.0. Procedia CIRP 2016 Dec;54:1-6 [FREE Full text] [doi: 10.1016/j.procir.2016.05.102]
21. Awami S. Computer competency as an indicator of healthcare institution readiness for health information systems: a study in Benghazi. In: Proceedings of the IADIS International Conference e-Health 2020. 2020 Presented at: IADIS 2020; July 21-23, 2020; Virtual Event. [doi: 10.33965/eh2020_202009l014]
22. Barbosa B, Oliveira C, Bravo I, Couto JG, Antunes L, McFadden S, et al. An investigation of digital skills of therapeutic radiographers/radiation therapists: a European survey of proficiency level and future educational needs. Radiography (Lond) 2023 May;29(3):479-488 [FREE Full text] [doi: 10.1016/j.radi.2023.02.009] [Medline: 36878157]
23. Brice S, Almond H. Health professional digital capabilities frameworks: a scoping review. J Multidiscip Healthc 2020;13:1375-1390 [FREE Full text] [doi: 10.2147/JMDH.S269412] [Medline: 33173300]
24. Brown J, Morgan A, Mason J, Pope N, Bosco AM. Student nurses' digital literacy levels: lessons for curricula. Comput Inform Nurs 2020 Mar 13;38(9):451-458. [doi: 10.1097/CIN.0000000000000615] [Medline: 33955370]
25. Burzyńska J, Bartosiewicz A, Januszewicz P. Dr. Google: physicians-the web-patients triangle: digital skills and attitudes towards e-health solutions among physicians in south eastern Poland-a cross-sectional study in a pre-COVID-19 era. Int J Environ Res Public Health 2023 Jan 05;20(2):978 [FREE Full text] [doi: 10.3390/ijerph20020978] [Medline: 36673740]

XSL•FO
RenderX

26. Butler-Henderson K, Dalton L, Probst Y, Maunder K, Merolli M. A meta-synthesis of competency standards suggest allied health are not preparing for a digital health future. Int J Med Inform 2020 Dec;144:104296. [doi: 10.1016/j.ijmedinf.2020.104296] [Medline: 33091830]
27. Cabero-Almenara J, Barroso-Osuna J, Gutiérrez-Castillo JJ, Palacios-Rodríguez A. The teaching digital competence of health sciences teachers. a study at Andalusian Universities (Spain). Int J Environ Res Public Health 2021 Mar 04;18(5):2552 [FREE Full text] [doi: 10.3390/ijerph18052552] [Medline: 33806483]
28. Cham K, Edwards ML, Kruesi L, Celeste T, Hennessey T. Digital preferences and perceptions of students in health professional courses at a leading Australian university: a baseline for improving digital skills and competencies in health graduates. Australas J Educ Technol 2021 Sep 20;38(1):69-86. [doi: 10.14742/ajet.6622]
29. Evangelinos G, Holley D. A qualitative exploration of the EU digital competence (DIGCOMP) framework: a case study within healthcare education. In: Proceedings of the E-Learning, E-Education, and Online Training. 2014 Presented at: eLEOT 2014; September 18-20, 2014; Bethesda, MD. [doi: 10.1007/978-3-319-13293-8_11]
30. Faihs V, Figalist C, Bossert E, Weimann K, Berberat PO, Wijnen-Meijer M. Medical students and their perceptions of digital medicine: a question of gender? Med Sci Educ 2022 Oct 02;32(5):941-946 [FREE Full text] [doi: 10.1007/s40670-022-01594-x] [Medline: 36276758]
31. Golz C, Peter KA, Müller TJ, Mutschler J, Zwakhalen SM, Hahn S. Technostress and digital competence among health professionals in swiss psychiatric hospitals: cross-sectional study. JMIR Ment Health 2021 Nov 04;8(11):e31408 [FREE Full text] [doi: 10.2196/31408] [Medline: 34734840]
32. Hallit S, Tawil S, Sacre H, Rahme C, Hajj A, Salameh P. Lebanese pharmacists' confidence and self-perceptions of computer literacy: scale validation and correlates. J Pharm Policy Pract 2020 Aug 24;13(1):44 [FREE Full text] [doi: 10.1186/s40545-020-00246-y] [Medline: 32855813]
33. Hilty DM, Torous J, Parish MB, Chan SR, Xiong G, Scher L, et al. A scoping review to develop a framework of asynchronous technology competencies for psychiatry and medicine. J Technol Behav Sci 2021 Jan 07;6(2):231-251. [doi: 10.1007/s41347-020-00185-0]
34. Holt KA, Overgaard D, Engel LV, Kayser L. Health literacy, digital literacy and eHealth literacy in Danish nursing students at entry and graduate level: a cross sectional study. BMC Nurs 2020 Apr 10;19(1):22 [FREE Full text] [doi: 10.1186/s12912-020-00418-w] [Medline: 32308559]
35. Jarva E, Oikarinen A, Andersson J, Tuomikoski AM, Kääriäinen M, Meriläinen M, et al. Healthcare professionals' perceptions of digital health competence: a qualitative descriptive study. Nurs Open 2022 Mar 30;9(2):1379-1393 [FREE Full text] [doi: 10.1002/nop2.1184] [Medline: 35094493]
36. Jarva E, Oikarinen A, Andersson J, Tomietto M, Kääriäinen M, Mikkonen K. Healthcare professionals' digital health competence and its core factors; development and psychometric testing of two instruments. Int J Med Inform 2023 Mar;171:104995 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.104995] [Medline: 36689840]
37. Jimenez G, Spinazze P, Matchar D, Koh Choon Huat G, van der Kleij RM, Chavannes NH, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020 Nov;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
38. Jose A, Tortorella GL, Vassolo R, Kumar M, Mac Cawley AF. Professional competence and its effect on the implementation of healthcare 4.0 technologies: scoping review and future research directions. Int J Environ Res Public Health 2022 Dec 28;20(1):478 [FREE Full text] [doi: 10.3390/ijerph20010478] [Medline: 36612799]
39. Kaihlanen AM, Gluschkoff K, Kinnunen UM, Saranto K, Ahonen O, Heponiemi T. Nursing informatics competences of Finnish registered nurses after national educational initiatives: a cross-sectional study. Nurse Educ Today 2021 Nov;106:105060 [FREE Full text] [doi: 10.1016/j.nedt.2021.105060] [Medline: 34315050]
40. Kayser L, Karnoe A, Duminski E, Jakobsen S, Terp R, Dansholm S, et al. Health professionals' eHealth literacy and system experience before and 3 months after the implementation of an electronic health record system: longitudinal study. JMIR Hum Factors 2022 Apr 29;9(2):e29780 [FREE Full text] [doi: 10.2196/29780] [Medline: 35486414]
41. Kim S, Jeon J. Factors influencing eHealth literacy among Korean nursing students: a cross-sectional study. Nurs Health Sci 2020 Sep 24;22(3):667-674. [doi: 10.1111/nhs.12711] [Medline: 32154981]
42. Konttila J, Siira H, Kyngäs H, Lahtinen M, Elo S, Kääriäinen M, et al. Healthcare professionals' competence in digitalisation: a systematic review. J Clin Nurs 2019 Mar;28(5-6):745-761. [doi: 10.1111/jocn.14710] [Medline: 30376199]
43. MacLure K, Stewart D. A qualitative case study of ehealth and digital literacy experiences of pharmacy staff. Res Social Adm Pharm 2018 Jun;14(6):555-563. [doi: 10.1016/j.sapharm.2017.07.001] [Medline: 28690128]
44. MacLure K, Stewart D. Digital literacy knowledge and needs of pharmacy staff: a systematic review. J Innov Health Inform 2016 Oct 07;23(3):840 [FREE Full text] [doi: 10.14236/jhi.v23i3.840] [Medline: 28059697]
45. Matthews B. Digital literacy in UK health education: what can be learnt from international research? Contemp Educ Technol 2021;13(4):ep317. [doi: 10.30935/cedtech/11072]
46. McGregor D, Keep M, Brunner M, Janssen A, Quinn D, Avery J, et al. Preparing e-health ready graduates: a qualitative focus group study. Stud Health Technol Inform 2017;239:91-96. [Medline: 28756442]
47. Montebello V. Digital literacy in post-certification healthcare education. J Perspect Appl Acad Pract 2016 Jul 13;4(1). [doi: 10.14297/jpaap.v4i1.185]

XSL•FO
RenderX

48. O'Connor S, LaRue E. Integrating informatics into undergraduate nursing education: a case study using a spiral learning approach. Nurse Educ Pract 2021 Jan;50:102934. [doi: 10.1016/j.nepr.2020.102934] [Medline: 33278702]
49. Oo HM, Htun YM, Win TT, Han ZM, Zaw T, Tun KM. Information and communication technology literacy, knowledge and readiness for electronic medical record system adoption among health professionals in a tertiary hospital, Myanmar: a cross-sectional study. PLoS One 2021 Jul 1;16(7):e0253691 [FREE Full text] [doi: 10.1371/journal.pone.0253691] [Medline: 34197506]
50. Poncette AS, Glauert DL, Mosch L, Braune K, Balzer F, Back DA. Undergraduate medical competencies in digital health and curricular module development: mixed methods study. J Med Internet Res 2020 Oct 29;22(10):e22161 [FREE Full text] [doi: 10.2196/22161] [Medline: 33118935]
51. Pontefract SK, Wilson K. Using electronic patient records: defining learning outcomes for undergraduate education. BMC Med Educ 2019 Jan 22;19(1):30 [FREE Full text] [doi: 10.1186/s12909-019-1466-5] [Medline: 30670000]
52. Rachmani E, Hsu CY, Chang PW, Fuad A, Nurjanah N, Shidik GF, et al. Development and validation of an instrument for measuring competencies on public health informatics of primary health care worker (PHIC4PHC) in Indonesia. Prim Health Care Res Dev 2020 Jul 06;21:e22. [doi: 10.1017/s1463423620000018]
53. Reixach E, Andrés E, Sallent Ribes J, Gea-Sánchez M, Àvila López A, Cruañas B, et al. Measuring the digital skills of Catalan health care professionals as a key step toward a strategic training plan: digital competence test validation study. J Med Internet Res 2022 Nov 30;24(11):e38347 [FREE Full text] [doi: 10.2196/38347] [Medline: 36449330]
54. Shiferaw KB, Tilahun BC, Endehabtu BF. Healthcare providers' digital competency: a cross-sectional survey in a low-income country setting. BMC Health Serv Res 2020 Nov 09;20(1):1021 [FREE Full text] [doi: 10.1186/s12913-020-05848-5] [Medline: 33168002]
55. Skiba DJ. Nursing informatics education: from automation to connected care. Stud Health Technol Inform 2017;232:9-19. [Medline: 28106577]
56. Tegegne MD, Tilahun B, Mamuye A, Kerie H, Nurhussien F, Zemen E, et al. Digital literacy level and associated factors among health professionals in a referral and teaching hospital: an implication for future digital health systems implementation. Front Public Health 2023 Apr 11;11:1130894 [FREE Full text] [doi: 10.3389/fpubh.2023.1130894] [Medline: 37113180]
57. Värri A, Tiainen M, Rajalahti E, Kinnunen UM, Saarni L, Ahonen O. The definition of informatics competencies in Finnish healthcare and social welfare education. Stud Health Technol Inform 2020 Jun 16;270:1143-1147. [doi: 10.3233/SHTI200341] [Medline: 32570560]
58. Vehko T, Hyppönen H, Puttonen S, Kujala S, Ketola E, Tuukkanen J, et al. Experienced time pressure and stress: electronic health records usability and information technology competence play a role. BMC Med Inform Decis Mak 2019 Aug 14;19(1):160 [FREE Full text] [doi: 10.1186/s12911-019-0891-z] [Medline: 31412859]
59. Virtanen L, Kaihlanen AM, Laukka E, Gluschkoff K, Heponiemi T. Behavior change techniques to promote healthcare professionals' eHealth competency: a systematic review of interventions. Int J Med Inform 2021 May;149:104432 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104432] [Medline: 33684712]
60. Vissers D, Rowe M, Islam MS, Taeymans J. Ownership and attitudes towards technology use in physiotherapy students from seven countries. Health Prof Educ 2018 Sep;4(3):198-206. [doi: 10.1016/j.hpe.2017.12.003]
61. Whittaker R, Dobson R, Hopley L, Armstrong D, Corning-Davis B, Andrew P. Training clinicians to lead clinical IT projects. N Z Med J 2020 Dec 18;133(1527):116-122. [Medline: 33332334]
62. Wubante SM, Tegegne MD, Melaku MS, Mengiste ND, Fentahun A, Zemene W, et al. Healthcare professionals' knowledge, attitude and its associated factors toward electronic personal health record system in a resource-limited setting: a cross-sectional study. Front Public Health 2023 Mar 15;11:1114456 [FREE Full text] [doi: 10.3389/fpubh.2023.1114456] [Medline: 37006546]
63. Punie Y, Redecker C. European framework for the digital competence of educators: DigCompEdu. Publications Office of the European Union. 2017. URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC107466 [accessed 2024-03-09]
64. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
65. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
66. What is ACTIC. Gencat. URL: https://tinyurl.com/2xbr3b5r [accessed 2024-03-09]
67. Dahlstrom E, Bichsel J. ECAR Study of undergraduate students and information technology, 2014. EDUCAUSE. 2014. URL: https://tinyurl.com/ez62fb7s [accessed 2024-03-09]
68. Fogarty G, Cretchley P, Harman C, Ellerton N, Konki N. Validation of a questionnaire to measure mathematics confidence, computer confidence, and attitudes towards the use of technology for learning mathematics. Math Ed Res J 2001 Sep;13(2):154-160. [doi: 10.1007/bf03217104]
69. Kayser L, Karnoe A, Furstrand D, Batterham R, Christensen KB, Elsworth G, et al. A multidimensional tool based on the eHealth literacy framework: development and initial validity testing of the eHealth Literacy Questionnaire (eHLQ). J Med Internet Res 2018 Feb 12;20(2):e36 [FREE Full text] [doi: 10.2196/jmir.8371] [Medline: 29434011]
70. Furstrand D, Kayser L. Development of the eHealth Literacy Assessment Toolkit, eHLA. Stud Health Technol Inform 2015;216:971. [Medline: 26262273]

XSL•FO
RenderX

71. Yoon S, Yen PY, Bakken S. Psychometric properties of the self-assessment of nursing informatics competencies scale. Stud Health Technol Inform 2009;146:546-550 [FREE Full text] [Medline: 19592902]
72. Kinnunen UM, Heponiemi T, Rajalahti E, Ahonen O, Korhonen T, Hyppönen H. Factors related to health informatics competencies for nurses-results of a national electronic health record survey. Comput Inform Nurs 2019 Aug;37(8):420-429. [doi: 10.1097/CIN.0000000000000511] [Medline: 30741730]
73. Ng W. Can we teach digital natives digital literacy? Comput Educ 2012 Nov;59(3):1065-1078. [doi: 10.1016/j.compedu.2012.04.016]
74. Ferrari A. Digital competence in practice an analysis of frameworks. Publications Office of the European Union. 2012. URL: https://data.europa.eu/doi/10.2791/82116 [accessed 2024-03-09]
75. Bawden D. Information and digital literacies: a review of concepts. J Document 2001;57(2):218-259. [doi: 10.1108/EUM0000000007083]
76. Sharpe R, Beetham H. Understanding students' uses of technology for learning: towards creative appropriation. In: Sharpe R, Beetham H, de Freitas S, editors. Rethinking Learning for a Digital Age: How Learners are Shaping their Own Experiences. New York, NY: Routledge; 2010.
77. Gretton C, Honeymen M. The digital revolution: eight technologies that will change health and care. The King's Fund. 2016. URL: https://www.kingsfund.org.uk/publications/digital-revolution [accessed 2024-03-09]
78. A health and care digital capabilities framework. National Health Service. 2017 Dec. URL: https://tinyurl.com/etnhfswp [accessed 2024-03-09]
79. Watson P. Electronic health records: manual for developing countries. World Health Organization. 2006. URL: http://www.wpro.who.int/publications/docs/EHRmanual.pdf [accessed 2024-03-09]
80. Skiba DJ, Barton AJ, Estes K, Gilliam E, Knapfel S, Lee C, et al. Preparing the next generation of advanced practice nurses for connected care. Stud Health Technol Inform 2016;225:307-313. [Medline: 27332212]
81. Hurst A. Qualitativ orientierte evaluationsforschung im kontext virtuellen lehrens und lernens. Ludwigsburg University of Education. 2006. URL: https://phbl-opus.phlb.de/frontdoor/index/index/year/2008/docId/19 [accessed 2024-03-09]

Abbreviations

DigComp 2.2: updated version of the digital competence framework for citizens
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by T Leung; submitted 22.12.23; peer-reviewed by J Castro, I Said-Criado; comments to author 31.01.24; revised version received 26.02.24; accepted 29.02.24; published 29.03.24.

Please cite as:
Mainz A, Nitsche J, Weirauch V, Meister S
Measuring the Digital Competence of Health Professionals: Scoping Review
JMIR Med Educ 2024;10:e55737
URL: https://mededu.jmir.org/2024/1/e55737
doi:10.2196/55737
PMID:38551628

©Anne Mainz, Julia Nitsche, Vera Weirauch, Sven Meister. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.03.2024. 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.

XSL•FO
RenderX

Review

Global Rate of Willingness to Volunteer Among Medical and Health Students During Pandemic: Systemic Review and Meta-Analysis

Mahsusi Mahsusi[1], PhD; Syihaabul Hudaa[2], MPD; Nuryani Nuryani[3], PhD; Mustofa Fahmi[4], PhD; Ghina Tsurayya[5], MD; Muhammad Iqhrammullah[6], PhD

[1]Department of Islamic Education Management, Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia

[2]Department of Management, Institut Teknologi dan Bisnis Ahmad Dahlan Jakarta, Banten, Indonesia

[3]Department of Indonesian Language and Literature Education, Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia

[4]Ministry of Religious Affairs of the Republic of Indonesia, Jakarta, Indonesia

[5]Medical Research Unit, School of Medicine, Universitas Syiah Kuala, Banda Aceh, Indonesia

[6]Postgraduate Program of Public Health, Universitas Muhammadiyah Aceh, Banda Aceh, Indonesia

Corresponding Author:
Mahsusi Mahsusi, PhD
Department of Islamic Education Management
Faculty of Tarbiyah and Teacher Training
Universitas Islam Negeri Syarif Hidayatullah Jakarta
Jl Ir H Djuanda No 95
Tangerang Selatan, 15412
Indonesia
Phone: 62 83806254803
Email: mahsusi@uinjkt.ac.id

Abstract

Background: During health crises such as the COVID-19 pandemic, shortages of health care workers often occur. Recruiting students as volunteers could be an option, but it is uncertain whether the idea is well-accepted.

Objective: This study aims to estimate the global rate of willingness to volunteer among medical and health students in response to the COVID-19 pandemic.

Methods: A systematic search was conducted on PubMed, Embase, Scopus, and Google Scholar for studies reporting the number of health students willing to volunteer during COVID-19 from 2019 to November 17, 2023. The meta-analysis was performed using a restricted maximum-likelihood model with logit transformation.

Results: A total of 21 studies involving 26,056 health students were included in the meta-analysis. The pooled estimate of the willingness-to-volunteer rate among health students across multiple countries was 66.13%, with an I2 of 98.99% and *P* value of heterogeneity (*P*-Het)<.001. Removing a study with the highest influence led to the rate being 64.34%. Our stratified analyses indicated that those with older age, being first-year students, and being female were more willing to volunteer (*P*<.001). From highest to lowest, the rates were 77.38%, 77.03%, 65.48%, 64.11%, 62.71%, and 55.23% in Africa, Western Europe, East and Southeast Asia, Middle East, and Eastern Europe, respectively. Because of the high heterogeneity, the evidence from this study has moderate strength.

Conclusions: The majority of students are willing to volunteer during COVID-19, suggesting that volunteer recruitment is well-accepted.

(JMIR Med Educ 2024;10:e56415) doi:10.2196/56415

KEYWORDS

COVID-19; education; health crisis; human resource management; volunteer

XSL•FO
RenderX

Introduction

The initial outbreak of COVID-19, an emerging and highly infectious respiratory illness, which originated in Wuhan City, Hubei Province, China, occurred in early December 2019 [1]. Subsequently, the situation escalated swiftly, leading to its declaration as a Public Health Emergency of International Concern by the World Health Organization (WHO) [1]. Some of its clinical presentations are fever, cough, dryness, fatigue, dyspnea, and myalgia [2]. The disease also induces prolonged anxiety, chest pain, persistent depression, dizziness, and other lingering symptoms after recovery [3]. In 2020, the WHO reported the rapid spread of the disease to various parts of the world, marking it as a global pandemic, with the number of confirmed cases and deaths escalating worldwide [4]. The pandemic not only affects health but also disrupts various aspects of life, including emotional stability, environmental quality, and the economy [5-7].

The high incidence of COVID-19 has led to an increased demand for health care services and workers [8]. Unlike many other sectors, jobs in health care were not temporarily halted during the COVID-19 pandemic, as health care professionals are essential in combating and preventing viral transmission [9]. However, infections among medical personnel have resulted in an acute shortage of workforce in this sector. Coupled with the increased workload of health care workers, this has led to inadequate patient management [8,10]. A previous study found that nearly half of health care workers exposed to COVID-19 experienced burnout and compassion fatigue, stemming from factors such as excessive workload, emotional exhaustion, personal infection risk, and fear of transmitting the virus to their families [5,11]. Consequently, hospitals faced the challenge of addressing staffing deficits [12].

During health emergencies, it is crucial to bolster the human resource capacity within the health care system. Among the various approaches available, recruiting volunteers is an option worthy of consideration [13]. Volunteering entails participating in activities where individuals dedicate their time to providing services to vulnerable populations without coercion [14,15]. Medical and health students can actively participate in volunteering activities to help manage the COVID-19 crisis. In certain countries and health care institutions, it is suggested that medical and health students voluntarily contribute to crisis management based on their competencies [13,16,17]. Collaborating with volunteers to provide community services could help bridge gaps in human resource capacity and decrease instances of burnout among health care workers during the COVID-19 crisis [18].

Volunteering among health care students has emerged as a valuable resource during outbreaks. A previous study has evaluated the willingness of medical students to volunteer during pandemics and disasters [19]. Furthermore, a previously published systematic review on the willingness of health students to volunteer for COVID-19 reported willingness-to-volunteer rates ranging from 19.5% to 91.5% [20]. Unfortunately, a meta-analysis was not conducted in that systematic review [20]. Data on the global rate of willingness to volunteer are necessary as a basis for evaluating the feasibility of recommending volunteering for health students. Moreover, it is crucial to observe feasibility across different populations, economies, and regions. Therefore, our aim is to conduct a new systematic review with a meta-analysis on the willingness-to-volunteer rate among medical and health students in response to the COVID-19 pandemic.

Methods

Study Design

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement was used as the guidance for this study (see Tables S1 and S2 in Multimedia Appendix 1) [21]. The research questions were formulated as follows: (1) What is the percentage of COVID-19 volunteer willingness among health care students? (2) What are the demographic factors associated with the willingness of health care students to volunteer? This review was not registered because it did not evaluate direct effects on human health.

Search Strategy

A systematic review search was carried out on PubMed, Scopus, and Embase up to December 10, 2023. Google Scholar was also included as a gray literature source in the search. The keywords used were "health students," "willingness," "volunteer OR volunteering OR volunteerism OR voluntary," AND "COVID-19 OR covid-19 OR SARS-CoV-2 OR COVID-19 pandemic." The complete search technique is outlined in Table S3 in Multimedia Appendix 2 and searches for the other databases were developed using the Embase search strategy.

Eligibility Criteria, Articles Selection, and Data Extraction

The inclusion criteria were cross-sectional studies aimed at evaluating the rate of willingness to volunteer among medical and health students (encompassing disciplines such as medical, nursing, pharmacy, dentistry, midwifery, public health, and other relevant fields) in response to the COVID-19 situation from December 2019 to December 2023. Willingness to volunteer was defined as a "yes" response to the question "Are you willing to volunteer?" or "Do you want to volunteer?" We only included studies involving undergraduate or diploma students; studies involving other levels of education were excluded. Medical and health students were defined as individuals pursuing higher education degrees (undergraduate or diploma) in medicine, nursing, dentistry, pharmacy, midwifery, public health, and related fields. Exclusion criteria were applied to studies meeting any of the following conditions:(1) qualitative analysis, (2) focused on postgraduate or professional students, (3) non–English language articles, (4) review articles, (5) case reports, (6) randomized controlled trials, (7) clinical trial proposals, and (8) case-control studies.

GT and MI independently screened all duplicate topics, titles, abstracts, and full texts using Zotero version 6.0.30 (Corporation for Digital Scholarship). Duplicate entries were removed, and title and abstract screening were conducted on the remaining records. Subsequently, the selected records were searched for full-text access, and further comprehensive screening was

XSL•FO
RenderX

conducted on the obtained full texts by applying the eligibility criteria. Data extraction was conducted using the tabulation method, covering details such as author and year of publication, country, student population, sample size, gender distribution, academic year of the students, health status, marital status, living arrangements, volunteer experience, and the proportion indicating willingness. GT and MI independently carried out the data extraction process. Continuous data were presented as mean (SD), with conversion from median performed when necessary using an online calculator [22]. Any discrepancies were resolved through consensus.

Quality Appraisal

The quality of individual studies was assessed by one reviewer (MM) and independently reviewed for agreement by a second reviewer (SH). In instances of disagreement, a third review author (NN) was consulted. The standardized Quality Assessment Checklist for Survey Studies in Psychology (Q-SSP) tool, consisting of 20 checklist items, was used for the quality assessment of the included studies [23]. High-quality articles were defined as those scoring 70% or higher. The score was determined by the percentage of "yes" responses on the checklist. The utilization of this tool aligns with a previous study [24].

Statistical Analysis

The proportion was initially transformed using the logit function (y=logit[x]) before being pooled for meta-analysis with a restricted maximum-likelihood model. The rate was then obtained by multiplying the pooled proportion, following the back transformation from the logit function (y=1/(1+exp[–x]), with 100%. A rate exceeding 50% was considered the threshold for determining the majority's willingness to volunteer during the pandemic. The CI was set at 95% (ie, 95% CI), with a *P* value of total effect (*P*-tot) <.05 indicating statistical significance. A value of I^2 greater than 50% or a *P* value of heterogeneity (*P*-Het) <.1 was used as the cutoff for determining data heterogeneity in the pooled analysis. Begg's funnel plot was used to assess the presence of publication bias. The meta-analysis was conducted using jamovi 2.3.21. A moderator analysis was conducted to examine the effects of sample size, age, gender (indicated as male-to-female ratio), academic year (indicated as the ratio of students in second to first year, third to first year, and so on), volunteer experience (indicated as the ratio of students with to without volunteer experience), type of academic program (indicated as the ratio of medical to nursing students and the ratio of medical to dentistry students), country income category, and continent. Country income was categorized based on the World Bank classifications (high income, upper middle income, lower middle income, and low income). The variables used in the moderator effect analysis were also used in the stratification analysis, with the following cutoffs: 22 years old for age, 1 for the male-to-female ratio, 15% for the proportion of first-year students, 1 for the ratio of students with to without volunteer experience, and 5 for the ratio of medical to nursing students. Statistical significance in the stratification analysis was determined using Z-statistics. The statistical analysis adhered to recommendations from previous studies [24-26].

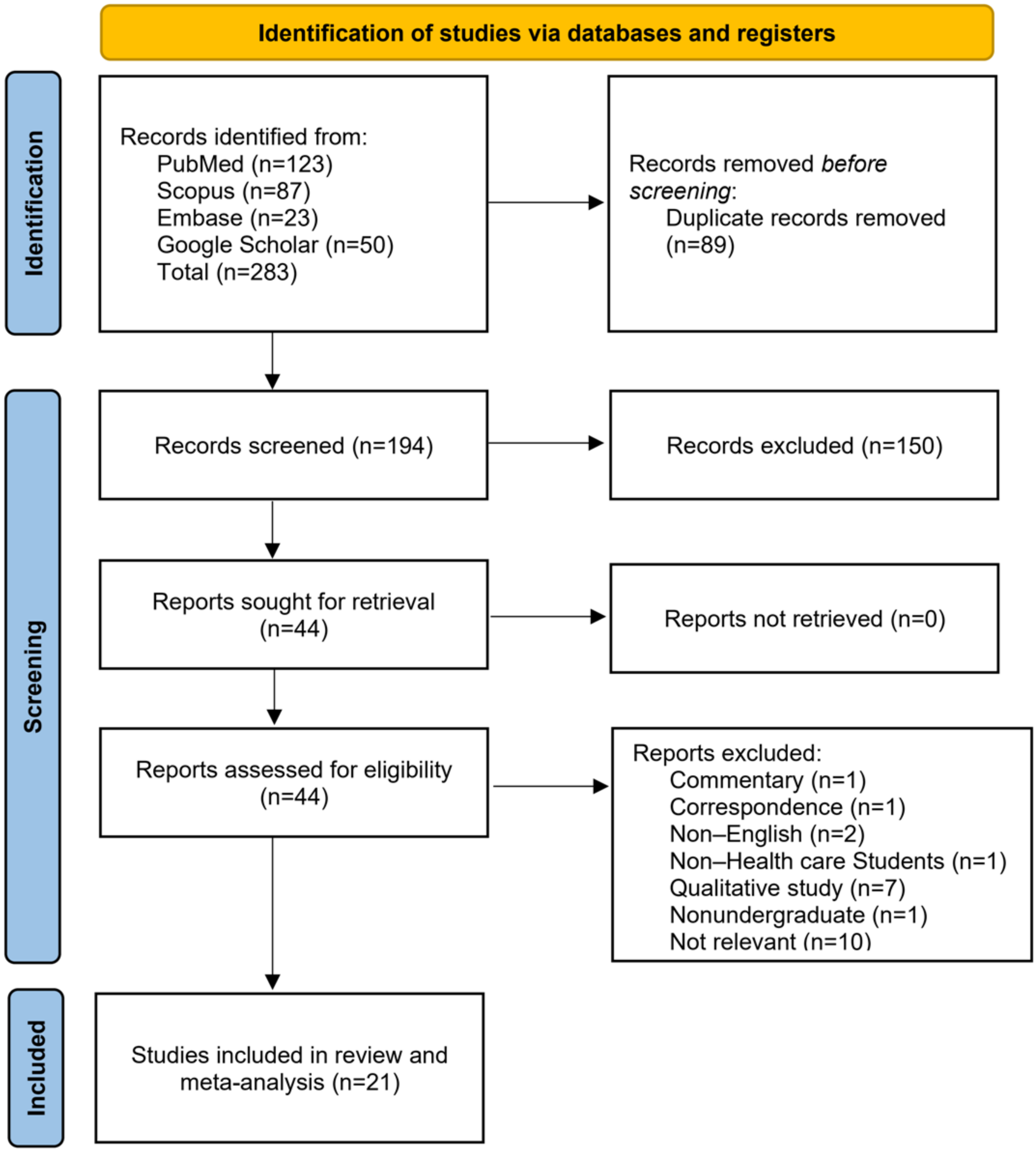
Results

Search Findings

Collectively, 283 records were identified from PubMed, Scopus, Embase, and Google Scholar in the initial stage. A total of 89 duplicates were automatically detected and subsequently removed. The remaining 194 records underwent screening for relevance based on the title and abstract. Forty records were then selected for full-text access and further thorough screening. During this stage, we identified 1 commentary [27], 1 correspondence [28], and 2 non-English articles [29,30], which were subsequently excluded. One study was excluded because the participants were not specified as medical or health students [31]. Seven studies were found to be qualitative, and therefore data extraction was not feasible; these studies were subsequently removed [32-38]. Additionally, 1 study was excluded because the participants were not pursuing undergraduate degrees [39]. Ten studies were deemed irrelevant to the objective of this review [40-49]. Finally, 21 studies were included in the systematic review and meta-analysis [50-70]. The screening and selection processes are depicted in Figure 1.

XSL•FO
RenderX

Figure 1. Schematic diagram for screening and selection of eligible studies following PRISMA guideline. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



Characteristics and Quality of the Included Studies

Characteristics of the included studies along with their quality are presented in Table 1. A total of 21 studies were included, with a combined sample size of 26,024 students [50-70]. The studies were conducted in various countries, including Nigeria (n=3), Pakistan (n=2), Saudi Arabia (n=2), Serbia (n=1), India (n=1), Bulgaria (n=1), Vietnam (n=1), Poland (n=1), Brunei Darussalam (n=1), Australia (n=1), Nepal (n=1), Indonesia (n=1), Romania (n=1), the United Kingdom (n=1), China (n=1), Syria (n=1), and Sudan (n=1). Eligible studies from South and North Americas (including the United States, Canada, and Mexico) were not identified in this systematic review. The average ages of the participants ranged from 22 to 24 years, whereas the proportion of female students varied considerably across studies. Nine studies exclusively recruited medical students [50-54,61,62,67,69], 3 focused on nursing students [55,60,65], and others included a mix of students from different departments. Thirteen studies were categorized as "high quality" based on the Q-SSP [50-62], while others had scores below 70% [63-67], and some even scored 50% or below [68-70]. Detailed assessment results based on the Q-SSP tool are presented in Table S4 in Multimedia Appendix 2 [50-70]. A total of 9/21 (43%) studies did not provide sufficient justification for the sample size. See Table S5 in Multimedia Appendix 2 for the 20 checklist items of Q-SSP and their respective code.

XSL•FO
RenderX

Table 1. Characteristics and quality of the included studies.

Author [reference]	Country	Geographic location	Female, n (%)	Age (years), mean (SD)	Department or faculty	Q-SSP[a], %
Byrne et al [67]	United Kingdom	Western Europe	835 (72.9)	22 (0.61)	Medicine	60
Gazibara and Pesakovic [62]	Serbia	Eastern Europe	247 (75.8)	23.0 (1.2)	Medicine	75
Yordanova et al [68]	Bulgaria	Eastern Europe	Not reported	Not reported	Medicine, nursing, physician assistant, medical rehabilitation and occupational therapy, and midwife	45
Adejimi et al [63]	Nigeria	Africa	211 (62.6)	23.4 (2.6)	Medicine and dentistry	65
Joseph and Manasvi [69]	Indian	South Asia	119 (58.3)	21.6 (1.1)	Medicine	50
Nazir et al [61]	Pakistan	South Asia	77 (27.3)	21.9 (1.26)	Medicine	70
Tran et al [64]	Vietnam	East and Southeast Asia	1192 (58.7)	22.8 (3.7)	General medicine, traditional medicine, pharmacy, medical technique, preventive medicine, nursing, dentistry, public health, midwifery, and medical imaging	65
Domaradzki and Walkowiak [59]	Poland	Western Europe	116 (27.7)	Not reported	Medicine, nursing, midwife, pharmacy, electroradiology, medical analytics, dentistry, medical rescue, and others	70
Hj Abdul Aziz et al [60]	Brunei Darussalam	East and Southeast Asia	16 (22.2)	Not reported	Nursing	80
Prisca et al [65]	Nigeria	Africa	598 (82.6)	21.5 (2.5)	Nursing	65
Adejimi et al [58]	Nigeria	Africa	257 (62.5)	23.26 (2.59)	Medicine and dentistry	80
Al Gharash et al [55]	Australian	Pacific	5 (5.6)	Not reported	Nursing	75
AlOmar et al [56]	Saudi Arabia	Middle East	3506 (58.3)	22.07 (1.84)	Medicine, nursing, dentistry, applied medical sciences, and public health	90
Karki et al [57]	Nepal	South Asia	152 (58.2)	Not reported	Medicine and nursing	90
Khalid et al [53]	Pakistan	South Asia	142 (71.0)	21.5 (1.4)	Medicine	85
Lazarus et al [54]	Indonesia	East and Southeast Asia	3399 (69.8)	20 (0.27)	Medicine	80
Magdas et al [70]	Romania	Eastern Europe	805 (78.8)	Not reported	Medicine and nursing	50
Feng et al [66]	China	East and Southeast Asia	3582 (66.6)	20 (1.5)	Medicine, nursing, public health, medical technology, and health and medical administrative services	60
AlSaif et al [50]	Saudi Arabia	Middle East	39 (29.1)	Not reported	Medicine	75
Alsuliman et al [51]	Syria	Middle East	589 (49.1)	Not reported	Medicine	95
Elsheikh et al [52]	Sudan	Africa	424 (68.2)	23 (2)	Medicine	80

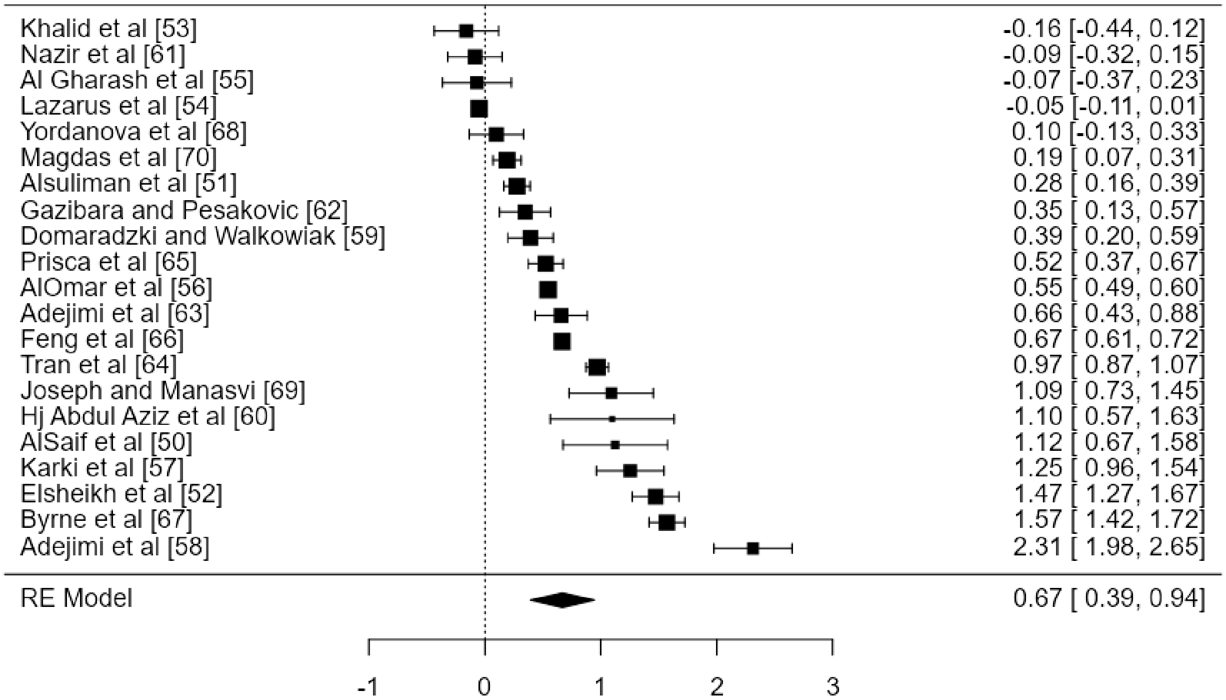
[a]Q-SSP: Quality Assessment Checklist for Survey Studies in Psychology.

Willingness-to-Volunteer Rate

The forest plot of the pooled analysis on the rate of willingness to volunteer is presented in Figure 2. After being transformed back from the logit function, the pooled proportion of willingness to volunteer was 66.13% (95% CI 56%-72%). The heterogeneity for this pooled estimate was high, with I^2=98.99% and *P*-Het<.001.

XSL•FO
RenderX

Figure 2. Forest plot for the pooled proportion of willingness to volunteer among health students. RE: random effect.



Sensitivity Analysis

To observe if a single study affects the entire pooled estimate, a sensitivity test based on a one-leave-out analysis was conducted. The pooled estimates for each study removed are presented in Table 2. The lowest logit proportion was obtained when Adejimi et al [58] was removed (0.59, 95% CI 0.35-0.82), with the I^2 value becoming relatively lower (72.06%), although the *P*-Het remained <.001. The overall rate of willingness to volunteer after the removal of Adejimi et al [58] was 64.34% (95% CI 59%-69%). It is noteworthy that the rate of willingness to volunteer reported by Adejimi et al [58] was the highest among all included studies, at 90.97%.

XSL•FO
RenderX

Table 2. Results from the one-leave-out analysis for the pooled willingness to volunteer among health students.

Study removed	Logit proportion	95% CI	*P* value of total effect	I^2, %	*P* value of heterogeneity
Hj Abdul Aziz et al [60]	0.65	0.37-0.94	<.001	99.06	<.001
Prisca et al [65]	0.68	0.39-0.97	<.001	99.05	<.001
Adejimi et al [58]	0.59	0.35-0.82	<.001	72.06	<.001
Adejimi et al [63]	0.67	0.38-0.96	<.001	99.07	<.001
Al Gharash et al [55]	0.71	0.43-0.98	<.001	99.01	<.001
AlOmar et al [56]	0.68	0.39-0.97	<.001	98.78	<.001
AlSaif et al [50]	0.65	0.36-0.93	<.001	99.06	<.001
Alsuliman et al [51]	0.69	0.40-0.98	<.001	99.00	<.001
Byrne et al [67]	0.62	0.35-0.89	<.001	98.93	<.001
Domaradzki and Walkowiak [59]	0.68	0.40-0.97	<.001	99.06	<.001
Gazibara and Pesakovic [62]	0.69	0.40-0.97	<.001	99.06	<.001
Joseph and Manasvi [69]	0.65	0.36-0.93	<.001	99.06	<.001
Karki et al [57]	0.64	0.36-0.92	<.001	99.03	<.001
Khalid et al [53]	0.71	0.43-0.99	<.001	98.99	<.001
Lazarus et al [54]	0.71	0.43-0.99	<.001	98.73	<.001
Magdas et al [70]	0.69	0.41-0.98	<.001	99.00	<.001
Nazir et al [61]	0.71	0.43-0.99	<.001	99.00	<.001
Tran et al [64]	0.65	0.37-0.94	<.001	98.99	<.001
Yordanova et al [68]	0.70	0.42-0.98	<.001	99.03	<.001
Feng et al [66]	0.67	0.38-0.96	<.001	98.82	<.001
Elsheikh et al [52]	0.63	0.35-0.90	<.001	98.97	<.001

Moderator Effect

The effects of moderators were analyzed, and the results are presented in Table 3. The ratio of third- to first-year students significantly affects the overall rate of willingness to volunteer with *P*=.02. Furthermore, the higher statistical significance of the moderator effect was observed on the ratios of fourth-, fifth-, or sixth- to first-year students (*P*<.001, respectively). However, other variables did not moderate the pooled estimate of the willingness-to-volunteer rate (*P* value ranged from .22 to .70).

XSL•FO
RenderX

Table 3. Moderator effect on the pooled estimates of willingness-to-volunteer proportion (N=21).

Moderator	Data type	Study, n (%)	Z	*P* value
Sample size	Continuous	21 (100)	–0.61	.54
Age	Continuous	14 (67)	0.79	.43
Male-to-female ratio	Continuous	20 (95)	–1.23	.22
Second-to-first-year students ratio	Continuous	7 (33)	–2.93	.003[a]
Third-to-first-year students ratio	Continuous	7 (33)	–2.28	.02[b]
Fourth-to-first-year students ratio	Continuous	7 (33)	–3.33	<.001[a]
Fifth-to-first-year students ratio	Continuous	6 (29)	–2.27	<.001[a]
Sixth-to-first-year students ratio	Continuous	4 (19)	–4.01	<.001[a]
Single-to-married ratio	Continuous	6 (29)	–0.923	.36
With-to-without volunteer experience ratio	Continuous	9 (43)	0.946	.34
Medical-to-nursing student ratio	Continuous	7 (33)	0.671	.50
Medical-to-dentistry student ratio	Continuous	5 (24)	–0.421	.67
Country income	Category	21 (100)	0.378	.70
Continent	Category	21 (100)	–1.16	.24

[a]Significant at *P*<.01.

[b]Significant at *P*<.05.

Stratification Analysis

We further stratified the pooled estimate of the willingness-to-volunteer rate based on several variables, and the results are presented in Table 4. According to Z-statistics, groups with older mean age, a higher number of male participants, and a higher number of first-year students had a significantly higher rate of willingness to volunteer (*P*<.001). Conversely, a higher number of participants from medical school contributed to a lower rate of willingness to volunteer (*P*<.001). As compared with the pooled rates of willingness to volunteer in high-income countries, those in upper-middle-income (*P*<.001) and low-income countries (*P*=.04) tend to be significantly lower, except for lower-middle-income countries (66.37% vs 69.42%; *P*<.001). Based on regions, rates of willingness to volunteer were the highest among African and Western European countries (77.38% and 77.03%, respectively). It is worth noting that the heterogeneity of a pooled estimate of studies from Eastern European countries was negligible (I^2=0.05%, *P*=.30), where the rate was 55.23%—the lowest among all regions. The number of samples recruited in studies according to regions and their corresponding rate of willingness to volunteer are presented in Multimedia Appendix 3.

XSL•FO
RenderX

Table 4. Stratification analysis based on the characteristics of participants.

Variable	Study, n (%)	Sample, n	Logit proportion	95% CI	Rate, %	*P*-Z	I^2, %	*P* value of heterogeneity
Mean age (years)								
≤22	7 (33.3)	12,757	0.51	0.02 to 0.99	62.41	<.001[a]	98.97	<.001
>22	6 (21.57)	9744	1.04	0.47 to 1.62	73.89	N/A[b]	99.08	<.001
Male-to-female ratio								
≤1	14 (66.66)	23,528	0.81	0.45 to 1.16	69.21	<.001[a]	99.31	<.001
>1	6 (21.57)	2246	0.42	0.01 to 0.83	60.34	N/A	94.12	<.001
Proportion of first-year students								
≤15	3 (14.28)	1665	0.16	–0.14 to 0.45	54	<.001[a]	86.11	.006
>15	4 (19.04)	3510	1.23	0.94 to 1.53	77.38	N/A	88.5	<.001
With-to-without volunteer experience ratio								
≤1	5 (23.80)	7524	0.56	–0.14 to 1.26	63.65	.156	99.22	<.001
>1	4 (19.04)	2833	0.63	0.17 to 1.10	65.25	N/A	95.25	<.001
Medical-to-nursing student ratio								
≤5	3 (14.28)	5922	0.67	0.03 to 1.34	66.15	<.001[a]	97.08	<.001
>5	4 (19.04)	9513	0.53	0.20 to 0.85	62.95	N/A	97.6	<.001
Country income								
High income	7 (33.33)	8974	0.68	0.24 to 1.12	66.37	Reference	98.15	<.001
Upper middle income	4 (19.04)	10,857	0.27	–0.05 to 0.59	56.71	<.001[a]	97.95	<.001
Lower middle income	8 (38.09)	4404	0.82	0.27 to 1.36	69.42	<.001[a]	97.94	<.001
Low income	2 (9.52)	1821	0.57	0.47 to 0.67	63.88	.041[c]	99.02	<.001
Regions								
Africa	4 (19.04)	2096	1.23	0.54 to 1.93	77.38	Reference	97.64	<.001
Eastern Europe	3 (14.28)	1656	0.21	0.12 to 0.30	55.23	<.001[a]	0.05	.30
Western Europe	2 (9.52)	1562	1.21	1.00 to 1.24	77.03	.802	98.84	<.001
East and Southeast Asia	4 (19.04)	12,353	0.64	0.20 to 1.08	65.48	<.001[a]	99.08	<.001
South Asia	4 (19.04)	898	0.52	–0.12 to 1.16	62.71	<.001[a]	94.97	<.001
Middle East	3 (14.28)	7317	0.58	0.24 to 0.91	64.11	<.001[a]	94.95	<.001

[a]Significant at *P*<.01.

[b]N/A: not applicable.

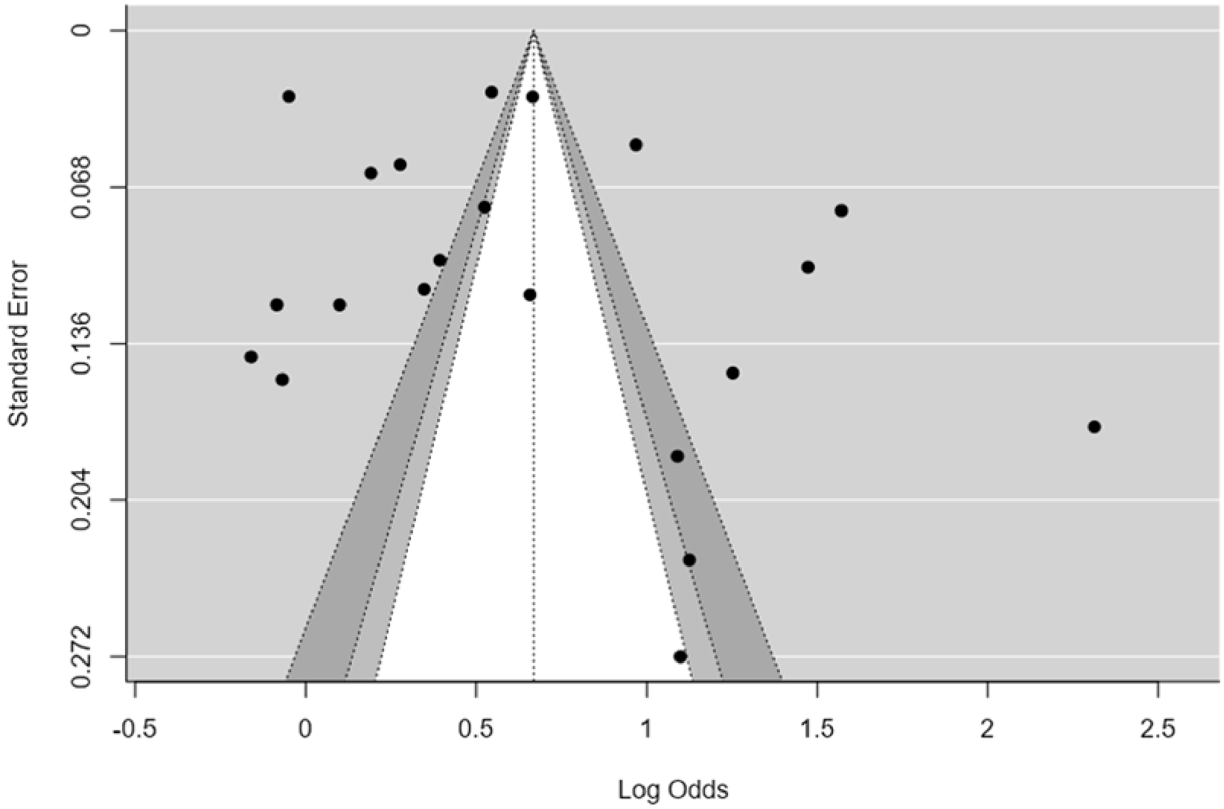
[c]Significant at *P*<.05.

Publication Bias

Begg's funnel plot for the overall pooled estimate is presented in Figure 3. The symmetrical shape of the funnel plot suggests that publication bias was not detected, with *P*-Begg=.14.

XSL•FO
RenderX

Figure 3. Begg's funnel plot for studies reporting willingness to volunteer during pandemic among health students. The shape is symmetrical with *P*-Begg's=.14.



Discussion

Rate of Willingness to Volunteer During the Pandemic

The pooled estimate herein revealed that the overall rate of willingness to volunteer was 66.13%, with the rate being over 50% in almost every individual study. When stratified based on regions, the highest rate of willingness to volunteer was found among students from African countries (77.38%), followed by Western European countries (77.03%). Furthermore, the stratified analyses indicate that being older or female was associated with a higher rate of willingness to volunteer (*P*<.001, respectively). In addition, students in the first academic years were more willing to volunteer compared with those in more senior years (*P*<.001). According to individual studies, students were willing to volunteer due to various motivations, particularly internal factors [50-70]. Studies reported that students are driven by altruistic and duty-driven reasons, where individuals, particularly those aspiring to become future health professionals, feel a sense of responsibility in assisting during the pandemic [35]. They also perceive volunteering activity as an opportunity to learn, gain clinical skills, and enhance personal growth [34]. Research suggests that medical students exhibiting high prosocial motivation are more likely to engage in volunteer activities and persist in such endeavors, even in the absence of prior experience [49]. Moreover, external motivations contribute to such willingness in various forms, notably compensation-related factors. These external motivations include the recognition of academic credit, achievements, receipt of scholarships, and provision of material compensation [49]. Importantly, the willingness of students to volunteer is heightened when such engagement aligns with governmental needs and is endorsed by universities. Students are more likely to engage in volunteering if it is needed by the government and recommended by universities [71].

Despite the positive aspects of volunteering, attention should also be given to the mental health of students. A previous study by Tempski and colleagues [16] revealed a negative association between the mental health problems of medical students (including stress, anxiety, and depression) and their participation as volunteers during the COVID-19 pandemic. Furthermore, many volunteer students express fear of getting infected and spreading the virus to their relatives or friends during their volunteer tasks [72,73]. Students also reported feeling unprepared to deal with the pandemic, citing issues such as personal protective equipment shortages, lack of training and knowledge, role confusion, insufficient information, and a lack of support from social or family networks [71]. As highlighted in a qualitative study, the well-being of volunteers was neglected due to a lack of access to psychological support [36]. Moreover, these challenges are further perpetuated by the academic workload and responsibilities as undergraduate students [74].

In this study, we found that the willingness-to-volunteer rate among first-year students was higher compared with students in the second to sixth year of education. As suggested by a previous study, most first-year students have not yet commenced their involvement in extracurricular activities and are still in the process of selecting the type of activities they wish to pursue

XSL•FO
RenderX

[75]. This further poses a challenge in using student volunteers to overcome the health care workforce shortage, as first-year students lack skills and experience, making them unsuitable for direct clinical assistance. For first-year students, community-based work is more suitable, including but not limited to childcare for health care workers, delivery of medicines to vulnerable populations, mental health checks on children, and other similar tasks. However, aiming to reduce the workload and burnout incidence among health care workers still necessitates recruiting students in higher academic years. Students in higher academic years are more involved in extracurricular activities and more occupied with lecture schedules, making them less willing to volunteer. Moreover, as students are exposed to more medical and health knowledge, they become more considerate of preventive measures. A previous study reported that students were more willing to volunteer if they were assured their grades would not suffer and be compensated, guaranteed coverage of treatment costs if they got infected while volunteering, offered separate accommodation during the duration of their volunteer work, and provided with psychological support [53]. Therefore, addressing these barriers is crucial to encourage students in higher academic years to volunteer during the pandemic.

Herein, we found that female students are more likely to volunteer than male students. This aligns with previous findings, showing a willingness proportion of 60.2% for females compared with 52.3% for males [20]. Consistent with a previous study, women were reported to be more willing to volunteer due to a nurturing inclination to help people in need and their empathetic nature, driven by personal and thoughtful motivations in the long term [76]. However, a study by Lazarus et al [54] stated that being male was one of the significant demographic factors influencing willingness to volunteer in Indonesia. The differing findings among these studies are indicative of the influence of sociocultural factors on students' willingness to volunteer during the pandemic.

Our findings indicate that the highest rate of willingness to volunteer is observed in Africa and Western Europe, while the lowest is in Eastern Europe. This aligns with a study revealing significantly lower volunteer rates in Eastern Europe compared with Western Europe, except for the trade union [77]. One of the primary factors contributing to the lower willingness-to-volunteer rate in Eastern Europe is the historical context—having been under communist rule for half a century, memories of mandatory volunteering have imbued the concept of volunteer work with a distinctly negative connotation. This negative perception is further compounded by a postcommunist lack of trust in any public activity [78]. Nowadays, Eastern European countries have progressed beyond acknowledging volunteering to establish a legal framework that actively promotes volunteering [78]. This indicates that certain countries might have to put extra effort into encouraging students to volunteer.

In pandemic settings such as COVID-19, health students play essential roles in addressing the shortage of health care workers and responding to health problems [71]. The students' activities can be placed into various categories, such as hospital works (triage, admission wards, and emergency rooms), call centers, administrative epidemiological aspects (contact tracing, testing), online or remote consultation (regarding COVID-19 or non–COVID-19 cases, using the phone or internet), laboratory-related works, food or personal protective equipment supply, mentoring juniors, providing childcare for health workers, public education (such as countering hoaxes), and research programs [20,45,71]. With the help of volunteers, health care providers express appreciation for their valuable contributions. There are many significant advantages to volunteering, including helping provide more services and clinical care, reducing the workload for local staff, improving the quality of care, and shortening waiting times for patients. In return, it enhances how the community views and uses health care services [79].

Recommendations and Considerations

Based on the findings of this study, we propose several recommendations to increase the willingness-to-volunteer rate among medical and health students during the pandemic. First, we suggest implementing a robust encouragement program that integrates volunteering activities into curricula and offers psychological and accommodation support. Second, schools should prioritize the provision of high-quality training, promote knowledge, ensure clear role distribution, and effectively disseminate information to enhance the overall volunteering experience. Third, it is imperative for schools to ensure the complete safety of health care students by implementing measures such as preventing shortages of personal protective equipment, facilitating grade conversion, and guaranteeing coverage of treatment costs in case of infections incurred during volunteering.

Last but not least, although deploying students as volunteers could help overcome the health care worker shortage during the pandemic, it is important to consider potential drawbacks. The risks of contracting the disease and consequently experiencing death or long COVID-19 symptoms are high, especially during the early stages of the pandemic when managing the disease is significantly challenging. This further implicates liability issues for universities, colleges, or academic health centers. Therefore, it is crucial to actively inform students who are willing to volunteer regarding the aforementioned risks. Moreover, during volunteering, students might not be able to study optimally. The additional burden on health care workers to supervise volunteers should also be considered, implying the necessity to prepare students with volunteering skills beforehand and to establish a specific body tasked with managing the volunteers.

Limitations

Our study is the first to calculate the global rate of willingness to volunteer during the pandemic among medical and health students. We obtained data from countries across different regions, namely, Africa, Eastern Europe, Western Europe, East and Southeast Asia, South Asia, and the Middle East. However, the study has several limitations, including being unable to retrieve data from sources other than scientific publications. We did not collect data from reports published by government or nongovernmental organizations. Additionally, we did not contact experts who might have unpublished data regarding the willingness-to-volunteer rate. The rate was calculated from

XSL•FO
RenderX

heterogeneous data, which indicates the moderate strength of evidence. More than 40% of the included studies did not sufficiently justify the sample size, raising caution about the representativeness of the data. Moreover, moderator effects might be influenced by confounding factors that could not be controlled in this study. For example, as the effect of gender was observed based on the male-to-female ratio, the numbers might be influenced by differences in baseline demographics across countries and the composition of medical and other health professional schools. It is, therefore, important to confirm the findings through primary research.

Conclusions

The overall rate of willingness to volunteer among medical and health students during COVID-19 was 66.13%. This number indicates that the recommendation for medical and health students to volunteer can be pursued, as the majority of students are willing to volunteer, although efforts to increase willingness remain necessary. Higher rates of willingness to volunteer were observed among studies with more first-year students and female participants. According to the region, students from African and Western European countries were more willing to volunteer during the pandemic. Unfortunately, the interpretation of the pooled estimate is limited by high heterogeneity, which is expected due to the variability in different countries, settings, and populations. However, this study can serve as a basis for managing medical and health students in volunteering during health crises.

Acknowledgments

We acknowledge the collaboration between Universitas Islam Negeri Syarif Hidayatullah Jakarta, Institut Teknologi dan Bisnis Ahmad Dahlan Jakarta, Universitas Syiah Kuala, and Universitas Muhammadiyah Aceh during the study and the preparation of this article.

Data Availability

All data underlying the results are available as part of the article and no additional source data are required.

Authors' Contributions

MM contributed to the conceptualization, investigation, resource allocation, and drafting of the original manuscript. SH, NN, and MF contributed to the validation, review and editing of the manuscript, and supervision. GT contributed to the investigation and drafting of the original manuscript. MI contributed to the conceptualization, validation, methodology development, and drafting of the original manuscript. All authors have reviewed and approved the final version of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 20-item checklist and PRISMA abstract checklist.
[DOCX File , 34 KB - mededu_v10i1e56415_app1.docx]

Multimedia Appendix 2
Search strategy and combination of keywords used in each database; detailed assessment of the included studies using the Q-SSP tool; and the 20 checklist items of Q-SSP and their respective code. Q-SSP: Quality Assessment Checklist for Survey Studies in Psychology.
[DOCX File , 26 KB - mededu_v10i1e56415_app2.docx]

Multimedia Appendix 3
(A) Number of samples and (B) rate of willingness-to-volunteer based on regions.
[PNG File , 194 KB - mededu_v10i1e56415_app3.png]

References

1. Harapan H, Itoh N, Yufika A, Winardi W, Keam S, Te H, et al. Coronavirus disease 2019 (COVID-19): a literature review. J Infect Public Health 2020 May;13(5):667-673 [FREE Full text] [doi: 10.1016/j.jiph.2020.03.019] [Medline: 32340833]
2. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020 Feb 15;395(10223):507-513 [FREE Full text] [doi: 10.1016/S0140-6736(20)30211-7] [Medline: 32007143]

XSL•FO
RenderX

3. Fahriani M, Ilmawan M, Fajar JK, Maliga HA, Frediansyah A, Masyeni S, et al. Persistence of long COVID symptoms in COVID-19 survivors worldwide and its potential pathogenesis - a systematic review and meta-analysis. Narra J 2021 Aug 01;1(2):e36 [FREE Full text] [doi: 10.52225/narraj.v1i2.36] [Medline: 38449463]
4. World Health Organization (WHO). WHO COVID-19 dashboard. WHO. 2023. URL: https://data.who.int/dashboards/covid19/cases?n=c [accessed 2024-01-09]
5. Hamdan A, Eastaugh J, Snygg J, Naidu J, Alhaj I. Coping strategies used by healthcare professionals during COVID-19 pandemic in Dubai: a descriptive cross-sectional study. Narra X 2023 Apr 30;1(1):e71 [FREE Full text] [doi: 10.52225/narrax.v1i1.71]
6. Wagner AL, Rajamoorthy Y, Taib NM. Impact of economic disruptions and disease experiences on COVID-19 vaccination uptake in Asia: a study in Malaysia. Narra J 2021 Aug 01;1(2):e42 [FREE Full text] [doi: 10.52225/narraj.v1i2.42] [Medline: 38449462]
7. Otolorin GR, Oluwatobi AI, Olufemi OT, Esonu DO, Dunka HI, Adanu WA, et al. COVID-19 pandemic and its impacts on the environment: a global perspective. Narra J 2022 Apr 01;2(1):e72 [FREE Full text] [doi: 10.52225/narra.v2i1.72] [Medline: 38450389]
8. Choi KR, Skrine Jeffers K, Cynthia Logsdon M. Nursing and the novel coronavirus: risks and responsibilities in a global outbreak. J Adv Nurs 2020 Jul 15;76(7):1486-1487 [FREE Full text] [doi: 10.1111/jan.14369] [Medline: 32202336]
9. Pogoy JM, Cutamora JC. Lived experiences of Overseas Filipino Worker (OFW) nurses working in COVID-19 intensive care units. Belitung Nurs J 2021 Jun 28;7(3):186-194 [FREE Full text] [doi: 10.33546/bnj.1427] [Medline: 37469346]
10. Kandel N, Chungong S, Omaar A, Xing J. Health security capacities in the context of COVID-19 outbreak: an analysis of International Health Regulations annual report data from 182 countries. The Lancet 2020 Mar;395(10229):1047-1053. [doi: 10.1016/s0140-6736(20)30553-5]
11. Ghahramani S, Lankarani KB, Yousefi M, Heydari K, Shahabi S, Azmand S. A systematic review and meta-analysis of burnout among healthcare workers during COVID-19. Front Psychiatry 2021 Nov 10;12:758849 [FREE Full text] [doi: 10.3389/fpsyt.2021.758849] [Medline: 34858231]
12. Shanafelt T, Ripp J, Trockel M. Understanding and addressing sources of anxiety among health care professionals during the COVID-19 pandemic. JAMA 2020 Jun 02;323(21):2133-2134. [doi: 10.1001/jama.2020.5893] [Medline: 32259193]
13. Hodge JG, Gable LA, Cálves SH. Volunteer health professionals and emergencies: assessing and transforming the legal environment. Biosecur Bioterror 2005 Sep;3(3):216-223. [doi: 10.1089/bsp.2005.3.216] [Medline: 16181044]
14. Heidarpoor P, Maniati M, Cheraghi M, Beheshtinasab M, Afshari P. Organization of volunteers in the healthcare system and the type of services provided by them during the COVID-19 pandemic. fmpcr 2021;23(2):169-173. [doi: 10.5114/fmpcr.2021.105909]
15. Wilson J. Volunteering. Annu Rev Sociol 2000 Aug;26(1):215-240. [doi: 10.1146/annurev.soc.26.1.215]
16. Tempski P, Arantes-Costa FM, Kobayasi R, Siqueira MAM, Torsani MB, Amaro BQRC, et al. Medical students' perceptions and motivations during the COVID-19 pandemic. PLoS One 2021;16(3):e0248627 [FREE Full text] [doi: 10.1371/journal.pone.0248627] [Medline: 33730091]
17. Chawłowska E, Staszewski R, Lipiak A, Giernaś B, Karasiewicz M, Bazan D, et al. Student volunteering as a solution for undergraduate health professions education: lessons from the COVID-19 pandemic. Front Public Health 2020 Jan 26;8:633888 [FREE Full text] [doi: 10.3389/fpubh.2020.633888] [Medline: 33575246]
18. Miao Q, Schwarz S, Schwarz G. Responding to COVID-19: community volunteerism and coproduction in China. World Dev 2021 Jan;137:105128 [FREE Full text] [doi: 10.1016/j.worlddev.2020.105128] [Medline: 32834397]
19. Byrne MHV, Ashcroft J, Alexander L, Wan JCM, Harvey A. Systematic review of medical student willingness to volunteer and preparedness for pandemics and disasters. Emerg Med J 2021 Oct 07;39(10):e6. [doi: 10.1136/emermed-2020-211052] [Medline: 34620625]
20. Umar TP, Samudra MG, Nashor KMN, Agustini D, Syakurah RA. Health professional student's volunteering activities during the COVID-19 pandemic: a systematic literature review. Front Med (Lausanne) 2022 Jul 19;9:797153 [FREE Full text] [doi: 10.3389/fmed.2022.797153] [Medline: 35928294]
21. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
22. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat Methods Med Res 2018 Jun 27;27(6):1785-1805. [doi: 10.1177/0962280216669183] [Medline: 27683581]
23. Protogerou C, Hagger M. A checklist to assess the quality of survey studies in psychology. Methods in Psychology 2020 Dec;3:100031 [FREE Full text] [doi: 10.1016/j.metip.2020.100031]
24. Wikurendra EA, Aulia A, Fauzi ML, Fahmi I, Amri I. Willingness-to-pay for urban green space: a meta-analysis of surveys across China. Narra X 2024 Jan 01;1(3):1-13 [FREE Full text] [doi: 10.52225/narrax.v1i3.105]
25. Iqhrammullah M, Duta TF, Alina M, Qanita I, Naufal MA, Henira N, et al. Role of lowered level of serum vitamin D on diabetic foot ulcer and its possible pathomechanism: a systematic review, meta-analysis, and meta-regression. Diabetes Epidemiology and Management 2024 Jan;13:100175. [doi: 10.1016/j.deman.2023.100175]

XSL•FO
RenderX

26. Sandeep M, Padhi BK, Yella SST, Sruthi K, Venkatesan RG, Krishna Sasanka KB, et al. Myocarditis manifestations in dengue cases: a systematic review and meta-analysis. J Infect Public Health 2023 Nov;16(11):1761-1768 [FREE Full text] [doi: 10.1016/j.jiph.2023.08.005] [Medline: 37738692]
27. Pelicic D. The importance of involving student volunteers and nursing students during the COVID-19 pandemic. BJSTR 2022 Feb 10;41(5):33023-33024 [FREE Full text] [doi: 10.26717/bjstr.2022.41.006657]
28. AlSaif HI, AlDhayan AZ, Alosaimi MM, Alanazi AZ, Alamri MN, Alshehri BA, et al. Medical students' response to: willingness and self-perceived competence of final-year medical students to work as part of the healthcare workforce during the COVID-19 pandemic [Response to Letter]. IJGM 2020 Oct;13:865-866. [doi: 10.2147/ijgm.s285816]
29. Anggraeni Y, Ninin RH, Astuti SR. Psychological preparedness for disasters and adaptive performance of student volunteers in handling the Covid-19 pandemic at Padjadjaran University. Psychology 2021 May 30;3(1):1-9 [FREE Full text] [doi: 10.35747/ph.v3i1.117]
30. Larina E, Shumskiy V. Professional motivation and existential fulfillment of medical university's students working with COVID-19 patients (in Russian). Organizational Psychology 2022 Apr 04;12(2):145-156 [FREE Full text]
31. Azar AJ, Khamis AH, Naidoo N, Lindsbro M, Boukhaled JH, Gonuguntla S, et al. Design, implementation and evaluation of a distance learning framework to expedite medical education during COVID-19 pandemic: a proof-of-concept study. J Med Educ Curric Dev 2021 Mar 31;8:23821205211000349 [FREE Full text] [doi: 10.1177/23821205211000349] [Medline: 35392266]
32. Badger K, Morrice R, Buckeldee O, Cotton N, Hunukumbure D, Mitchell O, et al. "More than just a medical student": a mixed methods exploration of a structured volunteering programme for undergraduate medical students. BMC Med Educ 2022 Jan 03;22(1):1 [FREE Full text] [doi: 10.1186/s12909-021-03037-4] [Medline: 34980091]
33. Patel T, Paudyal V, Hadi MA. A qualitative exploration of pharmacy students' opinions and experiences of volunteering during the COVID-19 pandemic. Curr Pharm Teach Learn 2022 Aug;14(8):1004-1014 [FREE Full text] [doi: 10.1016/j.cptl.2022.07.006] [Medline: 36055690]
34. Seah B, Ho B, Liaw SY, Ang ENK, Lau ST. To volunteer or not? Perspectives towards pre-registered nursing students volunteering frontline during COVID-19 pandemic to ease healthcare workforce: a qualitative study. Int J Environ Res Public Health 2021 Jun 21;18(12):6668 [FREE Full text] [doi: 10.3390/ijerph18126668] [Medline: 34205791]
35. Siqueira MAM, Torsani MB, Gameiro GR, Chinelatto LA, Mikahil BC, Tempski PZ, et al. Medical students' participation in the Volunteering Program during the COVID-19 pandemic: a qualitative study about motivation and the development of new competencies. BMC Med Educ 2022 Feb 19;22(1):111 [FREE Full text] [doi: 10.1186/s12909-022-03147-7] [Medline: 35183158]
36. Domaradzki J. 'Who else if not we'. Medical students' perception and experiences with volunteering during the COVID-19 crisis in Poznan, Poland. Int J Environ Res Public Health 2022 Feb 17;19(4):2314 [FREE Full text] [doi: 10.3390/ijerph19042314] [Medline: 35206496]
37. Patel J, Robbins T, Randeva H, de Boer R, Sankar S, Brake S, et al. Rising to the challenge: qualitative assessment of medical student perceptions responding to the COVID-19 pandemic. Clin Med (Lond) 2020 Nov 09;20(6):e244-e247 [FREE Full text] [doi: 10.7861/clinmed.2020-0219] [Medline: 33037028]
38. Siqueira MAM, Torsani MB, Gameiro GR, Chinelatto LA, Mikahil BC, Tempski PZ, et al. Medical students perceptions over their motivations and competencies to volunteer in COVID-19 pandemic - an opportunity to develop leadership skills during medical training. Research Square Preprint posted online on October 5, 2021 [FREE Full text] [doi: 10.21203/rs.3.rs-882146/v1]
39. Chung LYF, Han L, Du Y, Liu L. Reflections on volunteer nurses' work and caring experiences during COVID-19: a phenomenological study. J Res Nurs 2021 Aug 27;26(5):457-468 [FREE Full text] [doi: 10.1177/17449871211007529] [Medline: 35251276]
40. Alslamah T, Altuwaijri EA, Abalkhail A, Alwashmi ASS, Alannas SM, Alharbi AH, et al. Motivational factors influencing undergraduate medical students' willingness to volunteer during an infectious disease pandemic in Saudi Arabia, a cross-sectional study. Eur Rev Med Pharmacol Sci 2022 Sep;26(17):6084-6089 [FREE Full text] [doi: 10.26355/eurrev_202209_29624] [Medline: 36111908]
41. Büssing A, Lindeberg A, Stock-Schröer B, Martin D, Scheffer C, Bachmann HS. Motivations and experiences of volunteering medical students in the COVID-19 pandemic-results of a survey in Germany. Front Psychiatry 2021 Jan 4;12:768341 [FREE Full text] [doi: 10.3389/fpsyt.2021.768341] [Medline: 35058817]
42. Bellomo TR, Prasad S, Bhaumik D, Cartwright J, Zhang Y, Azzouz L, et al. Understanding motivations behind medical student involvement in COVID-19 pandemic relief efforts. BMC Med Educ 2022 Dec 05;22(1):837 [FREE Full text] [doi: 10.1186/s12909-022-03900-y] [Medline: 36471275]
43. Cerbin-Koczorowska M, Przymuszała P, Kłos M, Bazan D, Żebryk P, Uruski P, et al. Potential of volunteering in formal and informal medical education—a theory-driven cross-sectional study with example of the COVID-19 pandemic. Int J Environ Res Public Health 2022 Dec 16;19(24):16955 [FREE Full text] [doi: 10.3390/ijerph192416955] [Medline: 36554834]
44. Drexler R, Hambrecht JM, Oldhafer KJ. Involvement of medical students during the coronavirus disease 2019 pandemic: a cross-sectional survey study. Cureus 2020 Aug 30;12(8):e10147 [FREE Full text] [doi: 10.7759/cureus.10147] [Medline: 33014645]

XSL•FO
RenderX

45. Häikiö K, Andersen JV, Bakkerud M, Christiansen CR, Rand K, Staff T. A retrospective survey study of paramedic students' exposure to SARS-CoV-2, participation in the COVID-19 pandemic response, and health-related quality of life. Scand J Trauma Resusc Emerg Med 2021 Oct 18;29(1):153 [FREE Full text] [doi: 10.1186/s13049-021-00967-2] [Medline: 34663422]
46. Mühlbauer L, Huber J, Fischer M, Berberat P, Gartmeier M. Medical students' engagement in the context of the SARS-CoV-2 pandemic: The influence of psychological factors on readiness to volunteer. GMS J Med Educ 2021;38(6):Doc110 [FREE Full text] [doi: 10.3205/zma001506] [Medline: 34651068]
47. Passemard S, Faye A, Dubertret C, Peyre H, Vorms C, Boimare V, et al. Covid-19 crisis impact on the next generation of physicians: a survey of 800 medical students. BMC Med Educ 2021 Oct 13;21(1):529 [FREE Full text] [doi: 10.1186/s12909-021-02955-7] [Medline: 34645453]
48. Phillips HE, Jennings RB, Outhwaite IR, Grosser S, Chandra M, Ende V, et al. Motivation to impact: medical student volunteerism in the COVID 19 pandemic. Med Sci Educ 2022 Oct 19;32(5):1149-1157 [FREE Full text] [doi: 10.1007/s40670-022-01639-1] [Medline: 36160291]
49. Shi Y, Zhang S, Fan L, Sun T. What motivates medical students to engage in volunteer behavior during the COVID-19 outbreak? A large cross-sectional survey. Front Psychol 2020 Jan 15;11:569765 [FREE Full text] [doi: 10.3389/fpsyg.2020.569765] [Medline: 33519583]
50. AlSaif HI, AlDhayan AZ, Alosaimi MM, Alanazi AZ, Alamri MN, Alshehri BA, et al. Willingness and self-perceived competence of final-year medical students to work as part of the healthcare workforce during the COVID-19 pandemic. IJGM 2020 Sep;13:653-661. [doi: 10.2147/ijgm.s272316]
51. Alsuliman T, Alasadi L, Kasem RA, Hawat M, Almansour M, Al Khalaf R, et al. Assessment of medical students' preparedness and willingness for integration into a war-torn healthcare system: the example of COVID-19 pandemic scenario. Med Confl Surviv 2022 Mar 16;38(1):31-48. [doi: 10.1080/13623699.2021.2015828] [Medline: 34913769]
52. Ahmed Elsheikh EH, Mahjoub Saeed EA, Ahmed Hamed Saleh G, Azhari Gasmalla Gadeltayeb F, M MaliK E. Willingness of medical students to participate in the response to Covid-19 pandemic in Sudan, 2020. Arch Clin Biomed Res 2020;4(6):595-604 [FREE Full text] [doi: 10.26502/acbr.50170128]
53. Khalid M, Khalid H, Bhimani S, Bhimani S, Khan S, Choudry E, et al. Risk perception and willingness to work among doctors and medical students of Karachi, Pakistan during the COVID-19 pandemic: a web-based cross-sectional survey. RMHP 2021 Aug;Volume 14:3265-3273. [doi: 10.2147/rmhp.s310453]
54. Lazarus G, Findyartini A, Putera AM, Gamalliel N, Nugraha D, Adli I, et al. Willingness to volunteer and readiness to practice of undergraduate medical students during the COVID-19 pandemic: a cross-sectional survey in Indonesia. BMC Med Educ 2021 Mar 01;21(1):138 [FREE Full text] [doi: 10.1186/s12909-021-02576-0] [Medline: 33648516]
55. Al Gharash H, Smith M, Cusack L. Nursing students' willingness and confidence to volunteer in a pandemic. SAGE Open Nurs 2021;7:23779608211044615 [FREE Full text] [doi: 10.1177/23779608211044615] [Medline: 34692997]
56. AlOmar RS, AlShamlan NA, AlAmer NA, Aldulijan F, AlMuhaidib S, Almukhadhib O, et al. What are the barriers and facilitators of volunteering among healthcare students during the COVID-19 pandemic? A Saudi-based cross-sectional study. BMJ Open 2021 Feb 18;11(2):e042910 [FREE Full text] [doi: 10.1136/bmjopen-2020-042910] [Medline: 33602709]
57. Karki P, Budhathoki L, Khadka M, Maharjan S, Dhakal S, Pokharel S, et al. Willingness of Nepalese medical and nursing students to volunteer during COVID-19 pandemic: a single-centered cross-sectional study. Ann Med Surg (Lond) 2021 Dec;72:103056. [doi: 10.1016/j.amsu.2021.103056] [Medline: 34812288]
58. Adejimi A, Odugbemi B, Odukoya O, Okunade K, Taiwo A, Osibogun A. Volunteering during the COVID-19 pandemic: attitudes and perceptions of clinical medical and dental students in Lagos, Nigeria. Niger Postgrad Med J 2021;28(1):1-13. [doi: 10.4103/npmj.npmj_379_20] [Medline: 33642318]
59. Domaradzki J, Walkowiak D. Medical students' voluntary service during the COVID-19 pandemic in Poland. Front Public Health 2021;9:618608 [FREE Full text] [doi: 10.3389/fpubh.2021.618608] [Medline: 33928061]
60. Hj Abdul Aziz AAH, Abdul-Mumin KH, Abdul Rahman H. Willingness of university nursing students to volunteer during the COVID-19 pandemic in Brunei Darussalam. Belitung Nurs J 2021;7(4):285-293 [FREE Full text] [doi: 10.33546/bnj.1518] [Medline: 37484895]
61. Nazir M, Ashraf A, Hamid S, Rizvi Z. Willingness of medical students to volunteer for assisting frontline doctors during The COVID-19 pandemic: a cross-sectional study. Journal of Islamic International Medical College (JIIMC) 2022;17(1):46-50 [FREE Full text]
62. Gazibara T, Pesakovic M. Understanding attitudes and willingness to volunteer in COVID-19 hospitals in a setting where medical students were not deployed. BLL 2023;124(05):387-393. [doi: 10.4149/bll_2023_059]
63. Adejimi AA, Okunade KS, Odukoya OO, Roberts AA, Odugbemi BA, Osibogun A. Willingness and motivations towards volunteering during the COVID-19 pandemic: a cross-sectional survey among final year medical students in Lagos, Nigeria. Dialogues Health 2022 Dec;1:100038 [FREE Full text] [doi: 10.1016/j.dialog.2022.100038] [Medline: 36785628]
64. Tran VD, Pham DT, Dao TNP, Pham KAT, Ngo PT, Dewey RS. Willingness of healthcare students in Vietnam to volunteer during the COVID-19 pandemic. J Community Health 2022 Feb;47(1):108-117 [FREE Full text] [doi: 10.1007/s10900-021-01030-y] [Medline: 34468931]

XSL•FO
RenderX

65. Prisca OA, Olawale AM, Adeniyi FO, Adelani WT, Ijeoma LO, Simeon KO, et al. Knowledge, attitude and willingness of Nigerian nursing students to serve as volunteers in covid-19 pandemic. Int J Nurs Midwifery 2021 Jan 31;13(1):1-10. [doi: 10.5897/ijnm2020.0448]
66. Feng Y, Zhang X, Wang H, Pan Y, Chen S, Chen Z. Willingness of medical undergraduates to work as volunteers for anti-epidemic of COVID-19: reflections on medical education in the post-epidemic era of China. Research Square. Preprint posted online on March 1, 2021 2021 [FREE Full text] [doi: 10.21203/rs.3.rs-360416/v1]
67. Byrne M, Ashcroft J, Wan J, Alexander L, Harvey A, Arora A, et al. Examining medical student volunteering during the COVID-19 pandemic as a prosocial behaviour during an emergency. Postgrad Med J 2023 Jul 21;99(1174):883-893. [doi: 10.1093/postmj/qgad015] [Medline: 37002858]
68. Yordanova R, Kyuchukova S, Andonova A, Nikolova M, Platikanova M. Voluntary behavior of students studying medical specialties in the context of a pandemic. Journal of IMAB - Annual Proceeding 2023 Oct 16;29(4):5163-5167 [FREE Full text] [doi: 10.5272/jimab.2023294.5163]
69. Joseph N, Manasvi M. Perception of medical undergraduate students regarding their readiness to volunteer in relief activities during the COVID-19 pandemic: a multi-institutional study carried out in South India. fmpcr 2022;24(2):120-125. [doi: 10.5114/fmpcr.2022.115872]
70. Magdas T, Jolobai A, Simonescu-Colan R, Mosteanu EO, Pop TA. Practical experience as a determining factor of preparedness of medical and nursing students in Romania during COVID-19 pandemic. Med Pharm Rep 2022 Jan;95(1):54-58 [FREE Full text] [doi: 10.15386/mpr-1963] [Medline: 35720243]
71. Susanti RD, Yudianto K, Mulyana AM, Amalia IN. A systematic scoping review of motivations and barriers in COVID-19 volunteering among health students: the potential for future pandemic volunteers. JMDH 2023 Jun;Volume 16:1671-1681. [doi: 10.2147/jmdh.s411896]
72. Mihatsch L, von der Linde M, Knolle F, Luchting B, Dimitriadis K, Heyn J. Survey of German medical students during the COVID-19 pandemic: attitudes toward volunteering versus compulsory service and associated factors. J Med Ethics 2022 Sep 21;48(9):630-636 [FREE Full text] [doi: 10.1136/medethics-2020-107202] [Medline: 34021060]
73. Zhang K, Peng Y, Zhang X, Li L. Psychological burden and experiences following exposure to COVID-19: a qualitative and quantitative study of Chinese medical student volunteers. Int J Environ Res Public Health 2021 Apr 13;18(8):4089 [FREE Full text] [doi: 10.3390/ijerph18084089] [Medline: 33924448]
74. Gómez-Durán EL, Fumadó CM, Gassó AM, Díaz S, Miranda-Mendizabal A, Forero CG, et al. COVID-19 pandemic psychological impact and volunteering experience perceptions of medical students after 2 years. Int J Environ Res Public Health 2022 Jun 20;19(12):7532 [FREE Full text] [doi: 10.3390/ijerph19127532] [Medline: 35742780]
75. Achar Fujii RN, Kobayasi R, Claassen Enns S, Zen Tempski P. Medical students' participation in extracurricular activities: motivations, contributions, and barriers. A qualitative study. AMEP 2022 Sep;Volume 13:1133-1141. [doi: 10.2147/amep.s359047]
76. Ljubetić M. On volunteering – mostly from female perspective. Przegląd Badań Edukacyjnych (Educational Studies Review) 2023;42:189-215 [FREE Full text]
77. Voicu B, Voicu M. Volunteering in Eastern Europe: one of the missing links? Institutul de Cercetare a Calită ii Vie ii. Bucharest, Romania: Romanian Academy of Sciences; 2003. URL: https://www.iccv.ro/valori/texte/04.Bogdan&Malina%20Voicu.pdf [accessed 2024-03-21]
78. Enjolras B. Explaining the varieties of volunteering in Europe: a capability approach. Voluntas 2021 Apr 05;32(6):1187-1212. [doi: 10.1007/s11266-021-00347-5]
79. Hayes F, Clark J, McCauley M. Healthcare providers' and managers' knowledge, attitudes and perceptions regarding international medical volunteering in Uganda: a qualitative study. BMJ Open 2020 Dec 12;10(12):e039722 [FREE Full text] [doi: 10.1136/bmjopen-2020-039722] [Medline: 33310799]

Abbreviations

***P*-Het:** *P* value of heterogeneity
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
P-tot: *P* value of total effect
Q-SSP: Quality Assessment Checklist for Survey Studies in Psychology
WHO: World Health Organization

XSL•FO
RenderX

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 16.01.24; peer-reviewed by M Mahmic Kaknjo, C Gibson; comments to author 14.02.24; revised version received 20.02.24; accepted 23.02.24; published 15.04.24.

Please cite as:
Mahsusi M, Hudaa S, Nuryani N, Fahmi M, Tsurayya G, Iqhrammullah M
Global Rate of Willingness to Volunteer Among Medical and Health Students During Pandemic: Systemic Review and Meta-Analysis
JMIR Med Educ 2024;10:e56415
URL: https://mededu.jmir.org/2024/1/e56415
doi:10.2196/56415
PMID:38621233

©Mahsusi Mahsusi, Syihaabul Hudaa, Nuryani Nuryani, Mustofa Fahmi, Ghina Tsurayya, Muhammad Iqhrammullah. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.04.2024. 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.

XSL•FO
RenderX

Review

Curriculum Frameworks and Educational Programs in AI for Medical Students, Residents, and Practicing Physicians: Scoping Review

Raymond Tolentino[1], BHSc, MSc; Ashkan Baradaran[1], MSc, MD; Genevieve Gore[2], BA, MLIS; Pierre Pluye[1†], MD, PhD; Samira Abbasgholizadeh-Rahimi[1,3,4,5], BEng, PhD

[1]Department of Family Medicine, McGill University, Montreal, QC, Canada
[2]Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montreal, QC, Canada
[3]Mila - Quebec AI Institute, Montreal, QC, Canada
[4]Lady Davis Institute for Medical Research, Herzl Family Practice Centre, Jewish General Hospital, Montreal, QC, Canada
[5]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, QC, Canada
[†]deceased

Corresponding Author:
Samira Abbasgholizadeh-Rahimi, BEng, PhD
Department of Family Medicine
McGill University
5858 Chemin de la Côte-des-Neiges
Montreal, QC, H3S 1Z1
Canada
Phone: 1 514 399 9218
Email: samira.rahimi@mcgill.ca

Abstract

Background: The successful integration of artificial intelligence (AI) into clinical practice is contingent upon physicians' comprehension of AI principles and its applications. Therefore, it is essential for medical education curricula to incorporate AI topics and concepts, providing future physicians with the foundational knowledge and skills needed. However, there is a knowledge gap in the current understanding and availability of structured AI curriculum frameworks tailored for medical education, which serve as vital guides for instructing and facilitating the learning process.

Objective: The overall aim of this study is to synthesize knowledge from the literature on curriculum frameworks and current educational programs that focus on the teaching and learning of AI for medical students, residents, and practicing physicians.

Methods: We followed a validated framework and the Joanna Briggs Institute methodological guidance for scoping reviews. An information specialist performed a comprehensive search from 2000 to May 2023 in the following bibliographic databases: MEDLINE (Ovid), Embase (Ovid), CENTRAL (Cochrane Library), CINAHL (EBSCOhost), and Scopus as well as the gray literature. Papers were limited to English and French languages. This review included papers that describe curriculum frameworks for teaching and learning AI in medicine, irrespective of country. All types of papers and study designs were included, except conference abstracts and protocols. Two reviewers independently screened the titles and abstracts, read the full texts, and extracted data using a validated data extraction form. Disagreements were resolved by consensus, and if this was not possible, the opinion of a third reviewer was sought. We adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist for reporting the results.

Results: Of the 5104 papers screened, 21 papers relevant to our eligibility criteria were identified. In total, 90% (19/21) of the papers altogether described 30 current or previously offered educational programs, and 10% (2/21) of the papers described elements of a curriculum framework. One framework describes a general approach to integrating AI curricula throughout the medical learning continuum and another describes a core curriculum for AI in ophthalmology. No papers described a theory, pedagogy, or framework that guided the educational programs.

Conclusions: This review synthesizes recent advancements in AI curriculum frameworks and educational programs within the domain of medical education. To build on this foundation, future researchers are encouraged to engage in a multidisciplinary

XSL•FO
RenderX

approach to curriculum redesign. In addition, it is encouraged to initiate dialogues on the integration of AI into medical curriculum planning and to investigate the development, deployment, and appraisal of these innovative educational programs.

International Registered Report Identifier (IRRID): RR2-10.11124/JBIES-22-00374

(JMIR Med Educ 2024;10:e54793) doi:10.2196/54793

KEYWORDS

artificial intelligence; machine learning; curriculum; framework; medical education; review

Introduction

The field of medicine is constantly evolving with new technologies and discoveries [1]. One of the emerging technologies is artificial intelligence (AI), a simulation of human intelligence powered by machines, specifically computer systems that use machine learning and deep learning [2]. AI allows for complex decision-making and the ability for human capabilities such as tasks done by physicians and other health care providers [2]. Through recent advancements, AI has begun to become an innovation to be adopted in the field of medicine [3]. Current fields using this type of technology are radiology [4], pathology [5], dermatology [6], primary care [7], and surgery [8], among other fields of medicine [9]. These AI-related medical innovations can be seen through different ways, including robot-assisted surgical procedures, diagnosis and risk assessments, as well as the development and customization of drugs [3,10]. However, to move forward with the implementation of AI in clinical practice, physicians need to have a better understanding of AI and how to use it in clinical practice [11].

Although medicine has seen major changes over the last decades, medical education is still largely based on traditional curricula [12]. It often lacks fundamental concepts and even basic familiarization with AI and other emerging technologies [13]. A recent survey by Stanford Medicine found that 44% (230/523) of physicians and 23% (48/210) of medical students and residents reported that their education had not been helpful in preparing for new technologies in health care [14]. Currently, there are no accreditation requirements related to AI [15]. The knowledge gap between engineers, clinicians, and scientists continue to grow as health care moves to a more digital environment, which will ill-prepare young physicians who will work with AI-enabled tools and technologies [16,17].

At the moment, AI is beginning to enter the field of medical education through its uses in learning support, assessments of students' learning, and curriculum review [2]; however, there are several publications urging institutes and clinical educators to begin integrating AI educational concepts into their medical curricula [12,13,15-20]. There have been efforts to include AI education globally within each level of medical training. These efforts are led by national medical associations such as the UK National Health Service [21], the US American Medical Association [22], and Canada's Royal College of Physicians and Surgeons [23]. They have released documents recommending policies for integrating AI within their respective medical educational institutions [21-23]. This highlights the importance of the work on the intersection of medical education and AI around the world. Surveys of medical trainees have also supported the need to incorporate the teaching of AI in the undergraduate medical curriculum [24]. To our knowledge, there are no medical schools with formal required courses on AI in health care. While still uncommon, the importance of AI medical education has been identified and acted on at some institutions, such as Duke University, which offers a training course called *Machine Learning School for the School of Medicine* [12]. Other institutions have also developed elective courses to teach AI to residents, such as in radiology [25]. As AI is being used in a variety of fields within medicine [9], it is important to have a structured and validated curriculum framework because future medical providers will be exposed to these types of technologies depending on their chosen fields.

A curriculum framework is a document which describes "the educational environment in which syllabuses (or subject-specific outlines of objectives, outcomes, content and appropriate assessment and teaching methodologies) can be developed" [26]. Curriculum frameworks can be described as educational road maps to teaching and learning. For example, a curriculum framework was created for global health concepts in family medicine education [27]. Medical educators work regularly with frameworks to inform the appropriate learning, assessment, and performance of the health care workforce [28]. Frameworks are tools that can inform the delivery of teaching and curricula development as well as inspire innovation in health care education. There are various aspects that can be included in curriculum frameworks and how they may be used for other disciplines. Obadeji [29] clearly describes the common elements of curriculum frameworks for health professional education, which include (1) the need and the purpose of a curriculum or a program, (2) learning objectives and outcomes, (3) course content that will facilitate the accomplishment of the objectives or learning outcomes, (4) organization of the content, and (5) implementation of curriculum—educational strategies and methods of assessment.

Due to the broad nature of this topic and its prospective limited data, a scoping review is the most appropriate method. Previous reviews exploring topics surrounding AI and medical education have focused on the application of AI in medical education [2,30], attitudes of medical students toward AI [31], and gaps of AI learning within medical education [32]. A recent review of AI educational programs and competencies for health care professionals was published [33]; however, due to the increase in attention on this topic, further reviews must be conducted. Furthermore, the previous reviews had some limitations, such as the exclusion of continuing professional education and the lack of investigating learning theories, pedagogies, and frameworks of their identified AI educational programs. Our review will cover these limitations by focusing on the medical

XSL•FO
RenderX

education continuum as the developed AI educational programs for medical students, residents, and practicing physicians can help medical educators navigate the learning pathway for current and future physicians. Moreover, no review has focused on examining curriculum frameworks that guide AI concepts within medical education.

Thus, we conducted a scoping review of published literature on AI curricula being used in medical education. Overall, the main aim of this scoping review is to synthesize knowledge from the literature on curriculum frameworks and current educational programs that focus on the teaching and learning of AI for medical students, residents, and practicing physicians. More specifically, we aim to investigate the details of the current educational programs including (1) the framework, pedagogy, or theory used; (2) the delivery of the educational program; (3) the curricular content; and (4) the evaluation of the program, to inform future research on developing or adopting AI curriculum frameworks for use in medical educational institutions.

Methods

Protocol and Registration

The protocol for this review was developed in accordance with the Joanna Briggs Institute (JBI) Reviewers Manual for Evidence Synthesis [34] and guided by the methodological framework developed by Arksey and O'Malley [35], supplemented by Levac et al [36]. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [37] was used when reporting results, and is reported in Multimedia Appendix 1. The protocol was registered on Open Science Framework Registries and published on JBI Evidence Synthesis [38].

Eligibility Criteria

Participants

To be eligible for inclusion, the participants of the studies had to fall under the population that provided medical education or received medical education, which includes medical students. This includes undergraduate medical education (UME), residents or postgraduate medical education (PGME), and practicing physicians (continuing medical education [CME]) at any health care setting (ie, primary, secondary, and tertiary care).

Exposure

Included studies must describe either a curriculum framework or programs for AI education within medicine. The frameworks and programs must focus on learning about AI and how to use AI-specific tools for the medical profession.

Outcome

For the purpose of this review, all elements of a curriculum framework described by Obadeji [29], either in part or as a whole, were considered and reported. Included papers may also describe current and developed educational programs for AI training in medicine. These educational programs have already been developed or evaluated, and papers describing recommendations of what to teach or programs not yet developed were not considered. This review focused on any framework, theory, or pedagogy mentioned within the program; the delivery of the educational program (eg, course and workshop); and curricular content (eg, learning topics and learning objectives); if the educational program was evaluated, it was described according to the model of training evaluation developed by Kirkpatrick et al [39].

Information Sources

All types of studies were included, such as theoretical work, program descriptions, and empirical studies. Commentaries, reviews, perspectives, opinions, as well as position papers and any companion papers associated were also included. All study designs for empirical studies using qualitative, quantitative, or mixed methods studies were eligible for inclusion. These include experimental and quasi-experimental studies (such as randomized controlled trials, quasi-randomized controlled trials, nonrandomized clinical trials, interrupted time series, and controlled before-and-after studies), observational studies (such as cohort, case control, cross-sectional, and case series studies), qualitative studies (such as ethnography, narrative, phenomenological, grounded theory, and case studies), and mixed methods studies. Conference abstracts and protocols were excluded. Conference abstracts often contain preliminary findings that may not be as comprehensive or validated as full-text articles. As they are brief summaries of studies, they may lack the detailed methodology and results needed for a thorough understanding and synthesis in our scoping review. Furthermore, as protocols are plans of how to conduct the research, they do not provide findings or results that are necessary for a scoping review's goal to map the extent, range, and nature of research activity in a given field. Therefore, considering the provided justifications, we decided to exclude conference abstracts and protocols.

Search Strategy

The following search strategy has been developed by a specialized librarian. The text words contained in the titles and abstracts of relevant papers and the index terms used to describe the papers were used to develop a full search strategy. The search strategy took an iterative approach, initially using general terms such as "artificial intelligence," with the later addition of variations and synonyms such as "deep learning" and "machine learning." In addition, terms for the concepts of medical education and curriculum were added. An initial limited search of MEDLINE (PubMed) was conducted to identify relevant papers on this topic. An information specialist (GG) performed a comprehensive search in the following bibliographic databases: Ovid MEDLINE, Ovid Embase, CENTRAL (Cochrane Library), CINAHL, and Scopus. To identify any unpublished frameworks, web searches of Google, New York Academy of Medicine Grey Literature Report, and medical learning institutional websites were searched. Reference lists of all included research papers and all relevant reviews were back searched, and Google Scholar was used for forward citation tracking to identify further studies.

Papers were restricted to English and French due to the constraints of the research team. Papers were also restricted by date beginning in the year of 2000, as during the 1950s to the late 1990s AI was in its early phase with reduced funding and

XSL•FO
RenderX

interest of AI in medicine [40]. The initial search was conducted in November 2021 and later updated in May 2023.

Selection of Sources of Evidence

Following the search, all identified records were collated and uploaded into a reference management system, EndNote (version 20.3; Clarivate Analytics), where duplicates were removed. Following a pilot test with 2 reviewers (RT and AB) using 10% (510/5104) of the studies, titles and abstracts were then screened using Rayyan, a web-based research platform, by 2 independent reviewers (RT and AB) for assessment against the inclusion criteria for the review. The full text of selected citations was assessed in detail against the inclusion criteria by 2 independent reviewers (RT and AB). Any disagreements that arose between the 2 reviewers were resolved by a third reviewer (SAR).

Data Extraction

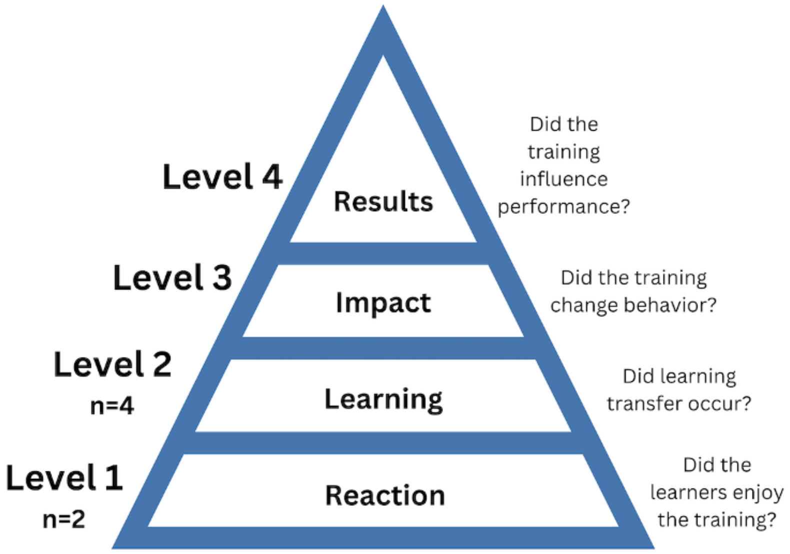
Data were extracted by 2 reviewers (RT and AB) using a data extraction tool on an Excel (Microsoft Corp) sheet developed and validated by the team. The data extraction tool was created and validated using previously validated data extraction tools [32-34] and input from experts in the field. It focuses on key characteristics related to curriculum framework elements and educational program details. Any disagreements that arose between the 2 reviewers were resolved by a third reviewer (SAR). Data on paper characteristics (eg, authors, title, country of origin, type of study, and year of publication), curriculum framework elements, and educational program details were extracted.

Synthesis of Results

The results of the review are presented as a table of the data extracted from the included literature to highlight the key findings with respect to the aims of this scoping review. Descriptive statistics (eg, frequency) was used when reporting paper characteristics and education program details. For curriculum frameworks described, reviewers presented main elements, including (1) the need and purpose of curriculum, (2) the learning objectives and outcomes, (3) course content that will facilitate the accomplishment of the objectives or learning outcomes, (4) the organization of the content, and (5) implementation of curriculum. For current educational programs described, reviewers independently recorded and presented data on the framework, theory, or pedagogy that may have been used; the delivery of the educational program; and curricular content; and if the educational program was evaluated, it was described according to the model of training evaluation developed by Kirkpatrick et al [39].

The model of training evaluation developed by Kirkpatrick et al [39] was used to categorize educational outcomes evaluations (Figure 1 [39]). Level 1 describes the degree to which learners find the training favorable, engaging, and relevant; level 2 describes the degree to which learners acquire the intended knowledge, skills, confidence, and commitment based on their participation in the training; level 3 describes the degree to which learners apply what they learned during training when they are back to work; and level 4 describes the degree in whether the targeted outcomes resulted from the training program at an organizational level [39]. A narrative summary accompanied [41] the charted results and described what and how AI curriculum content is being delivered to trainees of various medical education stages.

Figure 1. Outcomes (and their meaning) of the 4-level training evaluation developed by Kirkpatrick and Kirkpatrick [39].



Quality Appraisal of Included Studies

Due to the nature of this review, the methodological quality or risk of bias of the included papers was not appraised, which is consistent with scoping reviews guidelines [34,37].

Results

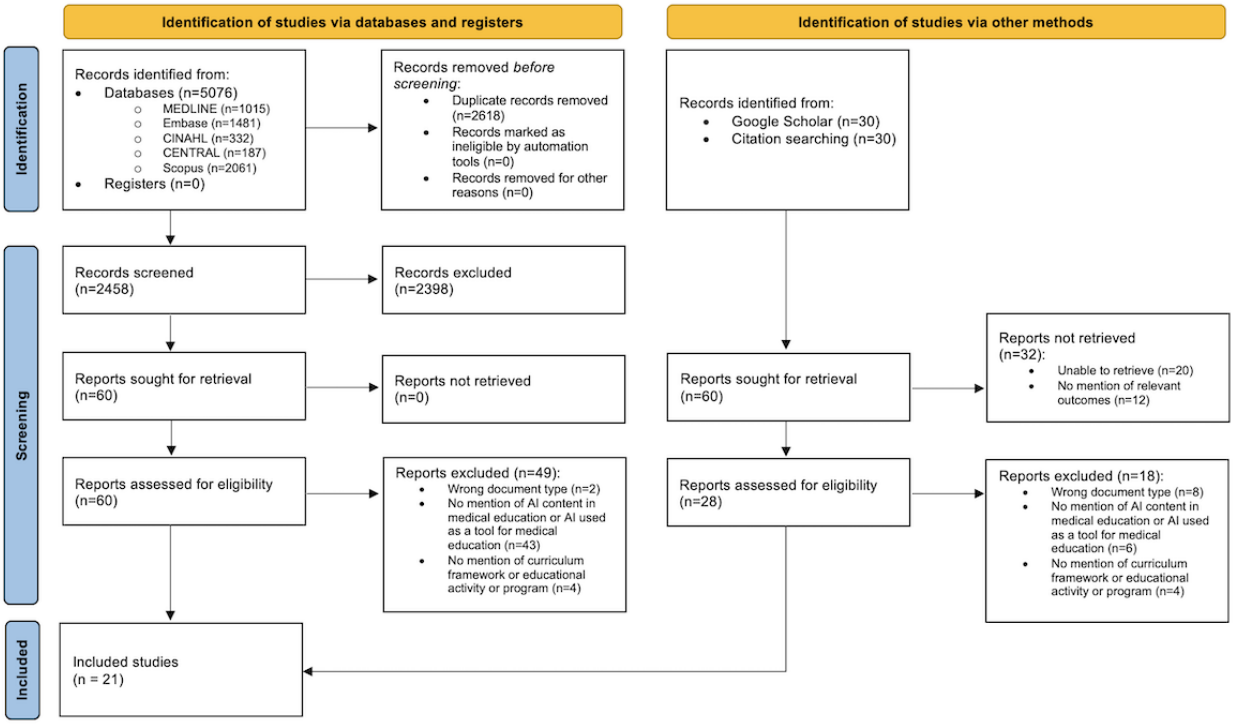
Search Results

From the systematic search, 5076 total papers were identified. These papers were extracted from web-based databases, and the computer software EndNote was used to manage these

XSL•FO
RenderX

references. Following removal of duplicates on EndNote, 2458 papers were uploaded to Rayyan software and screened by title and abstract. After abstract and title screening, 60 papers remained for full-text screening. A gray literature search identified 60 papers from Google Scholar and reference lists, from which 28 (47%) papers were retrieved for full-text screening, and 32 (53%) papers were not retrieved or were irrelevant. Following full-text screening of databases and gray literature, 21 papers were included for further analysis [12,25,31-33,42-57]. Refer to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram (Figure 2) [58].

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.



Characteristics of the Included Studies

Data was collected from 21 included studies and summarized in Multimedia Appendix 2 [12,25,31-33,42-57]. A total of 12 studies were published in the United States [12,31,32,42-45,48,51,52,54,57]; 6 in Canada [33,46,47,49,50,55]; and 1 each in Germany [25], Korea [53], and Oman [56] (Multimedia Appendix 3). The earliest publication retrieved was from 2016, with 77% (15/21) of the papers [25,31-33,42,43,45,47,49-51,54-57] published in the last 3 years since the pandemic began (Multimedia Appendix 3). From the 21 studies, 6 (29%) were reviews [31-33,45,53,54], 4 (19%) were commentaries [44,47,50,51], 4 (19%) were opinions [12,48,52,56], 3 (14%) were perspectives [43,55,57], 3 (14%) were empirical studies using a cross-sectional survey design [25,42,49], and 1 (5%) was a position paper [46].

In terms of setting, 43% (9/21) of the papers mentioned multiple levels of education ranging from UME, PGME, to CME [12,31-33,46,50,51,54,56], while 24% (5/21) of the papers specified on UME [42,44,47,53,55], 19% (4/21) of the papers specified on PGME [48,49,52,57], and 14% (3/21) of the papers were focused on CME [25,43,45]. Across the 21 included studies, 19 (90%) altogether described 30 current or previously offered educational programs [12,25,31-33,42-55] and 2 (10%) described elements of a curriculum framework [56,57].

Curriculum Framework Elements

Two papers described the main elements of a curriculum framework (Table 1) [56,57]. The first paper was an opinion paper by Masters [56], which mentions 3 of the 5 elements of a curriculum framework. The paper describes the need and purpose of a curriculum, course content, and brief descriptions in terms of organization of content. The second paper to describe elements of a curriculum framework was the study by Valikodath et al [57], which provides information for all 5 elements. This includes the main purpose of an ophthalmology AI curriculum, the learning objectives, course content topics, a 4-year resident organization plan, and implementation of the curriculum, as outlined in Table 1. We noticed similarities in relation to what medical trainees should learn, as emphasized in Figure 3 [56,57].

XSL•FO
RenderX

Table 1. Curriculum framework studies' characteristics (n=2).

	Masters [56]	Valikodath et al [57]
Program audience	• Multiple (undergraduate medical education, PGME[a], and continuing medical education; general)	• PGME; ophthalmology
Need or purpose	• This general framework will • allow medical schools to assess their own position in relation to AI[b] projects • place these projects within that framework to better understand them • develop new projects based on their needs	• The goals of a core AI curriculum in ophthalmology include the following: • recognizing major studies and discoveries of AI with regard to ophthalmology • identifying the limitations of AI • learning about potential applications in clinical practice
Learning objectives	—[c]	• Learning objective 1: To understand the basic components of AI • Learning objective 2: To identify the limitations of AI, especially in health care and research • Learning objective 3: To summarize current uses of AI in ophthalmology and evaluate the primary literature • Learning objective 4: To know how to potentially apply AI into clinical practice, including telemedicine and web-based visits
Course content	• Topic 1. AI as AI • Option A: the basics • "...we need now to teach AI literacy and a basic understanding of Data Management and AI concepts, models and terminology (such as big data (and the growing number of Vs), data mining, machine learning, deep learning, supervised and unsupervised learning, natural language processing and neural networks) [...]" • Option B: more advanced • "...the curriculum will need to be adjusted, and electives, projects dealing with AI applications in solving medical problems, and assessing AI evaluations would be a starting point [...]" • Option C: common for all • "In all cases where AI is taught, the current limitations of AI need to be identified [...] Understanding these systems will be necessary to evaluate the applicability and appropriateness of solutions. [...]" • Topic 2. AI in medical systems • "Students will need to know the mechanics and processes of AI systems that they will be expected to use [...]" • Topic 3. Self-awareness • "There needs to be a self-awareness, in which the doctor is not merely using the tool, but is engaged in a cooperative exercise with the tool. This co-operation does not imply compliance, but rather operating together [...]" • Topic 4. Ethical, legal, and social implications • "Related to the health professionals' perception of themselves and their role in healthcare, a host of Ethical, Legal and Social Implications emerge, and medical students will need to consider these and the questions they raise [...]"	• Topic 1. Basic mathematics and statistics • Topic 2. Fundamentals of AI, machine learning, deep learning • Topic 3. How to evaluate AI literature • Topic 4. Review of seminal articles • Topic 5. Clinical applications • Topic 6. Surgical applications • Topic 7. Ethics • Topic 8. Medicolegal implications • Topic 9. Health disparities • Topic 10. Humanization of medicine

XSL•FO
RenderX

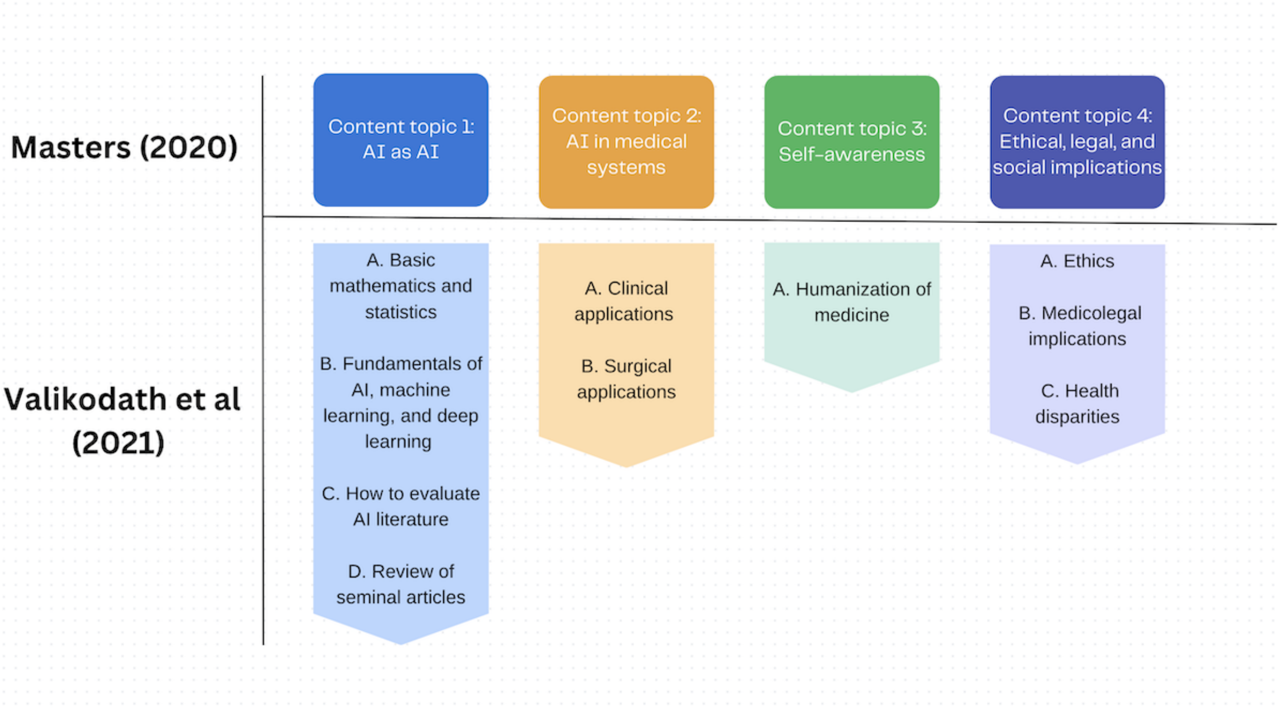
	Masters [56]	Valikodath et al [57]
Organization of content	—	• Year 1 and 2: understand basic statistics and mathematics • Year 1-3: become familiar with components and functions of AI • Year 1-4: use web-based learning tools (articles, lectures, modules, and case-based learning) • Year 2-4: assess primary literature on current AI systems in ophthalmology • Year 3 and 4: understand integration of AI into clinical practice
Implementation of content	—	• Teaching tools (curriculum delivery and assessment methods) • background reading: articles on concepts in AI • case studies • web-based lecture series from experts in the field (regularly updated) • interactive webinars and modules • surgical simulation-based training • standardized tests

[a]PGME: postgraduate medical education.

[b]AI: artificial intelligence.

[c]Not applicable.

Figure 3. The comparison between the course content described by Masters [56] and Valikodath et al [57].



From our comparisons, we found that the main curricular topics presented by Masters [56] appropriately corresponded to the curricular topics presented by Valikodath et al [57], for example, a main curricular topic of "AI in Medical Systems," which describes the way in which students should learn the structures and processes of AI systems that they will be using in the future. This corresponds to "Clinical Applications" and "Surgical Applications" in which the content is targeted into learning how to use AI applications for ophthalmology. It appears that Masters' [56] framework on course content can work as the foundation on what curricular concepts a program should include. This is because previous reviews have detailed similar curricular topics currently being taught.

Current Educational Programs

From the 19 papers that describe an educational program, 30 current or previously offered educational programs were identified (Table 2) [12,25,31-33,42-55]. A total of 13 papers

XSL•FO
RenderX

described, mentioned, or presented 24 educational programs [12,31-33,43,45-47,50-54], while 6 papers described and assessed 6 educational programs using evaluation methods (eg, survey and test scores) [25,42,44,48,49,55]. No papers described a theory, pedagogy, or framework that guided the educational program.

Table 2. Educational program characteristics (n=30 educational programs described in 19 papers).

Characteristic	Frequency, n (%)
Type of educational program	
Course	15 (50)
Project	4 (13)
Lecture (dedicated to artificial intelligence)	4 (13)
Webinar	3 (10)
Educational summit or conference	2 (7)
Workshop	2 (7)
Pathway of education and program audience	
Undergraduate medical education	17 (57)
General topics	16 (94)
Radiology	1 (6)
Postgraduate medical education	5 (17)
Radiology	5 (100)
Continuing medical education, n (%)	8 (27)
General topics	4 (50)
Radiology	3 (34)
Cardiology	1 (13)
Delivery setting	
Medical school	23 (77)
National or international medical association	7 (23)

Educational Program Delivery

Of the 30 educational programs described collectively in the 19 remaining papers, 15 (50%) programs were courses, 4 (13%) were project-related initiatives, 4 (13%) were didactic lectures dedicated to AI, 3 (10%) were webinars, 2 (7%) were an educational summit or conference, and 2 (7%) were 1-day workshops. “AI courses were defined as elective courses focused on AI-based education. Didactic lectures dedicated to AI are 1 or 2 lectures that mention AI education but not a full course. There were 77% (23/30) educational programs delivered from a medical school, while 23% (7/30) were delivered from recognized national or international medical associations. Furthermore, it is important to clarify that some papers used multiple educational program delivery approaches. For example, an included paper explained their educational intervention was a course, but this course included didactic lectures, mentorship, and a final project. However, the reporting of this educational program’s delivery is classified as only a course and not counted as another delivery approach to minimize confusion.

Of the 30 educational programs described collectively in the 19 remaining papers, 17 (57%) UME educational programs were targeted toward medical students. Of these 17 programs, 16 (94%) were UME educational programs focused on general topics of AI in medicine and 1 (6%) was an UME educational program focused on radiology concepts. In total, 17% (5/30) of the postgraduate educational programs were for residents who were in the radiology specialty. Of the 30 educational programs, 8 (26%) were specified for practicing physicians (n=4, 50% were CME educational programs focused on general topics of AI in medicine, n=3, 37% were radiology for CME education, and n=1, 13% was in cardiology for CME). The educational program characteristics are provided in Table 2.

Curricular Content

The following curricular concepts were adapted and framed from previous similar reviews [32,33]. The curricular content and concepts were divided into 2 types: theoretical curricular concepts and application-based curricular concepts. The subcategories and their descriptions are outlined in Table 3. The following describe the theoretical curricular concepts: fundamental of AI for using AI systems (15/19, 79%) [12,25,31-33,42-47,49,51-53]; fundamentals of health care data science for using AI systems (10/19, 53%) [12,25,31-33,45,47,49-51]; strengths and limitations of AI (9/19, 47%) [31-33,45-49]; and ethical, legal, and economic considerations of AI systems (11/19, 58%) [12,25,31-33,42,45-48,52]. The following describe the application-based curricular concepts: applications of AI systems

XSL•FO
RenderX

(19/19, 100%) [12,25,31-33,42-55], operating AI systems in health care settings (10/19, 53%) [12,25,31-33,43,46,47,52,55], impact of AI on clinical reasoning and medical decision-making (7/19, 37%) [12,25,31-33,43,55], communication of AI results to patients (4/19, 21%) [12,31-33], and critical appraisal of AI systems (7/19, 37%) [12,31-33,50,53,54].

Table 3. Curricular concepts mentioned in the educational program papers (n=19).

AI[a] curricular concept	Description of curricular concept	Reference
Theoretical curricular concepts (learning what is AI in medicine)[b]		
Fundamental of AI for using AI systems	Providing an overview of AI definitions and concepts, including machine learning; natural language processing; and the basics of data acquisition, cleaning, analysis, and visualization	[12,25,31-33,42-47,49,51-53]
Fundamentals of health care data science for using AI systems	Providing an overview of the environment supporting AI, which includes biostatistics, big data, and the use and processing of data by algorithms and machine learning	[12,25,31-33,45,47,49-51]
Strengths and limitations of AI	Promoting learners' comprehension of the advantages and limitations of various AI systems, such as factors that affect AI accuracy (eg, sources of error and bias)	[31-33,45-49]
Ethical, legal, and economic considerations of AI systems	Developing a comprehensive understanding of ethics, equity, inclusion, patient rights, and confidentiality, alongside regulatory frameworks, policy considerations, liability, and intellectual property issues related to using AI systems as well as grasping the potential alterations to business or clinical processes resulting from the integration of AI technologies	[12,25,31-33,42,45-48,52]
Application-based curricular concepts (learning how to use AI for clinical practice)[c]		
Applications of AI systems	Familiarizing with clinical application of AI systems in clinical practice to understand how they are used	[12,25,31-33,42-55]
Operating AI systems in health care settings	Understanding how to embed and engage with AI tools into clinical settings and workflows (eg, learning to engage in data mining tools or how to properly communicate with AI systems to receive meaningful results)	[12,25,31-33,43,46,47,52,55]
Impact of AI on clinical reasoning and medical decision-making	Having the ability to understand, interpret, and apply results of AI systems in clinical practice	[12,25,31-33,43,55]
Communication of AI results to patients	Communicate findings to patients in a personalized and meaningful manner and engage in discussions regarding the use of AI in the medical decision-making process	[12,31-33]
Critical appraisal of AI systems	Acquiring proficiency in assessing diagnostic and therapeutic algorithms powered by AI to ensure safe and effective integration and use in clinical practice	[12,31-33,50,53,54]

[a]AI: artificial intelligence.

[b]The mentioned concepts encompass foundational learning that serves as the basis of medical artificial intelligence educational philosophy and clinical practice.

[c]The mentioned concepts prioritize the practical applications of artificial intelligence knowledge and skills in a clinical context.

Assessment of Educational Outcomes

Of the 19 papers, 6 (32%) presented the results of their evaluation of an educational program (Table 4) [25,42,44,48,49,55]. Two papers described only level 1 evaluation outcomes (eg, learner reaction and satisfaction with the educational program) in which participants were overall very satisfied with the AI content learned [42,48]. Four papers described level 2 evaluation outcomes (eg, change in attitude, knowledge, or skill) in which learners demonstrated acquisition of a variety of competencies (linear algebra pertaining to AI and basics of AI) and skills (eg, incorporate medical decisions given by an algorithm and implementing AI in clinical practice) [25,44,49,55] where two of these papers also evaluated level 1 outcomes [25,49]. There were no outcomes that could be categorized as level 3 or level 4; thus, the program evaluations did not comment on the change in behavior or affect at the organizational level or on patient outcomes.

XSL•FO
RenderX

Table 4. Studies describing evaluation outcomes (n=6).

Study	Educational program	Levels and outcomes of the model of training evaluation developed by Kirkpatrick and Kirkpatrick [39]
Alderson et al [42], 2021	Course	• Level 1: "...satisfaction scores of 4.4/5.0 (n=13) [...]"
Barbour et al [44], 2019	Educational summit	• Level 2: "...there was a general belief [about 70% from the figures] that AI would make health care less humanistic." • Level 2: "...did not observe a meaningful shift in attitudes regarding the desire to take a leadership role in developing or implementing AI [...]" • Level 2: "Attendees arrived believing they had a poor baseline understanding of AI's role in health care, and left the summit with an enhanced understanding of the topic [...]"
Hedderich et al [25], 2021	Course	• Level 1: "The participants were overall very satisfied with the study material and the organization of the course, and deemed the content of the course important for their work as a clinician or scientist." • Level 2: "...self-perceived skills improved in all areas, for understanding Python code as well as for understanding concepts of linear algebra pertaining to AI." • Level 2: "...participants felt more confident to analyze a research paper in the field, to implement an AI algorithm in a clinical environment, and to incorporate the decisions given by an algorithm into their clinical decision making." • Level 2: "Most of the participants felt more competent at dealing with AI in medical imaging after the course."
Kang et al [48], 2017	Workshop	• Level 1: "Ninety percent of the residents... reported that the course was helpful or very helpful […]" • Level 1: "...94% of the participants...felt that the lectures were of high or very high quality." • Level 1: "Eighty-two percent...reported that they planned to pursue additional educational or research training in CER or big data analytics after the course [...]" • Level 1: "[...] 98% of the respondents felt that health services and big data research are important or very important for the future of radiology.'
Lindqwister et al [49], 2021	Course	• Level 1: "Exit surveys demonstrated a high degree of learner satisfaction, with an aggregate rating of 9.8/10." • Level 2: "There is a statistically significant difference between all pre- and postlecture question results (*P*<.04) by Wilcoxon Sign-rank test."
Tschirhart et al [55], 2022	Workshop	• Level 2: "...considerable improvement in the first independent dataset, with further improvement in subsequent datasets [...]"

Discussion

Principal Findings

The development and implementation of AI in medical education has greatly increased within the last decade, specifically since the COVID-19 pandemic where there was a global shift into the digital world accelerating the development of AI technology [59]. This can be seen as the majority (15/21, 77%) of included papers within this review were published since COVID-19 pandemic. Although there is a growing field within research and practice, AI medical education, specifically within curricula development, is still limited. We found that the current curriculum frameworks for AI medical education are limited, indicating a need for further research. We also found that the current state of AI educational programs lack the use of a theory, framework, or pedagogy. In addition, we uncovered alternative methods and different levels of in-depth curriculum planning for AI in medical education.

Current State of Curriculum Frameworks for AI Medical Education

This is the first review to identify curriculum frameworks for AI medical education, and our findings demonstrate that they are very limited. Although the literature is abundant in terms of recommendations and potential plans of actions for integrating AI education within medical education, there is an inadequate amount of formal curricula or frameworks [20,60,61]. For example, curricular recommendations lack specific learning outcomes and are not based on a particular education theory, as they usually focus solely on the content or competencies that should be taught [32,56]. Although understanding what concepts should be taught in AI is important, curriculum frameworks must be as comprehensive as possible.

From the identified frameworks, Masters [56] outlines a broad framework for any level of education, while Valikodath et al [57] outlines a complete framework for ophthalmology residency education. Their frameworks remain dissimilar in all aspects, except in how their course content was described. As seen with these 2 papers, the lack of curriculum frameworks in the literature is staggering. Further studies should focus on the development of these frameworks and start thinking on how to plan for the impending changes in medical education. As Valikodath et al [57] demonstrated their AI curriculum framework for ophthalmology, other specialties should follow suit, as AI affects each specialty differently [9]. Overall, the current state of curriculum framework in medical education

XSL•FO
RenderX

appears to be far from sufficient in the existing literature, and further research is needed.

Current State of AI Medical Educational Programs

Overview

In comparison to curriculum frameworks, educational programs in this field have been reviewed recently, specifically in the past 3 years [31-33]. However, research in AI medical education evolves quickly, and thus, a further identification of programs was carried out. We specifically looked at the framework, pedagogy, or learning theory described; the content and its audience; and if the program was evaluated for outcomes, which were used to assess its effectiveness, according to the model developed by Kirkpatrick et al [39].

The Lack of Learning Theories and Pedagogies

There were no papers that referenced a framework, pedagogy, or learning theory that guided the existence of the educational program. However, the use of frameworks, pedagogies, or learning theories is important for informing the development of valid, accurate, and competent educational programs [62-64]. By using frameworks, pedagogies, or learning theories, educators can choose the most effective instructional tactics, learning objectives, assessment, and evaluation approaches that can best help their students to learn [65]. A recent paper that fell outside the scope of our search date describes the use of constructivist theory and backward design learning principles that guided the development of their AI course [66]. Further papers should implement and report on a learning theory, framework, or pedagogy, as they have a role in medical education [65].

The Generalized AI Medical Content

The integration of AI concepts and topics within medical education remains generalized throughout the different levels of medical education, as seen with the educational programs described in our review. A total of 20 educational programs were described as focusing on general topics such as introductions to AI or information on AI and its application to medicine. The only postgraduate and continuing educational programs that had an AI-specific educational material were radiology, ophthalmology, and cardiology. This can be attributed to various reasons, including the constant evolution and novelty of AI technology, which may describe why generalized educational programs for AI appear across the medical educational continuum [67]. Radiology had the highest number of educational programs and was seen in all levels of medical education because AI in medicine was first applied in the field of radiology as it detected microcalcifications in a mammography during the year of 1992, or it could be due to the field being highly technological [68]. It is encouraging to see that specialties such as ophthalmology and cardiology have increased interest in AI education; other specialties and medical institutions should begin to follow suit. This is encouraging as it demonstrates that other specialties besides the highly technological field of radiology have been learning AI within medical education. This is especially important as more fields of medicine besides radiology are integrating AI within their practice, such as cardiology, pathology, and ophthalmology [3]. Furthermore, most of the educational programs were found in UME and within medical schools, which is ideal as it introduces a large audience of medical students to the concept of AI and its applications early in their careers.

The Success of Current AI Educational Programs

The included literature demonstrates that current efforts are being made to evaluate the outcomes of AI-related educational initiatives. According to the model developed by Kirkpatrick et al [39], an internationally recognized tool for evaluating and analyzing the results of educational, training, and learning programs, current AI programs have overall been positively received by medical learners. This was represented by the positive reactions, opinions, and attitudes toward AI after completing an educational program (level 1) as well as the acquisition of AI-related knowledge, skills, and confidence (level 2). These findings were also presented in a similar review in which the AI educational programs they identified also had positive outcomes, which were categorized as level 1 or level 2 [33]. However, further studies must assess educational programs for outcomes in relation to behavioral changes (level 3), specifically if there has been a transfer of AI-related knowledge, skills, and abilities into their daily work.

Further studies should also assess how the acquisition and application of these AI-related knowledge, skills, and abilities has affected the organization as a whole (eg, Has the increase in AI-educated physicians improved overall efficiency at the hospital?) or on patient outcomes (eg, Has there been an improvement in the patient's functional status or safety because of AI-educated physicians? [level 4]). By assessing for these additional outcomes, educators and medical organizations can understand how current AI educational programs have affected physician performance with AI technology. Increased research on the evaluations of educational programs can help further validate current educational tools and be used as inspiration for other institutions to create their own educational material. As seen in the review [33], there is a lack of consistency in the measures of these outcomes, as self-constructed and nonvalidated instruments were also used. Future studies should develop a validated tool to evaluate educational outcomes for a comprehensive synthesis.

Curriculum Planning and Framework Development of AI Medical Education

Curriculum planning of AI educational initiatives within medical education is insufficient. Although limited studies of curriculum frameworks were published, other forms of curriculum planning can be seen in the literature. Some medical institutions have conducted AI perception surveys [69,70], curriculum needs assessment surveys [71-73], and an interview [74] to understand what should be integrated into the AI medical curriculum. These studies are promising and contribute to the overall efforts to understanding how current educators, medical students, residents, and physicians consider AI within their educational system.

The absence of curriculum frameworks is staggering, especially given that AI competence is likely to become a required skill for medical graduates [75]. The development of AI curricula

XSL•FO
RenderX

and frameworks have already been gaining traction across other fields of education and levels. This can be seen as early as childhood education; for example, Su and Zhong [76] present their own curriculum framework, which outlines their concepts, teaching methods, teaching activities, projects, and assessment suggestions for AI education.

From a global perspective, the United Nations Educational, Scientific, and Cultural Organization, a specialized agency of the United Nations, released a document outlining the current practices of developing and implementing AI curricula in primary and secondary school education (K-12) [77]. From their report, several types of frameworks for AI literacy have been suggested, such as the AI Literacy Competency Framework, the AI4K12: 5 Big Ideas Framework, and the Machine Learning Education Framework. These recent reports and papers suggest increased efforts to integrate AI education before postsecondary school, which further stresses the importance of developing AI curricula and frameworks in medical education. Although there are current educational frameworks for AI education, each target audience must have their own specialized curricula to tailor the educational needs of the learners.

Medical educators can develop their curriculum through several different methodologies, such as the 10 key questions to be addressed while developing a curriculum [78] and the 6-step approach for curricular design [79]. However, curriculum frameworks allow a visual and detailed road map to implement a curriculum. Through this detailed format, educators are able to easily navigate the curriculum and its implementation, especially for new concepts in medicine, such as AI. To develop curriculum frameworks for AI in medicine, there must be an interdisciplinary team consisting of medical educators, AI experts and users, researchers, and curriculum designers due to the multiple fields incorporated.

The introduction of AI in medicine must be properly structured and organized within UME, PGME, and CME. Therefore, curriculum frameworks should be properly established through different levels of education and specialties. This has been emphasized by other reviews that call for integration of AI education in all levels and, thus, all specialties of medicine [17,33]. For example, a curriculum framework for UME will be different than a curriculum framework for PGME in dermatology. Curriculum frameworks can be adapted and they most likely will be, especially since AI education in medical education is still in its infancy. This is where leaders in UME, PGME, and CME organizations (eg, policy makers, medical educators, and researchers) must communicate effectively to eliminate any crossover education and repeated information. New technology and innovations in relation to AI and medicine will inevitably occur; however, it is important to be cognizant of the fundamentals of AI and how it will affect a physician's practice at the time. Sufficient planning of an AI curricula will deliver effective education for physicians who will increasingly be using AI technology in the near future; therefore, medical educators and institutions must begin to consider curriculum planning.

Incorporating and Advocating for AI Into the Medical Curriculum

The literature emphasizes the need to introduce AI in the medical education curriculum [12,13,15-20]; however, there are several challenges that have been discussed in terms of implementing this type of education. This includes insufficient time, insufficient resources (eg, lack of teaching staff or knowledge), and variable aptitude and interest in AI [80-82]. However, this review details several approaches to implementation as well as 6 studies that have evaluated their educational program. These successful educational programs can provide medical schools and national and international medical organizations with examples of current AI content topics and implementation methods that have worked for others. These medical education institutions can view how AI-based medical education is currently being offered around the world and understand any challenges, opportunities, and strengths about these programs. Although the content and provision of AI education is heterogenous, this heterogeneity can allow educators and students to view the many types of programs that were offered. As AI education for medicine is still in its infancy, educators should explore these programs where they can then potentially modify an educational program that best suits their needs. As seen in this review, there are several ways to incorporate AI material into the current curriculum seamlessly, such as an AI fundamentals lecture or module, an AI elective, or a research project.

Medical students, residents, and practicing physicians also have the opportunity to advocate for the inclusion of AI education at their respective institutions [46]. For example, there are several North American university chapters of the Artificial Intelligence in Medicine Student Society, such as the University of Toronto and University of Alberta, which organizes workshops, conferences, and multiple speaker sessions throughout the year [46]. These student interest groups demonstrate the increased interest for AI and can potentially build momentum and advocate for AI education at their respective institutions. As some of the offerings at these student interest groups include brief educational material for AI, medical institutions can work with these students as a starting point.

Strengths and Limitations

The strengths of this review include the comprehensive search strategies, the inclusion of a variety of information sources, and rigorous methodological approaches that are replicable. For example, study selection was completed by 2 reviewers, and disagreements were resolved by discussion or consensus involving a third investigator. Furthermore, a scoping review protocol was registered and published to improve transparency of the methodological process.

Although this study was conducted in a structured and systematic manner, there are some limitations that are important to consider. A limited number of papers were retrieved during the search and selection process. Only 2 papers reported having a curriculum framework, with 1 reporting a full curricula plan related to AI in medicine. This can be because AI technology is emerging and continuing to change within medicine and it has been limiting in terms of educational advances. Because of

XSL•FO
RenderX

the nature of the scoping review, the quality of each identified study was not assessed.

Conclusions

Medicine is rapidly evolving from the information age to the age of AI, where machines will become an integral part of medical practice. Thus, medical education needs to keep pace with changes in medical practice. This review synthesized knowledge from the literature on curriculum frameworks and current educational programs that focus on the teaching and learning of AI for medical students, residents, and practicing physicians. To better integrate AI curricula into the continuum of medical education, discussions surrounding curriculum planning of AI should begin where institutions are recommended to work collaboratively with teams of curriculum designers, data scientists, and medical educators to develop AI curricula and educational programs. There is a need to (1) develop a general AI education curriculum framework for UME; (2) develop a specific AI education curriculum framework for each specialty within PGME and CME; and (3) design, implement, and evaluate current educational programs. Overall, institutions must begin equipping current and future physicians with the knowledge, skills, and confidence to effectively use AI applications as it will continue to grow within the field of health care.

Acknowledgments

SAR is Canada Research Chair (Tier II) in Advanced Digital Primary Health Care, received salary support from a Research Scholar Junior 1 Career Development Award from the Fonds de Recherche du Québec-Santé (FRQS) during a portion of this study, and her research program is supported by the Natural Sciences Research Council (NSERC) Discovery (grant 2020-05246). The study was also supported by the *Fonds de recherche du Québec–Société et Culture* team grant to the McGill Family Medicine Education Research Group.

Authors' Contributions

RT, SAR, and PP conceived the idea, developed the research protocol and methods, and drafted and edited the final manuscript. GG helped develop and run the search strategy. AB, PP, and SAR helped to refine and develop the research question and study methods and helped with drafting and editing of the manuscript. All authors except PP approved the final manuscript submitted; however, the authors would like to acknowledge that the late PP provided many meaningful contributions to this work before his passing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 107 KB - mededu_v10i1e54793_app1.docx]

Multimedia Appendix 2
Study characteristics (N=21).
[DOCX File , 41 KB - mededu_v10i1e54793_app2.docx]

Multimedia Appendix 3
Countries and years of publications included in the review.
[PDF File (Adobe PDF File), 824 KB - mededu_v10i1e54793_app3.pdf]

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun 13;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
2. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
3. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
4. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
5. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Path Lab Med 2019 Jul;143(7):859-868. [doi: 10.5858/arpa.2018-0147-oa]

XSL•FO
RenderX

6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
7. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HT, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res 2021 Sep 03;23(9):e29839 [FREE Full text] [doi: 10.2196/29839] [Medline: 34477556]
8. Birkhoff DC, van Dalen AS, Schijven MP. A review on the current applications of artificial intelligence in the operating room. Surg Innov 2021 Oct 24;28(5):611-619 [FREE Full text] [doi: 10.1177/1553350621996961] [Medline: 33625307]
9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
10. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today 2021 Jan;26(1):80-93 [FREE Full text] [doi: 10.1016/j.drudis.2020.10.010] [Medline: 33099022]
11. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec 11;19(1):460 [FREE Full text] [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
12. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
13. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
14. Minor LB. Stanford medicine 2020 health trends report: the rise of the data-driven physician. Stanford Medicine. 2020. URL: https://med.stanford.edu/dean/healthtrends.html [accessed 2022-07-10]
15. Pucchio A, Papa JD, de Moraes FY. Artificial intelligence in the medical profession: ready or not, here AI comes. Clinics (Sao Paulo) 2022;77:100010 [FREE Full text] [doi: 10.1016/j.clinsp.2022.100010] [Medline: 35176642]
16. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018 Sep 27;1(1):54 [FREE Full text] [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
17. Mehta S, Vieira D, Quintero S, Bou Daher D, Duka F, Franca H, et al. Redefining medical education by boosting curriculum with artificial intelligence knowledge. J Cardiol Curr Res 2020 Oct 13;13(5):124-129. [doi: 10.15406/jccr.2020.13.00490]
18. Abdulhussein H, Turnbull R, Dodkin L, Mitchell P. Towards a national capability framework for artificial intelligence and digital medicine tools – a learning needs approach. Intell Based Med 2021;5:100047. [doi: 10.1016/j.ibmed.2021.100047]
19. James CA, Wheelock KM, Woolliscroft JO. Machine learning: the next paradigm shift in medical education. Acad Med 2021 Jul 01;96(7):954-957. [doi: 10.1097/ACM.0000000000003943] [Medline: 33496428]
20. Lomis K, Jeffries P, Palatta A, Sage M, Sheikh J, Sheperis C, et al. Artificial intelligence for health professions educators. NAM Perspect 2021 Sep 8;2021:202109a [FREE Full text] [doi: 10.31478/202109a] [Medline: 34901780]
21. Topol E. The Topol review: preparing the health care work- force to deliver the digital future. National Health Service, UK. 2019. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2023-04-25]
22. AMA passes first policy recommendations on augmented intelligence internet. American Medical Association. 2018. URL: https://www.ama-assn.org/press-center/press-releases/ama-passes-first-policy-recommendations-augmented-intelligence [accessed 2023-04-25]
23. Reznick RK, Harris K, Horsley T, Hassani MS. Artificial intelligence (AI) and emerging digital technologies. The Royal College of Physicians and Surgeons of Canada. URL: https://www.royalcollege.ca/en/health-policy/initiatives-driven-by-research/ai-task-force.html [accessed 2022-06-18]
24. Pinto Dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr 6;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
25. Hedderich DM, Keicher M, Wiestler B, Gruber MJ, Burwinkel H, Hinterwimmer F, et al. AI for doctors-a course to educate medical professionals in artificial intelligence for medical imaging. Healthcare (Basel) 2021 Sep 28;9(10):1278 [FREE Full text] [doi: 10.3390/healthcare9101278] [Medline: 34682958]
26. Stabback P. Guidelines for constructing a curriculum framework for basic education. International Bureau of Education, UNESCO. 2007. URL: http://www.ibe.unesco.org/fileadmin/user_upload/COPs/News_documents/2007/0709Kigal [accessed 2022-07-10]
27. Redwood-Campbell L, Pakes B, Rouleau K, MacDonald CJ, Arya N, Purkey E, et al. Developing a curriculum framework for global health in family medicine: emerging principles, competencies, and educational approaches. BMC Med Educ 2011 Jul 22;11(1):46 [FREE Full text] [doi: 10.1186/1472-6920-11-46] [Medline: 21781319]
28. Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. Lancet Digit Health 2020 Mar;2(3):e111-e112. [doi: 10.1016/s2589-7500(20)30023-6]
29. Obadeji A. Health professions education in the 21st century: a contextual curriculum framework for analysis and development. J Contemp Med Edu 2019;9(1):34. [doi: 10.5455/jcme.20181212085450]
30. Iqbal S, Ahmad S, Akkour K, Wafa AN, AlMutairi HM, Aldhufairi AM. Review article: impact of artificial intelligence in medical education. MedEdPublish 2021;10(1):41. [doi: 10.15694/mep.2021.000041.1]

XSL•FO
RenderX

31. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. J Med Educ Curric Dev 2021 Sep 06;8:23821205211036836 [FREE Full text] [doi: 10.1177/23821205211036836] [Medline: 34778562]
32. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 01;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
33. Charow R, Jeyakumar T, Younus S, Dolatabadi E, Salhia M, Al-Mouaswas D, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021 Dec 13;7(4):e31043 [FREE Full text] [doi: 10.2196/31043] [Medline: 34898458]
34. Peters MD, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2010.
35. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
36. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
37. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
38. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in artificial intelligence for medical students, residents, and practicing physicians: a scoping review protocol. JBI Evid Synth 2023;21(7):1477-1484. [doi: 10.11124/jbies-22-00374]
39. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels. Oakland, CA: Berrett-Koehler Publishers; 2006.
40. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc 2020 Oct;92(4):807-812. [doi: 10.1016/j.gie.2020.06.040] [Medline: 32565184]
41. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme version. Lancaster University. 2006. URL: https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf [accessed 2024-04-29]
42. Alderson PO, Donlin MJ, Morrison LA. A model to introduce medical students to the use of artificial intelligence and genomics for precision medicine. medRxiv Preprint posted online May 17, 2021 [FREE Full text] [doi: 10.1101/2021.05.13.21255493]
43. Balthazar P, Tajmir SH, Ortiz DA, Herse CC, Shea LA, Seals KF, et al. The artificial intelligence journal club (#RADAIJC): a multi-institutional resident-driven web-based educational initiative. Acad Radiol 2020 Jan;27(1):136-139. [doi: 10.1016/j.acra.2019.10.005] [Medline: 31685386]
44. Barbour AB, Frush JM, Gatta LA, McManigle WC, Keah NM, Bejarano-Pineda L, et al. Artificial intelligence in health care: insights from an educational forum. J Med Educ Curric Dev 2019 Jan 28;6:2382120519889348 [FREE Full text] [doi: 10.1177/2382120519889348] [Medline: 32064356]
45. Forney MC, McBride AF. Artificial intelligence in radiology residency training. Semin Musculoskelet Radiol 2020 Feb;24(1):74-80. [doi: 10.1055/s-0039-3400270] [Medline: 31991454]
46. Harish V, Bilimoria K, Mehta N, Morgado F, Aissiou A, Eaton S, et al. Preparing medical students for the impact of artificial intelligence on healthcare. Canadian Federation of Medical Students. 2019. URL: https://www.cfms.org/files/position-papers/AGM_2020_CFMS_AI.pdf [accessed 2022-09-10]
47. Hu R, Fan KY, Pandey P, Hu Z, Yau O, Teng M, et al. Insights from teaching artificial intelligence to medical students in Canada. Commun Med (Lond) 2022 Jun 03;2(1):63 [FREE Full text] [doi: 10.1038/s43856-022-00125-4] [Medline: 35668847]
48. Kang SK, Lee CI, Pandharipande PV, Sanelli PC, Recht MP. Residents' introduction to comparative effectiveness research and big data analytics. J Am Coll Radiol 2017 Apr;14(4):534-536 [FREE Full text] [doi: 10.1016/j.jacr.2016.10.032] [Medline: 28139415]
49. Lindqwister AL, Hassanpour S, Lewis PJ, Sin JM. AI-RADS: an artificial intelligence curriculum for residents. Acad Radiol 2021 Dec;28(12):1810-1816 [FREE Full text] [doi: 10.1016/j.acra.2020.09.017] [Medline: 33071185]
50. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020 Jun 19;3(1):86 [FREE Full text] [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
51. Nagy M, Radakovich N, Nazha A. Why machine learning should be taught in medical schools. Med Sci Educ 2022 Apr 24;32(2):529-532 [FREE Full text] [doi: 10.1007/s40670-022-01502-3] [Medline: 35528308]
52. Nguyen GK, Shetty AS. Artificial intelligence and machine learning: opportunities for radiologists in training. J Am Coll Radiol 2018 Sep;15(9):1320-1321. [doi: 10.1016/j.jacr.2018.05.024] [Medline: 29941242]
53. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? J Educ Eval Health Prof 2019 Jul 03;16:18 [FREE Full text] [doi: 10.3352/jeehp.2019.16.18] [Medline: 31319450]

XSL•FO
RenderX

54. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
55. Tschirhart J, Woolsey A, Skinner J, Ahmed K, Fleming C, Kim J, et al. Introducing medical students to deep learning through image labelling: a new approach to meet calls for greater artificial intelligence fluency among medical trainees. Can Med Educ J 2023 Jun 21;14(3):113-115 [FREE Full text] [doi: 10.36834/cmej.75074] [Medline: 37465748]
56. Masters K. Artificial intelligence developments in medical education: a conceptual and practical framework. MedEdPublish (2016) 2020;9(1):239 [FREE Full text] [doi: 10.15694/mep.2020.000239.1] [Medline: 38058891]
57. Valikodath NG, Cole E, Ting DS, Campbell JP, Pasquale LR, Chiang MF, American Academy of Ophthalmology Task Force on Artificial Intelligence. Impact of artificial intelligence on medical education in ophthalmology. Transl Vis Sci Technol 2021 Jun 01;10(7):14 [FREE Full text] [doi: 10.1167/tvst.10.7.14] [Medline: 34125146]
58. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
59. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res 2023;15(7):4820-4828 [FREE Full text] [Medline: 37560249]
60. Nagy M, Radakovich N, Nazha A. Machine learning in oncology: what should clinicians know? JCO Clin Cancer Inform 2020 Nov(4):799-810. [doi: 10.1200/cci.20.00049]
61. Ngo B, Nguyen D, van Sonnenberg E. Artificial intelligence: has its time come for inclusion in medical school education? Maybe…maybe not. MedEdPublish 2021;10(1):131. [doi: 10.15694/mep.2021.000131.1]
62. Tredinnick-Rowe J. The role of pedagogy in clinical education. In: Cavero OB, Llevot-Calvet N, editors. New Pedagogical Challenges in the 21st Century - Contributions of Research in Education. Rijeka, Croatia: InTech; 2018:6-85.
63. Khalil MK, Elkhider IA. Applying learning theories and instructional design models for effective instruction. Adv Physiol Educ 2016 Jun;40(2):147-156 [FREE Full text] [doi: 10.1152/advan.00138.2015] [Medline: 27068989]
64. Fuller JC, Woods ME. The science of learning: why learning theories matter in graduate medical education. HCA Healthc J Med 2021 Aug 31;2(4):247-250 [FREE Full text] [doi: 10.36518/2689-0216.1203] [Medline: 37424848]
65. Mukhalalati BA, Taylor A. Adult learning theories in context: a quick guide for healthcare professional educators. J Med Educ Curric Dev 2019 Apr 10;6:2382120519840332 [FREE Full text] [doi: 10.1177/2382120519840332] [Medline: 31008257]
66. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R. Grounded in reality: artificial intelligence in medical education. JAMIA Open 2023 Jul;6(2):ooad037 [FREE Full text] [doi: 10.1093/jamiaopen/ooad037] [Medline: 37273962]
67. Grassini S. Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. Educ Sci 2023 Jul 07;13(7):692. [doi: 10.3390/educsci13070692]
68. Driver CN, Bowles BS, Bartholmai BJ, Greenberg-Worisek AJ. Artificial intelligence in radiology: a call for thoughtful application. Clin Transl Sci 2020 Mar 30;13(2):216-218 [FREE Full text] [doi: 10.1111/cts.12704] [Medline: 31664767]
69. Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. MedEdPublish 2021;10(1):75. [doi: 10.15694/mep.2021.000075.1]
70. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. J Med Educ Curric Dev 2021 Jun 23;8:23821205211024078 [FREE Full text] [doi: 10.1177/23821205211024078] [Medline: 34250242]
71. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 09;22(1):772 [FREE Full text] [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
72. Gray K, Slavotinek J, Dimaguila GL, Choo D. Artificial intelligence education for the health workforce: expert survey of approaches and needs. JMIR Med Educ 2022 Apr 04;8(2):e35223 [FREE Full text] [doi: 10.2196/35223] [Medline: 35249885]
73. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: a global mixed-methods study of medical students' perspectives. Digit Health 2022 May 02;8:20552076221089099 [FREE Full text] [doi: 10.1177/20552076221089099] [Medline: 35521511]
74. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. JMIR Med Educ 2023 Apr 24;9:e46428 [FREE Full text] [doi: 10.2196/46428] [Medline: 36946094]
75. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS One 2022 Jul 21;17(7):e0271872 [FREE Full text] [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
76. Su J, Zhong Y. Artificial Intelligence (AI) in early childhood education: curriculum design and future directions. Comput Educ Artif Intell 2022;3:100072. [doi: 10.1016/j.caeai.2022.100072]
77. Miao F, Shiohira K. K-12 AI curricula: a mapping of government-endorsed AI curricula. United Nations Educational, Scientific and Cultural Organization. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380602.2022;3:1144399 [accessed 2023-01-02]

XSL•FO
RenderX

78. Harden RM. Ten questions to ask when planning a course or curriculum. Med Educ 1986 Jul;20(4):356-365. [doi: 10.1111/j.1365-2923.1986.tb01379.x] [Medline: 3747885]
79. Thomas PA, Kern DE, Hughes MT, Tackett SA, Chen BY. Curriculum Development for Medical Education – A Six–Step Approach. Baltimore, MD: Johns Hopkins University Press; 2022.
80. Azer SA, Guerrero AP. The challenges imposed by artificial intelligence: are we ready in medical education? BMC Med Educ 2023 Sep 19;23(1):680 [FREE Full text] [doi: 10.1186/s12909-023-04660-z] [Medline: 37726741]
81. Grunhut J, Marques O, Wyatt AT. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 07;8(2):e35587 [FREE Full text] [doi: 10.2196/35587] [Medline: 35671077]
82. Ng FY, Thirunavukarasu AJ, Cheng H, Tan TF, Gutierrez L, Lan Y, et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med 2023 Oct 17;4(10):101230 [FREE Full text] [doi: 10.1016/j.xcrm.2023.101230] [Medline: 37852174]

Abbreviations

AI: artificial intelligence
CME: continuing medical education
PGME: postgraduate medical education
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
UME: undergraduate medical education

Edited by T de Azevedo Cardoso; submitted 22.11.23; peer-reviewed by SA Steindal, K Latulippe; comments to author 30.12.23; revised version received 26.03.24; accepted 29.04.24; published 18.07.24.

Please cite as:
Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S
Curriculum Frameworks and Educational Programs in AI for Medical Students, Residents, and Practicing Physicians: Scoping Review
JMIR Med Educ 2024;10:e54793
URL: https://mededu.jmir.org/2024/1/e54793
doi:10.2196/54793
PMID:39023999

©Raymond Tolentino, Ashkan Baradaran, Genevieve Gore, Pierre Pluye, Samira Abbasgholizadeh-Rahimi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.07.2024. 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.

XSL•FO
RenderX

Review

Identifying Learning Preferences and Strategies in Health Data Science Courses: Systematic Review

Narjes Rohani[1], MSc; Stephen Sowa[2], PhD; Areti Manataki[3], PhD

[1]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

[2]Moray House School of Education and Sport, University of Edinburgh, Edinburgh, United Kingdom

[3]School of Computer Science, University of St Andrews, St Andrews, United Kingdom

Corresponding Author:
Narjes Rohani, MSc
Usher Institute
University of Edinburgh
Old Medical School
Teviot Place
Edinburgh, EH8 9AG
United Kingdom
Phone: 44 131 650 3
Email: Narjes.rohani@ed.ac.uk

Abstract

Background: Learning and teaching interdisciplinary health data science (HDS) is highly challenging, and despite the growing interest in HDS education, little is known about the learning experiences and preferences of HDS students.

Objective: We conducted a systematic review to identify learning preferences and strategies in the HDS discipline.

Methods: We searched 10 bibliographic databases (PubMed, ACM Digital Library, Web of Science, Cochrane Library, Wiley Online Library, ScienceDirect, SpringerLink, EBSCOhost, ERIC, and IEEE Xplore) from the date of inception until June 2023. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and included primary studies written in English that investigated the learning preferences or strategies of students in HDS-related disciplines, such as bioinformatics, at any academic level. Risk of bias was independently assessed by 2 screeners using the Mixed Methods Appraisal Tool, and we used narrative data synthesis to present the study results.

Results: After abstract screening and full-text reviewing of the 849 papers retrieved from the databases, 8 (0.9%) studies, published between 2009 and 2021, were selected for narrative synthesis. The majority of these papers (7/8, 88%) investigated learning preferences, while only 1 (12%) paper studied learning strategies in HDS courses. The systematic review revealed that most HDS learners prefer visual presentations as their primary learning input. In terms of learning process and organization, they mostly tend to follow logical, linear, and sequential steps. Moreover, they focus more on abstract information, rather than detailed and concrete information. Regarding collaboration, HDS students sometimes prefer teamwork, and sometimes they prefer to work alone.

Conclusions: The studies' quality, assessed using the Mixed Methods Appraisal Tool, ranged between 73% and 100%, indicating excellent quality overall. However, the number of studies in this area is small, and the results of all studies are based on self-reported data. Therefore, more research needs to be conducted to provide insight into HDS education. We provide some suggestions, such as using learning analytics and educational data mining methods, for conducting future research to address gaps in the literature. We also discuss implications for HDS educators, and we make recommendations for HDS course design; for example, we recommend including visual materials, such as diagrams and videos, and offering step-by-step instructions for students.

(JMIR Med Educ 2024;10:e50667) doi:10.2196/50667

KEYWORDS

health data science; bioinformatics; learning approach; learning preference; learning tactic; learning strategy; interdisciplinary; systematic review; medical education

XSL•FO
RenderX

Introduction

Background

In the era of artificial intelligence, big data, and digitalization of health care, there is a growing demand for educating specialists in analyzing health data [1-3]. The integration of IT into health care has undergone significant evolution in recent decades that has led to a change in the definition of health informatics. The current definition of health informatics encompasses the interdisciplinary study of designing, developing, adopting, and applying IT-based innovations in health care service delivery, management, and planning. By contrast, health care data analytics, a nascent subfield within health informatics, specifically addresses methods and techniques for analyzing, integrating, and interpreting health care data. Health data analytics, or health data science (HDS), as it can also be understood, involves data manipulation, mining, and statistical analysis to gain valuable insights from health, medical, or biological data. In other words, while health informatics encompasses noncomputational aspects, such as system development and maintenance, health data analytics or HDS concentrates only on using computational tools and methods for analyzing data [4].

However, given the novel and interdisciplinary nature of HDS, learning and teaching HDS is highly challenging [1,5,6]. Students and teachers are often faced with a lack of common language and prior knowledge in health or computational sciences, thus making it hard to learn and teach HDS concepts effectively [1,7]. In postgraduate study, in particular, students who enroll in HDS courses have diverse academic backgrounds, including computational and medical backgrounds (but rarely a combination of the two); therefore, traditional learning and teaching approaches in biology, medicine, or computer sciences may not be effective for HDS training [1,8].

Shedding light on HDS students' learning preferences and strategies is particularly important in this context and can help address some of these challenges [7,9-12]. There is heterogeneous literature around the definitions of *learning strategy*, *learning tactic*, *learning approach*, *learning style*, and other related terms [10,13,14,15]. In this paper, we view learning strategy as the approach that students use to manage their learning processes.

Similar to recent studies [16-19], we also understand learning preference as the perceived tendency of learners regarding the presentation of learning materials, types of learning activities, and the organization of their learning process, while learning strategy or learning approach is the actual way in which students manage their learning process [19].

We also recognize that the learning preferences that students exhibit within the HDS field inform the strategies they use to support their learning [20,21]. We decided to focus on learning preferences and strategies from the aforementioned perspectives because these field-specific preferences and strategies can offer insights into HDS education, which are useful for personalized learning [17,22-24].

Given the aforementioned definition of learning preference, research studies about learning styles in HDS-related fields touch upon HDS-specific learning preferences and can thus be used to identify students' tendencies in the field regarding information presentation, learning activities, and learning organization. However, it should be mentioned that the term *learning style* has been consistently misinterpreted [18] and defined variably across numerous studies in the literature [18,25]. In recent years, several research studies [26,27] have criticized the claim that each individual student has a dominant learning style, which is a stable neurological, psychological, or innate learning preference. Nonetheless, these and other studies [10,12,16,18,26,27] have also acknowledged that students in each field of study, specific to the nature of the discipline, might exhibit some preferences regarding course materials and activities and the way in which they approach these materials and activities [10,12,16,18,26]. As mentioned in a previous study [26], while the concept of stable learning styles for students is considered a myth, there are preferences that students exhibit within each field that informs the strategies they use to support their learning, which can in turn support personalized learning [10,11,16,18,20,26,28-30].

Given this context, gaining knowledge about learners' preferences and strategies in HDS courses can help course designers create optimized courses or redesign existing courses [10,31,32], creating a positive impact on student interest, engagement, and performance [16,32]. In addition, informing teachers about students' learning preferences and strategies in HDS courses can assist them not only in selecting appropriate teaching methods but also in providing personalized feedback to students [10,30,33,34].

Although several systematic reviews have been conducted to investigate the learning preferences of nurses [35,36] and physiotherapists [37], none of them are related to interdisciplinary programs in the realm of HDS. To fill this gap, and following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [38], we conducted a systematic review to present the current state of knowledge on learning strategies and preferences in HDS.

There are important aspects of learning strategies and preferences that are of interest in this systematic review because they are useful for implementing personalized learning in the HDS field [11,21]. The types of multimedia resources in a course are important because they significantly influence engagement, understanding, and the overall learning experience of students [39,40]. Each discipline has its unique nature [10,26], and presenting concepts in an effective way that is aligned with students' preferences in the discipline can improve students' satisfaction [41]. Therefore, insight into preference regarding the types of multimedia resources used for information delivery can enhance course design and student satisfaction.

Collaborative learning is one of the popular strategies in education, but it is not always easy to implement it successfully because engaging all students in teamwork is challenging [42-44]. Therefore, understanding students' collaboration preferences in HDS can facilitate the integration of both peer learning and independent study within a course to improve

XSL•FO
RenderX

collaborative skills, support diverse perspectives, and help students to develop self-directed learning skills [42-44].

In addition, understanding whether HDS students prefer a global or sequential approach when studying topics can inform both teachers and students about effective learning strategies to enhance the student educational journey; for example, course designers can arrange topics in more effective sequences that align better with students' preferences, thereby improving the overall learning experience [45].

Moreover, understanding the preferred focus granularity of students, such as their inclination toward details or abstract concepts, assists in prioritizing topics for teaching and determining effective teaching strategies [46,47]; for example, identifying whether HDS students prefer applied topics or theoretical aspects helps educators decide the level of details to include in the course materials [47]. These are all important topics related to learning strategies and preferences, which are worth shedding light on in the context of HDS education.

Research Questions

Therefore, this systematic review focuses on the following research questions (RQs), which were selected based on available literature and their potential benefits for personalized learning [20,21]:

- RQ1: What types of information presentation do students prefer in HDS?
- RQ2: Do students prefer team-based learning over independent learning in HDS?
- RQ3: How do students organize their learning process (global vs sequential) in HDS?
- RQ4: Do students in HDS prefer abstract concepts over factual concepts?

Our goal with this systematic review is not only to present and analyze research findings on learning strategies and preferences in HDS but also to discuss their implications for future course design in HDS. This way, we can help HDS educators make informed decisions about teaching methods and assist them with developing effective courses. To the best of our knowledge, this is the first systematic review that discusses learning strategies and preferences in HDS-related disciplines. The contributions of this study are as follows:

- It consolidates the heterogeneous knowledge available in the literature and presents it in 4 categories, that is, information presentation (RQ1), collaboration preference (RQ2), organization strategy (RQ3), and focus granularity (RQ4).
- It provides suggestions to assist course designers and teachers in delivering more effective HDS-related courses.
- It provides suggestions for future research in HDS education, which can help researchers conduct better informed investigations in this area.

Methods

Overview

This systematic review was conducted to understand what learning strategies and preferences are used by students in HDS-related fields. To this end, we followed all steps outlined in the PRISMA guidelines [38] except for the meta-analysis step because, given the diversity of the included papers, the narrative data synthesis approach [48] was deemed more appropriate for combining the findings from the different studies. Therefore, we used narrative data synthesis to report our findings. The PRISMA checklists for abstracts and articles are available in Tables S1 and S2 in Multimedia Appendix 1, respectively. We also used the Mixed Methods Appraisal Tool (MMAT) [49] to assess the quality of the included articles. The MMAT allows the assessment of the quality of studies with different methodological designs, such as quantitative, qualitative, and mixed designs. The protocol used in this study is available in Multimedia Appendix 2 [38,48-51], and the PICO (Population, Intervention, Comparison, and Outcomes) components of the review question are presented in Table S1 in Multimedia Appendix 2.

Types of Studies and Participants

In this systematic review, we considered various types of primary studies, including both quantitative and qualitative journal or conference papers, all of which focused on exploring learning preferences or strategies in HDS-related courses. We did not apply any restrictions regarding participants' academic degrees; therefore, all high school, undergraduate, and postgraduate students as well as nontraditional learners (eg, health care professionals) were included.

Study Eligibility

This systematic review focuses on courses and programs falling within the scope of HDS (using data analytics methods to analyze biological, medical, and health data) [4,7]. Studies focusing on non–data analytics aspects of health informatics were not considered in this systematic review.

The inclusion criteria are presented in Textbox 1 (for more study eligibility details, refer to Table S2 in Multimedia Appendix 2).

XSL•FO
RenderX

Textbox 1. Criteria used to select the studies.

Inclusion criteria

- Language of publication: English
- Year of publication: no restriction applied regarding the year of publication
- Participants: students in fields highly relevant to health data science (HDS; using computational methods for medical, biological, or health data analysis), such as bioinformatics, biostatistics, computational biology, neuroinformatics, biomedical science, precision medicine, HDS, and HDS courses
- Participants' academic level: high school, undergraduate, and postgraduate students in any relevant course; nontraditional learners, such as health care professionals, also included
- Type of publication: conference and journal papers; primary research articles
- Subject: papers discussing learning preferences, strategies, tactics practices, or styles of the aforementioned learners
- Analysis type: both quantitative and qualitative methods included

Study Identification

The literature search was carried out on June 15, 2023. The PubMed, ACM Digital Library, Web of Science, Cochrane Library, Wiley Online Library, ScienceDirect, SpringerLink, EBSCOhost, ERIC, and IEEE Xplore databases were searched independently. We supplemented the literature search by using Google Scholar manually to find potentially missed articles. Given the interdisciplinary nature of HDS, these databases were selected to cover literature across computer science, education, and medicine. We used a combination of terms to identify papers about students' learning preferences and strategies in a variety of courses and programs related to HDS. The keywords used for searching the literature are presented in Table S3 in Multimedia Appendix 2, and the queries used for each database are presented in Table S4 in Multimedia Appendix 2.

Study Selection

The title, abstract, and full-text screening were carried out independently by 2 reviewers: NR, who has an academic background in HDS; and SS, who has a background in education. They screened the titles and abstracts of all extracted articles, followed by a full-text review of eligible studies (Cohen κ agreement index=0.95). In cases of disagreement, a third screener, AM, was involved to resolve conflicts. The screening questions are presented in Table S5 in Multimedia Appendix 2.

Data Extraction

Both NR and SS used a standardized Microsoft Word form for extracting and documenting data (for details, refer to Multimedia Appendix 2). The data they extracted included the following categories: publication characteristics (this included details such as the publication title, journal or conference, authors, and publication year); methodological features (the reviewers recorded various methodological aspects, such as the participants' field and course name, the number of participants, the method of analysis used, the type of input data used, the students' degree level, the study subject, and any learning inventory used); and learning preference or learning strategy (information regarding reported learning preferences or strategies was collected, along with the corresponding percentage of students exhibiting each learning preference or strategy).

After the initial extraction, both reviewers cross-checked the extracted data to ensure accuracy. In addition, both reviewers assessed the quality of the included articles independently by using the MMAT [49]. Finally, any discrepancies or inconsistencies were independently resolved by the third reviewer, AM.

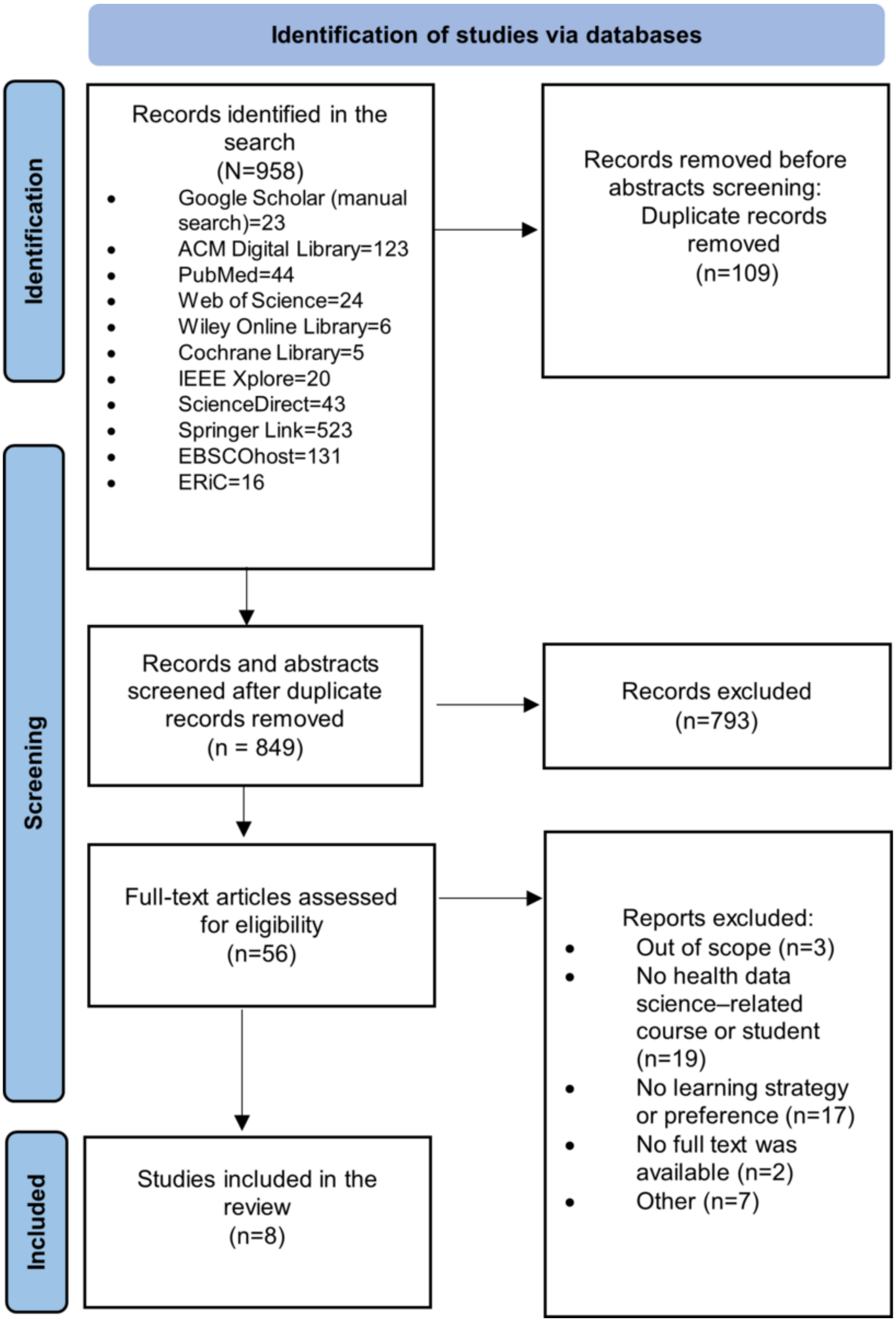
Results

Search Results

The literature search resulted in 958 articles, which were reduced to 849 (88.6%) studies after removing 109 (11.4%) duplicates (for details, refer to Figure 1). Of these 849 articles, after full-text review, 8 (0.9%) studies that were published between 2005 and 2021 were included in the synthesis. The reasons for excluding papers during full-text screening are presented in Table S6 in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection process. Full texts could not be found for 2 (4%) of the 56 papers considered for full-text review after abstract screening.



Characteristics of the Included Studies

As shown in Table 1, most of the articles (7/8, 88%) were published between 2017 and 2021. Of the 8 studies, 2 (25%) were conducted in the United States; 2 (25%) in Malaysia (2/8, 25%); and 1 (12%) each in Denmark, India, Sweden, and Israel.

Of the 8 studies, 3 (38%) [45,46,52] focused on undergraduate students, and 3 (38%) focused on postgraduate students [53-55], while high school learners were investigated by 1 (12%) study [56], and health care professionals were the focus of 1 (12%) study [41].

XSL•FO
RenderX

Table 1. Summary of the included studies.

Study, year; country	Sample size, n	Participants' field	Course	Course delivery type	Participants' academic level	Study subject	Learning inventory	Results
Holtzclaw et al [45], 2017; United States	28	Genetics	Bioinformatics	Face-to-face with web-based materials	Undergraduate student	Learning style	FSILS[a]	Procession: active=54%[b], reflective=46%; input: visual=82%, verbal=18%; perception: sensing=67%, intuitive=33%; understanding: sequential=79%, global=21%
Micheel et al [41], 2017; United States	751	Oncology	Precision medicine	Web based	HCP[c]	Learning style	Custom survey with 1 question	Multimodal (80%): watching, listening, and reading=39%, watching and reading=19%, listening and reading=12%, watching and listening=10%; unimodal (20%): reading=15%, watching=3%, listening=2%
Nielsen and Kreiner [53], 2017; Denmark	57	Public health	Advanced statistics	Face-to-face	Postgraduate student	Learning style	D-SA-LSI[d] and qualitative analysis	Function: executive=5.42[e] (strong), legislative=4.59 (strong), judicial=4.41 (medium); form: democratic=4.62 (strong), anarchic=4.34 (medium), monarchic=3.68 (medium), hierarchic=4.12 (medium), oligarchic=2.65 (weak); learning: conservative=4.54 (strong), progressive=4.83 (strong); level: global=3.97 (medium), local=3.59 (medium); scope: external=5.43 (strong), internal=3.53 (medium)
Diwakar et al [54], 2018; India	84	Biotechnology, microbiology, and bioinformatics	Bioinformatics and biotechnology	Web based	Postgraduate student	Learning style	Kolb learning style inventory	Assimilators=60%, divergers=20%, convergers=16%, accommodators=4%
Sani Ibrahim [46], 2020; Malaysia and Nigeria	2 data sets were used: procession data set=95, perception data set=2168	Bioinformatics	Genomics technology	Web based	Undergraduate student	Learning style	FSILS and data mining	Procession: active=70%, reflective=30%; perception: intuitive=94%, sensing=6%
Abrahamsson and Dávila Lopez [55], 2021; Sweden	65	Bioinformatics	Bioinformatics	Web based and face-to-face	Postgraduate student	Learning style	Custom survey and qualitative analysis	Lecture format: real-time Zoom sessions=64%, offline as a video=27%, offline as reading=9%; synchronize work preference: alone=50%, alone and then in group=12%, same group=19%, different group=19%
Gelbart et al [56], 2009; Israel	4	Biology	Bioinformatics	Face-to-face with web-based materials	High school student	Learning strategy or approach	Custom survey and qualitative analysis	1 pair research oriented and 1 pair task oriented

XSL•FO
RenderX

Study, year; country	Sample size, n	Participants' field	Course	Course delivery type	Participants' academic level	Study subject	Learning inventory	Results
Li and Abdul Rahman [52], 2018; Malaysia	46	Bioinformatics	Genomics technology	Web based	Undergraduate student	Learning style	FSILS and data mining	Procession: active=55%, reflective=24%, neutral=21%; input: visual=66%, verbal=18%, neutral=16%; perception: sensing=31%, intuitive=48%, neutral=21%; understanding: sequential=62%, global=12%, neutral=26%

[a]FSILS: Felder and Soloman Index of Learning Styles.

[b]The numbers denoted by a percentage sign in the Results column represent the percentage of learners who have declared the corresponding learning preference among all learners.

[c]HCP: health care professional.

[d]D-SA-LSI: Danish Self-Assessment Learning Styles Inventory.

[e]The scores indicate the strength of students' inclination toward the corresponding preference and were calculated based on the D-SA-LSI (range 0-7).

Of the 8 included studies, 6 (75%) [45,46,52,54-56] explored the learning strategies and preferences of bioinformatics students or courses, 1 (12%) investigated a precision medicine course [41], and 1 (12%) investigated an advanced statistics [53] course. It is worth noting that none of the included studies focused on courses specifically labeled as "health data science" courses.

Of the 8 included studies, 7 (88%) explored learning preferences, while 1 (12%) [56] analyzed students' learning strategies. Slightly more than one-third of the studies (3/8, 38%) [41,55,56] used a custom survey to measure students' learning preferences or strategies, while the rest (5/8, 62%) [45,46,52-54] used learning inventories, which are questionnaires that categorize students into different groups based on various learning dimensions (for a detailed description, refer to Multimedia Appendix 3 [57-63]). Of the 8 included studies, 3 (38%) [45,46,52] used the Felder and Soloman Index of Learning Styles (FSILS) [58,59], 1 (12%) [54] used the Kolb learning style inventory [57], and 1 (12%) [53] used the Danish Self-Assessment Learning Styles Inventory based on the theory propounded by Sternberg [60,64,65].

Regarding the analysis approach and data, most of the articles (6/8, 75%) performed only a qualitative analysis using a questionnaire and simple quantitative methods, such as statistical descriptive techniques applied to questionnaires (3/6, 50%) [53,55,56]. Three of them also supplemented their studies with a qualitative method. However, 2 (25%) of the 8 papers [46,52] used advanced data mining methods, such as k-means, and analyzed log data alongside self-reported data. Nevertheless, these studies [46,52] did not use log data to identify students' learning preferences; instead, they relied on self-reported inventories to train their models. Furthermore, all included studies except that by Micheel et al [41] had sample sizes of <100 (average 65) participants. The characteristics of the included articles are illustrated using various visualizations in Figures S1 and S2 in Multimedia Appendix 2.

The studies' quality, assessed using the MMAT, ranged between 73% and 100%, indicating excellent quality overall. None of the studies were excluded based on the MMAT score. Further details regarding the quality of the included articles and the MMAT checklists can be found in Multimedia Appendix 4 [41,45,46,52-56].

We used the narrative data synthesis approach [48] to combine the included studies to identify the learning preferences and strategies used in HDS. The studies were synthesized and narrated across different aspects, including information presentation preference (RQ1), collaboration preference (RQ2), preferred organization of learning process (RQ3), and preferred focus granularity (RQ4).

Proxies Used for Synthesis

Due to the heterogeneity among the included studies in terms of the measurements used to determine learning preferences and strategies in HDS courses, we found it necessary to define specific proxies for each learning preference. These proxies help in making connections between the results presented in the different studies. Table 2 displays the proxies associated with each RQ in this systematic review. More information about the learning inventories discussed in the included studies is available in Multimedia Appendix 3.

XSL•FO
RenderX

Table 2. Proxies used to connect the included studies' results to the research questions (RQs). The supporting evidence column provides available evidence in the literature about the association between the learning preference, style, or strategy and the proxies used.

Learning preference or strategy	Proxy	Source of the learning preference or strategy	Supporting evidence	RQ
Watching	Visual	Customized survey designed by Micheel et al [41]	Tendency toward watching lectures can be equivalent to a preference for visuals [41]	RQ1
Lecture	Visual	Customized survey designed by Abrahamsson and Dávila Lopez [55]	Tendency toward watching lectures can be equivalent to a preference for visuals [55]	RQ1
Assimilator	Visual	Kolb learning style inventory [57]	Assimilators are interested in learning through visual materials, such as videos and figures [57]	RQ1
Active	Teamwork	Felder and Soloman Index of Learning Styles [58,59]	Active students tend to work as a group and discuss learning materials with others [58,59]	RQ2
External	Teamwork	Danish Self-Assessment Learning Styles Inventory based on the theory propounded by Sternberg [60,65]	External students tend to work in a team and collaborate with others to solve problems [60,65]	RQ2
Internal	Independent work	Danish Self-Assessment Learning Styles Inventory based on the theory propounded by Sternberg [60,65]	Internal students prefer to work alone without communication with others [60,65]	RQ2
Reflective	Independent work	Felder and Soloman Index of Learning Styles [58,59]	Reflective learners are inclined to work alone or communicate with a close friend instead of a large group [58,59]	RQ2
Sequential	Sequential	Felder and Soloman Index of Learning Styles [58,59]	Sequential students have a linear learning process, which means they prefer to gain knowledge by following incremental and logical steps [58,59]	RQ3
Assimilator	Sequential	Kolb learning style inventory [57]	Assimilator students can organize the gained knowledge in a logical and clear format [57]	RQ3
Sensing	Factual information	Felder and Soloman Index of Learning Styles [58,59]	Sensing learners are interested in facts and concrete concepts, and they prefer exploring detailed information and intend to solve problems with standard approaches rather than innovative ones [58,59]	RQ4
Intuitive	Abstract information	Felder and Soloman Index of Learning Styles [58,59]	Intuitive learners are enthusiastic about abstract information, such as theories, and the deep meaning of learning materials [58,59]	RQ4
Global	Abstract information	Danish Self-Assessment Learning Styles Inventory based on the theory propounded by Sternberg [60,65]	Global students have the desire to solve abstract and huge problems [60,65]	RQ4
Local	Factual information	Danish Self-Assessment Learning Styles Inventory based on the theory propounded by Sternberg [60,65]	Local students prefer problems that need detailed and realistic solutions [60,65]	RQ4
Assimilator	Abstract information	Kolb learning style inventory [57]	Assimilators tend to prefer abstract ideas and concepts and are capable of perceiving a diverse range of information [57,66]	RQ4

XSL•FO
RenderX

Learning preference or strategy	Proxy	Source of the learning preference or strategy	Supporting evidence	RQ
Task oriented	Factual information	Customized survey designed by Gelbart et al [56]	The task-oriented student pair preferred specific tasks, and they did not always stay involved in all research steps; therefore, they only got a basic idea of what the research was about. They concentrated more on learning the details [56]	RQ4
Research oriented	Abstract information	Customized survey designed by Gelbart et al [56]	Research-oriented students are high achievers who are highly motivated to learn concepts with a deep understanding. They focus on generating abstract ideas and explanations that are connected to theoretical concepts [56]	RQ4

Information Presentation Preference (RQ1): Multimodal With Higher Tendency Toward Visual Presentation

Of the 8 studies included in this systematic review, 5 (62%) explored the preference of students regarding the type of presentation [41,45,52,54,55]. All these studies reported that students in HDS-related courses prefer visual presentations and benefit more from visualizations than from audio or reading types of presentations. However, all articles also acknowledge that students are multimodal learners and do not have only 1 preference regarding information presentation. In other words, if students prefer visual presentations, such as watching videos, it does not necessarily mean that they do not have any tendency toward reading or other types of presentations; for instance, Micheel et al [41] investigated the learning styles of oncology health care professionals learning precision medicine from web-based educational materials, and their research study showed that 80% of the learners had multimodal learning styles: the majority of the learners (39%) preferred watching, listening, and reading, while the next largest group (19%) preferred watching and reading. Abrahamsson and Dávila Lopez [55] analyzed the learning preferences of graduate students in 5 web-based bioinformatics-related courses and found that 91% of the students preferred synchronous and asynchronous lectures, which include visual presentations, while only 9% favored reading materials. Li and Abdul Rahman [52] analyzed the learning styles of bioinformatics students using the FSILS and found that the majority of the students were visual learners (66%). Holtzclaw et al [45] investigated the learning styles of undergraduate genetics students in a bioinformatics module and reported that the most dominant learning style among the students was visual (82%) compared to verbal (18%). The results from these studies are consistent with other research [41,52,55] highlighting that the majority of students prefer visual presentations. Finally, the study by Diwakar et al [54] also found that HDS students prefer visual presentations. The authors used the Kolb learning style inventory to classify bioinformatics students into multiple learning preferences and found that the majority of learners were classified as assimilators (60%) [54]. Assimilators tend to learn visually and prefer to observe a clear explanation [57]. A summary of the results of the studies is presented in Table 1.

Collaboration Preference (RQ2): Inconclusive Evidence

Of the 8 included studies, 5 (62%) [45,46,52,53,55] focused on the collaboration preferences of HDS students, and the results were inconclusive (Table 1). Most of these studies (3/5, 60%) [45,52,55] demonstrated that approximately half of the students preferred teamwork, while the other half preferred to work individually. Conversely, 2 (40%) of these 5 studies [46,53] indicated that HDS students had a preference for working in groups.

The study by Holtzclaw et al [45] is 1 (33%) of the 3 studies that show no clear student preference regarding collaboration in HDS. In particular, the authors reported that 54% of the bioinformatics students were found to be active learners, who typically prefer collaborating with peers, and 46% were found to be reflective learners, who have a tendency to work independently [45]. The difference between the 2 groups was not significant enough to conclude that there was a clear preference for collaboration or individual work. Similarly, Li and Abdul Rahman [52] found that more than half (55%) of the undergraduate bioinformatics students in their study were categorized as active learners (a tendency to collaborate with others), with the rest being categorized as reflective learners (a preference to work alone) or neutral. Abrahamsson and Dávila Lopez [55] reported that approximately 50% of the bioinformatics students in their study preferred to work alone on course assignments, while the other half preferred to work in groups (19% preferred to study with the same group for all sessions, 19% preferred to study with different groups, and 12% preferred to work individually in the first sessions and then study in groups).

The study by Sani Ibrahim [46] is 1 (50%) of the 2 studies indicating HDS students' preference for working in groups. The author reported that 70% of the bioinformatics students participating in the study were active learners who performed better in groups. In addition, the findings from the study by Nielsen and Kreiner [53], who used the Danish Self-Assessment Learning Styles Inventory, demonstrated that students enrolled in an advanced health statistics course had a strong tendency to be external, which shows their preference toward teamwork, with 89.3% of the students scoring as strong or very strong in this dimension (Table 1). This strong preference for external scope style suggests that students are willing to work as a team and communicate with others.

XSL•FO
RenderX

Overall, no consistent conclusion can be drawn based on the studies regarding HDS students' preference for working individually or in a group. Abrahamsson and Dávila Lopez [55] discuss several possible reasons for this inconsistency: first, the academic level of students may influence their preferences—postgraduate students have a higher research workload and are busier, which may lead to a higher tendency to work alone. Second, the type of assignment can influence students' working preferences; for example, the authors encouraged students to adopt paired programming for their programming assignments, and this optional approach was adopted by 85% of the bioinformatics students in their study, highlighting the effect of including activities in course design to promote student interactions. Finally, according to the authors, another possible reason could be the course platform because collaboration can be difficult in web-based courses.

Learning Process Organization Preference (RQ3): Sequential Learning Is More Popular

According to 3 (38%) of the 8 studies [45,52,54], the majority of HDS learners tend to have a sequential learning preference for organizing their learning process. Li and Abdul Rahman [52] found that 62% of their study participants had a sequential learning preference, while Holtzclaw et al [45] reported an even higher percentage of 75% (Table 1). Diwakar et al [54] also supported this conclusion, with 60% of their student population being assimilators, who tend to organize information logically and with clear order [57]. However, we should note that the number of studies that explored this dimension of preference is low, and further research is required to draw strong conclusions.

Focus Granularity Preference (RQ4): Higher Preference Toward Abstract Information

Of the 8 papers included in this systematic review, 5 (62%) provide evidence regarding the focus of students on abstract versus detailed information [45,46,52-54], with the majority of these papers (4/5, 80%) [46,52-54] agreeing that HDS students prefer main and abstract ideas (refer to Table 1 for further details).

The evidence regarding students' preferences for detailed or abstract information can be identified from the different learning styles reported (eg, intuitive or sensing, global or local, assimilator, executive, and research or task oriented) in the learning inventories used by these 5 studies. The study by Li and Abdul Rahman [52] found that the percentage of intuitive students (48%) was higher than that of sensing students (approximately 30%), while approximately 20% of the students were neutral in this dimension. Intuitive students prefer to focus on abstract ideas rather than detailed and factual knowledge, and they use a creative approach to problem-solving [58]. Similarly, Sani Ibrahim [46] expanded on the findings of Li and Abdul Rahman [52] and after using their data in addition to Moodle data, concluded that 94% of the bioinformatics students were intuitive. In the study by Diwakar et al [54], students were mostly assimilators (60%), who typically focus on abstract ideas and concepts. In addition, Nielsen and Kreiner [53] showed that HDS students tend to be slightly more global (ie, have the intention to solve abstract problems) rather than local (ie, have the desire to address detailed and realistic problems). Although the difference in the average scores for the 2 groups is small (Table 1), a much higher percentage of students (approximately 30%) scored strongly or very strongly as global compared to local (approximately 11%).

In contrast to the aforementioned studies that indicate a preference for abstract information, Holtzclaw et al [45] found that most students (67%) had a preference for sensing learning, preferring to focus on factual and detailed information.

In addition to the aforementioned 5 studies, Gelbart et al [56] identified 2 learning approaches among high school biology students in a bioinformatics-related course: research oriented (where abstract ideas are valued more highly) and task oriented (where there is attention to detail and focus on factual knowledge). However, this study included only 4 participants (research oriented: 2 and task oriented: 2), with insufficient evidence for addressing the particular RQ.

In conclusion, there is some evidence supporting the inference that HDS students prefer abstract information. However, it should be noted that there are also contradictory findings, and further research is needed to arrive at a more solid conclusion.

Discussion

Overview

A total of 8 articles that were published between 2005 and 2021 were included in the synthesis step. The synthesized results show that most HDS learners prefer visual presentations as their learning input. Regarding learning process and organization, they mostly prefer to follow logical, linear, and sequential steps. In addition, they focus more on abstract information, rather than detailed information. In terms of collaboration, HDS students prefer a mix of teamwork and independent work. On the basis of the findings of this systematic review, we provide herein some suggestions for future research and some recommendations for improving the design of HDS courses.

Recommendations for Course Design

It is known that student preferences can guide course instructors in designing more effective courses [10,22,24]. On the basis of HDS students' preference for visual presentation of information, it would be beneficial to include more attractive plots, flowcharts, and visual graphics within the course materials to make them more visually impressive.

Given HDS students' inclination toward sequential learning, where they organize their learning process in logical and clear steps, it would be advantageous to consider a stepwise approach in course design. Including step-by-step instructions for practical implementations or dividing concepts into meaningful sequential parts, may also benefit students; for example, Holtzclaw et al [45] designed a bioinformatics module based on students' learning styles, containing highly visual components and facilitating sequential learning. On the basis of postcourse feedback, students rated this module as valuable for their educational goals.

In terms of collaboration preferences, there is no consistent conclusion based on existing studies. Therefore, we recommend

XSL•FO
RenderX

designing HDS courses in such a way that students can choose freely between individual work and teamwork. This includes coursework where both types of assignments are offered.

Our final suggestion is that, given the evidence regarding the higher focus of HDS students on main and abstract ideas (as opposed to detailed information) and their tendency to apply a creative approach to problem-solving, it would be advantageous to reduce the details in the main course materials and instead include them in an appendix. In addition, creating challenging assignments that prompt reflection on abstract concepts and encourage the use of intuitive approaches for problem-solving can be beneficial for HDS students.

Although the aforementioned recommendations are based on the preferences of the majority of students in the reviewed studies, it is essential for educators to be aware of the heterogeneity of students' learning preferences and diversify HDS course design accordingly [53]. As the suggestions presented in this systematic review are based on a limited number of available studies, it is essential for educators to carefully consider the context of their specific course and student population when integrating these suggestions into their course design.

Guidelines for Future Studies

Additional research is needed to explore learning preferences and strategies in HDS courses, especially considering the conflicting findings in certain learning preferences (eg, collaboration preference and preferred focus granularity). In this subsection, we provide some suggestions for future studies.

First, we recommend the use of log data and data mining methods to analyze learning preferences and strategies in HDS courses. The majority of the included studies (6/8, 75%) entirely relied on self-reporting questionnaires or think-aloud protocols [41,53,56]. However, several studies have shown that self-reported inventories may not accurately reflect the actual behavior of learners because the learners may over- or underestimate their learning preferences or learning strategies [10,67]. To avoid this bias, we suggest using log data from learning platforms and data mining methods to accurately analyze the actual behaviors of students and uncover their learning preferences and strategies [10,68,69]. Applying data mining tools on log data can also help to analyze the temporal and dynamic behavior of students over time [70]. Recent studies [10,71] have demonstrated that using data mining tools uncovers students' preferences or strategies, which are dynamic and highly correlated with their performance [72]. As students may change their learning preferences and strategies throughout their interaction with a course [10,73], it is important to shed light on such changes. In this review, data-driven methods were used only by 2 (25%) of the 8 studies [46,52], which, however, were not well designed because they did not identify students' learning preferences based on the log data. Instead, they applied the FSILS to identify students' learning styles and then used the identified learning styles based on self-reported data as labels to train a model using log data; for example, Li and Abdul Rahman [52] only trained a computational model based on self-reported data instead of finding students' learning preferences using an unsupervised approach.

Second, it is necessary to analyze larger samples to strengthen the results and increase the generalizability of the findings. As mentioned earlier, all included studies except for that by Micheel et al [41] analyzed courses with <100 learners, which can be a limitation depending on the type of analysis conducted. The study by Gelbart et al [56] had a sample size of only 2 pairs of students. Although the study used qualitative analysis, the number of students considered and the information reported about them seem insufficient to support the authors' conclusion regarding the learning approaches of students [56]. Therefore, researchers, depending on the type of analysis (quantitative or qualitative), should be aware of the importance of having a suitable sample size to minimize the risk of bias in their conclusions [74,75].

Third, most of the included studies (5/8, 62%) did not report the demographic information of students. This is an important omission because students' nationality, race, and culture may affect their learning preferences [52]. To minimize the impact of other factors on the students' preferences and capture the preferences related solely to the HDS discipline, future research needs to include a diverse range of learners in terms of nationality, race, and other demographic characteristics. It is worth mentioning that in this systematic review, we examined the learning strategies and preferences of students across different academic levels, but no statistically significant differences were found between the different levels. Nevertheless, it is important to note that students' academic levels may influence their learning strategies and preferences. This aspect requires further investigation in future studies.

Finally, future studies should focus on students' learning strategies rather than learning styles because learning strategies are known to provide more useful information about a field in comparison with learning styles [10,13,76]. In addition, previous research has shown that learning strategies are highly associated with students' academic performance [77,78], while the association between learning styles and performance is controversial [76,79]. Among the 8 included studies, only 1 (12%) [56] discussed the learning strategies of HDS students, which was limited to self-reported data and had a very small sample size. Overall, much more needs to be done to gain comprehensive knowledge about HDS students. We encourage researchers to explore learning strategies in HDS courses using both log data and self-reported data.

Limitations

A limitation of this systematic review concerns the small number of studies included (n=8). Although we were systematic in our review and synthesis of these 8 studies, we acknowledge that it is a small number of studies, and therefore the results should be interpreted with caution.

Second, the heterogeneity among the included studies required the use of proxies to synthesize the results, and using meta-analysis was impossible due to the diverse measurements used across the studies. Although this systematic review defined meaningful and valid proxies to connect the heterogeneous pieces of evidence in the included studies, the use of different inventories in the studies to measure learning preferences and strategies can affect the accuracy of our findings.

XSL•FO
RenderX

It is worth mentioning that none of the included studies labeled their courses as "health data science" courses; the majority (6/8, 75%) referred to them as bioinformatics courses. It is important to note that in this systematic review, HDS is defined as a discipline in which students use computational methods and tools to analyze biological, health, or medical data. We did not include courses that focus on non–data analytics aspects, such as mobile health or electronic health records. Therefore, the findings of this systematic review may not apply to non–data analytics courses in health informatics.

Regarding the search queries and inclusion criteria, our study only included primary research studies in English published in journal and conference formats. In addition, due to the wide range of terminologies used in the literature to describe learning preferences and strategies, some relevant studies might have been overlooked given the search keywords used in this review; for example, we did not use the keyword "learning approach" in our search query, which could have resulted in additional studies for inclusion.

Moreover, due to the high occurrence of false positives in the search results obtained through SpringerLink and Wiley Online Library, our query for these 2 databases was restricted to studies that included the keyword "student" in their abstracts, which could have led to studies involving health care professionals being overlooked.

Regarding the quality of the included studies, while the MMAT serves as a powerful tool with low bias in assessment [80], it should be acknowledged that the assessment of the quality of included papers can be subjective. However, the 3 reviewers who assessed the quality of the included articles have different academic backgrounds and levels of expertise, which can potentially mitigate the associated bias.

Finally, students' learning preferences and strategies can be influenced by the mode of course delivery (eg, web based or face-to-face) and course design [11]; therefore, teachers and course designers should not solely rely on the findings of this study without considering other factors that might influence students' learning strategies and preferences. In addition, some suggestions within this review may specifically apply to web-based courses; for instance, the recommendation to use learning analytics to analyze students' learning behavior to identify dynamic learning strategies is not feasible for face-to-face courses.

Conclusions

We reviewed the literature to identify student learning preferences and strategies in HDS courses. The PRISMA guidelines were used, and, as a result, 8 papers were included for narrative synthesis. The synthesis of these studies provided evidence that most HDS students are visual and prefer learning through visual materials, such as videos, diagrams, plots, and so on, as part of their learning process. They also tend to follow logical and sequential steps in their learning process, and they are inclined to focus more on abstract information than on factual and detailed information. Moreover, there is no agreement among existing studies regarding students' collaboration preferences (teamwork vs independent work). HDS students might prefer to work alone on some assignments, while sometimes they prefer to work as part of a team.

On the basis of the reviewed studies, we recommend including more visual and less detailed materials in HDS courses, accompanied by stepwise instructions.

Furthermore, to address the limitations of existing studies, future research should consider using log data instead of self-reported questionnaires to capture the actual HDS learning experience. Including a large sample of students from different backgrounds and races can also strengthen research results and reduce the impact of other cofactors unrelated to the HDS discipline.

In addition, analyzing the learning strategies of students, rather than their learning preferences, has the potential to yield deep insights into HDS education because learning strategies are more associated with student performance. Overall, because a small number of studies have investigated learning preferences and strategies in HDS courses, further research is needed to draw definitive conclusions.

Acknowledgments

This work was supported by the Medical Research Council (grant MR/N013166/1). The authors would like to thank the funder and the Precision Medicine Doctoral Training Programme of the University of Edinburgh for their support for this research study. The authors also thank Dr Michael Gallagher and Dr Kobi Gal for their support and constructive discussions about students' learning strategies.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

NR and AM conceptualized and designed the project. NR and SS screened and reviewed the papers, while AM served as the third screener to resolve any conflicts between the 2 screeners. NR synthesized and wrote the first draft of the paper, and AM and SS assisted in enhancing the written work. AM supervised this project. All authors have read and approved the final version of the paper.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklists for abstracts and articles.
[PDF File (Adobe PDF File), 272 KB - mededu_v10i1e50667_app1.pdf]

Multimedia Appendix 2
Study protocol, search queries, excluded and included studies, and characteristics of the included studies.
[PDF File (Adobe PDF File), 559 KB - mededu_v10i1e50667_app2.pdf]

Multimedia Appendix 3
Summary of the learning inventories used by the included studies.
[PDF File (Adobe PDF File), 179 KB - mededu_v10i1e50667_app3.pdf]

Multimedia Appendix 4
Mixed Methods Appraisal Tool checklist results, showing the quality of the included studies.
[XLSX File (Microsoft Excel File), 69 KB - mededu_v10i1e50667_app4.xlsx]

References

1. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018 Sep 27;1(1):54 [FREE Full text] [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
2. Jidkov L, Alexander M, Bark P, Williams JG, Kay J, Taylor P, et al. Health informatics competencies in postgraduate medical education and training in the UK: a mixed methods study. BMJ Open 2019 Mar 30;9(3):e025460. [doi: 10.1136/bmjopen-2018-025460] [Medline: 30928942]
3. Spekowius G, Wendler T. Advances in Healthcare Technology: Shaping the Future of Medical Care. Volume 6. Cham, Switzerland: Springer; 2006.
4. Wan T, Gurupur V. Understanding the difference between healthcare informatics and healthcare data analytics in the present state of health care management. Health Serv Res Manag Epidemiol 2020;7:2333392820952668 [FREE Full text] [doi: 10.1177/2333392820952668] [Medline: 32923520]
5. Doudesis D, Manataki A. Data science in undergraduate medicine: course overview and student perspectives. Int J Med Inform 2022 Mar;159:104668. [doi: 10.1016/j.ijmedinf.2021.104668] [Medline: 35033982]
6. Wee R, Soh E, Giles D. Teaching data science to medical trainees. Clin Teach 2021 Aug;18(4):384-385. [doi: 10.1111/tct.13391] [Medline: 34101355]
7. Işık EB, Brazas MD, Schwartz R, Gaeta B, Palagi PM, van Gelder CW, et al. Grand challenges in bioinformatics education and training. Nat Biotechnol 2023 Aug;41(8):1171-1174. [doi: 10.1038/s41587-023-01891-9] [Medline: 37568018]
8. Walpole S, Taylor P, Banerjee A. Health informatics in UK medical education: an online survey of current practice. JRSM Open 2016 Jan;8(1):2054270416682674 [FREE Full text] [doi: 10.1177/2054270416682674] [Medline: 28210492]
9. Theobald M. Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: a meta-analysis. Contemp Educ Psychol 2021 Jul;66:101976. [doi: 10.1016/j.cedpsych.2021.101976]
10. Matcha W, Gašević D, Ahmad Uzir N, Jovanović J, Pardo A, Lim L, et al. Analytics of learning strategies: role of course design and delivery modality. J Learn Anal 2020 Sep 19;7(2):45-71. [doi: 10.18608/jla.2020.72.3]
11. Wang HC, Huang TH. Personalized e-learning environment for bioinformatics. Interact Learn Environ 2013 Feb;21(1):18-38. [doi: 10.1080/10494820.2010.542759]
12. Jones C, Reichard C, Mokhtari K. Are students' learning styles discipline specific? Community Coll J Res Pract 2010 Dec 15;27(5):363-375. [doi: 10.1080/713838162]
13. Oxford RL. Language learning styles and strategies: concepts and relationships. IRAL Int Rev Appl Linguist Lang Teach 2003;41(4):06. [doi: 10.1515/iral.2003.012]
14. Schmeck RR. Learning Strategies and Learning Styles. New York, NY: Springer Science & Business Media; 2013.
15. Entwistle NJ. Approaches to learning and perceptions of the learning environment. High Educ 1991 Oct;22(3):201-204. [doi: 10.1007/bf00132287]
16. Deale CS. Learning preferences instead of learning styles: a case study of hospitality management students' perceptions of how they learn best and implications for teaching and learning. Int J Scholarsh Teach Learn 2019 May 29;13(2):1-7. [doi: 10.20429/ijsotl.2019.130211]
17. Zaric N, Roepke R, Lukarov V, Schroeder U. Gamified learning theory: the moderating role of learners' learning tendencies. Int J Serious Games 2021 Sep 17;8(3):71-91. [doi: 10.17083/ijsg.v8i3.438]

XSL•FO
RenderX

18. Felder RM. Opinion: uses, misuses, and validity of learning styles. Adv Eng Educ 2020;8(1):1-16 [FREE Full text]
19. Tsingos C, Bosnic-Anticevich S, Smith L. Learning styles and approaches: can reflective strategies encourage deep learning? Curr Pharm Teach Learn 2015 Jul;7(4):492-504. [doi: 10.1016/j.cptl.2015.04.006]
20. Fariani RI, Junus K, Santoso HB. A systematic literature review on personalised learning in the higher education context. Tech Know Learn 2022 Nov 17;28(2):449-476. [doi: 10.1007/S10758-022-09628-4]
21. Bernacki ML, Greene MJ, Lobczowski NG. A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose(s)? Educ Psychol Rev 2021 Apr 27;33(4):1675-1715. [doi: 10.1007/S10648-021-09615-8]
22. Goodyear P, Carvalho L, Yeoman P. Activity-centred analysis and design (ACAD): core purposes, distinctive qualities and current developments. Educ Technol Res Dev 2021 Jan 11;69(2):445-464 [FREE Full text] [doi: 10.1007/s11423-020-09926-7] [Medline: 33456288]
23. Goodyear P. Educational design and networked learning: patterns, pattern languages and design practice. Australas J Educ Technol 2005 Mar 24;21(1):82-101. [doi: 10.14742/ajet.1344]
24. Young CP, Perović N. ABC LD – a new toolkit for rapid learning design. Eur Distance Educ Network 2020 Jun 22(1):426-437. [doi: 10.38069/edenconf-2020-ac0041]
25. Nancekivell SE, Shah P, Gelman SA. Maybe they’re born with it, or maybe it’s experience: toward a deeper understanding of the learning style myth. J Educ Psychol 2020 Feb;112(2):221-235. [doi: 10.1037/edu0000366]
26. Kirschner PA. Stop propagating the learning styles myth. Comput Educ 2017 Mar;106:166-171. [doi: 10.1016/j.compedu.2016.12.006]
27. Newton PM, Miah M. Evidence-based higher education - is the learning styles 'Myth' important? Front Psychol 2017;8:444 [FREE Full text] [doi: 10.3389/fpsyg.2017.00444] [Medline: 28396647]
28. Bani Baker Q, Nuser MS. Design bioinformatics curriculum guidelines: perspectives. In: Suravajhala PN, editor. Your Passport to a Career in Bioinformatics. Cham, Switzerland: Springer; 2021:91-102.
29. Bartle E. Personalised learning: an overview. The University of Queensland. URL: https://espace.library.uq.edu.au/view/UQ:357951 [accessed 2024-04-29]
30. Matcha W, Gašević D, Uzir NA, Pardo A. A systematic review of empirical studies on learning analytics dashboards: a self-regulated learning perspective. IEEE Trans Learning Technol 2020 Apr 1;13(2):226-245. [doi: 10.1109/tlt.2019.2916802]
31. Andres HP, Akan OH. A test of the teaching-learning style mesh hypothesis in a Chinese MBA. J Int Educ Bus 2015;8(2):02-163. [doi: 10.1108/jieb-12-2014-0021]
32. Andrews JD. Teaching format and student style: their interactive effects on learning. Res High Educ 1981;14(2):161-178. [doi: 10.1007/bf00976292]
33. Zajac M. Using learning styles to personalize online learning. Campus Wide Inf Syst 2009;26(3):19-265. [doi: 10.1108/10650740910967410]
34. Matcha W, Gašević D, Uzir NA, Jovanović JM, Pardo A. Analytics of learning strategies: associations with academic performance and feedback. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. 2019 Presented at: LAK '19; March 4-8, 2019; Tempe, AZ p. 461-470. [doi: 10.1145/3303772.3303787]
35. O'Shea E. Self-directed learning in nurse education: a review of the literature. J Adv Nurs 2003 Jul 11;43(1):62-70. [doi: 10.1046/j.1365-2648.2003.02673.x] [Medline: 12801397]
36. Shumba TW, Iipinge SN. Learning style preferences of undergraduate nursing students: a systematic review. Afr J Nurs Midwifery 2019 Aug 12;21(1):1-25. [doi: 10.25159/2520-5293/5758]
37. Stander J, Grimmer K, Brink Y. Learning styles of physiotherapists: a systematic scoping review. BMC Med Educ 2019 Jan 03;19(1):2 [FREE Full text] [doi: 10.1186/s12909-018-1434-5] [Medline: 30606180]
38. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. J Clin Epidemiol 2021 Jun;134:103-112. [doi: 10.1016/j.jclinepi.2021.02.003] [Medline: 33577987]
39. Chen CM, Sun YC. Assessing the effects of different multimedia materials on emotions and learning performance for visual and verbal style learners. Comput Educ 2012 Dec;59(4):1273-1285. [doi: 10.1016/j.compedu.2012.05.006]
40. Lee YH, Hsiao C, Ho CH. The effects of various multimedia instructional materials on students’ learning responses and outcomes: a comparative experimental study. Comput Human Behav 2014 Nov;40:119-132. [doi: 10.1016/j.chb.2014.07.041]
41. Micheel CM, Anderson IA, Lee P, Chen S, Justiss K, Giuse NB, et al. Internet-based assessment of oncology health care professional learning style and optimization of materials for web-based learning: controlled trial with concealed allocation. J Med Internet Res 2017 Jul 25;19(7):e265 [FREE Full text] [doi: 10.2196/jmir.7506] [Medline: 28743680]
42. Jiang D, Dahl B, Du X. A systematic review of engineering students in intercultural teamwork: characteristics, challenges, and coping strategies. Educ Sci 2023 May 24;13(6):540. [doi: 10.3390/educsci13060540]
43. Kanevsky L, Lo CO, Marghelis V. Individual or collaborative projects? Considerations influencing the preferences of students with high reasoning ability and others their age. High Ability Studies 2021 May 26;33(1):87-119. [doi: 10.1080/13598139.2021.1903842]
44. Weisman D. Incorporating a collaborative web-based virtual laboratory in an undergraduate bioinformatics course. Biochem Mol Biol Educ 2010 Jan;38(1):4-9 [FREE Full text] [doi: 10.1002/bmb.20368] [Medline: 21567782]

XSL•FO
RenderX

45. Holtzclaw JD, Eisen A, Whitney EM, Penumetcha M, Hoey JJ, Kimbro KS. Incorporating a new bioinformatics component into genetics at a historically black college: outcomes and lessons. CBE Life Sci Educ 2006 Mar;5(1):52-64 [FREE Full text] [doi: 10.1187/cbe.05-04-0071] [Medline: 17012191]
46. Sani Ibrahim M. Learning style detection using k-means clustering. FUDMA J Sci 2020 Sep 24;4(3):375-381. [doi: 10.33003/fjs-2020-0403-351]
47. Via A, Blicher T, Bongcam-Rudloff E, Brazas MD, Brooksbank C, Budd A, et al. Best practices in bioinformatics training for life scientists. Brief Bioinform 2013 Sep;14(5):528-537 [FREE Full text] [doi: 10.1093/bib/bbt043] [Medline: 23803301]
48. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme. Lancaster University. 2006. URL: https://tinyurl.com/a3fsm46a [accessed 2024-04-29]
49. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
50. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015 Jan 01;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
51. Stovold E, Beecher D, Foxlee R, Noel-Storr A. Study flow diagrams in Cochrane systematic review updates: an adapted PRISMA flow diagram. Syst Rev 2014 May 29;3(1):54. [doi: 10.1186/2046-4053-3-54] [Medline: 24886533]
52. Li LX, Abdul Rahman SS. Students' learning style detection using tree augmented naive Bayes. R Soc Open Sci 2018 Jul 25;5(7):172108 [FREE Full text] [doi: 10.1098/rsos.172108] [Medline: 30109052]
53. Nielsen T, Kreiner S. Course evaluation for the purpose of development: what can learning styles contribute? Stud Educ Eval 2017 Sep;54:58-70. [doi: 10.1016/j.stueduc.2016.10.004]
54. Diwakar S, Radhamani R, Nizar N, Kumar D, Nair B, Achuthan K. Using learning theory for assessing effectiveness of laboratory education delivered via a web-based platform. In: Proceedings of the 15th International Conference on Remote Engineering and Virtual Instrumentation. 2018 Presented at: REV '18; March 21-23, 2018; Düsseldorf, Germany p. 639-648. [doi: 10.1007/978-3-319-95678-7_70]
55. Abrahamsson S, Dávila López M. Comparison of online learning designs during the COVID-19 pandemic within bioinformatics courses in higher education. Bioinformatics 2021 Jul 12;37(Suppl_1):i9-15 [FREE Full text] [doi: 10.1093/bioinformatics/btab304] [Medline: 34252967]
56. Gelbart H, Brill G, Yarden A. The impact of a web-based research simulation in bioinformatics on students' understanding of genetics. Res Sci Educ 2008 Sep 26;39(5):725-751. [doi: 10.1007/s11165-008-9101-1]
57. Kolb DA, Kolb AY. Learning style inventory. Experience Based Learning Systems Inc. 1999. URL: https://learningfromexperience.com/downloads/research-library/the-kolb-learning-style-inventory-4-0.pdf [accessed 2024-03-29]
58. Felder RM, Spurlin J. Index of learning styles. Int J Eng Educ 1991 [FREE Full text] [doi: 10.1037/t43782-000]
59. Soloman BA, Felder RM. Index of learning styles questionnaire. North Carolina State University. 2010. URL: https://learningstyles.webtools.ncsu.edu/ [accessed 2024-04-29]
60. Sternberg RJ. Thinking Styles. Cambridge, UK: Cambridge University Press; 2018.
61. Freedman RD, Stumpf SA. What can one learn from the learning style inventory? Acad Manag Ann 1978 Jun;21(2):275-282. [doi: 10.5465/255760]
62. Kolb DA, Kolb AY. The Kolb learning style inventory 4.0: a comprehensive guide to the theory, psychometrics, research on validity and educational applications. Experience Based Learning Systems. URL: https://www.researchgate.net/publication/303446688_The_Kolb_Learning_Style_Inventory_40_Guide_to_Theory_Psychometrics_Research_Applications [accessed 2024-04-29]
63. Pintrich PR, De Groot EV. Motivated strategies for learning questionnaire (MSLQ). J Educ Psychol 1991 Mar;82(1):33-40. [doi: 10.1037/0022-0663.82.1.33]
64. Sternberg RJ. Mental self-government: a theory of intellectual styles and their development. Hum Dev 1988;31(4):197-224. [doi: 10.1159/000275810]
65. Nielsen T, Kreiner S. Mental Self-government: development and validation of a Danish self-assessment learning styles inventory using Rasch measurement models. Nielsen. 2005. URL: https://www.academia.edu/download/42852050/Mental_Self-Government_Development_and_v20160219-24651-1dthttps.pdf [accessed 2024-04-29]
66. Akella D. Learning together: Kolb's experiential theory and its application. J Manag Organ 2015 Feb 02;16(1):100-112. [doi: 10.1017/s1833367200002297]
67. Hadwin AF, Nesbit JC, Jamieson-Noel D, Code J, Winne PH. Examining trace data to explore self-regulated learning. Metacognition Learn 2007 Nov 29;2(2-3):107-124. [doi: 10.1007/s11409-007-9016-7]
68. Rohani N, Gal K, Gallagher M, Manataki A. Discovering students' learning strategies in a visual programming MOOC through process mining techniques. In: Proceedings of the 2022 Conference on Process Mining Workshops. 2022 Presented at: ICPM '22; October 23-28, 2022; Bozen-Bolzano, Italy p. 539-551. [doi: 10.1007/978-3-031-27815-0_39]
69. Rohani N, Gal K, Gallagher M, Manataki A. Early prediction of student performance in a health data science MOOC. In: Proceedings of the 16th International Conference on Educational Data Mining. 2023 Presented at: EDM '23; July 11-14, 2023; Bengaluru, India p. 2023 URL: https://zenodo.org/records/8115721 [doi: 10.5281/zenodo.8115721]

XSL•FO
RenderX

70. Berland M, Baker RS, Blikstein P. Educational data mining and learning analytics: applications to constructionist research. Tech Know Learn 2014 May 3;19(1-2):205-220. [doi: 10.1007/s10758-014-9223-7]
71. Fan Y, Matcha W, Uzir NA, Wang Q, Gašević D. Learning analytics to reveal links between learning design and self-regulated learning. Int J Artif Intell Educ 2021 May 21;31(4):980-1021. [doi: 10.1007/s40593-021-00249-z]
72. Winne PH. Learning analytics for self-regulated learning. In: Lang C, Siemens G, Wise AF, Gašević D, Merceron A, editors. Handbook of Learning Analytics. New York, NY: Society for Learning Analytics Research (SoLAR); 2017.
73. Fan Y, Jovanović J, Saint J, Jiang Y, Wang Q, Gašević D. Revealing the regulation of learning strategies of MOOC retakers: a learning analytic study. Comput Educ 2022 Mar;178:104404. [doi: 10.1016/j.compedu.2021.104404]
74. Lund B. The questionnaire method in systems research: an overview of sample sizes, response rates and statistical approaches utilized in studies. VINE J Inf Knowl Manag Syst 2021 Jan 14;53(1):1-10. [doi: 10.1108/vjikms-08-2020-0156]
75. Creswell JW. Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research. New York, NY: Pearson Education; 2012.
76. Feeley AM, Biggerstaff DL. Exam success at undergraduate and graduate-entry medical schools: is learning style or learning approach more important? A critical review exploring links between academic success, learning styles, and learning approaches among school-leaver entry ("traditional") and graduate-entry ("nontraditional") medical students. Teach Learn Med 2015 Jul 09;27(3):237-244. [doi: 10.1080/10401334.2015.1046734] [Medline: 26158325]
77. Weinstein CE, Husman J, Dierking DR. Self-regulation interventions with a focus on learning strategies. In: Boekaerts M, Pintrich PR, Zeidner M, editors. Handbook of Self-Regulation. New York, NY: Elsevier Academic Press; 2000:727-747.
78. Alexander PA, Graham S, Harris KR. A perspective on strategy research: progress and prospects. Educ Psychol Rev 1998;10:129-154. [doi: 10.1023/A:1022185502996]
79. Jiraporncharoen W, Angkurawaranon C, Chockjamsai M, Deesomchok A, Euathrongchit J. Learning styles and academic achievement among undergraduate medical students in Thailand. J Educ Eval Health Prof 2015 Jul 08;12:38 [FREE Full text] [doi: 10.3352/jeehp.2015.12.38] [Medline: 26165948]
80. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot mixed methods appraisal tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012 Jan;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]

Abbreviations

FSILS: Felder and Soloman Index of Learning Styles
HDS: health data science
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question

Edited by T de Azevedo Cardoso; submitted 08.07.23; peer-reviewed by H Chen, A Kononowicz, R Evans; comments to author 20.10.23; revised version received 14.12.23; accepted 27.05.24; published 12.08.24.

Please cite as:
Rohani N, Sowa S, Manataki A
Identifying Learning Preferences and Strategies in Health Data Science Courses: Systematic Review
JMIR Med Educ 2024;10:e50667
URL: https://mededu.jmir.org/2024/1/e50667
doi:10.2196/50667
PMID:

©Narjes Rohani, Stephen Sowa, Areti Manataki. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.08.2024. 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.

XSL•FO
RenderX

Time for Medicine and Public Health to Leave Platform X

Toomas Timpka[1,2,3], MD, PhD

1

2

3

Corresponding Author:
Toomas Timpka, MD, PhD

Abstract

For more than 50 years, digital technologies have been employed for the creation and distribution of knowledge in health services. In the last decade, digital social media have been developed for applications in clinical decision support and population health monitoring. Recently, these technologies have also been used for knowledge translation, such as in the process where research findings created in academic settings are established as evidence and distributed for use in clinical practice, policy making, and health self-management. To date, it has been common for medical and public health institutions to have social media accounts for the dissemination of novel research findings and to facilitate conversations about these findings. However, recent events such as the transformation of the microblog Twitter to platform X have brought to light the need for the social media industry to exploit user data to generate revenue. In this viewpoint, it is argued that a redirection of social media use is required in the translation of knowledge to action in the fields of medicine and public health. A new kind of social internet is currently forming, known as the "fediverse," which denotes an ensemble of open social media that can communicate with each other while remaining independent platforms. In several countries, government institutions, universities, and newspapers use open social media to distribute information and enable discussions. These organizations control their own channels while being able to communicate with other platforms through open standards. Examples of medical knowledge translation via such open social media platforms, where users are less exposed to disinformation than in general platforms, are also beginning to appear. The current status of the social media industry calls for a broad discussion about the use of social technologies by health institutions involving researchers and health service practitioners, academic leaders, scientific publishers, social technology providers, policy makers, and the public. This debate should not primarily take place on social media platforms but rather at universities, in scientific journals, at public seminars, and other venues, allowing for the transparent and undisturbed communication and formation of opinions.

(JMIR Med Educ 2024;10:e53810) doi:10.2196/53810

KEYWORDS

internet; social media; medical informatics; knowledge translation; digital technology; clinical decision support; health services research; public health; digital health; perspective; medicine

Introduction

Digital technologies have been used for the creation and distribution of knowledge in medicine and public health for more than 50 years. In addition to their applications for clinical decision support and population health monitoring [1-5], in the last decade, digital social media have been used for knowledge translation [6] such as in the process of establishing research findings as scientific evidence and distributing these findings for use in clinical practice, policy making, and health self-management. However, in May 2023, an editorial in *Lancet Digital Health*, a leading digital health journal, stated that the social media industry serves as "a commercial determinant of health due to the indirect health consequences of its business practices and actions" [7]. Eight months earlier, the entrepreneur Elon Musk had taken control of the microblog Twitter with a US $44 billion deal. At the time, Twitter was the social media platform most frequently used for knowledge translation and opinion formulation in the fields of medicine and public health. In one of his first actions as Chief Operating Officer, Musk laid off approximately 80% of the company's employees [8] and phased out most of the content moderation work that countered disinformation and misbehavior [9]. In addition, Musk allowed previously suspended users such as Russian state actors to reactivate their accounts [10]. Changes to the platform also made it essentially impossible for researchers to study the activities occurring on Twitter [11]. Users with verified (paid) accounts were only allowed to read 6000 posts per day, those with unverified accounts could read 600 posts, and new users with unverified accounts could read 300 posts per day. This action exposed that the rationale for the verification and dissemination procedures on the platform is based on financial rather than social capital; consequently, the possibilities for external observers to inspect interactions had been made almost nonexistent. The dismantling of Twitter was manifested by renaming the service X in July 2023, thereby completing the transformation of the platform.

XSL•FO
RenderX

To date, it has been considered standard practice for academic and health service institutions as well as for individual researchers and practitioners to have social media accounts for the dissemination of research findings, access to topical information, and participation in debates about novel discoveries. However, the transformation of Twitter to platform X highlights new threats to the presence of health institutions on social media platforms; in particular, the rational translation of scientific knowledge to health action is currently at stake. In this viewpoint, it is argued that a redirection of social media use is needed with respect to the translation of knowledge to action in medicine and public health.

Social Media Platforms as Transformative Technologies

More than 60 years ago, mass communication researchers pointed out that the introduction of any novel information technology has the potential to change the shape and character of the affected community [12,13]. Without much forethought or debate, in many countries, social media platforms have become organic components of the information infrastructure. For instance, in Sweden, the proportion of the population regularly following national news broadcasts on the radio and television remained constant between 2009 and 2020, while the proportion of social media users increased from 33% to 78% and the proportion subscribing to printed morning newspaper decreased from 70% to 24% during this same time period [14].

Public institutions active on social media typically choose a platform that they consider generally agrees with the content they want to disseminate; universities and scientific journals share research findings on microblogs such as platform X, whereas cultural authorities tend to use TikTok to share visual content and elementary schools share information in Facebook groups. Even before Twitter was rebranded as platform X, the microblog was associated with several problems and challenges that could have a potential impact on knowledge translation, especially concerning the verification of posted and shared information [15]. The decision made by Elon Musk and his managerial team to reduce the possibility for researchers to analyze the content and interactions occurring on the platform brings to light the fact that a large share of these issues can be traced back to the need for the social media industry to exploit user data to generate revenue.

The logic underpinning such user and customer misuse can be explained by the "life cycle" model of social media platforms [16]. Initially, while financed by seeding financial capital, the platforms produce value for their primary users; the platforms then exploit the users to produce value for business customers and then finally also exploit their business customers to maximize the value for the owners, which eventually leads to death of the platform. The life cycle model thus highlights that small efforts needed to change the operation of social media platforms can serve to rapidly redistribute value between stakeholders in a "two-sided market," where the platforms sit between users and producers of information. For example, user value is downplayed to produce value for business customers when platform algorithms are tuned to reward conflict-making and fragmentation with the goal of generating more views of posts, which, by extension, increases the display of ads paid for by business customers [17]. A recent study of news dissemination on Facebook [18] reported the greater circulation of conservative than liberal news domains, and indicated that a larger share of the news content was labeled as "false" in the conservative domains than in the liberal domains. The explanation provided by the authors for the faster propagation of conservative content, which has also been observed in other studies [19,20], was that false news stories disseminate faster on social media than true stories because false news items have more "novelty" and tend to arouse more emotions such as fear, disgust, and surprise than true news [21]. The authors argued that the Facebook functions Pages and Groups constitute a news curation and dissemination machine [22], which could then be available to any interest group for manipulation of public opinion [19] or intentionally fracturing an information ecosystem [23].

The Fediverse and Open Social Media

The problems associated with current social media platforms indicate that if social technologies are to be used for knowledge translation in medicine and public health, this translation should only take place on digital platforms where users are not exploited to create value for the platform's business partners and investors. The fediverse concept denotes an ensemble of open social media that can communicate with each other while remaining independent platforms [24]. With the emergence of open microblogs such as Mastodon [25] and Bluesky [26], the photo-sharing service Pixelfed [27], and the video service PeerTube [28], users can choose how they want to participate and own their data. Users sign up to specific instances in the fediverse and these instances host their data. The instances are operated by various actors, ranging from the Mastodon project to public institutions, for-profit corporations, nonprofit organizations, and groups of individual users. Technically, the fediverse consists of interconnected network servers running software applications that can read and write the same content.

The open architecture of the fediverse can be compared to that used for email messaging. The email user does not have to compose, organize, and read messages in the same software application, and two users do not have to use the same tools to communicate. The underlying idea is that email simply represents a source of data, and thus many software applications should be able to understand and manipulate these data. In other words, although email applications can have different interfaces, privacy policies, and purposes, every email application can interpret the meaning of an email address and every email address can send messages to every other email address, regardless of the application used. Similarly, if a user posts on Mastodon in the fediverse, another user can see the post in their Pixelfed feed.

The key enabler of the fediverse has been ActivityPub, a communication protocol overseen by the World Wide Web Consortium. More recently, other similar protocols have appeared, including AT managed by Bluesky. In Germany, government institutions, universities, and newspapers have

XSL•FO
RenderX

already begun using such forms of open social media [29]. This enables these organizations to control their own channels while still being able to communicate with other platforms through the open standards. In several other countries, many institutions, ranging from public service organizations such as the BBC in England to civil society organizations, have chosen to establish themselves with official accounts at alternative digital platforms. However, reliable data on the magnitude of the migration to alternative digital platforms remain scarce. One early study estimated that approximately 2% of Twitter users deleted their accounts and left the platform for the Mastodon project within the first weeks following the Musk takeover [30]. Approximately 15% of the followers of these users migrated to the exact same Mastodon instance as that of the users they follow. While the larger Mastodon instances attracted more users (the 25% largest instances on Mastodon attract 96% of users), the smaller instances, directed toward specific topics, attracted the more active users.

Translation of Health Knowledge on Social Media

Based on their responsibility for the peer review, verification, and distribution of research findings, scientific journals play a central role in knowledge translation. Medical and public health journals currently use social media platforms for the promotion and dissemination of content, branding, and facilitating conversation [31,32]. Although the number of social media posts shows a positive correlation with journal Altmetric scores and impact factors [33,34], there is no evidence for causal associations between social media activity and improved knowledge translation. Typically, approximately 40% of all scientific literature is posted on social media [35]; however, half of these posts draw no clicks to the underlying research, whereas an additional 20% of the posts receive only one or two clicks [36]. Moreover, the citation of articles by other researchers will not benefit from social media posting [37]. Instead, the social media presence of scientific journals may indirectly impede rational knowledge translation by luring potential users of new health knowledge to digital environments where research evidence is not necessarily discriminated from unrestrained streams of disinformation. Therefore, medical and public health journals have multiple motivations for reevaluating their social media presence and considering movement to open platforms or even leaving social media altogether. Reconsidering their social media presence is also relevant for the academic and health service institutions that produce and use the knowledge managed by scientific publishers. Establishing accounts and developing the ability to communicate on open microblogs such as Mastodon and Bluesky have become a technically viable alternative along with the use of open standards and protocols such as ActivityPub and AT. By moving to open social media platforms, health institutions can create a digital community that owns and operates their own channels for communication with each other, policy makers, and the public. Open platforms are not susceptible to the capriciousness of private companies and can also provide channels for the dissemination of unaltered health knowledge required during contingencies such as a pandemic. However, moving only one or a few health institutions from platform X to open platforms will not suffice in creating such digital environments.

Toward a Social Internet for Medicine and Public Health

A new kind of social internet is currently forming. As of February 2024, approximately one-fifth of daily Twitter/X users had left the platform since the Musk takeover in 2022 [38,39]. Considering the status of the social media industry, a short-term goal of medical and public health institutions should begin with contemplating the purpose of their social media presence and explaining how they protect health science beneficiaries from being misled by disinformation (eg, whether and how they promote science literacy [40]) among their followers.

In parallel, a broad discussion is needed about the use of social technologies for knowledge translation in medicine and public health. Examples of translation of medical knowledge in social media platforms where users are less exposed to disinformation are beginning to appear [41]. However, the new social internet also offers possibilities for novel, innovative forms of knowledge translation (eg, in demanding settings such as global health contingencies). For instance, as grounding for a synchronized response to a future pandemic, social media instances with purposive interaction rules [42] can proactively be created in interpandemic periods. Here, instances can be created for separate professional disciplines (from virologists to modelers) and policy-maker categories (from public health officers to politicians) [43]. In parallel, a set of orthogonal instances, organized as multidisciplinary networks, can be created where the professionals and policy makers can collaborate in local and regional response programs [44].

Continued discussion about the use of social media for knowledge translation in medicine and public health should involve researchers and health service practitioners, academic leaders, scientific publishers, social technology providers, policy makers, and the public. This debate should not primarily take place on social media platforms but also at universities, in scientific journals, at public seminars, and other venues, enabling the transparent and undisturbed communication and formation of opinions.

Acknowledgments

The preparation of this viewpoint article was supported by grants from the Swedish Research Council (VR 2021–05608, VR 2022-05608), Region Östergötland (ALF–936190), and the Research Council of Southeast Sweden (FORSS–940915).

XSL•FO
RenderX

Authors' Contributions

TT conceptualized the idea and wrote the manuscript.

Conflicts of Interest

None declared.

References

1. Timpka T. Introducing hypertext in primary health care: a study on the feasibility of decision support for practitioners. Comput Methods Programs Biomed 1989 May;29(1):1-13. [doi: 10.1016/0169-2607(89)90084-9] [Medline: 2653718]
2. Timpka T. Proactive health computing. Artif Intell Med 2001 Aug;23(1):13-24. [doi: 10.1016/s0933-3657(01)00073-2] [Medline: 11470214]
3. Timpka T, Eriksson H, Gursky EA, et al. Population-based simulations of influenza pandemics: validity and significance for public health policy. Bull World Health Organ 2009 Apr;87(4):305-311. [doi: 10.2471/blt.07.050203] [Medline: 19551239]
4. Timpka T, Eriksson H, Gursky EA, et al. Requirements and design of the PROSPER protocol for implementation of information infrastructures supporting pandemic response: a nominal group study. PLoS One 2011 Mar 28;6(3):e17941. [doi: 10.1371/journal.pone.0017941] [Medline: 21464918]
5. Spreco A, Jöud A, Eriksson O, et al. Nowcasting (short-term forecasting) of COVID-19 hospitalizations using syndromic healthcare data, Sweden, 2020. Emerg Infect Dis 2022 Mar;28(3):564-571. [doi: 10.3201/eid2803.210267] [Medline: 35201737]
6. Fahim C, Courvoisier M, Somani N, De Matas F, Straus SE. Creation of a theoretically rooted workbook to support implementers in the practice of knowledge translation. Implement Sci Commun 2023 Aug 18;4(1):99. [doi: 10.1186/s43058-023-00480-w] [Medline: 37596659]
7. The Lancet Digital Health. Twitter, public health, and misinformation. Lancet Digit Health 2023 Jun;5(6):e328. [doi: 10.1016/S2589-7500(23)00096-1] [Medline: 37179158]
8. Toh M, Liu J. Elon Musk says he's cut about 80% of Twitter's staff. CNN Business. 2023 Apr 12. URL: https://edition.cnn.com/2023/04/12/tech/elon-musk-bbc-interview-twitter-intl-hnk/index.html [accessed 2023-08-26]
9. O'Kane C. Twitter is officially ending its old verification process on April 1. To get a blue check mark, you'll have to pay. CBS News. 2023 Mar 24. URL: https://www.cbsnews.com/news/twitter-blue-check-verification-ending-new-subscription-april-1-elon-musk/ [accessed 2023-08-26]
10. Good R. Russia mulls lifting Twitter ban after Musk reinstates Kremlin account. Euronews. 2023 Oct 4. URL: https://www.euronews.com/next/2023/04/10/russia-mulls-lifting-twitter-ban-after-musk-reinstates-kremlin-account [accessed 2023-08-26]
11. Malik Y. What does Twitter 'rate limit exceeded' mean for users? Reuters. 2023 Jul 4. URL: https://www.reuters.com/technology/what-does-twitter-rate-limit-exceeded-mean-users-2023-07-03/ [accessed 2023-08-10]
12. McLuhan M. Understanding Media: The Extensions of Man: McGraw-Hill Publishers; 1964.
13. Rogers EM. The extensions of men: the correspondence of Marshall McLuhan and Edward T. Hall. Mass Commun Soc 2000 Feb;3(1):117-135. [doi: 10.1207/S15327825MCS0301_06]
14. SOM Institute. Swedish trends 1986-2020. University of Gothenburg. URL: https://www.gu.se/sites/default/files/2021-04/9.%20Swedish%20trends%201986-2020_korrigerad%202021-04-26.pdf [accessed 2024-05-14]
15. Choo EK, Ranney ML, Chan TM, et al. Twitter as a tool for communication and knowledge exchange in academic medicine: a guide for skeptics and novices. Med Teach 2015 May;37(5):411-416. [doi: 10.3109/0142159X.2014.993371] [Medline: 25523012]
16. Doctorow C. The 'Enshittification' of Tiktok or how, exactly, platforms die. WIRED. 2023 Jan 23. URL: https://www.wired.com/story/tiktok-platforms-cory-doctorow/ [accessed 2024-05-03]
17. Fisher M. The Chaos Machine: The Inside Story of How Social Media Rewired Our Minds and Our World: Little, Brown and Company; 2021.
18. González-Bailón S, Lazer D, Barberá P, et al. Asymmetric ideological segregation in exposure to political news on Facebook. Science 2023 Jul 28;381(6656):392-398. [doi: 10.1126/science.ade7138] [Medline: 37499003]
19. Huszár F, Ktena SI, O'Brien C, Belli L, Schlaikjer A, Hardt M. Algorithmic amplification of politics on Twitter. Proc Natl Acad Sci U S A 2022 Jan 4;119(1):e2025334119. [doi: 10.1073/pnas.2025334119] [Medline: 34934011]
20. González-Bailón S, d'Andrea V, Freelon D, De Domenico M. The advantage of the right in social media news sharing. PNAS Nexus 2022 Jul;1(3):gac137. [doi: 10.1093/pnasnexus/pgac137] [Medline: 36741446]
21. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 9;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
22. Benkler Y, Faris R, Roberts H. Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics: Oxford University Press; 2018.
23. Bakshy E, Messing S, Adamic LA. Political science. exposure to ideologically diverse news and opinion on Facebook. Science 2015 Jun 5;348(6239):1130-1132. [doi: 10.1126/science.aaa1160] [Medline: 25953820]

XSL•FO
RenderX

24. Pierce D. 2023 in social media: the case for the fediverse. The Verge. 2023 Dec 19. URL: https://www.theverge.com/23990974/social-media-2023-fediverse-mastodon-threads-activitypub [accessed 2024-05-03]
25. Brembs B, Lenardic A, Chan L. Mastodon: a move to publicly owned scholarly knowledge. Nature 2023 Feb;614(7949):624. [doi: 10.1038/d41586-023-00486-3] [Medline: 36810883]
26. Silberling A, Stringer A, Corrall C. What is Bluesky? Everything to know about the app trying to replace Twitter. TechCrunch. 2024 Mar 19. URL: https://techcrunch.com/2024/02/08/what-is-bluesky-everything-to-know-about-the-app-trying-to-replace-twitter/ [accessed 2024-05-03]
27. Pixelfed. URL: https://pixelfed.org/ [accessed 2024-05-03]
28. PeerTube. URL: https://joinpeertube.org/en [accessed 2024-05-03]
29. Ottenheimer D. German government on Mastodon. Flying Penguin. 2022 Nov 9. URL: https://www.flyingpenguin.com/?p=41863 [accessed 2024-05-03]
30. He J, Zia HB, Castro I, Raman A, Sastry N, Tyson G. Flocking to Mastodon: tracking the great Twitter migration. Presented at: 2023 ACM Internet Measurement Conference (IMC '23); Oct 24 to 26, 2023; Montreal, QC, Canada. [doi: 10.1145/3618257.3624819]
31. Kim D, Jung W, Jiang T, Zhu Y. An exploratory study of medical journal's Twitter use: metadata, networks, and content analyses. J Med Internet Res 2023 Jan 19;25:e43521. [doi: 10.2196/43521] [Medline: 36656626]
32. Thamman R, Eshtehardi P, Narang A, Lundberg G, Khera A. Roles and impact of journal's social media editors. Circ Cardiovasc Qual Outcomes 2021 Nov;14(11):e007443. [doi: 10.1161/CIRCOUTCOMES.120.007443] [Medline: 34749514]
33. Han J, Ziaeian B. Social media usage, impact factor, and mean Altmetric attention scores: characteristics and correlates in major cardiology journals. J Am Coll Cardiol 2019 Mar;73(9):3027. [doi: 10.1016/S0735-1097(19)33633-2]
34. Erskine N, Hendricks S. The use of Twitter by medical journals: systematic review of the literature. J Med Internet Res 2021 Jul 28;23(7):e26378. [doi: 10.2196/26378] [Medline: 34319238]
35. Fang Z, Costas R, Tian W, Wang X, Wouters P. An extensive analysis of the presence of Altmetric data for web of science publications across subject fields and research topics. Scientometrics 2020;124(3):2519-2549. [doi: 10.1007/s11192-020-03564-9] [Medline: 32836523]
36. Fang Z, Costas R, Tian W, Wang X, Wouters P. How is science clicked on Twitter? Clickmetrics for Bitly short links to scientific publications. J Assoc Inf Sci Technol 2021 Jul;72(7):918-932. [doi: 10.1002/asi.24458]
37. Branch TA, C té IM, David SR, et al. Controlled experiment finds no detectable citation bump from Twitter promotion. PLoS One 2024;19(3):e0292201. [doi: 10.1371/journal.pone.0292201] [Medline: 38507397]
38. Ingram D. Fewer people are using Elon Musk's X as the platform struggles to attract and keep users, according to analysts. NBC News. 2024 Mar 22. URL: https://www.nbcnews.com/tech/tech-news/fewer-people-using-elon-musks-x-struggles-keep-users-rcna144115 [accessed 2024-05-03]
39. Hern A. Twitter usage in US 'fallen by a fifth' since Elon Musk's takeover. The Guardian. 2024 Mar 26. URL: https://amp.theguardian.com/technology/2024/mar/26/twitter-usage-in-us-fallen-by-a-fifth-since-elon-musks-takeover [accessed 2024-05-03]
40. Costa A, da Silva Loureiro M, Ferreira ME. Scientific literacy: the conceptual framework prevailing over the first decade of the twenty-first century. Rev Colomb Educ 2021 Jan;1(81):195-228. [doi: 10.17227/rce.num81-10293]
41. Schukow CP, Punjabi LS, Abdul-Karim FW. #PathX: #PathTwitter's transformation and a discussion on different social media platforms used by pathologists in 2024. Adv Anat Pathol 2023 Dec 4. [doi: 10.1097/PAP.0000000000000424] [Medline: 38047394]
42. Nicholson MN, Keegan BC, Fiesler C. Mastodon rules: characterizing formal rules on popular Mastodon instances. Presented at: The 26th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '23); Oct 14 to 18, 2023; Minneapolis, MN. [doi: 10.1145/3584931.3606970]
43. Saltelli A, Boulanger PM. Technoscience, policy and the new media. Nexus or vortex? Futures 2020 Jan;115:102491. [doi: 10.1016/j.futures.2019.102491]
44. Ansell C, Gash A. Collaborative governance in theory and practice. J Public Adm Res Theory 2008 Oct 1;18(4):543-571. [doi: 10.1093/jopart/mum032]

Edited by T Leung; submitted 19.10.23; peer-reviewed by C Lokker, S Ganesh; revised version received 31.03.24; accepted 31.03.24; published 24.05.24.

Please cite as:
Timpka T
Time for Medicine and Public Health to Leave Platform X
JMIR Med Educ 2024;10:e53810
URL: https://mededu.jmir.org/2024/1/e53810
doi:10.2196/53810

XSL•FO
RenderX

© Toomas Timpka. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.5.2024. 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.

XSL•FO
RenderX

Evidence-Based Learning Strategies in Medicine Using AI

Juan Pablo Arango-Ibanez[1,*], MD; Jose Alejandro Posso-Nuñez[1,*], MD; Juan Pablo Díaz-Solórzano[1,*], MD; Gustavo Cruz-Suárez[2,3], MD

1

2

3

*these authors contributed equally

Corresponding Author:
Juan Pablo Arango-Ibanez, MD

Abstract

Large language models (LLMs), like ChatGPT, are transforming the landscape of medical education. They offer a vast range of applications, such as tutoring (personalized learning), patient simulation, generation of examination questions, and streamlined access to information. The rapid advancement of medical knowledge and the need for personalized learning underscore the relevance and timeliness of exploring innovative strategies for integrating artificial intelligence (AI) into medical education. In this paper, we propose coupling evidence-based learning strategies, such as active recall and memory cues, with AI to optimize learning. These strategies include the generation of tests, mnemonics, and visual cues.

(JMIR Med Educ 2024;10:e54507) doi:10.2196/54507

KEYWORDS

artificial intelligence; large language models; ChatGPT; active recall; memory cues; LLMs; evidence-based; learning strategy; medicine; AI; medical education; knowledge; relevance

Introduction

e-Learning has revolutionized the way medicine is taught and learned through the use of different internet-based technologies that enhance education [1]. Among these technologies, artificial intelligence (AI) tools, especially large language models (LLMs), have notably garnered significant attention in recent years, given their promising implications for medical education. LLMs are algorithmic models that are trained by extensive data sets, and they have the capability to comprehend text and generate natural-language text in response to a given prompt (input). This allows for interactive engagement with these technologies in a conversational format akin to a “chat” [2,3]. One of the most known LLMs is ChatGPT (owned by OpenAI), and its latest version, ChatGPT-4, was recently released to the public.

Recent studies have demonstrated the great achievements of LLMs in relation to medical knowledge and reasoning, such as ChatGPT-4 scoring 90% when answering USMLE (United States Medical Licensing Examination)–type questions [4], ChatGPT-4 passing a neurosurgery written board examination [5], and ChatGPT outperforming physicians in terms of providing empathic responses [6]. The educational potential of this technology is immense, encompassing a wide variety of applications. These include but are not limited to tutoring (personalized learning), patient simulation, generation of examination questions, and streamlined access to information [7-9]. The revolutionary potential of LLMs has resulted in researchers and medical students exploring the integration of AI into medical school curricula [10,11]. The rapid advancement of medical knowledge and the need for personalized learning underscore the relevance and timeliness of exploring innovative strategies for integrating AI into medical education [12].

Although numerous publications have examined the implications of LLMs for medicine and medical education, few have explored, in detail, specific strategies whereby LLMs can be used to optimize learning. In this paper, we propose strategies based on active recall, mnemonics, and the use of ChatGPT-4 [13] and DALL·E 3 (through ChatGPT-4) for enhancing learning outcomes regarding factual knowledge and thus help fill this gap of information. These strategies include the creation of pretests and posttest quizzes, the development of mnemonics, and the use of visual cues and mnemonics. Pretests and posttests serve as effective tools for active recall—a proven method for improving memory retention. Mnemonics simplify complex information into more digestible and memorable formats. Visual cues provide a graphical representation of information, aiding in better understanding and recall.

Active Recall–Based Strategies

Medical school requires a significant amount of time spent on reading. Research indicates that medical students typically dedicate an average of 1.5 to 6 hours per day to reading [14,15]. Moreover, teacher-centered lectures, which predominantly focus

XSL•FO
RenderX

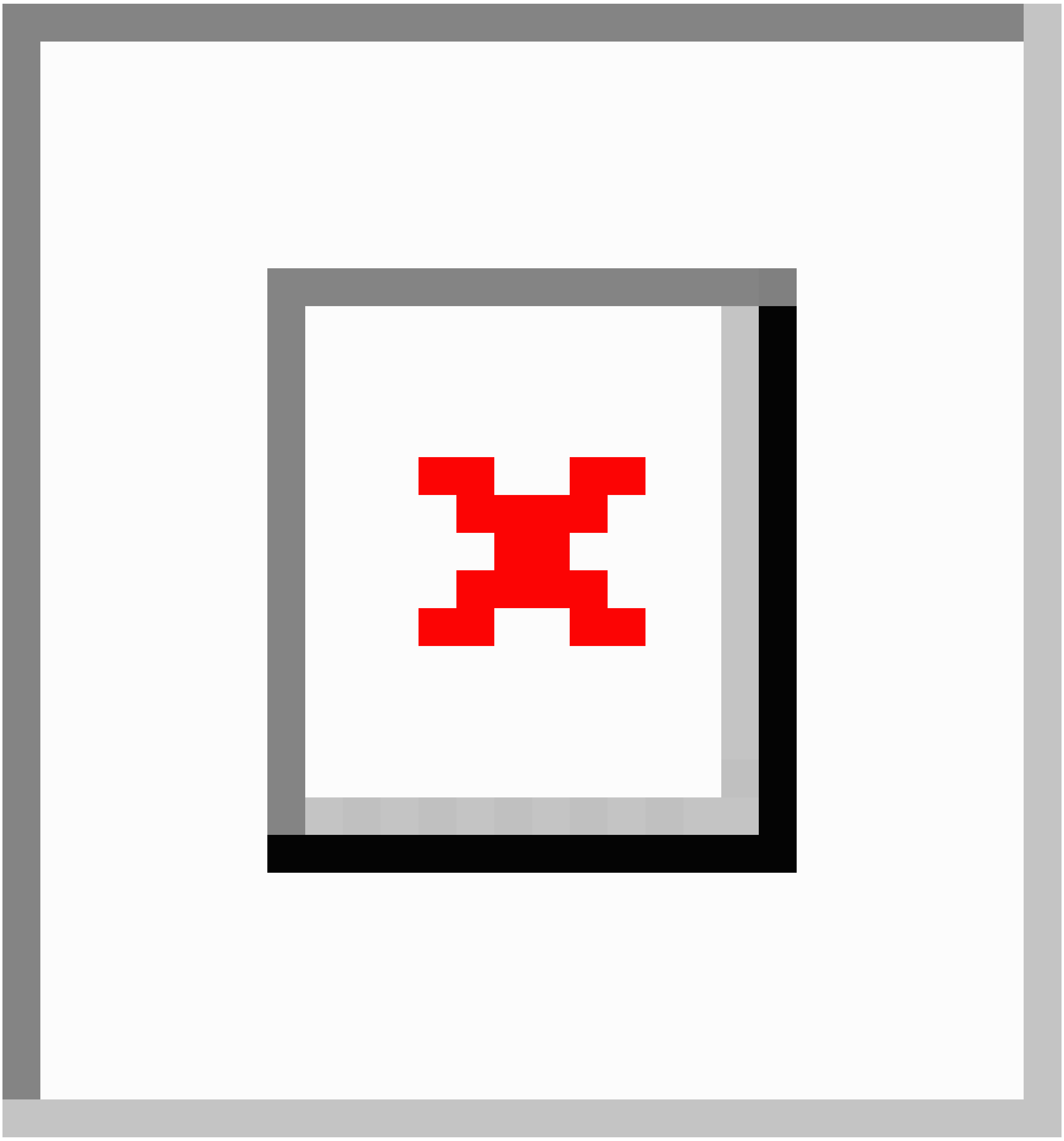
on passive learning, persist as one of the most used strategies despite challenges in the medical education community with regard to encouraging integration with active learning methods that enhance the retention and application of knowledge [16-18]. This may be because medical school students might prefer classic didactic lectures over demonstrations, small group discussions, feedback activities, group work (generating test questions and coming up with solutions to a problem), and other active learning methods that have been reported to better enhance memory and retention [19].

It is essential to adopt evidence-based strategies to enhance learning efficiency, especially considering the substantial academic workload and the ongoing reliance on passive learning methods. One such strategy is active recall, which involves actively retrieving information that was initially acquired passively through lectures, articles, or videos. This strategy is known to enhance learning significantly in comparison with passive learning strategies [20,21]. In this context, it is beneficial for readers to approach their reading proactively. This can be achieved by prefacing their reading with self-directed inquiries, understanding the main topics of the material, consistently formulating questions, and recognizing key concepts of high significance. Some related techniques include elaborative interrogation, which consists of answering "why" questions about a given concept to enhance medium- to long-term associative memory; self-explanation to relate new information to known information or explain steps taken for solving a problem to improve memory, comprehension, and transfer; and practice testing, which can improve memory, is not as time consuming, and can be applied at different times of the learning process [20].

One application of an active recall–based strategy involving the use of AI is illustrated in Figure 1, which shows ChatGPT being instructed to generate questions about cellulitis, as an example. Students are encouraged to attempt such questions before starting lectures or reading text. Answering questions before attending a lecture or reading a text (pretesting) is a strategy that enhances the learning process [22]. Interestingly, making mistakes during the study process can enhance learning by improving later memory; generating correct feedback; facilitating active learning; and stimulating the learner to redirect attention appropriately, especially when a mistake is followed by corrective feedback [23,24].

XSL•FO
RenderX

Figure 1. Example of the use of ChatGPT for pretesting.

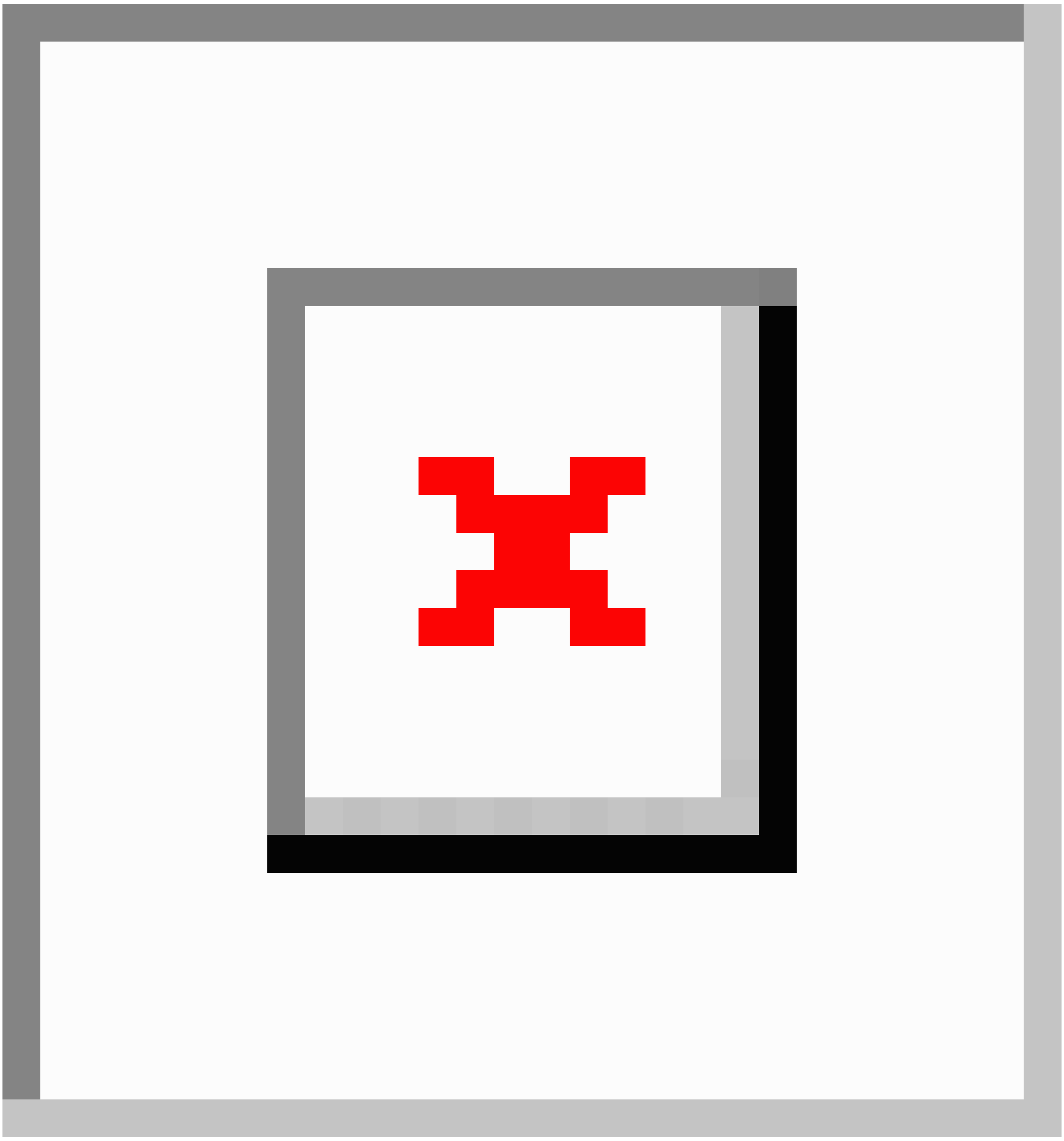


Taking tests has been proven to enhance learning in various studies [25-28]. Repeated test-taking increases the transfer of learning [25] and improves long-term recall [28], and it even outperformed concept mapping for long-term retention in a previous study [26]. This strategy can be integrated with AI, as shown in Figure 2, which depicts our attempt to extract information from a *StatPearls* article on cellulitis [29] and request ChatGPT to generate relevant questions. The AI system can produce various question formats, such as multiple-choice, true-false, and fill-in-the-blank questions, when given the appropriate prompts. These questions may be stored and reviewed days or weeks after the initial review to successfully apply spaced repetition, which has been demonstrated to improve learning and the consolidation of knowledge [21].

By using this method, one can input answers to questions and prompt ChatGPT to evaluate the answers' accuracy against the provided text. For instance, using questions from Figure 2, we tested ChatGPT's response by answering a query about common causative bacteria of cellulitis. We intentionally incorporated broad, correct concepts (gram-positive bacteria) and specific yet erroneous details (emphasizing staphylococci, particularly *Staphylococcus aureus*, as primary causatives instead of the correct streptococci) (Figure 3). ChatGPT feedback was tested again to contrast it with the feedback on a completely wrong answer (Figure 4).

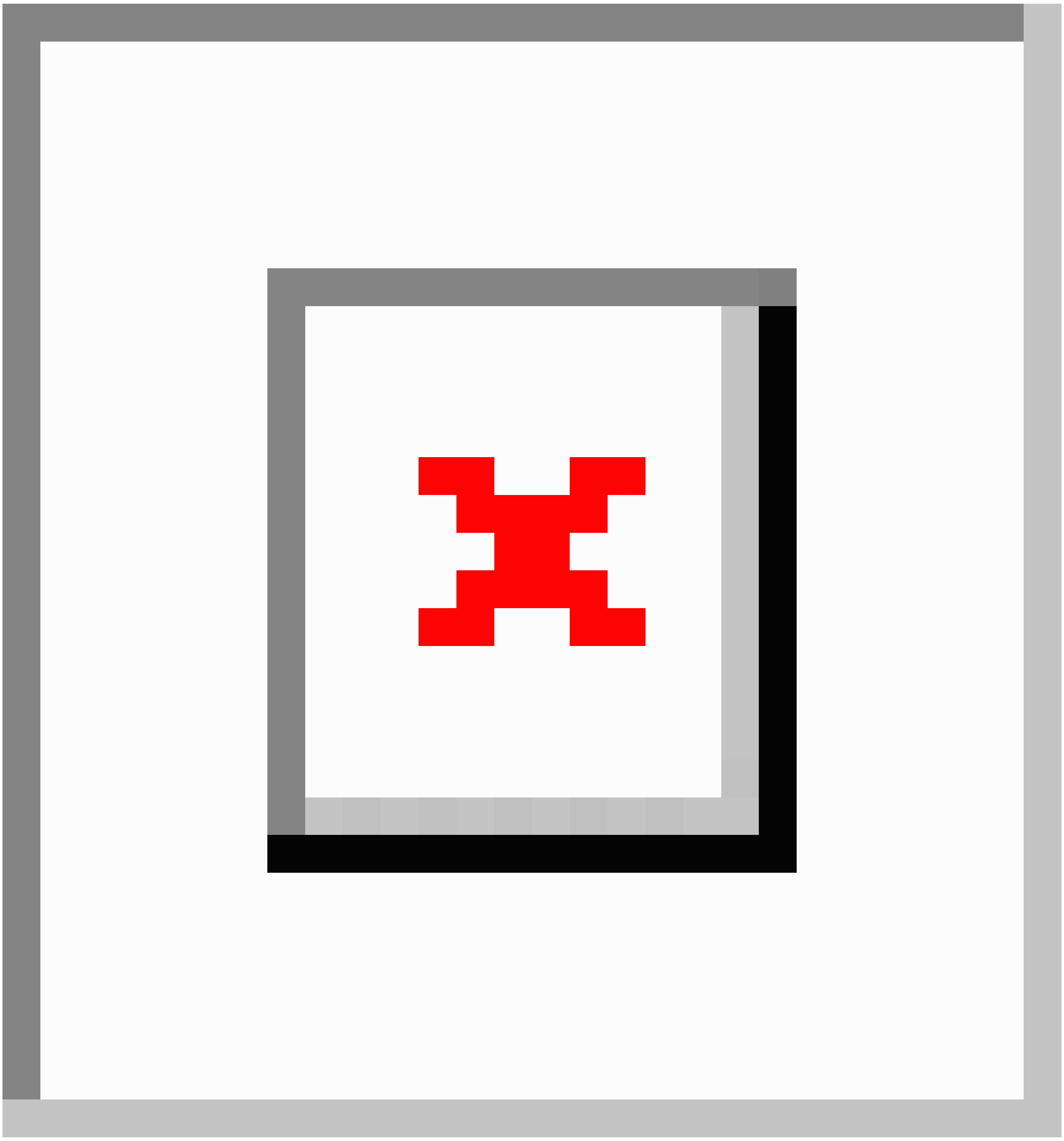
XSL•FO
RenderX

Figure 2. Example of the use of ChatGPT for creating a posttest.



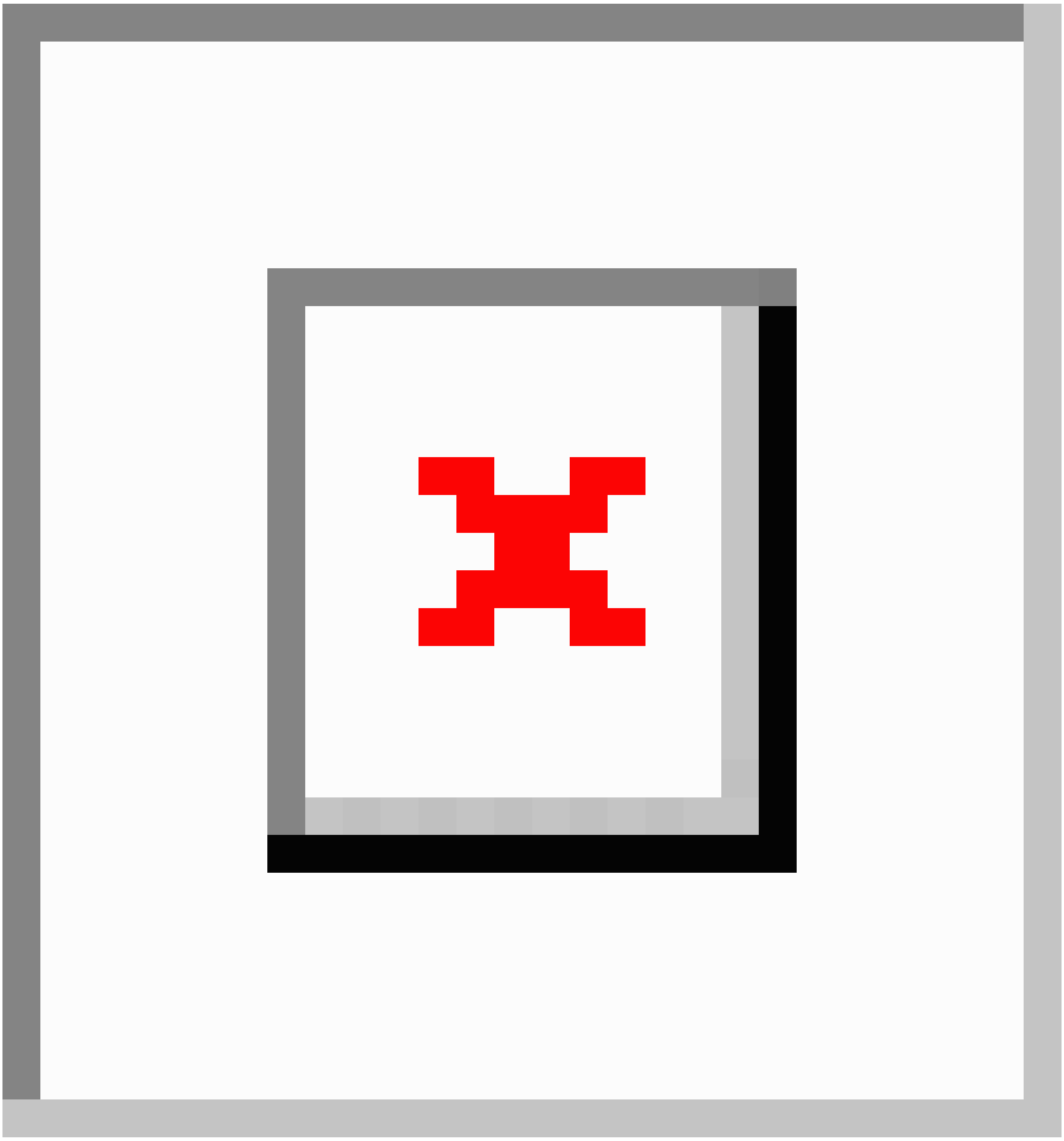
XSL•FO
RenderX

Figure 3. Feedback from ChatGPT on a partially correct answer to a question provided by ChatGPT.



XSL•FO
RenderX

Figure 4. Feedback from ChatGPT on a wrong answer to a question provided by ChatGPT.



Memory Cues

Memory cues are learning strategies in which a process of metacognition transforms information in a way that makes the information easier to recall or understand. Cues can be self-generated or generated by external agents, other people, or AI. Evidence has long suggested that self-generated cues are superior to cues generated by other people [30,31]. Nonetheless, there is available evidence that indicates that memory cues generated by others can still enhance recall [32].

Memory cues are effective because they make difficult-to-remember information into something simpler or meaningful, which facilitates recall [32,33]. For example, a classic memory cue in medical school for remembering descriptors of pain is the use of the mnemonic “SOCRATES” (site, onset, character, radiation, associations, time course, exacerbating factors, and severity). In this context, the name of the great philosopher is repurposed to recall how to properly assess pain in a patient, with the name becoming an acronym. In a recent meta-analysis, a statistically significant effect was found for cueing decreasing the learners’ perceived cognitive load and promoting learning outcomes, namely retention, and the transfer of knowledge [34].

Other modalities of memory cues that are commonly used include pictures, short stories, songs, and rhymes [32]. Evidence indicates conflicting conclusions regarding the superiority of a specific modality of cues over another. For instance, in a study

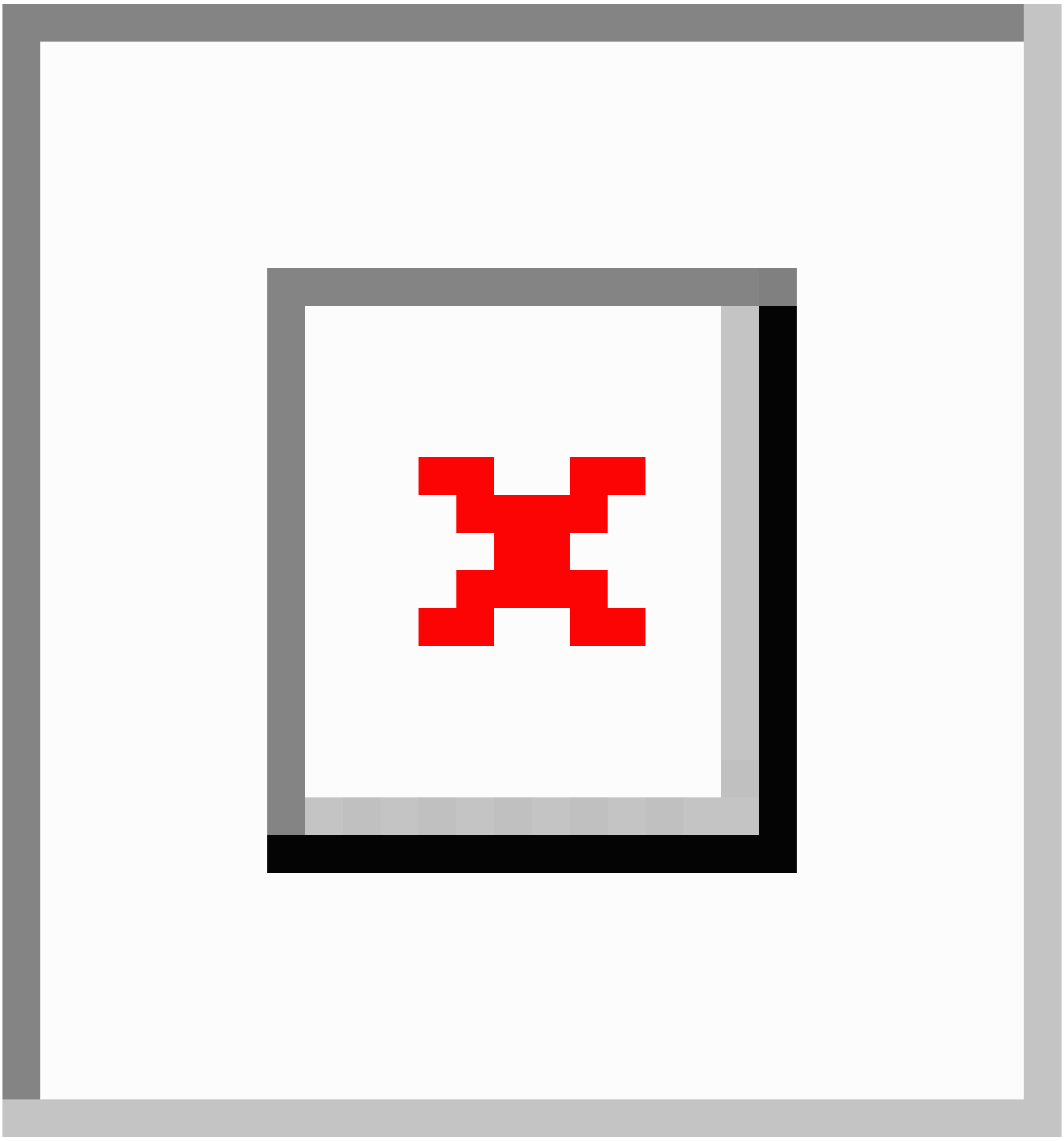
XSL•FO
RenderX

conducted by Pearson and Wilbiks [35], the authors attempted to evaluate the effect of the number of self-generated memory cues and aimed to test the findings of previous research that showed that the use of multisensory memory cues (ie, audiovisual cues) had a greater effect on recall than the use of one modality (ie, either visual cues [written words] or auditory cues [spoken words]). Their findings were that a greater number of cues led to higher recall, with statistical significance, but the modality of the cues did not have an effect on recall.

As previously indicated, one way to enhance the creation of an effective mnemonic is by using a common word as a cue to recall information [33]. This is one of the various ways that learners attempt to encode new vocabulary, abstract concepts, and master knowledge.

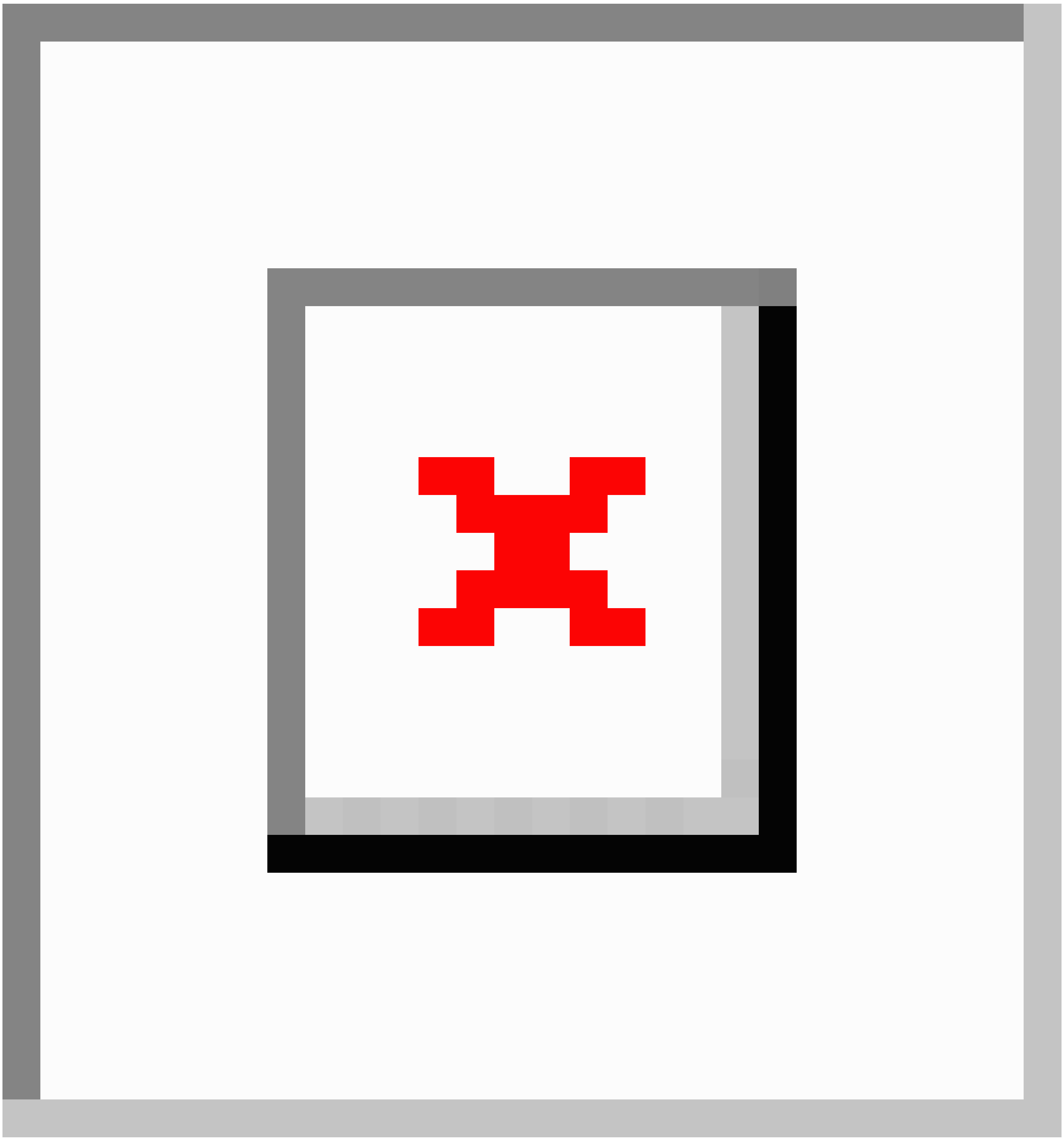
Figure 5 is an example of ChatGPT generating a mnemonic, using the word "brains" to recall the absolute contraindications of thrombolysis. Other examples are shown in Figure 6, in which ChatGPT creates a short story, and in Figure 7, in which ChatGPT creates a poem.

Figure 5. Acronym created by ChatGPT.



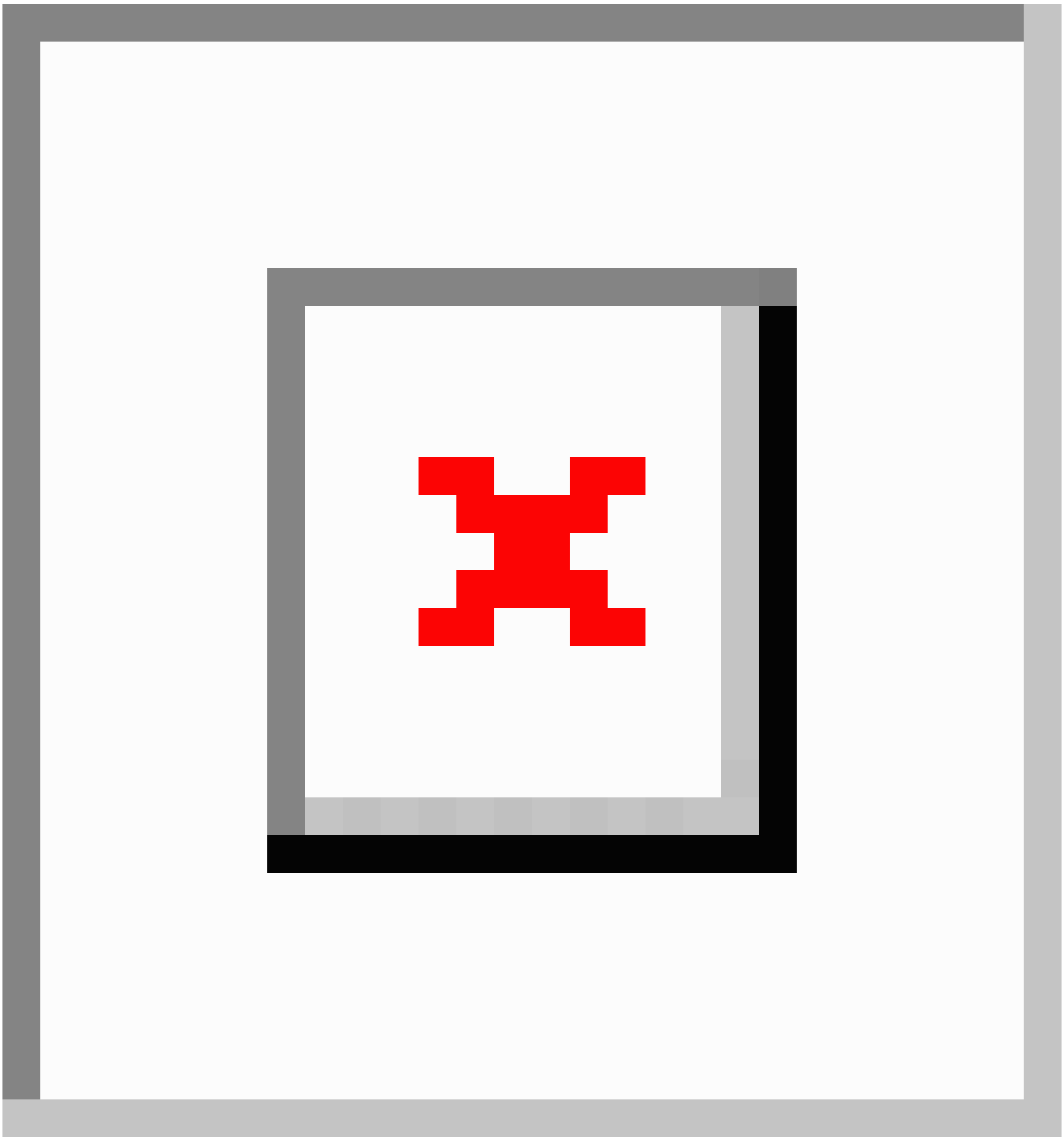
XSL•FO
RenderX

Figure 6. Short story created by ChatGPT.



XSL•FO
RenderX

Figure 7. Poem created by ChatGPT.



Visual Mnemonics

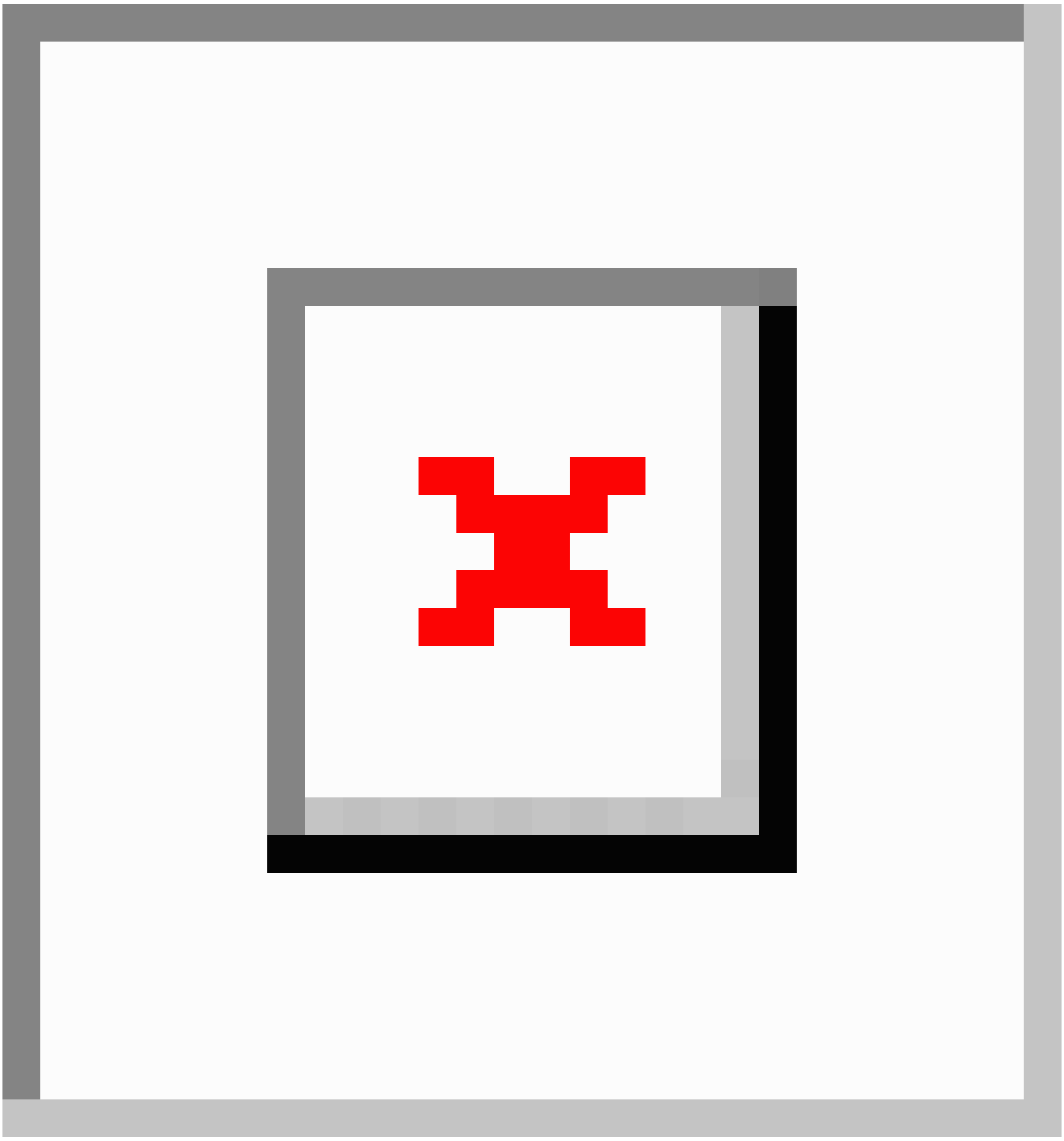
A visual mnemonic or cue is a tool that uses visual imagery to improve the recall of information. This differs from verbal mnemonics, which use words, phrases, or songs, as visual mnemonics use pictorial cues to forge memorable links. Their effectiveness stems from the incorporation of visual representations, analogies, or symbolism, which fortifies the associations and makes them more distinct. Visual mnemonics aid in recalling abstract or intricate information and facilitate both the sequential and the immediate retrieval of memorized material [36,37]. The use of mnemonics can be highly useful for learning difficult or abstract information [30], which is often found in the field of medicine [38]. Multiple studies have demonstrated that using visual or pictorial mnemonics can enhance learning outcomes [35,39].

DALL·E 3 is an AI system created by OpenAI that generates images based on prompts provided by the user and can be used for the creation of visual mnemonics. An example is given in Figure 8; a prompt was given to DALL·E 3 to create an image. For this example, which we created via DALL·E 3, the prompt "Fat purple man with long hair falling into a trap in a dry desert" was used to help recall some important features of hairy cell leukemia. "Fat" was used to recall the massive splenomegaly seen in patients with this condition; "purple" was used to make an association with lymphocytes, which are commonly seen as purple cells via hematoxylin and eosin staining and are involved in the pathogenesis of this neoplasm; "long hair" helps with

XSL•FO
RenderX

recalling the filamentous projections of cells in hairy cell leukemia; “trap” was used to remember that this disease stains positively in tartrate-resistant acid phosphatase staining; and “dry desert” was used to recall that bone marrow fibrosis leads to dry tap on aspiration.

Figure 8. DALL·E 3 creation with the prompt “Fat purple man with long hair falling into a trap in a dry desert.”



Discussion

Active Recall

Active recall is a highly effective learning strategy and significantly outperforms passive restudying when it comes to certain learning outcomes, such as conceptualization and long-term retention [21]. It has yielded better evaluation testing performance than traditional studying or rereading. Kornell et al [22] reviewed recall when participants were presented with fictional and nonfictional information, modifying the time for pretesting and read-only strategies. The testing strategy yielded a greater amount of correct answers than the read-only strategy, with statistical significance when equal or more time was allocated to the testing condition when the final test was performed more than 24 hours after the learning exercise, as well as in the fictional topic scenarios ($P<.01$). Other studies supporting the use of pretesting have been reported [22,40,41]. This highlights the role of pretesting in learning new information.

XSL•FO
RenderX

The benefits of active learning through testing have also been supported by other authors. Butler [25] tested students' recall ability when they were either passively restudying or studying via repeated testing. Butler [25] found that repeated testing resulted in better performance on a recall test than passive learning strategies and concluded that repeated test-taking increases the transfer of learning. In another study performed by Karpicke and Blunt [26], when retrieval practice (testing) was evaluated against passive learning strategies and even concept mapping, it proved to be better for verbatim and inference question answering, resulting in an improvement of about 50% in long-term retention scores (d=1.50; $F_{1,38}$=21.63; η_p^2=0.36). Additionally, the superiority of retesting over passive restudying for long-term retention has even been proven in a randomized controlled trial, wherein pediatric and emergency medicine residents were randomized to study the same text passages either via testing or repeated studying (ie, rereading). They were then tested on day 1, week 2, week 4, and month 6. The test results showed that the scores of participants who studied via testing were, on average, 13% higher than the scores of participants who performed repeated studying (P<.001), with an effect size of 0.91 [28]. Spaced testing (taking tests on different days between study sessions) has an even better effect on retention, long-term memory, and evaluation performance than repeated test-taking [27]. On the other hand, research indicates that spaced repetition (regardless of whether studying is done actively or passively) promotes more efficient and effective learning [42].

The previously mentioned studies highlight the importance of leveraging evidence-based techniques for studying rather than passive learning strategies. As we exemplified, these strategies can be coupled with AI. This approach addresses the limitation of relying solely on teacher-provided tests or textbook tests [20]. Moreover, ChatGPT is available on different platforms (web application and mobile app), and it can save chats (interactions) across these platforms. Therefore, students can easily access ChatGPT wherever it is needed and space their study sessions. Further, self-testing with AI reduces the pressure of graded assessments and leverages errors as learning opportunities [23,24], which research has shown to be particularly effective when the learner is confident in their incorrect answers [23]. This could be related to the effect of feedback.

Several articles on active recall and learning emphasize the role of feedback in enhancing learning processes, which is a characteristic that passive studying lacks. Roediger and Butler [27] compared traditional studying with test-taking studying without feedback and test-taking studying with differently timed feedback to determine whether test-taking studying results in better retention and whether retention is enhanced by feedback. They tested participants at different times and highlighted the efficiency of test-taking as a studying strategy, which was superior to that of traditional reading and restudying (22%, 32%, and up to a 43% difference between traditional studying and test-taking studying without feedback, test-taking studying with immediate feedback, and test-taking studying with delayed feedback, respectively). Since feedback has a clear impact on learning, especially when coupled with active recall strategies, non–AI-mediated active learning strategies could be limited by the lack of opportunities to offer feedback. Feedback generally comes from a reliable source, such as a teacher or an expert, or is obtained through an appropriate literature search, which can be time consuming. Sometimes, it is not possible to have the timely intervention of a teacher or an expert if there is a lot to study, and it may not be possible for a student to conduct a proper literature search if the student is new to a given topic. Furthermore, there are different types of feedback; some feedback is self-directed (ie, obtained through an introspective process). Feedback in the learning process can be used to enhance or develop skills for setting goals, monitoring one's own learning process, and assimilating input (feedback) toward enhancing performance [43]. All of these are important skills to have when one attempts to obtain feedback on their own, such as when using LLMs for feedback.

AI can enhance medical education by offering feedback and explanations to clarify incorrect responses, thereby increasing study efficiency. By using AI tools like ChatGPT, students can receive detailed feedback on their answers, including the identification of errors and the provision of correct information, as we have shown. ChatGPT presents promising implications in providing technically accurate medical feedback, given the exceptional knowledge it has exhibited, as we previously described. This process, however, should not replace thorough literature research or foundational knowledge acquisition. AI models can also serve as tutors to facilitate discussions on specific knowledge areas, similar to existing models in other fields, such as Khan Academy's Khanmigo, which serves as a fully personalized tutor [44]. One limitation of AI systems like ChatGPT is their character limit for inputs, which can be managed by breaking text into sections or using multiple prompts. Additionally, web searching is only available with ChatGPT's paid subscription; for the free version, one should provide ChatGPT with the reference text by copying and pasting it. Further research is needed to explore the potential of AI-assisted tutors in medical education, especially in education on basic subjects.

Memory Cues

Memory cues can be used as effective learning tools that ease the studying experience for a student attempting not only memorization but also mastery of complex concepts and new vocabulary. Evidence has described the superiority of self-generated memory cues over cues created by others [30,31]. The explanations behind the superiority of self-generated memory cues are (1) the generation effect of creating such cues, that is, the act of generation requires significant cognitive effort, which boosts memory, and (2) the cue selection process itself and the consequential metamnemonic effect. This means that students who identify the learning formats that work best for them are able to create their own memory cues by using a modality that is tailored to their needs [31]. For example, if a student prefers to have a visual representation of the ideas that they are attempting to memorize, they might lean toward the creation of mnemonics that create a mental picture to integrate information. Tullis and Fraundorf [31] proposed evidence that the benefits of self-generated cues come in great part from the correct selection of a cue from a list of candidates. If students

XSL•FO
RenderX

can create multiple cues, they can, with greater effectiveness, select the cue that best benefits retrieval. Tullis and Fraundorf [31] further suggested that allowing a learner to select from multiple options of cues requires less cognitive work, takes less time, and may not hinder memorization.

As we described, the act of generation is effortful and may be time consuming. AI tools like ChatGPT can help students as a result of their seemingly tireless and effortless generative capacity. In addition, these tools can create multiple cues with different modalities (textual-based cues and pictorial cues) when prompted to do so, thereby allowing learners to focus on understanding the material and selecting the most appropriate cue that fits their educational needs. The downside to the use of this method is that multiple attempts may be required for ChatGPT to produce a mnemonic that is subjectively good or fitting enough for a particular student. In addition, it is our opinion that these tools are best used when the user has an idea of what they should learn or memorize, and the user should prompt the AI tool to create a mnemonic device that facilitates the recall of the information they wish to encode. This is because there is abundant evidence of ChatGPT not only making errors but also blatantly providing false information [45], which is known as “artificial hallucination.”

Visual Mnemonics

Previous research has explored the effectiveness of visual mnemonics in improving learning outcomes. An experimental study that compared pictorial mnemonic use to traditional study methods found that pictorial mnemonics aid in learning from text passages by improving the recall of factual knowledge and long-term memory retention in college students [46]. Additionally, a randomized trial compared audiovisual mnemonics against traditional text-based learning for retaining medical knowledge; participants who used mnemonics demonstrated significant improvements in free-recall tests, with scores improving by 65%, 161%, and 208% immediately, after 1 week, and after 1 month, respectively, when compared to those who used text materials ($P<.001$). Moreover, the group that used mnemonics outperformed the group that used text materials by 55% in a 1-week–delayed multiple-choice test that focused on higher-order thinking ($P<.001$) [47]. In a comparative study of visual mnemonics versus traditional lectures for learning the porphyrin pathway, there was no significant difference in quiz scores immediately or 1 week after the intervention; however, the mnemonic group exhibited a 20% higher score 3 weeks later ($P=.02$) [48]. In another randomized trial that compared story-based audiovisual mnemonics with traditional text reading for memory retention among medical students, the audiovisual mnemonics group demonstrated significantly better performance in multiple-choice tests immediately after the intervention ($P=.04$), as well as at 1 week, 2 weeks, and 4 weeks after the intervention [49]. These results underscore story-based mnemonics’ superior effectiveness in enhancing immediate and long-term memory retention in medical education. Although there is some variation in the visual mnemonic techniques across studies (eg, the studies by Yang et al [47] and Abdalla et al [49] used some audiovisual mnemonics), they consistently demonstrated that factual knowledge can be represented visually and that the use of this type of mnemonic enhances both the recall and long-term retention of knowledge, with large effect sizes.

The visual mnemonic proposed in our study highly resembles the strategy used in the experiment by Rummel et al [46], in which visual mnemonics were created from texts about psychologists, incorporating elements for recalling both the psychologists’ names and the key aspects of their theories. In our mnemonic, “long hair” aids in recalling the name of the disease (hairy cell leukemia), and the other elements in the image are used to help recall the disease’s main features. The Picmonic System, which uses mnemonics from a web-based educational platform [50] that was used in the studies by Yang et al [47] and Abdalla et al [49], also adopts the visual mnemonic approach by combining visual elements and storytelling to enhance the recall of information; this is also highly similar to our approach. Thus, using DALL·E 3 for mnemonic generation shows promise for improving different learning outcomes, such as test performance, long-term retention, and free recall. Future studies should experimentally investigate the effectiveness of visual mnemonics generated by text-to-image models in learning processes. A significant limitation of using DALL·E 3 for medical mnemonic generation is its restriction on explicit content, prohibiting prompts with terms like “blood.” By recognizing this limitation, knowledge area–specific text-to-image models can be developed to more accurately describe the information needed and enable the use of words that are commonly used in a knowledge area but are censored in current models. Another limitation is that creating stories that accurately reflect the intended factual knowledge for mnemonic cues can be complex, particularly for certain subjects. Effective prompt engineering techniques could help in creating more relevant and coherent visual mnemonics.

Conclusions

LLMs, as a form of AI, are transforming the landscape of medical education. They offer a vast range of applications, and their potential has sparked discussions about integrating them into medical school curricula. Active recall–based learning strategies can be integrated with AI and can promisingly improve the recall and retention of information. This integration can be effectively applied by using AI to generate pretests and posttest quizzes. Memory cues, including self-generated mnemonics and mnemonics created by AI, can effectively simplify and transform complex information, thereby enhancing recall and optimizing learning. ChatGPT can create multiple types of memory cues, such as acronyms, short stories, and even poems. Moreover, AI tools, like DALL·E 3, can create images based on text and thus can be used to create visual mnemonics. However, crafting the right prompts can be challenging and time consuming, and results may vary. Thus, we believe that the use of new AI-based technologies, such as ChatGPT and DALL·E 3, is a highly useful strategy for learning, especially when these technologies are used with evidence-based principles. Further research is warranted to elucidate the impact of these strategies within the context of medical education.

XSL•FO
RenderX

Acknowledgments

JPAI designed and conceptualized the work. JPAI, JAPN, JPDS, and GCS gathered the data and performed the interpretation. JPAI, JAPN, and JPDS drafted the manuscript. JPAI, JAPN, JPDS, and GCS reviewed and corrected the final version of the manuscript. All authors have read and approved the final version of the paper.

Conflicts of Interest

None declared.

References

1. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of e-learning in medical education. Acad Med 2006 Mar;81(3):207-212. [doi: 10.1097/00001888-200603000-00002] [Medline: 16501260]
2. Naveed H, Khan AU, Qiu S, et al. A comprehensive overview of large language models. arXiv. Preprint posted online on Oct 5, 2023. [doi: 10.48550/arXiv.2307.06435]
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
4. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
5. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 1;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
6. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
7. Tsang R. Practical applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev 2023 May 24;10:23821205231178449. [doi: 10.1177/23821205231178449] [Medline: 37255525]
8. 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 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
9. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
10. Cooper A, Rodman A. AI and medical education - a 21st-century Pandora's box. N Engl J Med 2023 Aug 3;389(5):385-387. [doi: 10.1056/NEJMp2304993] [Medline: 37522417]
11. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
12. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285. [doi: 10.2196/19285] [Medline: 32602844]
13. ChatGPT. URL: https://chat.openai.com [accessed 2024-01-14]
14. Leff B, Harper GM. The reading habits of medicine clerks at one medical school: frequency, usefulness, and difficulties. Acad Med 2006 May;81(5):489-494. [doi: 10.1097/01.ACM.0000222273.90705.a6] [Medline: 16639211]
15. Klatt EC, Klatt CA. How much is too much reading for medical students? assigned reading and reading rates at one medical school. Acad Med 2011 Sep;86(9):1079-1083. [doi: 10.1097/ACM.0b013e31822579fc] [Medline: 21785317]
16. Prober CG, Heath C. Lecture halls without lectures--a proposal for medical education. N Engl J Med 2012 May 3;366(18):1657-1659. [doi: 10.1056/NEJMp1202451] [Medline: 22551125]
17. Prober CG, Khan S. Medical education reimagined: a call to action. Acad Med 2013 Oct;88(10):1407-1410. [doi: 10.1097/ACM.0b013e3182a368bd] [Medline: 23969367]
18. Bucklin BA, Asdigian NL, Hawkins JL, Klein U. Making it stick: use of active learning strategies in continuing medical education. BMC Med Educ 2021 Jan 11;21(1):44. [doi: 10.1186/s12909-020-02447-0] [Medline: 33430843]
19. Roffler M, Sheehy R. Self-reported learning and study strategies in first and second year medical students. Med Sci Educ 2022 Mar 18;32(2):329-335. [doi: 10.1007/s40670-022-01533-w] [Medline: 35528305]
20. Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT. Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. Psychol Sci Public Interest 2013 Jan;14(1):4-58. [doi: 10.1177/1529100612453266] [Medline: 26173288]
21. Augustin M. How to learn effectively in medical school: test yourself, learn actively, and repeat in intervals. Yale J Biol Med 2014 Jun 6;87(2):207-212. [Medline: 24910566]
22. Kornell N, Hays MJ, Bjork RA. Unsuccessful retrieval attempts enhance subsequent learning. J Exp Psychol Learn Mem Cogn 2009 Jul;35(4):989-998. [doi: 10.1037/a0015729] [Medline: 19586265]

XSL•FO
RenderX

23. Metcalfe J. Learning from errors. Annu Rev Psychol 2017 Jan 3;68:465-489. [doi: 10.1146/annurev-psych-010416-044022] [Medline: 27648988]
24. Mera Y, Rodríguez G, Marin-Garcia E. Unraveling the benefits of experiencing errors during learning: definition, modulating factors, and explanatory theories. Psychon Bull Rev 2022 Jun;29(3):753-765. [doi: 10.3758/s13423-021-02022-8] [Medline: 34820785]
25. Butler AC. Repeated testing produces superior transfer of learning relative to repeated studying. J Exp Psychol Learn Mem Cogn 2010 Sep;36(5):1118-1133. [doi: 10.1037/a0019902] [Medline: 20804289]
26. Karpicke JD, Blunt JR. Retrieval practice produces more learning than elaborative studying with concept mapping. Science 2011 Feb 11;331(6018):772-775. [doi: 10.1126/science.1199327] [Medline: 21252317]
27. Roediger HL 3rd, Butler AC. The critical role of retrieval practice in long-term retention. Trends Cogn Sci 2011 Jan;15(1):20-27. [doi: 10.1016/j.tics.2010.09.003] [Medline: 20951630]
28. Larsen DP, Butler AC, Roediger HL 3rd. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. Med Educ 2009 Dec;43(12):1174-1181. [doi: 10.1111/j.1365-2923.2009.03518.x] [Medline: 19930508]
29. Brown BD, Watson KLH. Cellulitis. In: StatPearls: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK549770/ [accessed 2024-01-15]
30. Tullis JG, Qiu J. Generating mnemonics boosts recall of chemistry information. J Exp Psychol Appl 2022 Mar;28(1):71-84. [doi: 10.1037/xap0000350] [Medline: 33939460]
31. Tullis JG, Fraundorf SH. Selecting effectively contributes to the mnemonic benefits of self-generated cues. Mem Cognit 2022 May;50(4):765-781. [doi: 10.3758/s13421-021-01245-3] [Medline: 34731430]
32. Tullis JG, Finley JR. Self-generated memory cues: effective tools for learning, training, and remembering. Policy Insights Behav Brain Sci 2018 Oct;5(2):179-186. [doi: 10.1177/2372732218788092]
33. Tullis JG, Finley JR. What characteristics make self-generated memory cues effective over time? Memory 2021 Nov;29(10):1308-1319. [doi: 10.1080/09658211.2021.1979585] [Medline: 34546833]
34. Xie H, Wang F, Hao Y, et al. The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: a meta-analysis and two meta-regression analyses. PLoS One 2017 Aug 30;12(8):e0183884. [doi: 10.1371/journal.pone.0183884] [Medline: 28854205]
35. Pearson HC, Wilbiks JMP. Effects of audiovisual memory cues on working memory recall. Vision (Basel) 2021 Mar 19;5(1):14. [doi: 10.3390/vision5010014] [Medline: 33808715]
36. Higbee KL. Mnemonics, psychology of. In: International Encyclopedia of the Social & Behavioral Sciences: Elsevier; 2001:9915-9918. [doi: 10.1016/B0-08-043076-7/01517-5]
37. O'Hanlon R, Laynor G. Responding to a new generation of proprietary study resources in medical education. J Med Libr Assoc 2019 Apr;107(2):251-257. [doi: 10.5195/jmla.2019.619] [Medline: 31019395]
38. Qiao YQ, Shen J, Liang X, et al. Using cognitive theory to facilitate medical education. BMC Med Educ 2014 Apr 14;14:79. [doi: 10.1186/1472-6920-14-79] [Medline: 24731433]
39. Reed LA, Hoffman LG. Pictorial cues and enhancement of patient recall of instructions or information. J Am Optom Assoc 1986 Apr;57(4):312-315. [Medline: 3700955]
40. Richland LE, Kornell N, Kao LS. The pretesting effect: do unsuccessful retrieval attempts enhance learning? J Exp Psychol Appl 2009 Sep;15(3):243-257. [doi: 10.1037/a0016496] [Medline: 19751074]
41. Latimier A, Riegert A, Peyre H, Ly ST, Casati R, Ramus F. Does pre-testing promote better retention than post-testing? NPJ Sci Learn 2019 Sep 24;4:15. [doi: 10.1038/s41539-019-0053-1] [Medline: 31583117]
42. Kang SHK. Spaced repetition promotes efficient and effective learning: policy implications for instruction. Policy Insights Behav Brain Sci 2016 Mar;3(1):12-19. [doi: 10.1177/2372732215624708]
43. Lüdeke ABEK, Olaya JFG. Effective feedback, an essential component of all stages in medical education. Universitas Medica 2020;61(3). [doi: 10.11144/Javeriana.umed61-3.feed]
44. Meet Khanmigo: Khan Academy's AI-powered teaching assistant & tutor. Khan Academy. URL: https://www.khanacademy.org/khan-labs [accessed 2024-01-19]
45. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb 19;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
46. Rummel N, Levin JR, Woodward MM. Do pictorial mnemonic text-learning aids give students something worth writing about? J Educ Psychol 2003;95(2):327-334. [doi: 10.1037/0022-0663.95.2.327]
47. Yang A, Goel H, Bryan M, et al. The Picmonic(®) Learning System: enhancing memory retention of medical sciences, using an audiovisual mnemonic web-based learning platform. Adv Med Educ Pract 2014 May 8;5:125-132. [doi: 10.2147/AMEP.S61875] [Medline: 24868180]
48. De Moll EH, Routt E, Heinecke G, Tsui C, Levitt J. The use of an imagery mnemonic to teach the porphyrin biochemical pathway. Dermatol Online J 2015 Apr 16;21(4):13030/qt4j51j8wt. [Medline: 25933070]
49. Abdalla MMI, Azzani M, Rajendren R, et al. Effect of story-based audiovisual mnemonics in comparison with text-reading method on memory consolidation among medical students: a randomized controlled trial. Am J Med Sci 2021 Dec;362(6):612-618. [doi: 10.1016/j.amjms.2021.07.015] [Medline: 34606752]

XSL•FO
RenderX

50. Picmonic® picture mnemonics - medical school, nursing school and more!. Picmonic. URL: https://www.picmonic.com/ [accessed 2024-01-03]

Abbreviations

AI: artificial intelligence
LLM: large language model
SOCRATES: site, onset, character, radiation, associations, time course, exacerbating factors, and severity
USMLE: United States Medical Licensing Examination

Edited by G Eysenbach; submitted 12.11.23; peer-reviewed by L Jantschi, S Ganesh; revised version received 20.01.24; accepted 23.03.24; published 24.05.24.

Please cite as:
Arango-Ibanez JP, Posso-Nuñez JA, Díaz-Solórzano JP, Cruz-Suárez G
Evidence-Based Learning Strategies in Medicine Using AI
JMIR Med Educ 2024;10:e54507
URL: https://mededu.jmir.org/2024/1/e54507
doi:10.2196/54507

© Juan Pablo Arango-Ibanez, Jose Alejandro Posso-Nuñez, Juan Pablo Díaz-Solórzano, Gustavo Cruz-Suárez. Originally published in JMIR Medical Education (https://mededu.jmir.org), 24.5.2024. 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.

XSL•FO
RenderX

Development of a Novel Web-Based Tool to Enhance Clinical Skills in Medical Education

Ayma Aqib[1], MBT; Faiha Fareez[2], MSc, MD; Elnaz Assadpour[1], MD; Tubba Babar[1], BSc; Andrew Kokavec[3], MD; Edward Wang[1], BSc; Thomas Lo[4], BMath; Jean-Paul Lam[4,5], MA, PhD; Christopher Smith[4], PhD

1
2
3
4
5

Corresponding Author:
Christopher Smith, PhD

Abstract

A significant component of Canadian medical education is the development of clinical skills. The medical educational curriculum assesses these skills through an objective structured clinical examination (OSCE). This OSCE assesses skills imperative to good clinical practice, such as patient communication, clinical decision-making, and medical knowledge. Despite the widespread implementation of this examination across all academic settings, few preparatory resources exist that cater specifically to Canadian medical students. MonkeyJacket is a novel, open-access, web-based application, built with the goal of providing medical students with an accessible and representative tool for clinical skill development for the OSCE and clinical settings. This viewpoint paper presents the development of the MonkeyJacket application and its potential to assist medical students in preparation for clinical examinations and practical settings. Limited resources exist that are web-based; accessible in terms of cost; specific to the Medical Council of Canada (MCC); and, most importantly, scalable in nature. The goal of this research study was to thoroughly describe the potential utility of the application, particularly its capacity to provide practice and scalable formative feedback to medical students. MonkeyJacket was developed to provide Canadian medical students with the opportunity to practice their clinical examination skills and receive peer feedback by using a centralized platform. The OSCE cases included in the application were developed by using the MCC guidelines to ensure their applicability to a Canadian setting. There are currently 75 cases covering 5 specialties, including cardiology, respirology, gastroenterology, neurology, and psychiatry. The MonkeyJacket application is a web-based platform that allows medical students to practice clinical decision-making skills in real time with their peers through a synchronous platform. Through this application, students can practice patient interviewing, clinical reasoning, developing differential diagnoses, and formulating a management plan, and they can receive both qualitative feedback and quantitative feedback. Each clinical case is associated with an assessment checklist that is accessible to students after practice sessions are complete; the checklist promotes personal improvement through peer feedback. This tool provides students with relevant case stems, follow-up questions that probe for differential diagnoses and management plans, assessment checklists, and the ability to review the trend in their performance. The MonkeyJacket application provides medical students with a valuable tool that promotes clinical skill development for OSCEs and clinical settings. MonkeyJacket introduces a way for medical learners to receive feedback regarding patient interviewing and clinical reasoning skills that is both formative and scalable in nature, in addition to promoting interinstitutional learning. The widespread use of this application can increase the practice of and feedback on clinical skills among medical learners. This will not only benefit the learner; more importantly, it can provide downstream benefits for the most valuable stakeholder in medicine—the patient.

(JMIR Med Educ 2024;10:e47438) doi:10.2196/47438

KEYWORDS

medical education; objective structured clinical examination; OSCE; e-OSCE; Medical Council of Canada; MCC; virtual health; exam; examination; utility; usability; online learning; e-learning; medical student; medical students; clinical practice; clinical skills; clinical skill; OSCE tool

Introduction

In 2020 and 2021, over 5000 final-year medical students graduated from a Canadian medical program and were matched to a residency program [1]. For these cohorts, portions of in-person clinical learning were limited due to the COVID-19 pandemic. Alongside clinical learning, the COVID-19 pandemic also caused numerous academic and health care institutions to

XSL•FO
RenderX

adopt more web-based learning platforms [2], thus emphasizing the importance of remote learning in the current day.

Prior to 2021, final-year Canadian medical students were required to pass an objective structured clinical examination (OSCE) held by the Medical Council of Canada (MCC) in order to progress to a residency training program [3]. Although this requirement has ceased for Canadian medical graduates, OSCEs remain integral within the medical education curriculum by serving as assessment tools for clinical skills. The goal of these OSCEs is to assess the candidate's clinical judgment, reasoning, knowledge, and skills. The examination is typically divided into twelve 11-minute–long stations, with a 2-minute break between each station. Stations can include clinical problems within the following fields: internal medicine, surgery, pediatrics, obstetrics and gynecology, psychiatry, and preventative medicine and public health [3].

The resources available to medical students for OSCE preparation and the real-world clinical setting are few and far between. Although such resources exist, they are limited by one or more factors. One of the biggest limitations for existing OSCE resources is that they are not specific to the MCC objectives, thus restricting their use in a Canadian medical education setting. Another major limitation is that they are often not directed at medical students but rather at students in other health care disciplines, such as pharmacy students and nursing students. Although these resources are beneficial for practice purposes, other professions have different scopes of practice, and the OSCE feedback generated for students via such resources may not always be translatable. Additionally, many of the existing OSCE preparation tools require user setup with platforms such as Zoom or Microsoft Teams; there are few that exist as stand-alone applications through which students can access feedback, clinical prompts, and OSCE assessments within a single centralized platform.

Another important limitation of existing resources is the inability to provide users with feedback regarding their clinical performance, specifically through formative learning experiences. Clinical educators often utilize quantitative scores and feedback in the form of checklists in order to provide students with assessments of their performance. However, this may not always be possible, given the time constraints of clinicians and staff. A possible solution to this is the utilization of peer feedback through formative learning experiences [4]. Unlike summative assessments and examinations, formative learning experiences provide students with opportunities in which they are able to focus on skill development as opposed to percentages and grades. Several studies have demonstrated the benefits of formative experiences, such as encouraging reflective review, reducing test anxiety, and advancing the learners' self-regulation skills [5,6]. Moreover, the remote nature of web-based platforms for formative learning can contribute to interinstructional learning, in which peers who have additional knowledge or exposure within certain medical fields can enhance the clinical skills of those whose training lacks in these areas.

Given the emphasis on web-based learning and the fact that few formative learning experiences exist for students, it is evident that there is a need for an electronic OSCE (e-OSCE) preparation tool that fills the aforementioned gaps in the medical education system. Thus, the beta version of the MonkeyJacket application for OSCE practice was developed with these gaps in mind [7]. The e-OSCE tool was piloted among a group of 6 medical students and resident physicians at Western University and McMaster University, with the goal of providing direct feedback to the software development team to refine the utility of the application. The primary research objective of this study was to describe the approach to the development and dissemination of the MonkeyJacket e-OSCE application tool. This paper also aims to describe the platform itself, the potential utility of the application as a tool that provides scalable formative feedback for learners, and how the application serves as a valuable tool in Canadian undergraduate medical education.

Development

Purpose of Development

The MonkeyJacket platform was built for the purpose of developing a formative learning experience (ie, rather than a summative one) in which the goals are to practice with various clinical cases and receive feedback through peer evaluations.

Tool Development

The backend of the MonkeyJacket platform was developed by a team of software engineers, project managers, and data scientists. The platform, including the video chat functionality, was custom coded by using a combination of Jitsi (8x8 Inc) and JavaScript Node.js (OpenJS Foundation). Through numerous rounds of user testing and quality control, the application was consistently reviewed and improved by the development team to ensure a smooth experience for users.

Development and Testing of the Application

The cases for the MonkeyJacket application were created by medical students and resident physicians. The trialing and testing of the application were conducted by a group of 6 medical students and resident physicians over a span of 3 months. Group members were encouraged to practice with everyone in the group to allow for diversity in perspectives and promote intragroup learning during the testing period. In addition to seeking group feedback regarding the practice cases and feedback checklists, the user study group was encouraged to provide feedback regarding functionality and ease of use. Comments were then relayed to the development team, and appropriate changes to the application were made.

Inclusion of Cases

The goal was to build practice cases that address CanMED (communicator, collaborator, leader, health advocate, scholar, professional, and medical expert) roles and provide formative feedback in the following disciplines: cardiology, respirology, gastroenterology, neurology, and psychiatry [8-10]. Within each discipline, cases were developed based on common and vital red-flag clinical presentations across patient demographics. Additionally, some uncommon and highly fatal conditions were also included within the data set to represent the diversity of cases seen in clinical settings. There are a total of 75 cases in

XSL•FO
RenderX

the data set, with no repeated diagnoses. All aspects of OSCEs, except the physical examination, were assessed. The cases were based on a composite of patient cases, of which some were created based on real-life deidentified scenarios, and others were adapted from an existing repertoire of cases from resources geared toward medical students, such as *OSCE and Clinical Skills Handbook* and other web-based resources [11-13]. Table 1 presents the number of cases per discipline.

Table . Breakdown of cases within the data set by medical discipline.

Medical discipline	Cases, n
Cardiology	14
Respirology	15
Gastroenterology	16
Neurology	15
Psychiatry	15

Building the Physician Candidate Prompts

The next step was developing the clinical prompt and task for each case, for both the student presenting as the "patient" and the student practicing as the "physician." We followed the MCC guidelines in ensuring that prompts were written in a clear and unambiguous manner and tasks could be completed in real time. For example, we avoided prompts such as "explore this further with the patient" and instead replaced them with prompts such as "take a thorough history, with a focus on GI symptoms and summarize your findings." We also avoided time-defining phrases, such as "the symptoms started at 9am," and instead replaced them with more definite timelines, such as "2 hours ago." All clinical stems included the patient's name, age, gender, and presenting symptoms and the task(s) that must be completed by the physician. The cases were framed such that it was the candidate's first time assessing the patient, rather than assuming that they had a pre-existing relationship with the patient.

Compiling Information for the Standardized Patient and Trainers

All patient case stems included the following demographic data: the patient's name; age; occupation; opening statement or history of the presenting illness, including symptoms with qualifications (onset, duration, quality, severity, timeline, alleviating factors, etc); associated symptoms; past medical history; medication history; family history; and social history. For the latter items, only positive histories (eg, if the patient has a history of past illnesses or a family history) were given. Nonverbal cues were also indicated on the patient's prompt so that they could be communicated to the physician, especially in psychiatry stations (eg, "I avoid eye contact, either looking at the ground or focusing on my hands. I give limited information making it obvious that I'm holding something back.").

Developing the Feedback Checklists

In deciding the number of checklist items for each clinical prompt, we included items that were relevant to assessing the candidate's abilities and ensured that the checklists were not exhaustive. The number of items on each checklist depended on the complexity of the case, but most checklists consisted of 30 to 40 items. The checklist items all began with an action verb to guide the standardized patient, who is also the examiner, on what was expected from the physician.

Using the MCC guidelines, we ensured that the items were discrete, observable, and dichotomous. Toward ensuring that items were discrete, each checklist item assessed for 1 concept or grouped concepts together; the candidate could get the full score even if they asked about 1 concept within the group. For example, a checklist item for qualifying pain was "Elicits character of pain – sharp, dull." For this checklist item, the candidate would get full marks for asking about any character of pain. In ensuring that items were observable, we avoided terminology including "understands" and "appreciates" and instead used terms like "asks about" and "gives reasonable differential diagnoses." Toward ensuring that items were dichotomous, the candidate either received the full mark for the item or did not; the checklist did not have any rating scales or instructions regarding part marks.

Review, Revise, and Pilot

The MCC states that case development is an iterative process requiring thought, review, and revision, and thus one should be open to feedback. The first step of the review involved the medical development team, which consisted of medical students and resident doctors, piloting the application in an iterative process to continue to refine the platform. This allowed us to identify missing information from the patient script and review the checklist to reduce ambiguity. Additionally, the cases were also reviewed by attending physicians in order to increase the validity of the clinical situations.

Ethical Considerations

This study did not contain or capture any human information or data. Therefore, as per Article 2.4 from the Tri-Council Policy Statement Research Ethics Board, this study was exempt from research and ethics review and did not require research ethics board approval [14].

Application Interface and Features

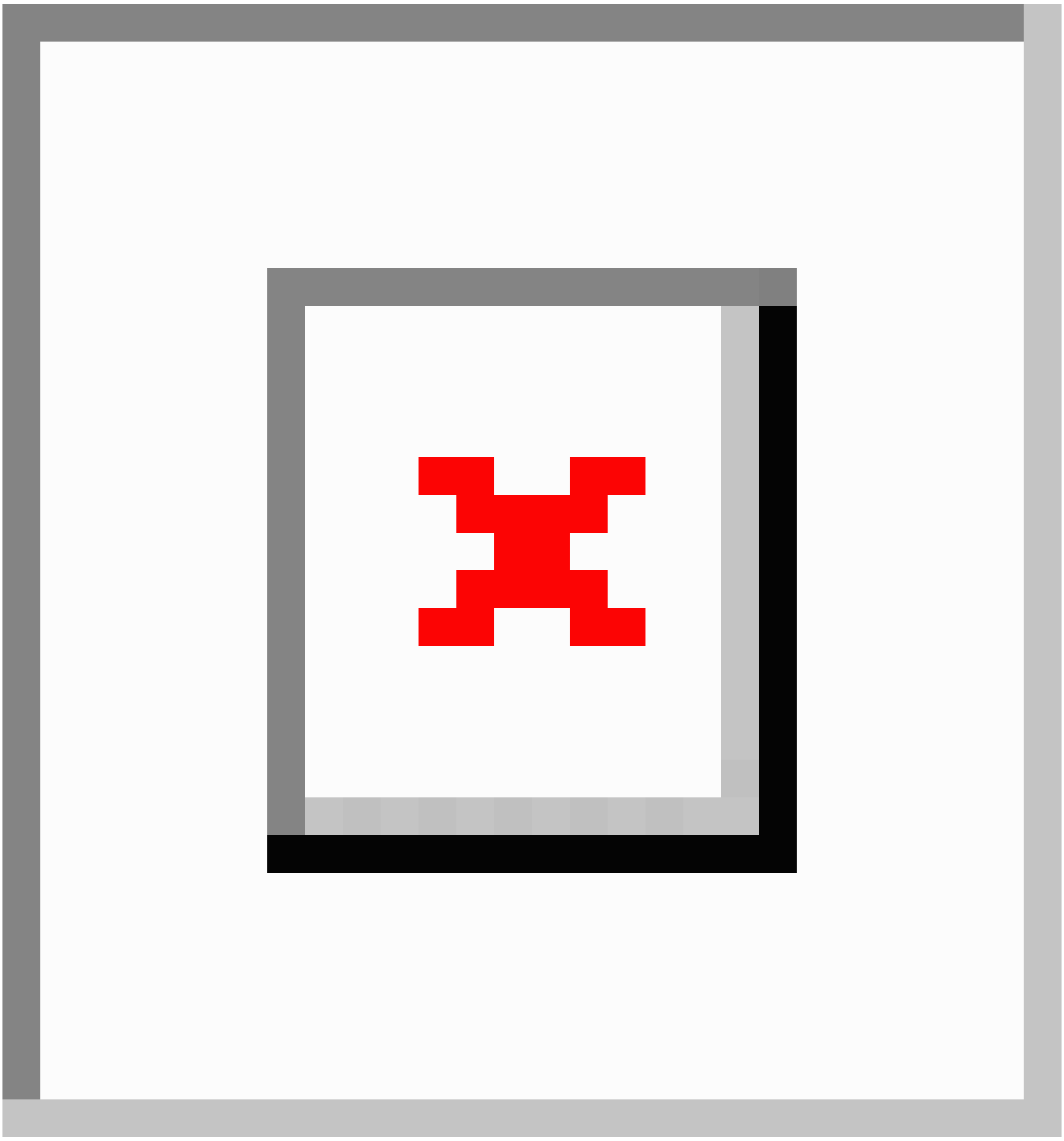
Description of the Application

Upon entry into the platform, students land on a home page in which they are able to enter their email and password credentials (Figure 1). Prior to the start of an OSCE station, the student completing the station as the acting physician receives a brief prompt that introduces the patient's name, age, and chief complaint (Figure 2).

XSL•FO
RenderX

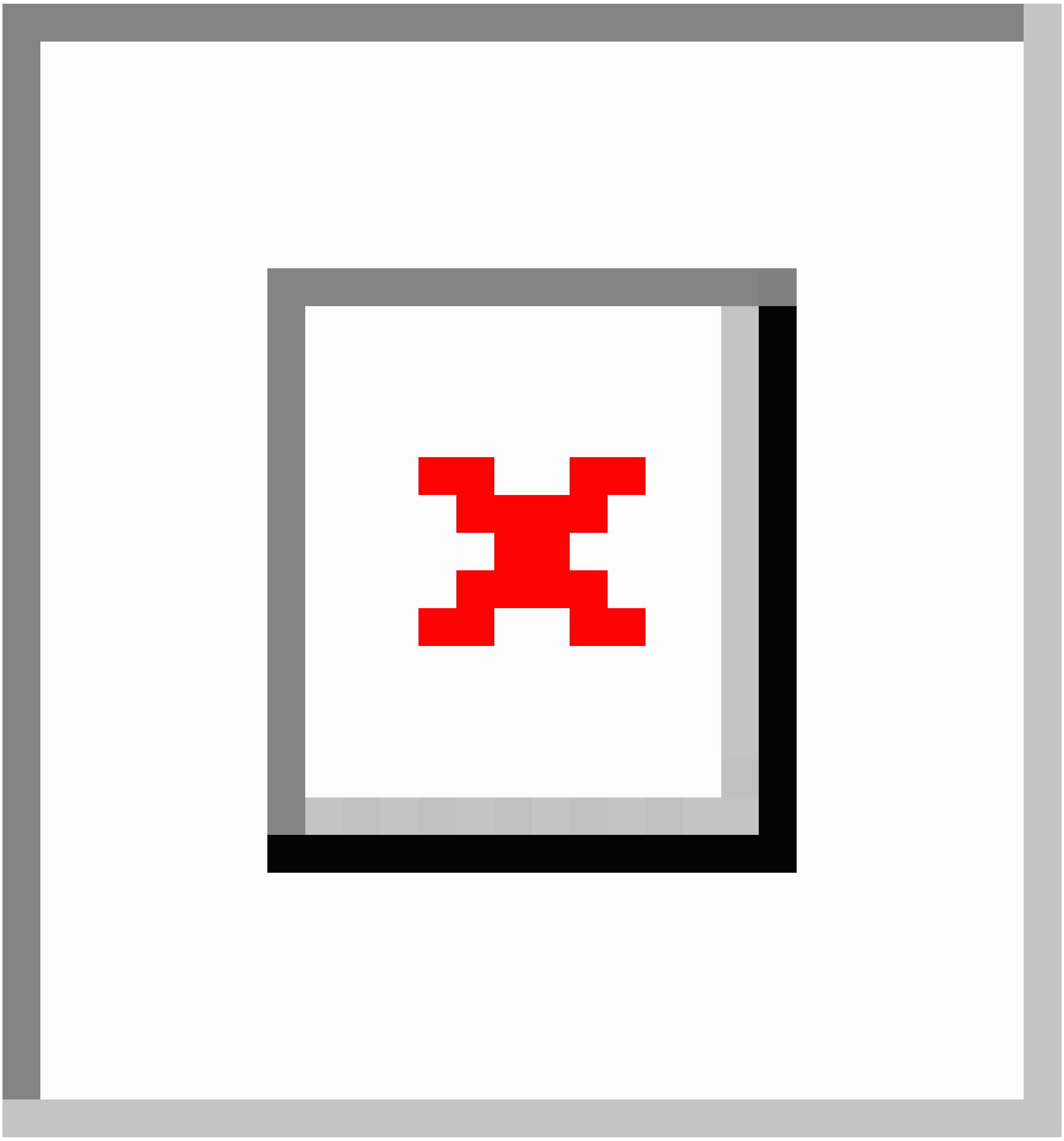
Once both students press "Begin station," the practice OSCE station starts, and the session begins. In this example, student A is practicing their skills as the "physician," and student B is providing feedback as the "patient." During this time, student A is only able to see the brief clinical prompt entailing the chief complaint. However, student B is able to view a more extensive patient history, along with behavioral cues, and the feedback checklist for items that student A should inquire about during the patient interview. While student A takes the history, student B is responsible for completing the checklist along with answering clinical questions, which are asked by student A, based on the history provided (Figure 3). At the end of the practice OSCE station, student B is responsible for completing the assessment checklist for student A in order to successfully save and submit the practice session.

Figure 1. Main log-in screen of the MonkeyJacket platform. OSCE: objective structured clinical examination.



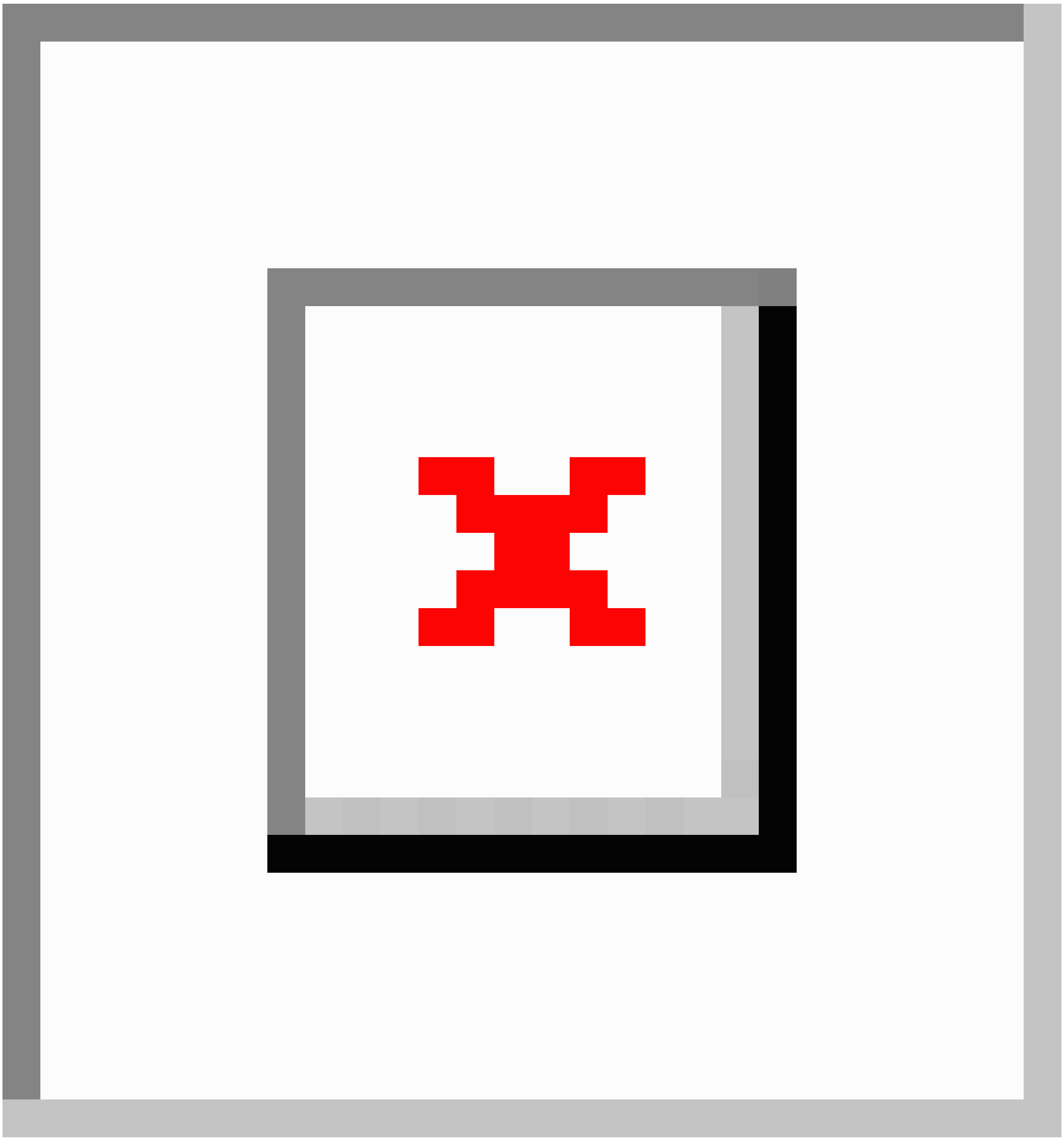
XSL•FO
RenderX

Figure 2. Example screen of the student in the role of the physician. The student physician is able to see the student patient on the left side of the screen and a blank clinical note that may be filled during the encounter.



XSL•FO
RenderX

Figure 3. Example of the MonkeyJacket platform screen as seen by the student in the role of the standardized patient. The case details are shown on the left, and the checklist assessment is shown on the right.



Feedback Checklists

Checklist items can be divided into two categories: (1) generic items and (2) items relevant to the presenting concern. Examples of general checklist items can be found in Textbox 1.

Relevant checklist items are those that are pertinent to the primary presenting concerns of the patient. For example, if the patient presents with shortness of breath, some relevant checklist items could include those listed in (Table 2).

At the end of all assessment checklists, the student is also asked to state the top 2 or 3 differential diagnoses based on the history presented. After stating the differential diagnoses, the student is asked for their top diagnosis. There are also other pertinent clinical questions that the student must answer. Examples of other clinical questions include questions about deciding on the most appropriate imaging modality, other diagnostic tests, and the initial management of the clinical presentation.

After assessment checklists are completed and submitted on the platform, a percentage score is calculated based on the total number of check marks received. The score is recorded and stored within the MonkeyJacket platform. Students are able to review all personal case attempts that they have completed within the platform. Additionally, audio files are also captured so that students can later review the session and reflect on not just their medical expert knowledge but also the soft skills of

XSL•FO
RenderX

communication and rapport building that they must demonstrate (Figure 4).

Textbox 1. Examples of general objective structured clinical examination checklist items.

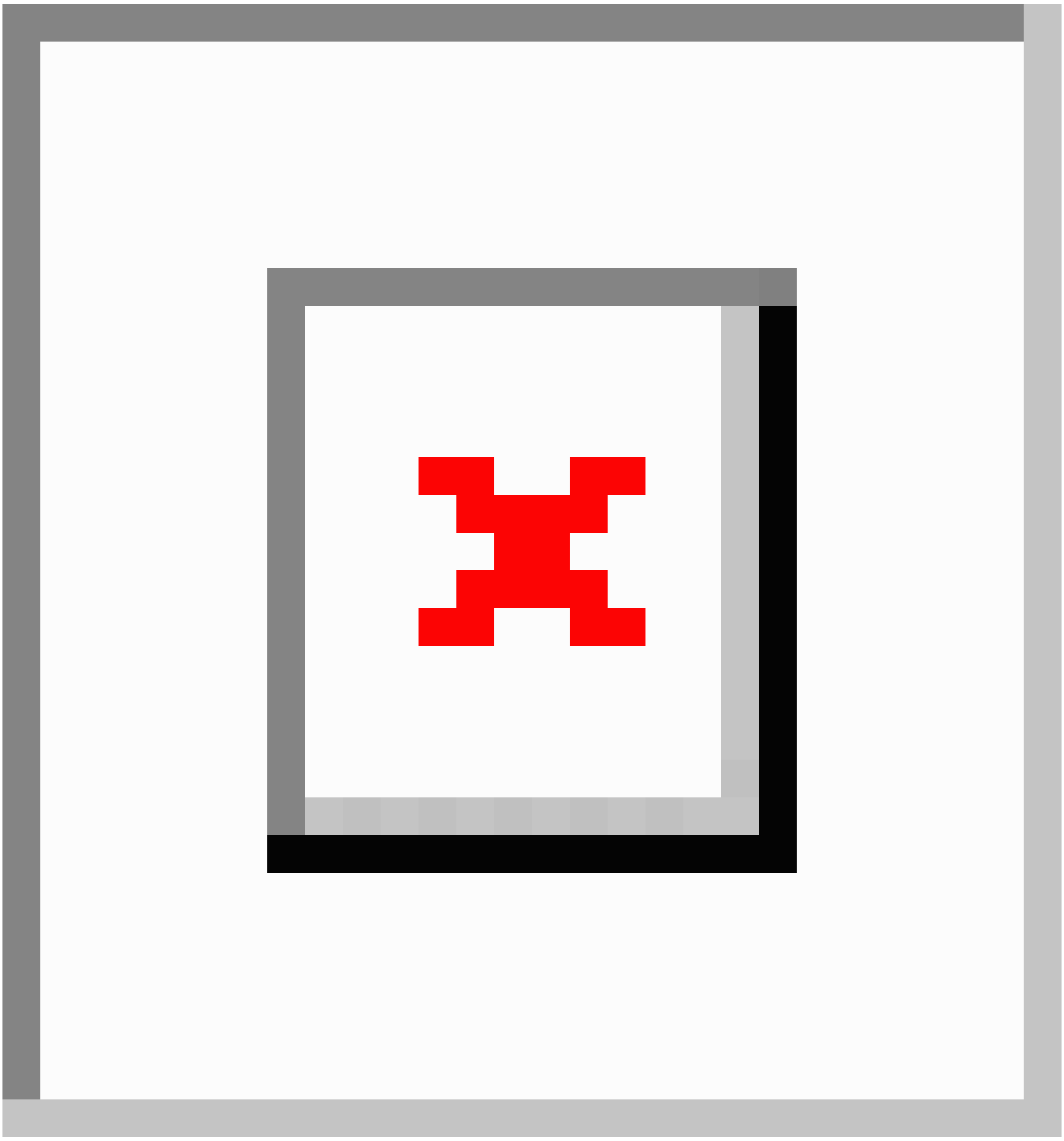
- Introducing self
- Confirming patient’s name and age
- Explaining reason for consult
- Building initial rapport
- Gaining consent
- Asking open-ended questions
- Asking about medications and allergies
- Exploring social history (including cigarettes, alcohol, recreational drugs, diet, occupation, and physical activity)
- Exploring and responding to ideas, concerns, and expectations
- Showing empathy
- Avoiding jargon
- Summarizing issues back to patient
- Global score
- Answering follow-up questions correctly

Table . Examples of relevant objective structured clinical examination checklist items, with the primary presenting concern being shortness of breath.

Assessment checklist items	Examples of what should be asked about
Asking qualifying questions about presenting symptoms	Onset, duration, site, character, severity, duration, and timeline of pain
Asking about relevant associated symptoms	Coughing, recent calf pain, palpitations, fever, and chest pain
Asking about recent illnesses and past medical history	Heart disease, stroke, diabetes, hypertension, etc
Asking about relevant family history	Heart disease among family members aged younger than 55 y, diabetes, high cholesterol, autoimmune disease, history of atopy, etc

XSL•FO
RenderX

Figure 4. Example of the review screen, through which students may access their scores, their clinical notes, comments from their peers, and an audio file of the encounter. SOAP: Subjective, Objective, Assessment, Plan.



Discussion

Overview

The MonkeyJacket application is a novel, innovative, and unique tool for medical students seeking additional practice regarding the development of clinical skills. The overarching goal of the MonkeyJacket application is to fill the gap that exists within medical education—a lack of scalable formative feedback for clinical skill development for learners. The MonkeyJacket application addresses this gap through the focus on peer feedback and the technological features built within the platform. Additionally, the application keeps track of participants' scores so that individuals may review the trend in and learn from their performance after practice sessions.

The biggest advantage of this platform is the potential for scalability it provides for medical learners. According to Medical Education Statistics 2020, there were 14,967 faculty members and 11,865 medical learners across Canadian medical schools by the beginning of 2020 [15]. On top of the clinical responsibilities of faculty members, they are also responsible for fulfilling teaching and academic requirements. As such, it is not feasible for faculty members to provide additional feedback to learners outside of the designated OSCE preparation time. The MonkeyJacket platform allows students to receive an abundance of feedback from peers, should they wish for

XSL•FO
RenderX

additional practice. The scalability of the platform also decreases the administrative load on medical schools, as students would have simple access to additional clinical skills feedback that does not require constant faculty supervision.

Another significant advantage of the MonkeyJacket application is the remote nature of the web-based platform. Traditionally, practice OSCE examinations have been conducted in person, often with a student's peer or friend. The utilization of the MonkeyJacket application is simple, in that it allows a student to share the link with anyone that has access to a computer and internet connection, thus allowing students to practice regardless of their geographical location. Moreover, medical students would be able to practice with students from other schools, thus promoting interinstitutional learning. A medical student residing in British Columbia could easily practice history-taking skills with a fellow student in Ontario, thus allowing both students to learn from each other and teach each other strategies that they have learned within their respective curricula. It is known that medical education institutions across Canada place emphasis on different areas of focus. For example, it was found that preclerkship pediatric clinical skills training greatly varied across the 17 Canadian medical schools, with 6 schools dedicating less than 7 hours and 8 schools dedicating over 10 hours—a total difference of 30% [16]. The development of a remote-based platform allows medical students to learn from their peers, who may have had more exposure within certain areas when compared to students' own training, thus enhancing their knowledge.

In addition to the remote nature of the application, it also poses a great advantage in terms of its accessibility with respect to cost. A significant barrier to finding accessible practice resources for medical students is the cost associated with purchasing resources. It was found that, on average, osteopathic medical students spend US $4129 on resources exclusively in preparation for their board examinations [17]. Although this finding is specific to medical students in the United States, where there are different board examinations, Canadian medical students are not exempt from such costs. Canadian medical graduates, on average, finish medical school with CAD $164,688 (US $846,612 as of the time of writing) of debt, including education-related and non–education-related expenses [18]. Although numerous companies offer preparation courses, these can vary in cost from a few hundred dollars to several thousands of dollars. Thus, costs associated with expensive preparation courses and resources can be a significant barrier for students seeking resources. The MonkeyJacket platform is completely open access and free of charge. For medical students looking to gain extra practice, the MonkeyJacket platform provides a simple and accessible option, with multiple opportunities for peer evaluation and progress tracking.

Limitations

To ensure that the MonkeyJacket web application was serving its intended population, relevant feedback from medical students and residents was taken into consideration when developing the functions and design of the web application. Nonetheless, there were some limitations to this study.

One limitation of this study is the sample size of students included in the feedback process. In this study, there were 5 medical students and 1 medical resident involved throughout the testing process. At the time of writing, the 6 participants have completed over 200 practice case scenarios via the MonkeyJacket platform. Future studies should include a larger sample size of participants in order to obtain more diverse feedback regarding the functionality and usability of the application.

Another limitation of this study is that all participants were from either Western University or McMaster University. This application originated from researchers based in Western University, and thus all students were recruited from the same institution for ease of organization and planning. Although this was advantageous, as the knowledge and OSCE skills were standardized among study participants, this can also reflect a lack of diversity in perspectives with respect to OSCE skills.

Lastly, traditional OSCE examinations are extensive, in that they also evaluate a candidate's ability to perform relevant physical examination and procedural skills in response to a primary patient concern. Given the web-based nature of the MonkeyJacket platform, it was not possible to integrate such assessments. However, one way to assess a candidate's knowledge regarding relevant physical examination skills is to add it to the checklist and ensure that the candidate knows the rationale for why certain physical examination components would be used.

Future Directions

In the future, the MonkeyJacket application will be preparing for extensive nationwide deployment across Canadian medical institutions. Through partnership with major Canadian medical student groups, the application will be disseminated for widespread use. This will allow us to collect a vast amount of quality improvement feedback. Ideally, we will be able to test if the use of the application leads to improved medical examination scores.

At the time of writing, the cases included within the platform are tailored toward scenarios that can help medical learners, who will become competent resident physicians, develop clinical skills. The expansion of the application in the future can include more specialized cases for specific residency subspecialties. In addition, MonkeyJacket is useful not only for Canadian medical students but also for medical trainees globally, as clinical skills examinations are part of many international medical education programs. This can be explored in the future, once the application is successfully deployed in Canada.

Conclusions

The MonkeyJacket OSCE tool is a comprehensive and accessible learning resource for medical learners. This innovative tool offers medical learners a solution that addresses the lack of practice tools and formative feedback within the realm of clinical skill development. As medical students proceed through their training, OSCEs remain an integral component of assessments ensuring that learners are demonstrating required competencies for safely practicing medicine upon graduation. The development of comprehensive and accessible OSCE

XSL•FO
RenderX

practice tools with built-in evaluations eases the stress associated with preparation for clinical examinations and promotes a more competent medical workforce, with the latter benefiting the most important stakeholders in medicine—the patients.

Acknowledgments

We are grateful for funding through Canada's Department of National Defence Innovation for Defence Excellence and Security (IDEaS) COVID-19 Challenge. No staff from the Department of National Defence participated directly in this research.

Authors' Contributions

CS oversaw the direction of the publication and was the senior author and organizer of the project. AA and FF wrote the manuscript. AA created the figures. EA provided a summary of Medical Council of Canada (MCC) objectives. AA, FF, EA, TB, AK, and EW conducted the practice OSCE sessions. All authors reviewed the final manuscript.

Conflicts of Interest

The MonkeyJacket application is owned by GoodLabs Studio. CS and TL are part of the GoodLabs Studio development team; however, there is no conflict of interest that affected this work.

References

1. Quota and applications by discipline. Canadian Resident Matching Service (CaRMS). URL: https://www.carms.ca/data-reports/r1-data-reports/r-1-match-interactive-data [accessed 2023-01-10]
2. Almarzooq ZI, Lopes M, Kochar A. Virtual learning during the COVID-19 pandemic: a disruptive technology in graduate medical education. J Am Coll Cardiol 2020 May 26;75(20):2635-2638. [doi: 10.1016/j.jacc.2020.04.015] [Medline: 32304797]
3. NAC Examination. Medical Council of Canada. URL: https://mcc.ca/examinations-assessments/nac-examination/ [accessed 2023-01-03]
4. Lerchenfeldt S, Mi M, Eng M. The utilization of peer feedback during collaborative learning in undergraduate medical education: a systematic review. BMC Med Educ 2019 Aug 23;19(1):321. [doi: 10.1186/s12909-019-1755-z] [Medline: 31443705]
5. Ismail SM, Rahul DR, Patra I, Rezvani E. Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. Language Testing in Asia 2022;12(1):40. [doi: 10.1186/s40468-022-00191-4]
6. Arrogante O, González-Romero GM, López-Torre EM, Carrión-García L, Polo A. Comparing formative and summative simulation-based assessment in undergraduate nursing students: nursing competency acquisition and clinical simulation satisfaction. BMC Nurs 2021 Jun 8;20(1):92. [doi: 10.1186/s12912-021-00614-2] [Medline: 34103020]
7. Monkey Jacket: an OSCE training ground with your medical buddies. Monkey Jacket. URL: https://www.monkeyjacket.app/login [accessed 2024-06-14]
8. Jefferies A, Simmons B, Tabak D, et al. Using an objective structured clinical examination (OSCE) to assess multiple physician competencies in postgraduate training. Med Teach 2007 Mar;29(2-3):183-191. [doi: 10.1080/01421590701302290] [Medline: 17701631]
9. Varkey P, Natt N, Lesnick T, Downing S, Yudkowsky R. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. Acad Med 2008 Aug;83(8):775-780. [doi: 10.1097/ACM.0b013e31817ec873] [Medline: 18667895]
10. Frohna JG, Gruppen LD, Fliegel JE, Mangrulkar RS. Development of an evaluation of medical student competence in evidence-based medicine using a computer-based OSCE station. Teach Learn Med 2006;18(3):267-272. [doi: 10.1207/s15328015tlm1803_13] [Medline: 16776616]
11. Hurley KF. OSCE and Clinical Skills Handbook: Elsevier/Saunders; 2011.
12. OSCE stations. OSCE Stations. URL: http://osce-stations.blogspot.com [accessed 2022-03-08]
13. Primary care clerkship practice exams. UW Family Medicine & Community Health. URL: https://www.fammed.wisc.edu/files/webfm-uploads/documents/med-student/pcc/practice-osce-scenarios.pdf [accessed 2022-03-04]
14. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans: Government of Canada; 2022.
15. The Association of Faculties of Medicine of Canada. Canadian Medical Education Statistics 2020: The Association of Faculties of Medicine of Canada; 2020, Vol. 42. URL: https://www.afmc.ca/wp-content/uploads/2022/10/CMES2020-Complete_EN.pdf [accessed 2024-06-07]

XSL•FO
RenderX

16. Hudson A, Mclaughlin R, Miller S, Holland J, Blake K. Canadian medical schools' preclerkship paediatric clinical skills curricula: how can we improve? Paediatr Child Health 2019 Oct 6;25(8):505-510. [doi: 10.1093/pch/pxz106] [Medline: 33354259]
17. Bhatnagar V, Diaz SR, Bucur PA. The cost of board examination and preparation: an overlooked factor in medical student debt. Cureus 2019 Mar 1;11(3):e4168. [doi: 10.7759/cureus.4168] [Medline: 31086753]
18. Health, safety + well-being. University of Alberta, Faculty of Medicine & Dentistry. URL: https://www.ualberta.ca/medicine/programs/md/student-resources/health-safety-well-being/index.html [accessed 2023-01-04]

Abbreviations

CanMED: communicator, collaborator, leader, health advocate, scholar, professional, and medical expert
e-OSCE: electronic objective structured clinical examination
MCC: Medical Council of Canada
OSCE: objective structured clinical examination

Edited by TDA Cardoso; submitted 28.03.23; peer-reviewed by H Alshawaf, J Waechter; revised version received 16.05.24; accepted 24.05.24; published 20.06.24.

Please cite as:
Aqib A, Fareez F, Assadpour E, Babar T, Kokavec A, Wang E, Lo T, Lam JP, Smith C
Development of a Novel Web-Based Tool to Enhance Clinical Skills in Medical Education
JMIR Med Educ 2024;10:e47438
URL: https://mededu.jmir.org/2024/1/e47438
doi:10.2196/47438

© Ayma Aqib, Faiha Fareez, Elnaz Assadpour, Tubba Babar, Andrew Kokavec, Edward Wang, Thomas Lo, Jean-Paul Lam, Christopher Smith. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.6.2024. 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.

XSL•FO
RenderX

Data-Driven Fundraising: Strategic Plan for Medical Education

Alireza Jalali[1,*], MD; Jacline Nyman[2,*], DBA; Ouida Loeffelholz[3,*], BA; Chantelle Courtney[1,*], LLB

1

2

3

*all authors contributed equally

Corresponding Author:
Alireza Jalali, MD

Abstract

Higher education institutions, including medical schools, increasingly rely on fundraising to bridge funding gaps and support their missions. This paper presents a viewpoint on data-driven strategies in fundraising, outlining a 4-step approach for effective planning while considering ethical implications. It outlines a 4-step approach to creating an effective, end-to-end, data-driven fundraising plan, emphasizing the crucial stages of data collection, data analysis, goal establishment, and targeted strategy formulation. By leveraging internal and external data, schools can create tailored outreach initiatives that resonate with potential donors. However, the fundraising process must be grounded in ethical considerations. Ethical challenges, particularly in fundraising with grateful medical patients, necessitate transparent and honest practices prioritizing donors' and beneficiaries' rights and safeguarding public trust. This paper presents a viewpoint on the critical role of data-driven strategies in fundraising for medical education. It emphasizes integrating comprehensive data analysis with ethical considerations to enhance fundraising efforts in medical schools. By integrating data analytics with fundraising best practices and ensuring ethical practice, medical institutions can ensure financial support and foster enduring, trust-based relationships with their donor communities.

(JMIR Med Educ 2024;10:e53624) doi:10.2196/53624

KEYWORDS

fundraising; philanthropy; crowdfunding; funding; charity; higher education; university; medical education; educators; advancement; data analytics; ethics; ethical; education; medical school; school; support; financial; community

Introduction

Higher education institutions play a crucial role in shaping society's future by developing new ideas, advancing knowledge, and preparing future leaders. However, to fulfill this role, institutions need financial resources. Digital medicine, a rapidly evolving field of technology and health care, is revolutionizing how fundraising is conducted toward funding medical research and public health initiatives. Data, in conjunction with artificial intelligence, are transforming health care and paving the way for personalized medicine. By leveraging cutting-edge digital tools in data analysis, organizations can now gather, process, and interpret vast amounts of health-related data more efficiently [1]. This data-driven approach enables a deeper understanding of disease patterns, patient needs, and effective treatment strategies, facilitating more targeted and impactful fundraising efforts (by segmenting fundraising strategies in keeping with these new patient data streams). Furthermore, innovative fundraising platforms are emerging, harnessing the power of artificial intelligence and algorithmic data analysis that draws from and informs social media, mobile technology, and web-based crowdfunding [2]. These platforms expand the reach of fundraising campaigns and allow for real-time engagement with donors, aimed at enhancing transparency and trust. By integrating these digital advancements, fundraising with digital medicine is becoming more efficient and more personalized, aligning donor interests with specific health care projects and research endeavors.

Fundraising is vital to higher education because it helps organizations acquire some of the financial resources needed to achieve their mission, vision, and strategic goals. Fundraising can help institutions fund new academic programs, foster research and student learning, advance knowledge, and build links with the community [3,4]. Higher education and medical schools, in particular, are expensive [5], and many students require financial assistance to attend college or university. Fundraising can help institutions provide financial support for scholarships, grants, and other forms of financial aid to students who need it most. The paper is aimed at a novice audience, providing a basic framework for data-driven fundraising. It is set in a Canadian context, where resources such as wealth indicators are less prevalent than those in the United States.

Proceeds generated from fundraising endeavors can enhance and modernize campus infrastructure, encompassing vital domains such as classrooms, laboratories, libraries, dormitories, and athletic amenities. Increased financial resources from fundraising activities can be strategically allocated to recruiting and retaining distinguished faculty and researchers, establishing

XSL•FO
RenderX

scholarships and awards for students, and supporting community outreach activities.

To provide a broader context on fund development terminology, it is crucial to differentiate and link fund development, fundraising, and advancement:

- Fund development: This term encompasses the overarching process of creating and nurturing relationships that foster an organization's growth. It includes strategic planning, donor engagement, and stewardship, going beyond mere transactional activities to build a sustainable funding base.
- Fundraising: This is a subset of fund development focused on securing funds. It may involve direct or web-based solicitation, single-purpose or multiyear comprehensive campaigns, events, grant writing, and sponsorships. Fundraising is the operational action that stems from the broader fund development strategy.
- Advancement: In academic and nonprofit sectors, advancement refers to efforts that advance the institution's mission through fundraising, stakeholder or alumni relations, public affairs, and marketing.

Advancement offices play a pivotal role in the ecosystem of fundraising and donor engagement within institutions. They are responsible for leading an ongoing dialogue between the faculty and the prospective or current donors. Key to advancement work is seeking to understand the impact that the donors want to make through their philanthropic activities and connecting those goals with approved fundraising priorities when possible. Furthermore, advancement offices are instrumental in identifying and capitalizing on various funding opportunities, ensuring that the institution's projects and initiatives are adequately supported. They also provide crucial support for single-purpose campaigns, offering strategic guidance and resources to shape success. By doing so, these offices foster immediate financial support and build long-term relationships between the institution and its benefactors.

Universities increasingly establish "advancement" units integrating fund development, fundraising, alumni, and community outreach activities. The term "fund development" is preferred, as it emphasizes cultivating and maintaining long-term relationships and support. This approach is integral to fostering enduring commitments rather than transactional contributions. Within universities, fundraising can occur at various levels, including within smaller units, departments, or faculties, as well as at the university-wide level. However, consulting and collaborating with advancement or fund development professionals are crucial before initiating fundraising efforts. This ensures that fundraising activities are in harmony with the broader institutional plans and goals, as well as coordinated among the different units and subunits to ensure maximum impact.

Professionals in advancement offices play a pivotal role in managing potential donor relationships, understanding their interests, and aiding units in crafting strategic fundraising plans and proposals. They are skilled in identifying donor interests in the institution and its programs, often starting with smaller initiatives to support donor engagement. These experts often keep a comprehensive list of funding opportunities, including scholarships, bursaries, and specific project or equipment funding. They may support small-scale or single-purpose campaigns for various university needs, such as funding a simulation center or a theater. Advancement professionals also know about the institution's approach to endowment funds versus expendable or "flow-through" funds. Endowments are typically maintained in perpetuity, providing a unit with a percentage of the total endowed principal plus investment, whereas expendable funds are designated for use over a predetermined period until depleted. Advancement professionals' expertise is vital in aligning donor intentions with the vision, strategic plan, and priorities of the university. Furthermore, these fundraising professionals are known to play the role of networker, negotiator, and knowledge broker [6].

The aim of this paper is to explore a data-driven approach to fundraising in medical education, integrating comprehensive data analysis with ethical considerations.

A Fundraising Road Map

A fundraising plan or road map is aligned with an institution's strategic plan and is an essential component of a successful fundraising effort. It is a written document that sets Specific, Measurable, Achievable, Relevant, and Time-bound (SMART) goals [7] for the fundraising campaign and communicates these goals and objectives to internal and external stakeholders, including faculty, staff, donors, prospective donors, and the general public. The road map helps determine the resources (human, financial, material, etc) needed to successfully implement the campaign, including academic leadership, advancement staff, volunteers, campaign materials, and so on, as well as how best to allocate these resources to maximize the return on investment (ROI).

Establishing SMART objectives is crucial for the success of fundraising campaigns. These objectives provide a clear road map, ensuring that every aspect of the fundraising effort is intentional and efficient. A well-structured fundraising plan, guided by these SMART objectives, allows for a strategic approach, setting clear milestones and measurable goals. This approach facilitates focused efforts, efficient resource allocation, and the ability to track progress effectively. Such meticulous planning is essential for aligning the fundraising activities with the institution's overarching goals and ensuring the optimal use of resources for maximum impact. Although setting SMART objectives is foundational, the consistent application of workflows, standards, and daily execution ultimately determines the campaign's success.

The plan also aims to recognize the organization's most generous philanthropic contributors. In particular, it helps ensure that prospective supporters with the necessary financial capacity are engaged, including alumni and friends, corporations, foundations, and government agencies, and that they share an interest in furthering the organization's mission, strategic goals, and fundraising priorities. The road map often includes targeted cultivation, solicitation, and stewardship strategies by fundraising program, unit, geographic region, and source of funds (eg, cash, gifts of publicly traded securities) to ensure maximum ROI [6].

XSL•FO
RenderX

By establishing a framework that includes timelines for evaluating the campaign's success, the fundraising plan helps keep the campaign on track by checking the progress made against goals set on a monthly, quarterly, and annual basis. The plan also includes key performance indicators to measure each fundraising program and solicitor's success. Some key performance indicators may consist of the number and quality of meetings undertaken with potential donors, gift proposals submitted, gifts received, total funds raised, number of new donors, and gift size; ultimately, they impact the organization.

Fundraising activities should align strategically with an institution's broader goals and plans. This alignment ensures that fundraising efforts secure the necessary funds and support and advance the institution's mission and objectives. The fundraising plan is closely aligned with the institution's strategic priorities, ensuring that fundraising supports key teaching, learning, and research goals, establishing a long-term focus instead of a short-term financial solution. Collaborating with advancement and fund development professionals is crucial in this process. These experts bring invaluable insights and strategies that help identify and engage with potential donors whose interests and values resonate with the institution's goals. Such collaboration ensures that fundraising activities are not only successful in the short term but also contribute to the long-term growth and success of the institution.

Using Data in Fundraising

Data play a critical role in developing a successful fundraising plan. Data provide valuable insights into the past, current, and future donor areas of philanthropic interest and a measurement of the effectiveness of past fundraising campaigns, strategies, programs, and methods. Once analyzed, data can guide the fundraising efforts toward opportunities for growth and help offer the indicators of potential risks or barriers to success. Publicly available data can help drive donor strategies in keeping with their consumption interests. Algorithms are used to make sense of these masses of data and to help predict the greatest potential fundraising strategies and potential donors.

Medical schools (or any faculty) can develop a data-driven fundraising plan based on information and insights aligned with achieving their mission and fundraising goals. Using data to inform strategy and decision-making, medical schools can develop more effective and targeted fundraising plans that deliver better results and build stronger relationships with donors, prospective donors, and the intended beneficiaries [8].

Establishing a data-centered culture in fundraising involves a strategic shift toward using data analytics to guide decision-making processes. This approach emphasizes collecting, analyzing, and leveraging data to understand donor behaviors, preferences, and trends. By adopting a data-centered culture, institutions can make informed decisions about whom to approach, when, and how, thereby increasing the effectiveness and efficiency of their fundraising efforts. Guidance on this aspect includes investing in the right tools and training for data analysis, cultivating a mindset among staff that values data-driven insights, and continuously refining strategies based on data feedback.

Making the data actionable is crucial; it should directly inform and shape your fundraising strategies. MacLaughlin [9] highlights the importance of using data strategically in fundraising. Key points include prioritizing data quality, using data for effective segmentation and personalized communication, and fostering a culture of experimentation. Data should inform strategies, help identify potential donors, and predict giving patterns. Investment in data skills and tools is crucial for effective analysis. Data-driven storytelling can demonstrate donation impact, while good data governance ensures responsible data management. A growth mindset encourages learning from successes and failures. Understanding the difference between business intelligence (analyzing historical data) and predictive analysis (using data to predict future outcomes) can also enhance fundraising strategies. While this paper focuses on business intelligence, which involves diagnostic analytics, predictive analysis involves forecasting future outcomes and represents a more advanced stage of data-driven fundraising.

The 4 Pillars of Data-Driven Fundraising

There are four key steps to developing an end-to-end data-driven fundraising plan: (1) data collection; (2) data analysis; (3) establishing fundraising goals and objectives; and (4) formulating targeted fundraising strategies.

Data Collection

The first step for medical schools in developing a data-driven fundraising plan is collecting quantitative and qualitative data. External and internal research, both from primary and secondary sources, are critical. External information about current and potential donors is often publicly available, including secondary data sets such as demographic and career information, professional networks, and affiliations; giving history; and philanthropic interests. Institutional primary donor data can be analyzed to determine linkages to the medical school; affinity; and giving patterns, including gift amounts, giving frequency, past gift designations, and retention rates. These primary data sets can often be used in algorithmic analysis (in combination with secondary data available in the marketplace) to deepen our understanding of the donor landscape.

Additional information about donor experience, potential future giving interests, and preferred methods of communication can be gleaned from various data collection tools, such as surveys, focus groups, personal meetings, and social media. Internal and external data sets, as well as primary and secondary data sets, can be used together and separately to help organizations better understand donors' and prospective donors' interests, as well as the financial capacity and emotional links to the organization, thereby enabling medical schools to develop targeted outreach strategies that better meet donors' needs and preferences.

In addition to collecting data on donors, it is also important to have access to information about an institution's past fundraising campaigns, comparative data about other professional schools and peer organizations, and industry trends. These data provide benchmarks and guideposts for planning, measurement, and evaluation.

XSL•FO
RenderX

Data Analysis

Once data on donors, past campaigns, and industry trends have been collected, they can be analyzed to uncover patterns and trends that provide insights for future fundraising goals, objectives, strategies, and resource allocation. For example, by examining the giving patterns of alumni, friends, retirees, corporations, and foundations, analysts can pinpoint the organization's foremost donors in terms of total contributions. Data analysis can reveal which donors have made the most substantial gifts, including those with deep ties to the organization, based on their motivations, giving history, donation frequency, level of involvement, and satisfaction. Based on these data points, predictive analysis can help guide donor stewardship and retention strategies, as well as determine which donors are most likely to increase their contributions, both in terms of their capability and willingness to give more. Analysis allows fundraising campaign planners to highlight individuals and organizations currently not donating but with significant potential to support crucial future initiatives.

Data from past fundraising campaigns can be analyzed to determine what programs (annual, major, principal, and planned gifts) and strategies were the most successful against their respective goals. Analysis helps us understand performance trends (high and low) and which areas have the most potential for future growth. Similarly, fundraising tactics, such as direct mail, call centers, electronic solicitations, personal approach, social media, events, and so on, can be evaluated to see which methods have the greatest ROI and growth potential. Understanding that donor-acquisition methods are generally costlier than those used to upgrade existing donors, steps in donor acquisition, retention, and upgrading can be further analyzed to understand and guide the fundraiser's strategic planning and resource deployment. Data can also help determine whether some geographical regions, sources of revenue, fund designations, cohorts, and programs will provide the most predictable ROI for scarce faculty resources.

An analysis of data on current and prospective donors; past campaign successes and failures; peer organizations; and industry, socioeconomic, and technological trends are some of the factors to consider when determining where to allocate future resources for maximizing fundraising success. A thoughtful fundraising strategy can also further equity, diversity, and inclusion initiatives, inviting a broad range of Canadians and international citizens to invest in specific educational missions.

Establishing Fundraising Goals and Objectives

To establish an overall fundraising campaign with SMART goals, medical schools should start by determining the cost of delivering programs and services as per their strategic plan and expected revenue from nonfundraising sources. Once the revenue gap is determined, fundraising campaign planners can then examine the following: (1) past fundraising campaign results; (2) financial capacity and affinity as well as inclination of the faculty's donor pool, notably high–net worth individuals with the greatest capacity to give; (3) opportunities for growth based on institutional, industry, technological, and global trends; (4) social, economic, and political climate; (5) organizational reputation; and (6) potential risks. These can be used to discover a realistic overall fundraising goal depending on the time frame established.

However, of utmost importance is establishing a compelling case in support of a given program or project (gift designation). Donors give because they are deeply interested in supporting a given cause and because of the impact their gift will have on the beneficiaries. Research on high–net worth donors has concluded that donors give because they want to make transformational change, have a societal impact, and leave a personal legacy [10]. So, as we analyze the data to focus on our most promising potential donors and the most efficient fundraising strategies, a compelling case for support remains crucial to the success of any fundraising program.

In addition to the overarching fundraising goal, there are other subgoals that organizations can establish to align with the overarching goal. For example, the subgoals may be to increase the total number of donors to a medical school by a specific percentage (ie, the percentage of alumni who give back to their alma mater), as well as increase the average gift size per donor, the number of annual campaign donors upgraded to major gifts, the number of new monthly donors, and the number of new multiyear pledges. Subgoals are chosen from the data analysis as these data points align with increasing overarching fundraising success.

Subgoals can also be established by year or decade of graduation; region; fundraising program (eg, an annual campaign, major gifts, principal gifts, and planned giving); source (eg, alumni, friends, foundations, corporations, and other organizations); gift designation (eg, case for supporting students, research, infrastructure, and community engagement); and individual fundraiser or solicitor, team, and unit. Donor communication, engagement, and stewardship subgoals can also be set. Aligning subgoals and measurements with overarching goals associated with success metrics supports an integrated strategy.

Developing Targeted Fundraising Strategies

Once SMART fundraising goals have been established, medical schools can use data to develop targeted strategies and related activities, budgets, and timelines to reach their organizational goals and objectives.

For example, if the goal is to increase revenue to a specific fund designation or case for support, and the data reveal specific donor groups to be most supportive of these, then targeted communications, outreach, and solicitation strategies can be developed to better reach these donor groups, thereby increasing the chances of fundraising success.

For example, suppose past giving trends indicate that alumni are most likely to support scholarship funds. In that case, medical schools may focus their fundraising efforts and activities on reaching out to their alumni via various means of communication, highlighting the impact that scholarships have on student access and achievement, and inviting alumni to make or upgrade their commitment to scholarship funds. In other words, the fundraising ROI becomes greater by reinforcing the compelling case for support with donor groups most likely to respond.

XSL•FO
RenderX

In another example, if the data show that a certain percentage of a medical school's donors is in the 60+ years age range, the faculty can develop focused, planned giving strategies, communicating how legacy gifts (eg, deferred giving in the form of a bequest or life insurance) contribute to the long-term sustainability of the medical school's vision. If the data show that a significant number of high–net worth donors live in a particular geographical area, regional engagement, cultivation, and solicitation strategies can be developed accordingly better to meet the needs and interests of this group. This list of examples shows how targeted approaches can be developed through data analysis and planning.

Ethical Practice in Fundraising

Ethics in higher education fundraising are about conducting fundraising efforts in a manner that is consistent with ethical values and principles and that fosters trust and accountability. In a normative context, it can be said that "fundraising is ethical when it promotes and protects trust in fundraising and unethical when it harms trust." [11] Medical schools must ensure that fundraising practices are transparent, honest, and responsible, and that all parties involved in the fundraising process are treated with respect.

Ethical practices in fundraising are of utmost importance, particularly in maintaining trust and integrity in the relationship between institutions and donors. A critical aspect of this is developing and adhering to clear gift acceptance policies. These policies help ensure that all donations are aligned with the institution's mission and ethical standards. In addition, it is crucial to manage the extent of donor influence. While donor engagement is important, institutions must maintain autonomy and ensure that donations do not compromise their values or objectives. Upholding these ethical standards fosters a transparent and trustworthy environment and protects the institution's reputation and long-term sustainability.

Many fundraisers in Canada are guided by the Association of Fundraising Professionals Code of Ethical Principles, which encourages them to practice their profession with integrity, honesty, and truthfulness, as well as safeguard public trust [12]. The Association of Fundraising Professionals Donor Bill of Rights highlights the principle of philanthropy rooted in voluntary action for communal benefit, emphasizing transparency, trust, and responsible stewardship in nonprofit engagements. It outlines donors' rights, including being informed of the organization's mission, usage of donations, board identity, access to financial statements, assurance of gift use as intended, respectful and confidential handling of donation information, professional interactions, clarity on the status of solicitors (volunteers or employees), option to opt out from mailing lists, and the freedom to inquire and receive honest responses when donating [13].

The authors make the case that "fundraisers are unlike commercial marketers in that they arguably have two key constituencies—their donors and their beneficiaries through a transfer rather than an exchange" [11]; therefore, there must be balance in protecting both donors and beneficiaries and not applying ethics to one at the expense of another.

Fundraisers should be transparent about the intended purpose of the funds being raised, how the funds will be used, and any potential benefits or risks associated with donating. Furthermore, they should treat all potential donors and beneficiaries fairly and equitably, without discrimination or favoritism. Donors' data, including personal information and donation history, should be confidential and not shared with third parties without consent [8]. Moreover, fundraisers should avoid situations that could create conflicts of interest for themselves, their organizations, the donors, and the beneficiaries.

One area of importance is grateful patient fundraising (GPFR), a unique approach to charitable giving where patients, often touched deeply by the care they have received, choose to support their health care institutions financially. Although GPFR is widespread, it raises ethical issues for patients, physicians, development professionals, and institutions. In 2004, the American Medical Association Council on Ethical and Judicial Affairs acknowledged that philanthropic donations are essential to maintaining state-of-the-art medical facilities and conducting research. However, they discouraged physicians from directly soliciting from their patients, especially during a clinical encounter [14]. More recently, Collins et al [13] made a list of recommendations and stated, among others, that GPFR discussions must be avoided when patients are clinically vulnerable. Philanthropy does not justify a level of medical care not available to other patients, and institutions should recognize and take measures to mitigate the ethical risks inherent in wealth screening [15].

Educational institutions are also responsible for using donated funds wisely and effectively and ensuring that the intended purposes of the donations are fulfilled. Institutions are urged to have a fundraising committee of educators, physicians, and fundraisers to reduce safety concerns and prevent fraudulent behavior. It is crucial to establish clear gift acceptance policies and set limitations on donor influence to maintain transparency and ethical fundraising practices. Fundraisers are required to abide by the established rules of their fundraising organization [16].

Conclusions

Developing a fundraising plan or road map is crucial for medical education institutions to achieve their institutional vision and goals; allocate resources effectively; and raise the funds they need to make up for budget shortfalls, remain competitive, and transform their institutions. This is a partnership among the advancement office, practitioner fundraisers, and academic leadership.

A fundraising plan with clear goals, objectives, strategies, action plans, and timelines helps build stronger relationships with existing and prospective donors and volunteers by providing them with a clear understanding of the institution's needs and goals, the institution's plan to achieve these goals, and how their gifts can make a difference.

Data collection and analysis are essential for establishing SMART fundraising goals and developing strategies to yield the greatest results. By having a well-designed and data-driven

XSL•FO
RenderX

fund development plan, medical schools can ensure that they have the resources to support their students, research, and mission in the short and long term.

Acknowledgments

This paper has been enhanced not only through the use of generative AI for synthesizing insights from peer reviews but also for grammatical corrections and proofreading. The AI technology aided in refining the language and structure, ensuring clarity and coherence in the presentation of our ideas.

Conflicts of Interest

None declared.

References

1. Freitas AT. Data-driven approaches in healthcare: challenges and emerging trends. In: Sousa Antunes H, Freitas PM, Oliveira AL, Martins Pereira C, Vaz de Sequeira E, Barreto Xavier L, editors. Multidisciplinary Perspectives on Artificial Intelligence and the Law: Springer, Cham; 2024. [doi: 10.1007/978-3-031-41264-6]
2. Mora-Cruz A, Palos-Sanchez PR. Crowdfunding platforms: a systematic literature review and a bibliometric analysis. Int Entrep Manag J 2023 Sep;19(3):1257-1288. [doi: 10.1007/s11365-023-00856-3]
3. Ortiz RA, Witte S, Gouw A, et al. Engaging a community for rare genetic disease: best practices and education from individual crowdfunding campaigns. Interact J Med Res 2018 Feb 5;7(1):e3. [doi: 10.2196/ijmr.7176] [Medline: 29402763]
4. Thelin JR, Trollinger RW. Philanthropy and American Higher Education: Palgrave Macmillan; 2014.
5. Walsh K. Why is medical education so expensive? J Biomed Res 2014 Jul;28(4):326-327. [doi: 10.7555/JBR.28.20140040] [Medline: 25050117]
6. Jalali A, Nyman JA, Hamelin-Mitchell E. Fundraising in education: road map to involving medical educators in fundraising. JMIR Med Educ 2022 Apr 5;8(2):e32597. [doi: 10.2196/32597] [Medline: 35380542]
7. Doran GT. There's a S.M.A.R.T way to write management's goals and objectives. Manage Rev 1981;70:35-36.
8. Birkholz J. Fundraising Analytics: Using Data to Guide Strategy: John Wiley & Sons; 2008.
9. MacLaughlin S. Data-Driven Nonprofits: The Saltire Press; 2016.
10. Nyman J, Pilbeam C, Baines P, Maklan S. Identifying the roles of university fundraisers in securing transformational gifts: lessons from Canada. Stud Higher Educ 2018 Jul 3;43(7):1227-1240. [doi: 10.1080/03075079.2016.1242565]
11. MacQuillin I. Normative fundraising ethics: a review of the field. J Philanthr Mark 2023 Nov;28(4):e1740. [doi: 10.1002/nvsm.1740]
12. The donor bill of rights. Association of Fundraising Professionals. URL: https://afpglobal.org/donor-bill-rights [accessed 2023-10-10]
13. Collins ME, Rum S, Wheeler J, et al. Ethical issues and recommendations in grateful patient fundraising and philanthropy. Acad Med 2018 Nov;93(11):1631-1637. [doi: 10.1097/ACM.0000000000002365] [Medline: 30024472]
14. Soliciting charitable contributions from patients. American Medical Association Code of Medical Ethics. URL: https://code-medical-ethics.ama-assn.org/ethics-opinions/soliciting-charitable-contributions-patients [accessed 2023-10-10]
15. Prokopetz JJZ, Lehmann LS. Physicians as fundraisers: medical philanthropy and the doctor-patient relationship. PLoS Med 2014 Feb;11(2):e1001600. [doi: 10.1371/journal.pmed.1001600] [Medline: 24523665]
16. Caboni TC. The normative structure of college and university fundraising behaviors. J Higher Educ 2010 May;81(3):339-365. [doi: 10.1080/00221546.2010.11779056]

Abbreviations

GPFR: grateful patient fundraising
ROI: return on investment
SMART: Specific, Measurable, Achievable, Relevant, and Time-bound

XSL•FO
RenderX

Edited by TDA Cardoso; submitted 12.10.23; peer-reviewed by H Ali, J Lockyer, S Hoscheit; revised version received 01.03.24; accepted 21.05.24; published 22.07.24.

Please cite as:
Jalali A, Nyman J, Loeffelholz O, Courtney C
Data-Driven Fundraising: Strategic Plan for Medical Education
JMIR Med Educ 2024;10:e53624
URL: https://mededu.jmir.org/2024/1/e53624
doi:10.2196/53624

© Alireza Jalali, Jacline Nyman, Ouida Loeffelholz, Chantelle Courtney. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.7.2024. 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.

XSL•FO
RenderX

Can an Online Course, Life101: Mental and Physical Self-Care, Improve the Well-Being of College Students?

Mahtab Jafari, PharmD

Department of Pharmaceutical Sciences, School of Pharmacy and Pharmaceutical Sciences, University of California, Irvine, Irvine, CA, United States

Corresponding Author:
Mahtab Jafari, PharmD

Abstract

The COVID-19 pandemic has had a significant impact on the mental health of college students worldwide. As colleges shifted to online instruction, students faced disruptions and increased stressors, leading to a decline in mental health that appears to continue in the postpandemic era. To alleviate this problem, academic institutions have implemented various interventions to address mental health issues; however, many of these interventions focus on a single approach and lack diverse delivery methods. This viewpoint introduces the concept of a multimodal self-care online course, *Life101: Mental and Physical Self-Care*, and discusses the potential effectiveness of such an intervention in improving students' well-being. The course combines evidence-based interventions and incorporates interactive lectures, workshops, and guest speakers. Pre- and postcourse surveys were conducted over a span of 4 academic terms to evaluate the impact of this course on the well-being and self-care practices of students. The survey data suggest positive outcomes in students taking *Life101*, including the adoption of healthier habits, reduced stress levels, and increased knowledge and practice of self-care techniques. *Life101* represents a novel multimodality intervention to address the epidemic of mental health issues faced by students today. By implementing similar evidence-based multimodal didactic curricula across campuses, academic institutions may be able to better equip students to navigate challenges and promote their overall well-being.

(JMIR Med Educ 2024;10:e50111) doi:10.2196/50111

KEYWORDS

self-care course; stress management; student mental health; multimodal online course; mental health interventions

Background

The COVID-19 pandemic has presented numerous challenges for college students worldwide [1]. As universities transitioned to online instruction in response to the pandemic, students faced significant disruptions in their lives [2]. These changes have tested students' ability to adapt to unforeseen circumstances and underscored the importance of robust mental health resources within academic institutions [3,4]. Surveys conducted within 2 months of the pandemic indicated a substantial decline in college students' mental health [1,4]. New stressors emerged, creating uncertainty about students' academic futures. Unfamiliar learning environments, loss of access to academic resources, limited social interaction, and sudden lifestyle changes led to increased rates of mental disorders, including anxiety, alcohol/substance abuse, depression, attention-deficit hyperactivity disorder, eating disorders, self-injury, and even suicidal ideation [5]. Marginalized communities such as first-generation college students, ethnic minorities, and LGBTQ+ communities were disproportionately affected [4]. Consequently, it is crucial for academic institutions to develop evidence-based resources that address the mental and physical health needs of their students.

While the impact of COVID-19 on physical health may diminish over time, its effects on mental health may lead to a new concern: an epidemic of mental illnesses [6]. To combat the rising prevalence of mental health issues and psychological stress among students, academic institutions have adopted various solutions. Many campuses have introduced mindfulness-based interventions to assist students in managing stress, while others have implemented positive psychology practices aimed at enhancing self-confidence and happiness [3,7]. However, institutional approaches often focus on a single intervention rather than equipping students with a range of evidence-based stress management techniques. A study that exposed college students to a combination of evidence-based multimodal strategies demonstrated enhanced mental well-being compared to that observed in studies testing a single intervention [8]. Moreover, institutional approaches are often delivered through a singular mode that may not cater to students' diverse learning styles. Therefore, evidence-based multimodal approaches, incorporating multiple stress management techniques and healthy lifestyle habits, may hold greater potential for addressing the complex and diverse demands of college students. This viewpoint describes the impact of one such approach, a course titled *Life101: Mental and Physical Self-Care* (hereafter referred to as *Life101*), on students' lives during the COVID-19 pandemic, and explores strategies for

XSL•FO
RenderX

further improving this course and similar modalities to help students manage stress.

Course History, Structure, Content, and Assessment

Life101 is a 10-week course that uses a combination of asynchronous and synchronous components to fulfill various learning outcomes related to self-care. The first online version of *Life101* was offered in 2013. Students watched a 1-hour video lecture on their own and then took a quiz at the end of the lecture. During the summer of 2020, with the help of a grant from the University of California Office of the President, an updated version of *Life101* was developed with new lectures and new educational modalities. This new multimodal course incorporated interactive video lectures with evidence-based content, online group discussions, workshops, quizzes, and practical exercises to facilitate student learning and encourage lifestyle changes. This version was adopted by the University of California system to be offered to students on all 10 University of California campuses. The course also became available to the general public through Coursera and has garnered significant popularity, with current enrollment exceeding 16,500 students and an impressive rating of 4.9 stars out of 5 (as of January 1, 2024) [9]. The primary objective of *Life101* is to enhance college students' academic and personal successes by equipping them with the necessary lifestyle skills to navigate the numerous stressors typical of college life. While the specific content of self-care courses may vary across institutions, they have consistently demonstrated a significant influence on retention rates and other measures of academic success [10].

The course is divided into 10 modules, each focusing on a distinct self-care topic (outlined in Table 1). Every module includes 3-4 short lecture videos ranging from 10 to 15 minutes in length, and reflective questions and exercises are interspersed throughout the videos to encourage student introspection and active learning. After watching the videos, students participate in an online discussion forum in small groups and share what they have learned in the lectures and how they have practiced what they have learned. For each module, supplemental online resources are also made available, such as reading lists of the relevant scientific literature and motivational videos. At the end of each module, students take an online quiz to assess their understanding of the content presented in the module.

XSL•FO
RenderX

Table . *Life101* module topics and their content summaries.

Week	Module topic	Content summary
1	The Science of Adopting Good Habits for Self-care	Importance of developing healthy lifestyle habits, how to develop and maintain healthy habits, developing SMART[a] habits
2	The Etiology, Physiology, Symptoms, and Health Outcomes of Stress	Stress response and relaxation response, how to identify symptoms of stress, impact of stress on health, how to manage stress
3	Nutrition & Wellness	How to read food labels, physiological effects of sugar consumption on health, strategies to avoid harmful foods, the importance of cooking your own meals
4	Mindfulness & Emotional Intelligence	Importance of mindfulness for optimal self-care, practicing mindfulness and relaxation through breathing exercises, definition of emotional intelligence (EI) and how to use it to manage stress
5	The Many Mental and Physical Health Benefits of Exercise	Role of exercise in chronic disease prevention, mental and physical health outcomes of exercise, developing an exercise plan
6	The Impact of Sleep on Mental & Physical Wellness	Importance of sleep for mental and physical health, cognitive impairments caused by sleep deprivation, how to implement good sleep hygiene habits
7	The Health Benefits of Volunteering & Gratitude	Health benefits of volunteering and gratitude, definition of "helper's high," how to develop a habit for a gratitude journal
8	Bad Drugs on College Campus	Commonly used substances with high abuse potential, negative health effects of substances of abuse, impact of energy drink consumption on health outcomes
9	Managing Personal Finances	Importance of managing personal finances, how to develop a monthly budget and pay attention to personal finances, 5 money principles to manage personal finances
10	The Impact of Nature Therapy on Stress Management	Definition of nature therapy and its role in stress management, methods to practice nature therapy

[a]SMART: specific, measurable, achievable, realistic, and time-bound.

Survey Project

Survey Design

At the beginning and following the conclusion of the course, a self-assessment survey was conducted to evaluate students' understanding of the topics presented and their own self-care practices, as well as to gain qualitative insight into the impact of *Life101* on these parameters. Together with some requests for narrative answers, the survey posed 67 statements with responses provided on a 7-point Likert scale ranging from 1 (eg, "never" or "strongly disagree") to 7 (eg, "always" or "strongly agree"). As an aggregate measure of the impact of the course on student beliefs and practices across different self-care areas, the proportion of students who responded positively (ie, selected responses 5 - 7) to questions with a scalar answer was compared between the precourse and postcourse surveys. Because the responses of individual students on the two surveys were not tracked for reasons of confidentiality, formal statistical analysis of the results was not possible.

Ethical Considerations

As outlined by the guidelines of its Office of Research [11], the Institutional Review Board of the University of California, Irvine did not require a formal review of this survey project since the research was conducted in an educational setting, involving normal educational practices.

Impact of Life101 on Self-Care Knowledge and Practices

The impact of the original version of the *Life101* course (offered from 2013 to 2020) on the self-care knowledge and practices of prehealth care undergraduate students has been reported previously in a descriptive fashion [12]. Given the context of the pandemic, an objective evaluation of the revised course appeared to be necessary. Pre- and postcourse surveys were conducted over 4 academic quarters (summer 2020, winter 2021, spring 2021, and summer 2021). Out of 1548 students surveyed, 71% (n=1099) reported a negative impact of the pandemic on their mental health.

XSL•FO
RenderX

As presented in Table 2, upon completing the course, the proportion of students who were able to replace unhealthy habits with healthier habits increased by 14% when compared to the precourse responses. Those who took the course during the winter 2021 quarter reported an even greater impact, with an increase of 27%. Many students with high baseline stress levels (68% of respondents) experienced a decrease in their stress levels after completing *Life101*. The survey also provided deeper insights into students' success in learning and practicing new self-care techniques. For example, as students' overall knowledge about mindfulness increased by 12%, their practice of mindfulness also increased by 18%. Similarly, as students learned about stress management techniques, they not only demonstrated an increase (+27%) in knowledge of these strategies but also reported substantial changes in their practice of specific stress management techniques. There were decreases in the proportion of students who relied on alcohol consumption (−3%) or on the use of various types of media (eg, social media and TV) as means of destressing. Conversely, there were increases in the practice of the destressing techniques emphasized by *Life101*, such as exercise (+10%), nature therapy (+25%), and meditation (+5%). While self-reporting does not necessarily translate directly into an actual change of behavior, the collected data nevertheless depict an overall beneficial outcome of *Life101* on the self-care practices of students.

XSL•FO
RenderX

Table . Impact of *Life101* on the knowledge and practice of selected self-care topics over a span of 4 academic terms.

Survey question	Positive answers in precourse survey, n	Positive answers in postcourse survey, n	Change to positive, %[a]	Mean change %[b]
I am successful in replacing unhealthy habits with healthier ones				13.5
Summer 2020 (n=328)[c]	195	272	23.5	
Winter 2021 (n=487)	173	302	26.5	
Spring 2021 (n=447)	89	113	5.4	
Summer 2021 (n=155)	154	152	−1.3	
I feel stressed most of the time				−10.4
Summer 2020 (n=328)	187	140	−14.3	
Winter 2021 (n=487)	381	267	−23.4	
Spring 2021 (n=447)	309	222	−19.5	
Summer 2021 (n=155)	89	113	15.5	
I know how to destress				27.4
Summer 2020 (n=328)	224	305	24.7	
Winter 2021 (n=487)	310	425	23.6	
Spring 2021 (n=447)	262	397	30.2	
Summer 2021 (n=155)	99	147	31.0	
I destress by drinking alcohol				−3.2
Summer 2020 (n=328)	41	15	−7.9	
Winter 2021 (n=487)	7	1	−1.2	
Spring 2021 (n=447)	6	1	−1.1	
Summer 2021 (n=155)	5	1	−2.6	
I destress by exercising				10.4
Summer 2020 (n=328)	201	270	21.0	
Winter 2021 (n=487)	78	97	3.9	
Spring 2021 (n=447)	71	108	8.3	
Summer 2021 (n=155)	39	52	8.4	
I destress by meditating				5.3
Summer 2020 (n=328)	67	121	16.5	
Winter 2021 (n=487)	16	21	1.0	
Spring 2021 (n=447)	7	23	3.6	
Summer 2021 (n=155)	3	3	0.0	
I destress by using social media				−3.8
Summer 2020 (n=328)	199	165	−10.4	
Winter 2021 (n=487)	85	56	−6.0	
Spring 2021 (n=447)	70	58	−2.7	
Summer 2021 (n=155)	11	17	3.9	
I destress by watching television				−3.3
Summer 2020 (n=328)	179	166	−4.0	
Winter 2021 (n=487)	69	49	−4.1	
Spring 2021 (n=447)	52	29	−5.1	
Summer 2021 (n=155)	12	12	0.0	
I know what mindfulness is				11.7

XSL•FO
RenderX

Survey question		Positive answers in precourse survey, n	Positive answers in postcourse survey, n	Change to positive, %[a]	Mean change %[b]
	Summer 2020 (n=328)	270	324	16.5	
	Winter 2021 (n=487)	451	482	6.4	
	Spring 2021 (n=447)	368	440	16.1	
	Summer 2021 (n=155)	136	148	7.7	
I practice mindfulness					17.7
	Summer 2020 (n=328)	185	289	31.7	
	Winter 2021 (n=487)	279	422	29.4	
	Spring 2021 (n=447)	34	92	13.0	
	Summer 2021 (n=155)	151	146	−3.2	
I practice nature therapy on a weekly basis					24.8
	Summer 2020 (n=328)	—[d]	—	—	
	Winter 2021 (n=487)	60	299	49.1	
	Spring 2021 (n=447)	203	192	−2.5	
	Summer 2021 (n=155)	18	61	27.7	

[a]Percent of students changing to a positive response to this statement between the pre- and postcourse surveys, normalized based on the total number of students per semester.

[b]Calculated as sum of normalized percent changes of students changing to a positive response to this statement between the pre- and postcourse surveys for each semester/number of semesters.

[c]n values represent the total number of students responding to both pre-and postcourse surveys for that term.

[d]This measure was not collected for the summer 2020 course offering.

Life101 as a Model to Address Mental Health Challenges in College Students

Life101 was developed and launched a decade ago to attempt to address the plethora of challenges college students face with regard to personal and mental health, which have only increased since the COVID-19 pandemic. The course has been modified and refined every year based on student feedback. The data from pre- and postcourse surveys conducted during the COVID-29 pandemic suggest that the current version of *Life101* has the potential to improve the mental and physical well-being of college students. Two studies examining similar approaches to *Life101* have also reported positive outcomes [8,13]. Morton et al [8] found that students participating in a 10-week multimodal program with multiple strategies to improve their mental health experienced greater improvements in mental health compared to those focusing solely on a single strategy. Similarly, a recent study evaluating an 8-week multimodal stress management program demonstrated positive effects of the program on college students' psychological distress during the pandemic [13]. Although the published literature regarding programs and courses similar to *Life101* is limited, the favorable outcomes obtained are likely attributed to the multimodal nature and multi-interventional design of the course, which equip students with a repertoire of stress-coping strategies for different situations. Similarly, the overall positive impact of *Life101* might in part be attributed to its incorporation of multiple pedagogical methods, including interactive video lectures, embedded reflective questions and activities in the videos, assigned scientific readings, workshops, discussion forums, practical exercises, and quizzes. By using various modes of content delivery and assessment, the course can enhance student comprehension and retention while accommodating diverse learning styles. In addition, the comprehensive range of topics covered in *Life101* empowers students to address the typical stressors of college life. Definitive proof of the superiority of multimodality approaches to addressing student mental health and well-being will require larger comparative studies of different teaching approaches.

Future Directions

While self-care courses such as *Life101* have the potential to benefit students' psychological and physical health, it is important to continually improve these courses to meet the evolving mental health needs of students. In addition to the typical stressors faced by college students, such as academic pressure and financial burdens, research has highlighted the link between psychological stress and excessive use of social media platforms [14-17]. Excessive use of social media has been associated with declining mental health [15,16]. Today's college students, often referred to as "digital natives," heavily rely on their mobile devices for various purposes, including accessing health information, entertainment, and maintaining social connections [18]. To take advantage of this fact, the practice of mindfulness, as introduced in *Life101*, should be expanded to include its application during the use of social media apps. Studies have shown that mindful use of social media can lead to reduced stress and increased well-being compared to passive scrolling [16,19,20]. By incorporating mindfulness into social media use, self-care courses can promote intentional engagement

XSL•FO
RenderX

with digital platforms and address the mental health issues associated with excessive social media use and exposure.

Another area of opportunity for institutions developing self-care programs is the prevention of online misinformation among college students. The COVID-19 pandemic highlighted the role of social media in the dissemination of health-related information [21], resulting in a flood of both reliable and unreliable content online. Since college students often rely on online sources for obtaining and sharing health-related information, it is crucial for them to be able to differentiate between reliable and unreliable sources [18,21,22]. A study on the information-seeking behavior of college students found that nearly half of the students (50%) found it challenging to evaluate the credibility of information [23]. Recognizing the importance of addressing this issue, the Department of Health and Human Services released an advisory on "Confronting Health Misinformation" in 2021, emphasizing the need for individuals to develop skills in assessing the credibility of online sources [24]. In addition, according to a systematic review conducted in 2021 on the prevalence of health misinformation on social media, misinformation in various social media platforms had a high prevalence, especially for vaccines and diseases [25]. Although the importance of evidence-based health information is discussed in all of the *Life101* modules, we are considering incorporating a module in this course that covers the basics of evaluating source credibility. This new module can empower students to make informed choices and safeguard their mental and physical well-being.

In addition to the aforementioned recommendations, delivering self-care programs such as *Life101* through mobile apps can be beneficial. Given that most college students own a mobile phone, rely on online resources for support, and spend a significant amount of time using apps, delivering a self-care course through a smartphone-based app aligns with their preferences [26,27]. A systematic literature review conducted in 2021 demonstrated the effectiveness of mental health apps in preventing stress, anxiety, and depression, and recommended that universities adopt mobile apps designed to benefit student mental health [28]. By delivering self-care courses through a mobile app, institutions can increase accessibility among a wider student population and gather real-time data on students' stress levels, sleep patterns, mood changes, and physical activity levels. These data can be used to track and analyze students' well-being and provide tailored and personalized recommendations through in-app notifications. Delivering self-care courses such as *Life101* via a mobile app can be a transformative step in empowering students to actively engage in their own health.

Conclusion

Given the individual needs and diverse challenges students face, the incorporation of diverse evidence-based educational strategies in *Life101* provides students with opportunities to practice self-care and take greater personal responsibility, which are essential aspects of early adulthood [3]. If colleges adopt a multimodal approach in self-care courses across all campuses nationwide, students would be better equipped to navigate challenges, both during and outside of a pandemic period. Furthermore, colleges can develop targeted resources that focus on the mindful use of social media, identification of accurate health misinformation, and the creation of mobile phone apps that deliver self-care content tailored specifically to students' needs.

Acknowledgments

The author would like to acknowledge an Innovation Learning Technology (ILTI) University of California Online grant from the University of California, Office of The President, that supported the design of the revised online Life101 course.

Data Availability

The complete survey data summarized in Table 2 are available upon request to the corresponding author.

Conflicts of Interest

None declared.

References

1. Jafari M, De Roche M, Eshaghi MR. COVID-19, stress and mental health: what students expect from academic institutions during a pandemic. J Am Coll Health 2023 Oct;71(7):1976-1983. [doi: 10.1080/07448481.2021.1951740] [Medline: 34398699]
2. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Policy and Global Affairs; Board on Health Sciences Policy; Board on Higher Education and Workforce; Committee on Mental Health, Substance Use, and Wellbeing in STEMM Undergraduate and Graduate Education. In: Scherer LA, Leshner AI, editors. Mental Health, Substance Use, and Wellbeing in Higher Education: Supporting the Whole Student: National Academies Press; 2021.
3. Dye L, Burke MG, Mason CP. Mindful Strategies for Helping College Students Manage Stress: A Guide for Higher Education Professionals: Routledge; 2021.

XSL•FO
RenderX

4. Liu CH, Pinder-Amaker S, Hahm HC, Chen JA. Priorities for addressing the impact of the COVID-19 pandemic on college student mental health. J Am Coll Health 2022 Jul;70(5):1356-1358. [doi: 10.1080/07448481.2020.1803882] [Medline: 33048654]
5. Clabaugh A, Duque JF, Fields LJ. Academic stress and emotional well-being in United States college students following onset of the COVID-19 pandemic. Front Psychol 2021 Mar;12:628787. [doi: 10.3389/fpsyg.2021.628787] [Medline: 33815214]
6. Stress in America 2020: a national mental health crisis. American Psychological Association. 2020. URL: https://www.apa.org/news/press/releases/stress/2020/report-october [accessed 2024-02-05]
7. Smit B, Stavrulaki E. The efficacy of a mindfulness-based intervention for college students under extremely stressful conditions. Mindfulness 2021;12(12):3086-3100. [doi: 10.1007/s12671-021-01772-9] [Medline: 34642590]
8. Morton DP, Hinze J, Craig B, et al. A multimodal intervention for improving the mental health and emotional well-being of college students. Am J Lifestyle Med 2020 Mar;14(2):216-224. [doi: 10.1177/1559827617733941] [Medline: 32231487]
9. Life 101: Mental and Physical Self-Care. Coursera. URL: https://www.coursera.org/learn/life101 [accessed 2024-02-05]
10. Young DG. Is first-year seminar type predictive of institutional retention rates? J Coll Stud Dev 2020;61(3):379-390. [doi: 10.1353/csd.2020.0035]
11. Exempt self-determination & UROP. University of California, Irvine Office of Research. URL: https://research.uci.edu/human-research-protections/do-you-need-irb-review/self-exempt [accessed 2024-05-24]
12. Jafari M. Life101 enhances healthy lifestyle choices in pre-health undergraduate students. J Univ Teach Learn Pract 2017 Jul 1;14(3):41-58. [doi: 10.53761/1.14.3.4]
13. Theurel A, Witt A, Shankland R. Promoting university students' mental health through an online multicomponent intervention during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Aug 22;19(16):10442. [doi: 10.3390/ijerph191610442] [Medline: 36012078]
14. Wei XY, Ren L, Jiang HB, et al. Does adolescents' social anxiety trigger problematic smartphone use, or vice versa? A comparison between problematic and unproblematic smartphone users. Comput Human Behav 2023 Mar;140:107602. [doi: 10.1016/j.chb.2022.107602]
15. Primack BA, Shensa A, Escobar-Viera CG, et al. Use of multiple social media platforms and symptoms of depression and anxiety: a nationally-representative study among U.S. young adults. Comput Human Behav 2017 Apr;69:1-9. [doi: 10.1016/j.chb.2016.11.013]
16. Lin LY, Sidani JE, Shensa A, et al. Association between social media use and depression among U.S. young adults. Depress Anxiety 2016 Apr;33(4):323-331. [doi: 10.1002/da.22466] [Medline: 26783723]
17. Wolniewicz CA, Tiamiyu MF, Weeks JW, Elhai JD. Problematic smartphone use and relations with negative affect, fear of missing out, and fear of negative and positive evaluation. Psychiatry Res 2018 Apr;262:618-623. [doi: 10.1016/j.psychres.2017.09.058] [Medline: 28982630]
18. Montagni I, Tzourio C, Cousin T, Sagara JA, Bada-Alonzi J, Horgan A. Mental health-related digital use by university students: a systematic review. Telemed J E Health 2020 Feb;26(2):131-146. [doi: 10.1089/tmj.2018.0316] [Medline: 30888256]
19. Hong W, Liu RD, Ding Y, Fu X, Zhen R, Sheng X. Social media exposure and college students' mental health during the outbreak of COVID-19: the mediating role of rumination and the moderating role of mindfulness. Cyberpsychol Behav Soc Netw 2021 Apr;24(4):282-287. [doi: 10.1089/cyber.2020.0387] [Medline: 33050721]
20. Chan SS, Van Solt M, Cruz RE, et al. From the fear of missing out (FOMO) to the joy of missing out (JOMO). J Consumer Affairs 2011;56(3):1312-1331. [doi: 10.1111/joca.12476]
21. Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC. Managing the COVID-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation. World Health Organization. 2020. URL: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation [accessed 2024-02-05]
22. Zhang D, Zhan W, Zheng C, et al. Online health information-seeking behaviors and skills of Chinese college students. BMC Public Health 2021 Apr 15;21(1):736. [doi: 10.1186/s12889-021-10801-0] [Medline: 33858389]
23. Vrdelja M, Vrbovšek S, Klopčič V, Dadaczynski K, Okan O. Facing the growing COVID-19 infodemic: digital health literacy and information-seeking behaviour of university students in Slovenia. Int J Environ Res Public Health 2021 Aug 12;18(16):8507. [doi: 10.3390/ijerph18168507] [Medline: 34444255]
24. Confronting health misinformation: the US Surgeon's General advisory on building a healthy information environment. US Department of Health and Human Services. URL: https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf [accessed 2024-02-05]
25. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187. [doi: 10.2196/17187] [Medline: 33470931]
26. Fook CY, Narasuman S, Aziz NA, Mustafa SMS, Han CT. Smart phone use among university students. AJUE 2021 Mar;17(1):282. [doi: 10.24191/ajue.v17i1.12622]

XSL•FO
RenderX

27. Chen L, Li J, Huang J. COVID-19 victimization experience and college students' mobile phone addiction: a moderated mediation effect of future anxiety and Mindfulness. Int J Environ Res Public Health 2022 Jun 21;19(13):7578. [doi: 10.3390/ijerph19137578] [Medline: 35805232]
28. Oliveira C, Pereira A, Vagos P, Nóbrega C, Gonçalves J, Afonso B. Effectiveness of mobile app-based psychological interventions for college students: a systematic review of the literature. Front Psychol 2021 May;12:647606. [doi: 10.3389/fpsyg.2021.647606] [Medline: 34045994]

Edited by TDA Cardoso; submitted 19.06.23; peer-reviewed by P Kadandale, R Bluhm, S Arya; revised version received 16.05.24; accepted 29.05.24; published 22.07.24.

Please cite as:
Jafari M
Can an Online Course, Life101: Mental and Physical Self-Care, Improve the Well-Being of College Students?
JMIR Med Educ 2024;10:e50111
URL: https://mededu.jmir.org/2024/1/e50111
doi:10.2196/50111

© Mahtab Jafari. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.7.2024. 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.

XSL•FO
RenderX

Reforming China's Secondary Vocational Medical Education: Adapting to the Challenges and Opportunities of the AI Era

Wenting Tong[1], MS; Xiaowen Zhang[2], MS; Haiping Zeng[3,4], MD; Jianping Pan[5], PhD; Chao Gong[6], MS; Hui Zhang[2,7,8], MD, PhD

1
2
3
4
5
6
7
8

Corresponding Author:
Hui Zhang, MD, PhD

Abstract

China's secondary vocational medical education is essential for training primary health care personnel and enhancing public health responses. This education system currently faces challenges, primarily due to its emphasis on knowledge acquisition that overshadows the development and application of skills, especially in the context of emerging artificial intelligence (AI) technologies. This article delves into the impact of AI on medical practices and uses this analysis to suggest reforms for the vocational medical education system in China. AI is found to significantly enhance diagnostic capabilities, therapeutic decision-making, and patient management. However, it also brings about concerns such as potential job losses and necessitates the adaptation of medical professionals to new technologies. Proposed reforms include a greater focus on critical thinking, hands-on experiences, skill development, medical ethics, and integrating humanities and AI into the curriculum. These reforms require ongoing evaluation and sustained research to effectively prepare medical students for future challenges in the field.

(JMIR Med Educ 2024;10:e48594) doi:10.2196/48594

KEYWORDS

secondary vocational medical education; artificial intelligence; practical skills; critical thinking; AI

Introduction

A well-established medical education system is pivotal in training a sufficient number of high-quality professionals to meet societal health needs and tackle challenges. China's medical education structure encompasses secondary vocational medical education, undergraduate, master's, and doctoral degrees [1]. Specifically, secondary vocational medical education is a 3-year program for junior high school graduates [2]. Its origin traces back to the "barefoot doctors" of the 1960s. While not all were formal doctors, they underwent fundamental medical and health training, primarily serving rural areas. To address the medical service shortage in rural areas, the Chinese government trained a group of farmers with basic medical skills in the 1960s [3,4]. These barefoot doctors played a pivotal role in China's health care system, significantly alleviating rural medical service shortages and improving overall health standards. However, as the medical system and the economy evolved in the late 1970s and early 1980s, the barefoot doctor model gradually phased out [3]. Despite its development, China still exhibits a dual nature due to uneven progress. On one side, in economically developed coastal and major urban areas, medical resources are comparable with those in economically developed regions such as Europe and America. On the other side, similar to some regions in Asia or Africa, areas such as Qinghai, Tibet in the west of China and many rural locales experience a severe lack of medical resources. In some of these areas, the standards for practicing qualifications have even been lowered to meet basic health care needs. This stark contrast underscores the challenge of achieving equitable health care access across diverse geographic and economic landscapes [5-8].

Secondary vocational medical education can be seen as an evolved version of the barefoot doctor model, aiming to address medical resource shortages and service imbalances due to regional disparities [1]. The core objectives of this educational system are to enhance grassroots medical levels, nurture qualified medical personnel, and reinforce grassroots medical institution infrastructure, thus bolstering public health response capabilities [9]. Nevertheless, this system heavily relies on

XSL•FO
RenderX

traditional teaching methods, leading to a significant disconnect between theory and practice for students [10].

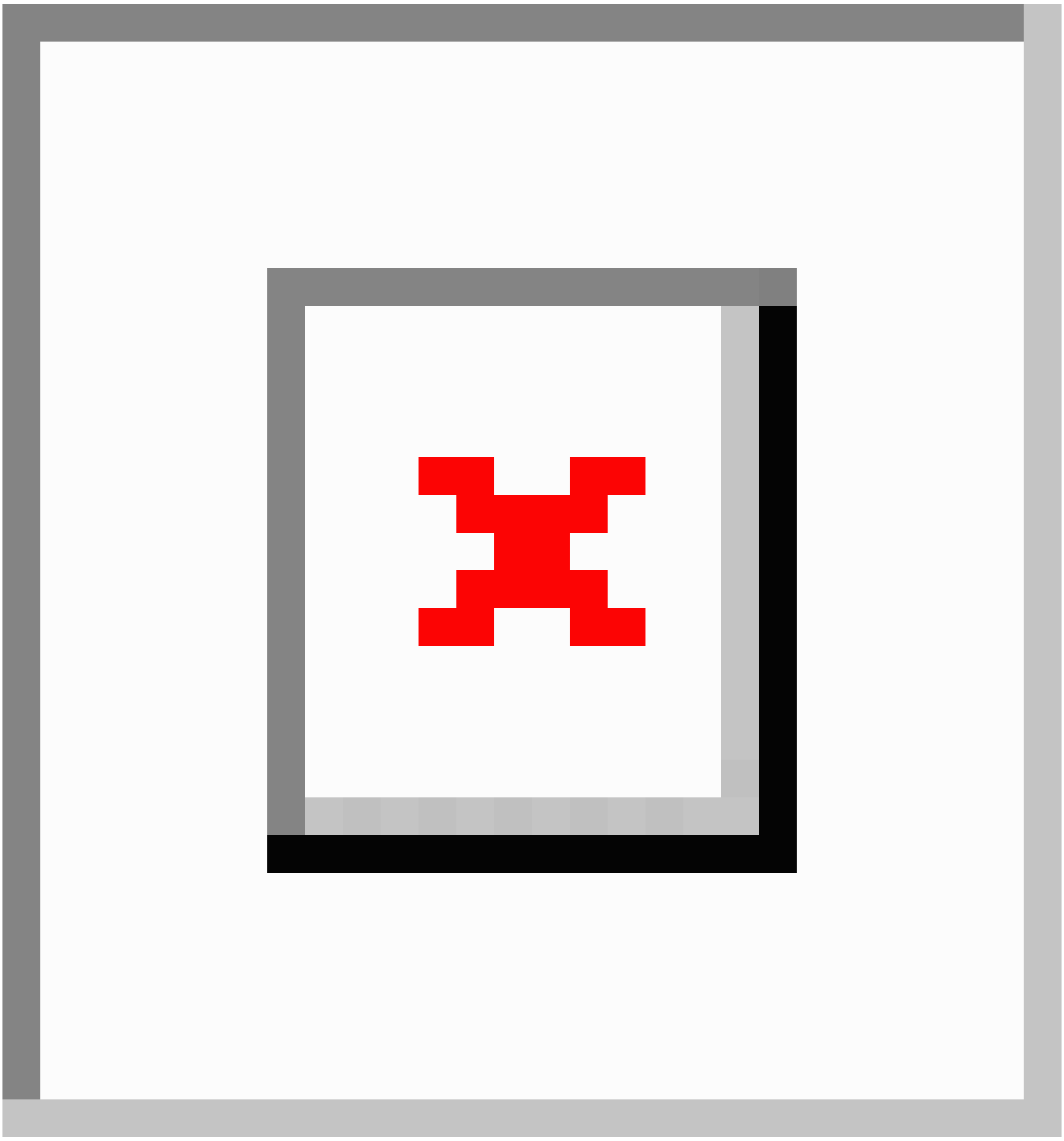
Artificial intelligence (AI) is profoundly reshaping the medical sector, but the current secondary vocational medical education system has not fully integrated AI technology into its teaching. To ensure that students can fully harness and address these technological revolutions, educators need to reconsider curriculum design, integrating these cutting-edge technologies and preparing students for future medical innovations [11-13]. Based on the long-term experience accumulated by the team in secondary vocational medical education and existing research, this study will delve into the challenges faced by secondary vocational medical education in the era of AI and potential strategies to address them, and this analysis will provide valuable insights and lessons that can be applied to similar countries and regions at various levels of economic development.

Opportunities and Challenges Faced by Secondary Vocational Education in the Era of AI

Over the past 70 years, the grassroots medical standard in China has been improving. However, constrained by economic development and population growth, the distribution of medical resources across the country remains limited and severely imbalanced. A small amount of high-quality medical resources is concentrated in economically developed areas (Figure 1), and medical staff generally bear a high workload. Research indicates that in rural grassroots areas, medical personnel work an average of about 8.9 hours daily, working at least 6 days a week [14]. Another study shows that in 2010, 2015, and 2016, health care workers had monthly workloads exceeding 30 days for 6, 5, and 9 months, respectively [15]. The massive patient flow due to the large number of patients places a heavy burden on doctors.

XSL•FO
RenderX

Figure 1. Distribution of the top 100 hospitals in China. Data were sourced from Fudan University [16]. Data from Taiwan, Hong Kong, and Macau are excluded from this analysis.



Concurrently, the 3-year fast-track training provided by secondary vocational medical education does not endow medical students with the plethora of skills they should ideally possess. Many doctors who enter grassroots work lack systematic and in-depth professional training. They might rely merely on basic medical knowledge and experience to diagnose and treat patients. While this may suffice for the majority of primary diagnoses and treatments, it still falls short when dealing with complex cases [17,18]. Combined with the heavy workload, this could potentially lead to misdiagnoses or overtreatment.

It is noteworthy that the overwhelming workload also results in doctors having little desire to communicate adequately with patients [19]. In many instances, the root cause of doctor-patient disputes is not merely the misdiagnosis itself but rather the lack of effective communication between the doctor and the patient.

In recent years, the rapid development of AI has demonstrated its potential impact on China's health care landscape. In the realm of direct patient care, the application of AI not only has the potential to enhance the efficiency of medical services [20-26] but also opens new opportunities for medical equity across different regions [27].

Before the widespread application of generative AI, scientists had already been using AI imaging systems and other technologies to address the disparities in medical standards between regions. In primary health care settings, AI is

XSL•FO
RenderX

progressively gaining prominence and is perceived as an auxiliary tool with immense potential [28-30]. Numerous studies have highlighted that particularly in areas lacking experienced radiologists, AI plays a pivotal role in medical imaging analysis, such as in x-rays or basic scans [31,32]. Furthermore, AI offers data-driven therapeutic suggestions not only to physicians, enhancing the accuracy and efficiency of treatments, but also in health care resource management, such as bed allocation. Its predictive models proficiently optimize resource allocation, ensuring that patients receive timely and appropriate care [12]. In the context of telemedicine, the integration of AI with wearable devices undoubtedly delivers more precise health information to medical practitioners, leading to more efficient health management. This would substantially alleviate the workload of health care providers, laying the groundwork for a more harmonious doctor-patient relationship [33].

Building on this foundation, generative AI technologies such as ChatGPT have opened up new possibilities for enhancing primary health care. Multiple studies have confirmed the significant potential and reliability of generative AI in the medical field, enhancing the decision-making capabilities of primary care physicians [20-24]. Notably, these technologies have also shown significant potential in improving doctors' communication skills.

Communication is one of the core skills for physicians, especially when dealing with a large number of patients. Effective communication skills can aid in better patient recovery and provide a harmonious practice environment [34,35]. Studies have demonstrated that ChatGPT exhibits a high level of empathy when addressing common queries [22,36]. In addition, research has explored the potential of generative AI to enhance the communication skills of emergency medical doctors, particularly in delivering bad news, by simulating patient reactions and dialogues during the disclosure of a cancer diagnosis [37]. With the aid of AI, we can better simulate clinical environments, thus improving the training of medical students in patient communication [38].

Furthermore, the future of medical practice will become more complex, requiring doctors to not only possess professional knowledge but also have a basic understanding of technologies related to health care, such as blockchain, cloud services, data quality, machine learning, electronic health records, and mobile health. Some of these technologies are straightforward, while others are complex; AI can help beginners simplify concepts and accelerate their learning [39].

However, despite the conveniences and advantages that AI brings to the training of medical students in primary care, we must still face several inherent challenges: the threat to employment, the necessity of skill updates, and the ongoing need for training. The automation capabilities of AI may gradually replace some basic and repetitive tasks, such as preliminary diagnosis and data entry, which could impact the job stability of medical personnel [12,13].

Moreover, the rapid development of AI technology may increase the obsolescence of certain traditional medical practices and skills. To keep pace with technological progress, health care workers may face more frequent training demands. Although the concept of lifelong learning is inherently positive, it could impose additional psychological stress on doctors [40,41].

Advancements in technology, while opening new treatment possibilities, also raise new ethical issues, such as data privacy and machine bias, which need to be addressed and resolved in medical education [42]. In addition, while problem-based learning (PBL) approaches attempt to bridge the gap between theory and practice, a lack of practical opportunities and the disconnection between theory and practice might leave students feeling unprepared when facing real medical challenges. These challenges need to be carefully considered and overcome in the AI-integrated educational process to ensure the quality of education and the professional development of students.

At this stage, AI primarily acts as an auxiliary tool [43], helping medical personnel solve problems more effectively and optimize services to patients. As educators, we have a responsibility to directly address any fears students may have [44] and to start popularizing and applying AI knowledge from the educational phase. This will help them more effectively use these technologies in their future practice and reduce resistance to new technologies.

The reform of AI in education will not happen overnight, as school reforms often depend on policy support and tend to lag [45]. Furthermore, the development of AI requires the integration of technology, which in turn necessitates significant resource investment, such as in hardware, software, and professional talent. Unequal resource distribution could make achieving this goal difficult. The rapidly changing technological environment also demands frequent updates to educational content, presenting ongoing challenges for educational institutions.

Recommendations for the Reform of Medical Education in Chinese Secondary Vocational Schools

Integration With Technology

In recent years, AI has taken a central role amidst the technological revolution in the medical field, particularly given its significant impact on enhancing diagnosis and treatment efficiencies. To adapt to this trend, secondary vocational medical education must adjust to ensure that students not just grasp traditional medical techniques but also intertwine with AI technology and applications. This encompasses understanding the significance of machine learning algorithms and data analysis techniques, as well as how to effectively use AI in real-world medical settings [46].

The preclinical teaching phase serves as an ideal starting point. Strengthening courses on health data management, integration, and governance and emphasizing foundational AI, ethics, and legal issues are paramount [46,47]. These courses can be offered independently, ensuring that students maintain foundational knowledge even if certain technologies or applications become obsolete [48].

In addition, students should comprehend the computer and software engineering principles behind AI applications. Insights

XSL•FO
RenderX

into hardware and software development and foundational knowledge of user experience design could be invaluable for their future medical careers [49-51]. Furthermore, during clinical rotations and residencies, students should focus on the practical application of AI, such as the widespread use of AI-based technologies for digital biomarkers and therapies in home settings, which offer large-scale diagnostic and therapeutic solutions [52,53]. In essence, tightly weaving AI into secondary vocational medical education will equip students to serve patients better, ensuring efficiency and accuracy in medical services.

Lifelong Learning

Given the constantly evolving nature of medicine, it is imperative for practitioners to adapt continually to its changing landscape. To ensure that medical students thrive in this dynamic environment, there should be a heightened emphasis on cultivating a growth mindset and fostering lifelong learning capabilities. This mindset encourages viewing challenges and failures as opportunities for learning and growth rather than end points [54,55]. With rapid advancements in medical technology and treatments, students need the awareness and ability to continually refresh their knowledge and skills, keeping them current [56,57]. Offering students exposure and hands-on experience with AI tools not only aligns them with current medical technology trends but also instills a strong adaptive and continuous learning culture—a key to success in a fast-evolving medical field [58,59].

Nurturing Ethical and Critical Thinking

There is a growing global focus on how medical curriculum design balances traditional medical education with the integration of emerging technologies [60-62]. Enhancing the medical curriculum should include not just traditional medical knowledge but also medical ethics and humanities [63,64]. Such a structure not just cultivates students' grasp of medical concepts but also strengthens their ethical foundations, critical thinking, and decision-making abilities [65]. With advancements in medical technology, especially the widespread adoption of AI, students must learn to balance technology use and ethics. Despite the unparalleled conveniences AI offers in health care, it has evident limitations. Students need sound judgment to ensure optimal treatment choices for patients [66-68]. Furthermore, as AI's role in medicine expands, solidifying foundational medical knowledge becomes even more crucial. Students require a robust medical foundation, providing them with a framework to make accurate judgments about AI technologies and ensuring their correct clinical application [69].

Emphasizing Practical Experience

Modern medical education is at a pivotal juncture, necessitating a closer alignment of profound theoretical knowledge with actual medical practice. To achieve this, there is a need to revisit and optimize the curriculum, placing hands-on experience and skill cultivation at its core. PBL offers a direction for this educational transformation. PBL not only stimulates students' proactivity, enabling them to devise solutions for real medical scenarios, but also nurtures their critical thinking abilities [70,71]. At the same time, the advent of AI will compel educators to abandon traditional teacher-centered instructional methods. With the assistance of AI, educators can facilitate active participation and personalized education for students. They can generate learning materials tailored to each student's learning status and needs, such as by simulating standard patients, providing diverse case studies, and offering brainstorming activities and practice exercises [72,73], thereby increasing clinical internship opportunities that allow students to delve deeper into and comprehend medical practice. Specific practical activities, such as internships and simulated diagnostics, not only deepen students' understanding of medical environments but also aid them in making wiser decisions when faced with intricate medical issues [74]. More crucially, such authentic clinical experiences bridge the gap between theoretical knowledge and practical operation, helping students foster a more professional demeanor, boosting their confidence, and ensuring superior performance in real medical settings [75].

Resource Allocation

To ensure the successful implementation of medical education reform, it is essential to focus on optimizing resource distribution [76]. Specifically, financial investments should be concentrated on upgrading educational infrastructure, acquiring and maintaining new technologies, and establishing a dedicated fund to support the technologization of medical education. In addition, the professional development of teachers is crucial. Systematic training must be provided on AI and related technologies to ensure that teachers possess the most advanced knowledge and skills. Schools should also be equipped with the necessary technical resources to access the latest medical databases and AI tools, such as high-speed internet, updated computer hardware, and software.

Policy Support

In terms of policy support, reforms in medical education should be facilitated through the development of policies that specifically support technology integration and lifelong learning. This includes establishing standards for educational quality and teacher evaluations while encouraging cross-sector collaboration between education departments and health, technology, and private sectors. Importantly, a comprehensive regulatory framework needs to be established to monitor the application of AI technology in medical education, ensuring that all activities comply with ethical and legal standards to protect the rights of students and patients.

Infrastructure Development

Modernizing educational infrastructure is key to enhancing teaching quality [77,78]. Relevant authorities should invest in upgrading traditional classrooms and laboratories to support applications such as virtual reality and augmented reality. Combined with AI, these technologies can be used to simulate complex medical scenarios and surgical procedures [79]. Developing or adopting advanced learning management systems to support web-based teaching and resource sharing, as well as constructing more modern clinical training facilities, can significantly enhance students' practical skills and lay a solid foundation for their future careers.

XSL•FO
RenderX

Discussion

As noted in earlier sections of this paper, China's secondary vocational medical education system, while comprehensive, still relies heavily on traditional teaching methods that emphasize rote memorization over practical application and critical thinking skills [80]. As we move further into the era of AI, these educational frameworks are becoming increasingly outdated. Technologies such as ChatGPT offer the potential to radically reform these traditional systems. By using AI tools and methods, we can address many of the current issues in our educational system, such as the disconnect between theoretical knowledge and practical application.

In the AI era, medical students are presented with unique opportunities to access a wealth of medical curricula previously unimaginable. For instance, AI integrated with augmented reality and virtual reality can greatly enhance interactivity, creating more engaging learning environments that allow students to practice skills in a risk-free setting [79]. Moreover, AI can facilitate personalized education, adjusting learning materials and pacing to meet individual student needs [72,73]. By leveraging these technologies, educational institutions can cultivate more skilled and versatile medical professionals who are well prepared to tackle the challenges of the modern medical environment.

Within the context of China's secondary vocational medical education, practical applications of AI should include the introduction of AI-driven diagnostic tools during clinical rotations, allowing students firsthand experience with their use. This exposure not only enhances their diagnostic capabilities but also enables them to critically understand the advantages and limitations of AI-assisted decision-making [81,82].

To effectively integrate AI into medical education, educational departments must revise curricula to include specialized courses in data science, machine learning, and the ethical considerations of using AI [63,64]. These courses should be designed to ensure students comprehend both the capabilities and limitations of AI technology. In addition, training for educators must also be undertaken to ensure that they possess the requisite up-to-date knowledge and concepts to teach these new modules.

While the benefits of integrating AI into medical education are clear, significant challenges and potential resistance exist. These challenges include transforming traditional educational paradigms, the high costs of technological integration, and the need for continual curriculum updates to keep pace with technological advancements. A crucial step in addressing these challenges involves engaging all stakeholders—including educators, students, and policy makers—in the educational reform process. Demonstrating the specific benefits of AI in enhancing student learning outcomes and patient care can help garner broader support to realize these changes.

Conclusions

The future of medical education in China, particularly at secondary vocational schools, will largely depend on the ability of educators, policy makers, and society to adapt and respond to technological advances. By embracing AI and incorporating it into curriculum design, we can train the next generation of health care professionals, equipping them not only with traditional medical knowledge but also with the skills to use technology to improve patient outcomes. Although challenges exist in the reform process, it is vital to ensure that medical students are well prepared for future medical practices.

Acknowledgments

This research was supported by the Scientific Research Project of Traditional Chinese Medicine Bureau of Guangdong Province (project number: 20221086).

Disclaimer

During the writing process of this study, we used ChatGPT for language polishing and editing, not for generating scientific content or data. We ensure that all content refined through ChatGPT fully aligns with the original authors' intent and expression, maintaining the accuracy and consistency of the information. The accuracy and originality of all scientific assertions, data analyses, and conclusions were independently carried out by our team. This disclosure is intended to maintain complete transparency in the research process.

Authors' Contributions

WT and H Zhang spearheaded the study, managing design, literature review, and manuscript drafting. XZ offered pivotal suggestions and comprehensive revisions, enriching the article's quality. H Zeng analyzed challenges in China's vocational medical education and proposed solutions. JP evaluated the influence of AI on medical education. CG discussed interdisciplinary collaboration and adaptability, while XZ assessed implementation challenges and resolutions. H Zhang supervised the project and refined the manuscript's intellectual content. All authors contributed to data interpretation and manuscript revision and approved the final submission.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Wang W. Medical education in China: progress in the past 70 years and a vision for the future. BMC Med Educ 2021 Aug 28;21(1):453. [doi: 10.1186/s12909-021-02875-6] [Medline: 34454502]
2. Reynolds TA, Tierney LM. STUDENTJAMA. Medical education in modern China. JAMA 2004 May 5;291(17):2141. [doi: 10.1001/jama.291.17.2141] [Medline: 15126446]
3. Dongyue S, Xiaoyan W, Chen W, et al. Management system during the barefoot doctor period and its implications for current rural health talent management [Article in Chinese]. Chin Gen Pract J 2011;7:3. [doi: 10.3969/j.issn.1007-9572.2011.07.009]
4. Fang X. A barefoot doctor's manual as a "Medical Bible": medical politics and knowledge transmission in China. Chin Ann Hist Sci Technol 2019;3(2):166-194. [doi: 10.3724/SP.J.1461.2019.02166]
5. Dingxiangyuan. About the current status and development of medical education in China [Article in Chinese]. Sohu. 2024. URL: https://www.sohu.com/a/775975626_296660 [accessed 2024-05-30]
6. Liu CY, Grant B, Ye L. Special issue: China's new urban realities and development policies. J Urban Aff 2019 Feb 17;41(2):149-149. [doi: 10.1080/07352166.2019.1565247]
7. The Writing Committee of the Report on Cardiovascular Health and Diseases in China, Hu SS. Report on cardiovascular health and diseases in China 2021: an updated summary. J Geriatr Cardiol 2023 Jun 28;20(6):399-430. [doi: 10.26599/1671-5411.2023.06.001] [Medline: 37416519]
8. Li Q, Han T, Zhang Y, et al. A nationwide survey on neonatal medical resources in mainland China: current status and future challenges. BMC Pediatr 2019 Nov 13;19(1):436. [doi: 10.1186/s12887-019-1780-4] [Medline: 31722687]
9. Zhang D, Unschuld PU. China's barefoot doctor: past, present, and future. Lancet 2008 Nov 29;372(9653):1865-1867. [doi: 10.1016/S0140-6736(08)61355-0] [Medline: 18930539]
10. Li XZ, Chen CC, Kang X. Research on the cultivation of sustainable development ability of higher vocational students by creative thinking teaching method. Front Psychol 2022;13:979913. [doi: 10.3389/fpsyg.2022.979913] [Medline: 36275280]
11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
12. Castelvecchi D. Are ChatGPT and AlphaCode going to replace programmers? Nature 2022 Dec 8. [doi: 10.1038/d41586-022-04383-z]
13. Harada Y, Katsukura S, Kawamura R, Shimizu T. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: an open-label randomized controlled study. Int J Environ Res Public Health 2021 Feb 21;18(4):2086. [doi: 10.3390/ijerph18042086] [Medline: 33669930]
14. Zhao T. Workload of primary care doctors in rural China and its influencing factors: an empirical analysis based on survey data from three provinces [Article in Chinese]. Econ Rev 2014;2014(1):12-24.
15. Jianhua SHI, Huining GU, Mengcen Q, et al. Workload of public health services in primary medical and health institutions: a seven-year trend analysis. Chin Gen Pract 2020;23(34):4291-4297. [doi: 10.12114/j.issn.1007-9572.2020.00.480]
16. 中国医院及专科声誉排行榜 复旦大学医院管理研究所 [Article in Chinese]. CN-Healthcare. URL: https://rank.cn-healthcare.com/fudan/national-general/year/2021 [accessed 2024-08-12]
17. Jili L, Lili G, Yongjun Q, et al. 2009年中山市基层医务人员糖尿病防治知识知晓情况调查 [Article in Chinese]. Prev Med Trib 2009;15(11):1082-1083. [doi: 10.16406/j.pmt.issn.1672-9153.2009.11.009]
18. Zhang R, He Q. Impact of continuing medical education on chronic obstructive pulmonary disease knowledge of medical doctors practicing at the grassroots [Article in Chinese]. Chin J Gen Pract 2009;8(5):320-322. [doi: 10.3760/cma.j.issn.1671-7368.2009.05.011]
19. Schillinger D, Piette J, Grumbach K, et al. Closing the loop: physician communication with diabetic patients who have low health literacy. Arch Intern Med 2003 Jan 13;163(1):83-90. [doi: 10.1001/archinte.163.1.83] [Medline: 12523921]
20. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the performance of ChatGPT versions 3.5, 4, and 4 With Vision in the Chilean medical licensing examination: observational study. JMIR Med Educ 2024 Apr 29;10:e55048. [doi: 10.2196/55048] [Medline: 38686550]
21. Noda M, Ueno T, Koshu R, et al. Performance of GPT-4V in answering the Japanese otolaryngology board certification examination questions: evaluation study. JMIR Med Educ 2024 Mar 28;10:e57054. [doi: 10.2196/57054] [Medline: 38546736]
22. Tong W, Guan Y, Chen J, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese National Medical Licensing Examination. Front Med 2023 Oct;10:1237432. [doi: 10.3389/fmed.2023.1237432] [Medline: 38020160]
23. Nakao T, Miki S, Nakamura Y, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: evaluation study. JMIR Med Educ 2024 Mar 12;10:e54393. [doi: 10.2196/54393] [Medline: 38470459]
24. Chen CW, Walter P, Wei JCC. Using ChatGPT-like solutions to bridge the communication gap between patients with rheumatoid arthritis and health care professionals. JMIR Med Educ 2024 Feb 27;10:e48989. [doi: 10.2196/48989] [Medline: 38412022]

XSL•FO
RenderX

25. Chen Y, Wu Z, Wang P, et al. Radiology residents' perceptions of artificial intelligence: nationwide cross-sectional survey study. J Med Internet Res 2023 Oct 19;25:e48249. [doi: 10.2196/48249] [Medline: 37856181]
26. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
27. Zhang H, Guan Y, Chen J, Tong W. Corrigendum: commentary: AI-based online chat and the future of oncology care: a promising technology or a solution in search of a problem? Front Oncol 2023;13:1334176. [doi: 10.3389/fonc.2023.1334176] [Medline: 38144532]
28. Pianykh OS, Langs G, Dewey M, et al. Continuous learning AI in radiology: implementation principles and early applications. Radiology 2020 Oct;297(1):6-14. [doi: 10.1148/radiol.2020200038] [Medline: 32840473]
29. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest 2021 Apr;101(4):412-422. [doi: 10.1038/s41374-020-00514-0] [Medline: 33454724]
30. Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur Radiol 2020 Jun;30(6):3576-3584. [doi: 10.1007/s00330-020-06672-5] [Medline: 32064565]
31. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. JMIR Med Educ 2023 Mar 8;9:e46876. [doi: 10.2196/46876] [Medline: 36867743]
32. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510. [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
33. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Front Med (Lausanne) 2020 Feb 5;7:27. [doi: 10.3389/fmed.2020.00027] [Medline: 32118012]
34. Frallicciardi A, Lotterman S, Ledford M, et al. Training for failure: a simulation program for emergency medicine residents to improve communication skills in service recovery. AEM Educ Train 2018 Oct;2(4):277-287. [doi: 10.1002/aet2.10116] [Medline: 30386837]
35. Adnan AI. Effectiveness of communication skills training in medical students using simulated patients or volunteer outpatients. Cureus 2022 Jul;14(7):e26717. [doi: 10.7759/cureus.26717] [Medline: 35967150]
36. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
37. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May;15(5):e38755. [doi: 10.7759/cureus.38755] [Medline: 37303324]
38. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
39. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024 Mar 20;10:e51151. [doi: 10.2196/51151] [Medline: 38506920]
40. Delgado Bolton RC, San-Martín M, Vivanco L. Role of empathy and lifelong learning abilities in physicians and nurses who work in direct contact with patients in adverse working conditions. Int J Environ Res Public Health 2022 Mar 4;19(5):3012. [doi: 10.3390/ijerph19053012] [Medline: 35270702]
41. Ding M, Babenko O, Koppula S, Oswald A, White J. Physicians as teachers and lifelong learners. J Contin Educ Health Prof 2019;39(1):2-6. [doi: 10.1097/CEH.0000000000000228] [Medline: 30394937]
42. Nicolaides A. Considering medical technology use, ethics and litigation. J Med Lab Sci Technol S Afr 2019;1(4):26-34.
43. Ran M, Banes D, Scherer MJ. Basic principles for the development of an AI-based tool for assistive technology decision making. Disabil Rehabil Assist Technol 2022 Oct;17(7):778-781. [doi: 10.1080/17483107.2020.1817163] [Medline: 33275457]
44. Schicktanz S, Welsch J, Schweda M, Hein A, Rieger JW, Kirste T. AI-assisted ethics? Considerations of AI simulation for the ethical assessment and design of assistive technologies. Front Genet 2023;14:1039839. [doi: 10.3389/fgene.2023.1039839] [Medline: 37434952]
45. Toh Y, Hung WLD, Chua PH, He S, Jamaludin A. Pedagogical reforms within a centralised-decentralised system: a Singapore's perspective to diffuse 21st century learning innovations. Int J Educ Manag 2016;30(7):1247-1267. [doi: 10.1108/IJEM-10-2015-0147]
46. Quinn TP, Coghlan S. Readying medical students for medical AI: the need to embed AI ethics education. arXiv. Preprint posted online on Sep 7, 2021. [doi: 10.48550/arXiv.2109.02866]
47. Price WN. 20—medical malpractice and black-box medicine. In: Cohen IG, Lynch HF, Vayena E, Gasser U, editors. Big Data, Health Law, and Bioethics: Cambridge University Press; 2018. [doi: 10.1017/9781108147972.027]
48. Shortliffe EH. Biomedical informatics in the education of physicians. JAMA 2010 Sep 15;304(11):1227-1228. [doi: 10.1001/jama.2010.1262] [Medline: 20841537]
49. Nosek TM, Bond GC, Ginsburg JM, et al. Using computer‐aided instruction (CAI) to promote active learning in the physiology classroom. Ann N Y Acad Sci 1993 Dec;701(1):128-129. [doi: 10.1111/j.1749-6632.1993.tb19792.x]
50. Ovsyanitskaya L, Yurasova E. Information technologies, mechatronics and robotics as a basis of an interdisciplinary approach to engineering and medical education. Вестник Южно-Уральского

XSL•FO
RenderX

государственного университета. Серия: Образование. Педагогические науки. Bulletin of the South Ural State University. Series: Education. Educational Sciences 2015;7(4):101-106. [doi: 10.14529/ped150414]
51. Khalid A, Mehmood A, Alabrah A, et al. Breast cancer detection and prevention using machine learning. Diagnostics (Basel) 2023 Oct 2;13(19):3113. [doi: 10.3390/diagnostics13193113] [Medline: 37835856]
52. Coravos A, Khozin S, Mandl KD. Erratum: author correction: developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2019;2(1):40. [doi: 10.1038/s41746-019-0119-8] [Medline: 31304386]
53. Sverdlov O, van Dam J, Hannesdottir K, Thornton-Wells T. Digital therapeutics: an integral component of digital innovation in drug development. Clin Pharmacol Ther 2018 Jul;104(1):72-80. [doi: 10.1002/cpt.1036] [Medline: 29377057]
54. Conway DL, Chang DA, Jackson JL. I don't think that means what you think it means: why precision in lifelong learning terminology matters to medical education. Med Teach 2022 Jul;44(7):702-706. [doi: 10.1080/0142159X.2022.2055456] [Medline: 35343869]
55. Richardson D, Kinnear B, Hauer KE, et al. Growth mindset in competency-based medical education. Med Teach 2021 Jul;43(7):751-757. [doi: 10.1080/0142159X.2021.1928036] [Medline: 34410891]
56. Bhuria M, Mangalesh S, Dudani S, Malik A. Learning approaches adopted by Indian medical students during distance learning: the revised two-factor study process questionnaire. BLDE Univ J Health Sci 2021 Jul;6(2):150-155. [doi: 10.4103/bjhs.bjhs_104_20]
57. Andryas FN, Irmarahayu A, Bustamam NM. Virtual learning environment and learning approach among pre-clinical undergraduate medical students during COVID-19 pandemic. J Pendidik Kedokt Indones 2022;11(1):10. [doi: 10.22146/jpki.63975]
58. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students’ education: scoping review. JMIR Med Educ 2022 Feb 2;8(1):e34860. [doi: 10.2196/34860] [Medline: 35107421]
59. Leung GM, Johnston JM, Tin KYK, et al. Randomised controlled trial of clinical decision support tools to improve learning of evidence based medicine in medical students. BMJ 2003 Nov 8;327:1090. [doi: 10.1136/bmj.327.7423.1090] [Medline: 14604933]
60. Pucchio A, Rathagirishnan R, Caton N, et al. Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study. BMC Med Educ 2022 Nov 28;22(1):815. [doi: 10.1186/s12909-022-03896-5] [Medline: 36443720]
61. Naseer F, Khan MN, Tahir M, Addas A, Aejaz SMH. Integrating deep learning techniques for personalized learning pathways in higher education. Heliyon 2024 Jun;10(11):e32628. [doi: 10.1016/j.heliyon.2024.e32628]
62. Wu Q, Wang Y, Lu L, Chen Y, Long H, Wang J. Virtual simulation in undergraduate medical education: a scoping review of recent practice. Front Med (Lausanne) 2022;9:855403. [doi: 10.3389/fmed.2022.855403] [Medline: 35433717]
63. Eno C, Piemonte N, Michalec B, et al. Forming physicians: evaluating the opportunities and benefits of structured integration of humanities and ethics into medical education. J Med Humanit 2023 Dec;44(4):503-531. [doi: 10.1007/s10912-023-09812-2] [Medline: 37526858]
64. Maramis WF. Medical humanities in medical schools. Jurnal Widya Medika 2015;3(1):1-10. [doi: 10.33508/jwm.v3i1.763]
65. Will ChatGPT transform healthcare? Nat Med 2023 Mar;29(3):505-506. [doi: 10.1038/s41591-023-02289-5] [Medline: 36918736]
66. D’Haese PF, Finomore V, Lesnik D, et al. Prediction of viral symptoms using wearable technology and artificial intelligence: a pilot study in healthcare workers. PLoS One 2021;16(10):e0257997. [doi: 10.1371/journal.pone.0257997] [Medline: 34648513]
67. Rahman A, Hossain MS, Muhammad G, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. Cluster Comput 2022 Aug 17:1-41. [doi: 10.1007/s10586-022-03658-4] [Medline: 35996680]
68. Ratti E, Graves M. Cultivating moral attention: a virtue-oriented approach to responsible data science in healthcare. Philos Technol 2021 Dec;34(4):1819-1846. [doi: 10.1007/s13347-021-00490-3]
69. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1:40. [doi: 10.1038/s41746-018-0048-y] [Medline: 31304321]
70. Arisa S, Sitinjak DS. Implementation of the STEM-PBL approach in online chemistry learning and its impact on students' critical thinking skills. Jurnal Pendidikan Kimia Indonesia 2022;6(2):88-96. [doi: 10.23887/jpki.v6i2.44317]
71. Setia Permana IPY, Nyeneng IDP, Distrik IW. The effect of science, technology, engineering, and mathematics (STEM) approaches on critical thinking skills using PBL learning models. Berkala Ilmiah Pendidikan Fisika 2021;9(1):1. [doi: 10.20527/bipf.v9i1.9319]
72. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
73. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023 Aug 14;9:e50945. [doi: 10.2196/50945] [Medline: 37578830]

XSL•FO
RenderX

74. Harrow S, Srithar K, Nadarajasundaram A, Mensah A. Collective medical student perspective on the importance of clinical placements in medical education. Med Sci Educ 2021 Dec;31(6):2161. [doi: 10.1007/s40670-021-01424-6] [Medline: 34631238]
75. Greenstone H, Wooding K. “It’s real life, isn’t it?” Integrated simulation teaching in undergraduate psychiatry education—a qualitative study. J Ment Health Train Educ Pract 2021 Aug 31;16(5):341-352. [doi: 10.1108/JMHTEP-09-2020-0067]
76. Greenwald L, Blanchard O, Hayden C, Sheffield P. Climate and health education: a critical review at one medical school. Front Public Health 2022;10:1092359. [doi: 10.3389/fpubh.2022.1092359] [Medline: 36711353]
77. Chengcai T, Zijie Z, Ling J, Limei L. Rural revitalization and high-quality development of culture and tourism: theoretical and empirical research. J Resour Ecol 2024;15(3):521-527. [doi: 10.5814/j.issn.1674-764x.2024.03.001]
78. Ruhiyat R, Saepudin D, Syafrin N, Handrianto B. Modernization of pesantren and graduate quality. Formosa J Multidiscip Res 2024;3(2):275-290. [doi: 10.55927/fjmr.v3i2.8227]
79. Eves J, Sudarsanam A, Shalhoub J, Amiras D. Augmented reality in vascular and endovascular surgery: scoping review. JMIR Serious Games 2022 Sep 23;10(3):e34501. [doi: 10.2196/34501] [Medline: 36149736]
80. Gao J, Yang L, Zou J, Fan X. Comparison of the influence of massive open online courses and traditional teaching methods in medical education in China: a meta-analysis. Biochem Mol Biol Educ 2021 Jul;49(4):639-651. [doi: 10.1002/bmb.21523] [Medline: 33894023]
81. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus 2023 Aug;15(8):e43271. [doi: 10.7759/cureus.43271] [Medline: 37692629]
82. Yang Z, Cao Z, Zhang Y, et al. MABEL: an AI-powered mammographic breast lesion diagnostic system. Presented at: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM); Mar 1 to 2, 2021:; Shenzhen, China p. 1-7. [doi: 10.1109/HEALTHCOM49281.2021.9398982]

Abbreviations

AI: artificial intelligence
PBL: problem-based learning

Edited by TDA Cardoso; submitted 29.04.23; peer-reviewed by N Domingues, W Yang, YD Cheng; revised version received 03.06.24; accepted 11.06.24; published 15.08.24.

Please cite as:
Tong W, Zhang X, Zeng H, Pan J, Gong C, Zhang H
Reforming China’s Secondary Vocational Medical Education: Adapting to the Challenges and Opportunities of the AI Era
JMIR Med Educ 2024;10:e48594
URL: https://mededu.jmir.org/2024/1/e48594
doi:10.2196/48594

© Wenting Tong, Xiaowen Zhang, Haiping Zeng, Jianping Pan, Chao Gong, Hui Zhang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.8.2024. 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.

XSL•FO
RenderX

The Digital Determinants of Health: A Guide for Competency Development in Digital Care Delivery for Health Professions Trainees

Katharine Lawrence, MPH, MD; Defne L Levine, MPH

Department of Population Health, New York University Grossman School of Medicine, 227 East 30th Street 6th Floor, New York, NY, United States

Corresponding Author:
Defne L Levine, MPH

Abstract

Health care delivery is undergoing an accelerated period of digital transformation, spurred in part by the COVID-19 pandemic and the use of "virtual-first" care delivery models such as telemedicine. Medical education has responded to this shift with calls for improved digital health training, but there is as yet no universal understanding of the needed competencies, domains, and best practices for teaching these skills. In this paper, we argue that a "digital determinants of health" (DDoH) framework for understanding the intersections of health outcomes, technology, and training is critical to the development of comprehensive digital health competencies in medical education. Much like current social determinants of health models, the DDoH framework can be integrated into undergraduate, graduate, and professional education to guide training interventions as well as competency development and evaluation. We provide possible approaches to integrating this framework into training programs and explore priorities for future research in digitally-competent medical education.

(JMIR Med Educ 2024;10:e54173) doi:10.2196/54173

KEYWORDS

digital health; digital determinants of health; digital health competencies; medical education curriculum; competency development; digital health education; training competencies; digital health skills; digital care delivery; health professions training

Introduction

The COVID-19 pandemic heralded a transformation in care delivery to virtual services and digital technologies such as telemedicine, remote patient monitoring, and asynchronous patient portal communications. This transition, coupled with the growing field of "Big Data" informatics and generative artificial intelligence ("GenAI"), has reinvigorated enthusiasm in the "digital transformation" of health care [1] and the use of novel digital technologies to provide personalized, convenient, and comprehensive care for all. It has also resulted in calls to improve the "digital health competencies" of clinicians, to help both current health care providers and trainees meet this transformative moment in care delivery [2-4].

Digital health tools—which include a wide range of "virtual" technologies such as telemedicine, remote sensors, and wearables, as well as medical "apps" and eHealth and mobile health tools, digitized health record and communications platforms (electronic health records [EHRs] and patient portals), clinical decision support systems, and personalized and predictive modeling technologies [5]—have been progressively integrated into mechanisms of care delivery over the last decade, with growing support from both patients and clinicians [3,6,7]. Patient empowerment and self-management are factors that contribute to patient use of digital health [8]. In the United States, 93% of physicians believe digital health tools are an advantage for patient care, with the majority citing a desire to provide competent remote care to patients as a significant motivator to adopt digital tools [6].

Among medical trainees, sentiments around use of digital health technologies are similar, with these technologies increasingly becoming inseparable from medical training [9,10]. This trend accelerated during the COVID-19 pandemic, as resident clinics pivoted to telemedicine and training shifted to virtual conferences, e-learning modules, and telesimulation [11,12]; nursing and other allied health professions saw similar shifts in their own education and care delivery experiences [13,14]. This dramatically shifted environment has created an appetite for both learning and teaching digital health skills among medical trainees, while also exposing gaps in current approaches to curricular development, implementation, and evaluation [3,4,15-17]. At the same time, there is growing recognition of the equity risks associated with digital health technology [18,19], particularly as the use of these tools was expanded during the pandemic and disparities in access and proficiency widened existing health care inequalities [18,20-22]. This reality underscores the need to cultivate a health care workforce that is both technically *and* culturally competent, as well as to better integrate health equity efforts in clinical training.

This paper explores the current state of digital health education and training competencies among medical and allied health

XSL•FO
RenderX

professions through a brief narrative review and identifies key limitations in these approaches. We then offer a novel framework—the digital determinants of health (DDoH)—that can help unify and direct ongoing competency development and evaluation efforts. The DDoH framework can also ensure that key equity considerations of digital health are incorporated into trainee competencies, thereby helping reduce disparities associated with this these technologies' use.

Current Digital Health Education and Training

At its core, the challenge of teaching digital health competencies to medical trainees lies in the lack of consensus regarding *what* those competencies are and *how* they should be taught. While several major medical organizations in the United States and internationally have released statements [23-25] regarding some element of digital health competencies at undergraduate, graduate, and professional continuing medical education levels, significant heterogeneity exists in these organizations' definitions, areas of focus, and evaluation tools and metrics (Table 1). A brief narrative review of the current medical literature on the topic of "digital health training" reveals both vagueness and variability in the definition of "digital health," with overrepresentation of language from biomedical informatics, health information technology, and telemedicine. Often, only general recommendations for training competency domains (eg, patient safety and medical knowledge) are offered, rather than any specific competencies. Existing instruments to measure competencies often focus on specific use cases (eg, EHR proficiency) rather than the broad-scope digital health tools and services that exist today [2]. Many instruments are not validated, being either adapted from previously developed tools or newly designed to meet the changing technological landscape and educational needs [2].

XSL•FO
RenderX

Table . Brief narrative review of digital health technology definitions, domains, competencies, and skills.

Source	Digital health definition	Main domains or technologies	Competencies and skills
Accreditation Council for Graduate Medical Education	Not defined: reviewed competencies relevant to digital health but not explicitly digital health specific	Specific domains and technologies are not recognized in core competencies	• Broad competencies encapsulate patient care, medical knowledge, practice-based learning and improvement, systems-based practice, interpersonal and communication, and professionalism. None are specific digital health competencies [26].
American Medical Association (AMA) [27]	Definition: "Digital health encompasses a broad scope of tools that can improve health care, enable lifestyle change and create operational efficiencies" [27]	Digital solutions: telemedicine and telehealth, mHealth[a], wearables, remote monitoring, and apps	• While specific competencies are not outlined, the AMA has been studying, since 2016, physicians' motivations for using digital clinical tools.
Association of American Medical Colleges (AAMC) [23]	Not defined	Telehealth competencies across 6 domains and 3 tiers	• Domains: "Patient safety and appropriate use of telehealth, access and equity in telehealth, communication via telehealth, data collection and assessment via telehealth, technology for telehealth, ethical practices and legal requirements for telehealth." • Competency tiers: "entry to residency or recent medical school graduate, entry to practice or recent residency graduate, experienced faculty physician or three to five years post-residency" [23].
Centers for Disease Control and Prevention (CDC)	Definition: "the systematic application of information and communications technologies, computer science, and data to support informed decision-making by individuals, the health workforce, and health systems, to strengthen resilience to disease and improve health and wellness" [28]	Outlines key competencies for public health professionals [29]	• The CDC uses the 10 essential public health services to guide its competencies. These 10 services do not include digital health–specific competencies [30].
The Standing Committee of European Doctors (CPME) [25]	Not defined: digital competency web page focuses on digital health literacy of health professionals	Calls on members to support investing in eHealth solutions to improve patient care and expand digital health literacy	• The CPME does not list specific digital competencies in this statement but outlines the importance of digital competencies, given the way digital health is transforming medicine and health care.

Royal Australasian College of Physicians (RACP); Scott et al [4]	Definition: digital health encompasses digital systems integrated in health care and "extends beyond electronic storage, retrieval or transmission of data to the active use of these data in quality improvement, service redesign and knowledge development" [4]	Digital systems: EMRs[b], e-ordering, e-prescribing, virtual care, e-messaging, e-consults, clinical decision support, mHealth, remote patient monitoring, and artificial intelligence	• 11 foundational digital competencies in knowledge and understanding outlined over 3 digital health capability horizons:
			  • "Horizon 1: Embedding safe, ethical, and effective use of systems if electronic records
			  • Horizon 2: Integrating new technologies and ways of working
			  • Horizon 3: Digital health transformation" [4].

XSL•FO
RenderX

Source	Digital health definition	Main domains or technologies	Competencies and skills
World Health Organization (WHO)	Definition: within global strategies for digital health, the WHO defines digital health as "the field of knowledge and practice associated with the development and use of digital technologies to improve health" [31]	Domains encompassed: eHealth, advanced computing, big data, and artificial intelligence	• The WHO proposes in their global strategy on digital health (2020 - 2025) to identify core digital health literacy competencies in short term for training of health professionals and ensure that digital health competencies are integrated into education [31].

Longhini et al [17]	Not defined: uses the WHO definition of digital health interventions, "discrete function of digital technology to achieve health care sector objectives" [32]	Digital health competencies including terms related to digital literacy, health informatics, and eHealth	• Four main categories of digital health competencies identified (with subcategories):
			  • Category 1: "self-rated competencies"
			    • Subcategories: "digital literacy," "eHealth literacy," "patient-oriented competencies," and "process of care-oriented competencies."
			  • Category 2: "psychological and emotional aspects towards digital technologies"
			    • Subcategories: "attitudes and beliefs," "confidence," and "awareness"
			  • Category 3: "use of digital technologies"
			    • Subcategory: "general use of digital technologies"
			  • Category 4: "knowledge about digital technologies"

Khurana et al [3]	Definition: "an umbrella term broadly defined as the use of digital technologies for health" and "a means by which to increase the delivery of and access to healthcare" [3]	Domains include EHRs[c], telehealth, mobile and wearable health technology, and artificial intelligence	• A total of 40 topics across 3 subcategories (digital health knowledge, digital health skills, and digital health attitudes) were identified.
Jimenez et al [33]	Definition: "digital health refers to a broad umbrella term encompassing eHealth...broadly defined as "the use of information and communications technology in support of health and health-related fields" as well as emerging areas of advance computing sciences" [33]	Domains include eHealth, genomics, and artificial intelligence	• Identified competency domains rather than competencies. Most prevalent digital health competency domains identified: electronic health/medical records, computer/tablet/app use and internet skills, practice administration/management, health information systems, and information literacy.
van Houwelingen et al [34]	Not defined: presents examples of telehealth and digital care: e-visits, devices for self-measurement, activity monitors, and personal alarms	52 competencies included for consideration: competencies focused on nursing curricula to adequately prepare nurses for the world of telehealth	• 32 competencies were specifically needed for telehealth provision. Competencies were identified and selected for each of the 14 nursing activities the authors included in the study.

Source	Digital health definition	Main domains or technologies	Competencies and skills
Health Information Technology Competencies (HITComp) database [17,35]	Not defined: HITComp does not define digital health but outlines technology competencies for health care professionals	5 competency domains: administration, direct patient care, engineering/information systems/ICT[d], informatics, and research/biomedicine	• 33 areas of competency are listed in the HITComp database, allowing users to select relevant areas. Competencies are defined for each domain. A total of 1025 competencies are included in the database [35].
Kinnuen et al [13]	Definition: authors quote digital health definition, "the field of knowledge and practice associated with the development and use of digital health technologies to improve health" [13,31]	Main competency domains: working in digital environment, nursing documentation, and ethics and data protection. Domains capture technologies for documenting nursing diagnosis, planned care, basic IT skills, and eHealth services	• 3 informatics competencies identified: ethics and data protection, nursing documentation, and digital environment.
Hübner et al [36]	Definition: defines informatics as focusing on data, information, knowledge, and user applications and defines information technology as addressing systems development and life cycle management. Health informatics described as comprised of informatics from multiple disciplines	TIGER[e] core competencies for nursing informatics: 24 core competency areas in nursing and nursing management in health informatics clustered in 6 domains. Questionnaire used by authors included 10 technological items, including eHealth, telematics, and telehealth	• The 6 domains for the TIGER competencies include "data, information, knowledge," "information exchange and information sharing," "ethical and legal issues," "systems life cycle management," "management," and "biostatistics and medical technology." • Results showed the top 10 core competency areas for 5 different roles: clinical nursing, quality management, coordination of interprofessional care, nursing management, and IT management in nursing.

[a]mHealth: mobile health.

[b]EMR: electronic medical record.

[c]EHR: electronic health record.

[d]ICT: information and communication technology.

[e]TIGER: Technology Informatics Guiding Education Reform.

This state of ambiguity has resulted in an uneven and ad hoc approach to digital health education programming in undergraduate and graduate training institutions. Since the pandemic, a growing number of medical schools have implemented digital health courses, consisting mostly of electives focused on biomedical informatics or (more recently) telemedicine [3,37]. Few of these programs are integrated into the larger medicine curriculum, however [38], in part because considerable knowledge gaps remain regarding the most effective ways to integrate them [33]. Even less work has been done at the graduate level, although several novel Objective Structured Clinical Examinations (OSCEs) have been developed to provide "hands-on" training to residents [26,38-40]. Overall, systematized approaches to understanding, defining, and building digital health curriculum for medical trainees are lacking, as are those for faculty development and practicing clinicians [41].

DDoH: A Novel Framework to Advance Digital Health Training and Competency Development

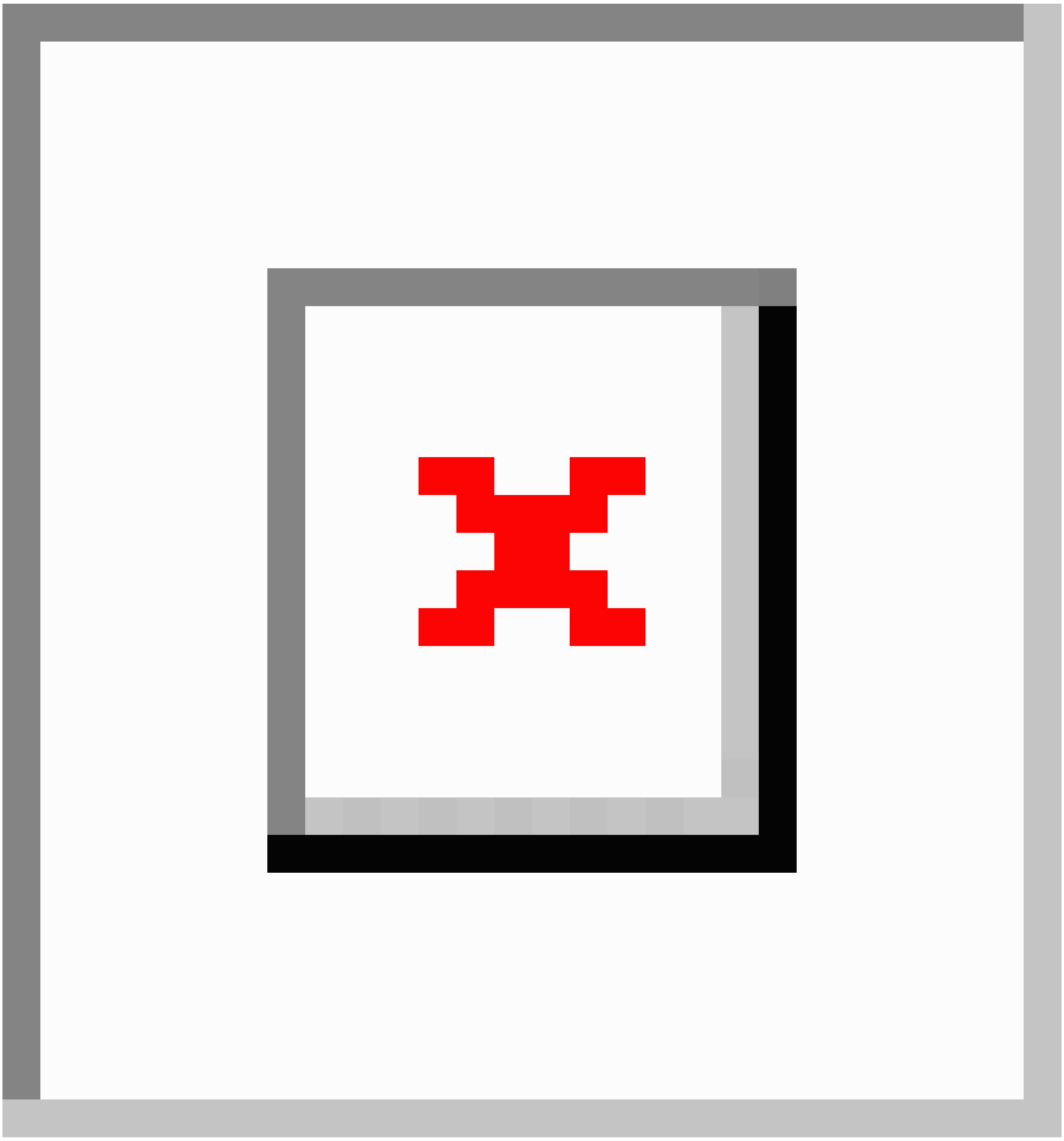
We argue that a comprehensive, multilevel approach to understanding, defining, and evaluating digital health skills for trainees is needed in order to properly prepare health professions students to meet the needs of patients in this new health care landscape. To accomplish this, we offer a model based on our growing understanding of "digital determinants of health"—the novel technological contexts and constructs that mediate an individual or community's interactions with the health care system—and their intersections with care delivery, innovations, education, and equity.

DDoH refer to antecedents within the digital environment that impact a patient's ability to access, use, and satisfactorily experience the health care system. DDoH exist at individual, community, and structural levels [18] and include a patient's personal experiences with digital health technology (eg, use patterns, preferences, and digital skills), communal attitudes (eg, perceptions of usefulness, trust, privacy, and surveillance),

XSL•FO
RenderX

cultural beliefs and social conditions (eg, the digital environments a community experiences, including “digital deserts”), and structural factors (eg, national technology policies, bias, and discrimination; Figure 1). DDoH can act as barriers or facilitators to effective health care and may disproportionately affect certain individuals or communities [18].

Figure 1. Mapping digital determinants of health to social determinants of health.



The DDoH framework is modeled on the well-established social determinants of health (SDoH) framework [42]. The SDoH framework consists of conditions that shape the lived experiences of individuals and environments that impact health [43]. SDoH include both place-based conditions and “political, socioeconomic, and cultural constructs” [43]. Some examples include income level, availability of transportation, and social support and community inclusivity [43]. SDoH affect populations in negative and positive ways and can both protect health and contribute to disparities [44]. Given the developments made in understanding SDoH in the last 2 decades and integrating them in health and medicine, a large body of literature now exists exploring socioeconomic factors and the ways they shape health outcomes [45,46]. As knowledge of SDoH has expanded to health care, it has been integrated into medical curricula, and SDoH training is now considered to be a core piece in medical education [47,48]. The success of SDoH competencies and curricula development can serve as a model for integrating DDoH into medical education. The DDoH model effectively transposes SDoH thinking into digital spaces and challenges us to think beyond individual characteristics (eg, digital health literacy) when considering a person, community,

XSL•FO
RenderX

or population's interactions with digital health tools (Figure 1) [19]. DDoH are valuable to effectively conduct virtual care delivery, which is becoming more prominent in health care today [49].

Why Teach DDoH?

Overall, we argue that digital care delivery cannot be successful or equitable without more attention to the DDoH that define it. Critical to this is the inclusion of DDoH thinking into training paradigms, programs, and resources at all levels.

Specifically, the DDoH framework can help improve digital health competency development through the following:

- *Ensures a standardized and comprehensive approach to curricular design* that would address digital health skill needs at individual, interpersonal, social, and structural levels: for example, when developing training tools to teach and evaluate a learner's ability to assess patient "readiness" for a telemedicine appointment, educators can use the DDoH framework to include not only screening for individual digital health literacy but also an evaluation of community and social factors such as access (eg, the "digital divide"). This approach allows for a better understanding of the specific barriers to an individual patient's use of health technologies, which can then be tailored to better meet that patient's needs. Systematically applying these layers across learning programs creates a shared mental model for digital health training that can unify language, competency domains, and evaluation tools.
- *Can be both technology* specific *and technology* agnostic: this means that the DDoH framework can be useful when developing both specific technical skills as well as universal competencies such as patient communication, education, and shared decision-making—all of which facilitate a patient's broader ability to successfully engage with the digital health ecosystem across devices and services. This can help avoid the ongoing challenge when developing digital health competencies, which is the tendency for technical skills to become outdated as the technologies themselves evolve (eg, computer-based web browsers vs smartphone apps vs smartwatches). For example, in teaching students EHR proficiency, using the DDoH layers can ensure that specific technical competence (eg, being able to log onto and successfully navigate the platform) is matched with interpersonal communication skills (eg, talking to the patient and *not* the EHR) and shared decision-making (eg, EHR screensharing with patients) that will serve learners regardless of the EHR platform they use.
- *Can build on existing successful SDOH-based curriculum and pedagogy,* allowing for more efficient program development and quick adaptation and validation of learning tools, rather than starting from scratch: existing SDoH programming that has been shown to be effective can quickly be adapted to DDoH contexts and tested in similar environments to assess for fidelity and effectiveness. There are now a variety of existing instructional frameworks for SDoH teaching [50] and curricula that are experiential, longitudinal, interprofessional, and community based that can be adapted [50,51]. This can also apply to preexisting SDoH evaluation tools, as well as common program requirements and other educational standards. For example, an undergraduate medical education OSCE designed to teach SDoH related to hypertension management can be quickly adapted to a case involving remote blood pressure monitoring, thereby exposing learners not only to well-known social factors related to hypertension control (eg, regular access to medications) but also unique technology-mediated factors such as access to reliable Wi-Fi for sending home values.
- *Supports digital health equity*: crucially, a DDoH-informed approach incorporates an equity-sensitive perspective into digital health training, by placing drivers of digital health disparities at the center of skills and competency development. Teaching trainees about the potential social biases of a piece of technology alongside the technology itself can help reduce the likelihood of producing clinicians that reinforce technology bias in their practice. This is particularly relevant given the growing literature exposing the relationships between digital health technology and health disparities, as well as the need for a workforce that is trained to address this even as the field expands and these tools become normalized as part of care delivery.

Applying DDoH to Health Professions Training

Overall, health professions educators should use the DDoH framework as a guide in creating robust educational programming and evaluation tools aimed at developing health professionals who understand and can competently use digital health tools to deliver care for diverse patients. Practically, the integration of DDoH in educational programming should leverage a mixed-pedagogical approach that extends beyond passive learning and includes applied learning strategies such as OSCEs and "flipped" classrooms and innovative technologies such as virtual reality. To accomplish this effort, educators can first identify existing spaces in didactic curriculum to infuse DDoH, including adding it to SDoH training. However, a DDoH-based approach can also be taught through problem-based learning, experiential and workplace-based learning, performance assessment, and continuing medical education. Experiential "hands-on training" within community-based and service-learning opportunities (eg, homeless shelters and community advocacy organizations) can imbue technical facility while also educating on social and structural contexts of care using these technologies. In particular, connecting trainees with lay community members such as community health workers or digital navigators can expose them to the common technical skills needed to support patients as well as the social and cultural nuances of a digital health technology's use in the real world. Incorporating trainees into health system IT efforts is another example that can provide unique administrative and regulatory contexts for learners. DDoH-sensitive learning can also complement quality improvement curricula, particularly as those programs increasingly involve EHRs, clinical decision support, informatics, and other digital health tools.

XSL•FO
RenderX

In keeping with a multilevel approach, digital health competency assessments should evaluate skills at technical, interpersonal, and structural levels, and educators applying the DDoH framework should consider stratifying their assessment domains based on these levels. For example, when developing a learning program on remote blood pressure monitoring in hypertension management, consider the following:

1. What *individual technical skills* are needed for both clinicians and patients to successfully install, set up, and transmit remote blood pressure data using currently available technologies?
2. What *immediate individual and/or community social contexts and barriers* might be relevant for patients being considered for a remote blood pressure monitoring program, and how would a clinician evaluate and address those?
3. What *larger national or structural social factors*—including scope of practice and device regulations—might impact a patient's ability to access and use a remote blood pressure monitor, or a clinician's ability to interpret and make medical decisions based on that data?

In this case, depending on the level of learner and training goals, while a teaching session dedicated purely to technical skill building (eg, training clinicians on the variety of remote blood pressure monitors and how to successfully take a home blood pressure measurement) may be the focus, using a DDoH-guided approach would allow educators to increase the value of the training by teaching clinicians to also address individual DDoH needs (eg, language preferences and physical or cognitive accessibility needs) and social layers (eg, local library Wi-Fi access) that may contribute to a blood pressure device's ultimately successful use.

There are some challenges to creating an educational system based on the DDoH framework. Designing robust experiences is time-consuming and often labor-intensive. Consequently, it is important to identify already developed programs that can be quickly adapted and evaluated; this can include existing SDoH programs, but it can also include the myriad of ad hoc telemedicine training tools that proliferated during the COVID -19 pandemic that can now be reworked to be more robust and structured. In general, flexible learning approaches that can respond to the short technology life cycles of many of these products are critical to ensure that skills remain relevant or can be quickly updated; this is challenging to keep on top of and may favor a longitudinal approach that offers multiple teaching touch points throughout a training program. Finding competent faculty to teach these skills may also be difficult, as practicing clinicians and educators are often learning about novel technologies alongside trainees. Finally, convincing health care stakeholders that DDoH are worth studying, learning, and evaluating in their own right take efforts, particularly given the other demands and competing priorities of health training. However, we strongly believe these technologies will only continue to proliferate and become further embedded in health care delivery, and ignoring their outsized and disruptive role in clinical care in critical training periods is ultimately a disservice to the health care workforce and patients.

Summary

There is growing need to develop unified digital health education and training competencies for health professions students. Efforts to cultivate a workforce adept in digital health tools must prioritize understanding and mitigating the digital determinants of health that shape individuals' interactions with health care technology. Using a DDoH framework in medical education—including not only didactic training but also hands-on skill building, as well as continuing education opportunities—can help guide robust educational programming and evaluation tools aimed at developing health professionals who understand and can competently use digital health tools to deliver care for diverse patients.

Acknowledgments

The authors of this paper would like to acknowledge Dr Safiya Richardson and Dr Devin Mann, as well as all the members of the HiBRID Lab at NYU Langone. This work was supported by grants from the Doris Duke Foundation and the National Science Foundation (NSF awards 1928614 and 2129076).

Conflicts of Interest

None declared.

References

1. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407. [doi: 10.3390/ijerph20043407] [Medline: 36834105]
2. Jarva E, Oikarinen A, Andersson J, Tomietto M, Kääriäinen M, Mikkonen K. Healthcare professionals' digital health competence and its core factors; development and psychometric testing of two instruments. Int J Med Inform 2023 Mar;171:104995. [doi: 10.1016/j.ijmedinf.2023.104995] [Medline: 36689840]
3. Khurana MP, Raaschou-Pedersen DE, Kurtzhals J, Bardram JE, Ostrowski SR, Bundgaard JS. Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Med Educ 2022 Feb 26;22(1):129. [doi: 10.1186/s12909-022-03163-7] [Medline: 35216611]
4. Scott IA, Shaw T, Slade C, et al. Digital health competencies for the next generation of physicians. Intern Med J 2023 Jun;53(6):1042-1049. [doi: 10.1111/imj.16122] [Medline: 37323107]

XSL•FO
RenderX

5. Ronquillo Y, Meyers A, Korvek SJ. Digital health. In: StatPearls: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK470260/ [accessed 2024-08-21]
6. Henry TA. 5 Insights into how physicians view, use digital health tools. American Medical Association. 2022 Oct 17. URL: https://www.ama-assn.org/practice-management/digital/5-insights-how-physicians-view-use-digital-health-tools#:~:text=80%25%20of%20physicians%20used%20televisits,use%20was%20at%20in%202019 [accessed 2024-08-21]
7. Diaz N. 96% of US hospitals have EHRs, but barriers remain to interoperability, ONC says. Becker's Healthcare. 2023 Mar 7. URL: https://www.beckershospitalreview.com/ehrs/96-of-us-hospitals-have-ehrs-but-barriers-remain-to-interoperability-onc-says.html#:~:text=As%20of%202021%2C%2096%20percent,implemented%20a%20certified%20EHR%20system [accessed 2024-08-21]
8. Madanian S, Nakarada-Kordic I, Reay S, Chetty T. Patients' perspectives on digital health tools. PEC Innov 2023 May 26;2:100171. [doi: 10.1016/j.pecinn.2023.100171] [Medline: 37384154]
9. Ma M, Li Y, Gao L, et al. The need for digital health education among next-generation health workers in China: a cross-sectional survey on digital health education. BMC Med Educ 2023 Jul 31;23(1):541. [doi: 10.1186/s12909-023-04407-w] [Medline: 37525126]
10. Machleid F, Kaczmarczyk R, Johann D, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827. [doi: 10.2196/19827] [Medline: 32667899]
11. Jeffries PR, Bushardt RL, DuBose-Morris R, et al. The role of technology in health professions education during the COVID-19 pandemic. Acad Med 2022 Mar 1;97(3S):S104-S109. [doi: 10.1097/ACM.0000000000004523] [Medline: 34789662]
12. Jumreornvong O, Yang E, Race J, Appel J. Telemedicine and medical education in the age of COVID-19. Acad Med 2020 Dec;95(12):1838-1843. [doi: 10.1097/ACM.0000000000003711] [Medline: 32889946]
13. Kinnunen UM, Kuusisto A, Koponen S, et al. Nurses' informatics competency assessment of health information system usage: a cross-sectional survey. Comput Inform Nurs 2023 Nov 1;41(11):869-876. [doi: 10.1097/CIN.0000000000001026] [Medline: 37931302]
14. Bouabida K, Lebouché B, Pomey MP. Telehealth and COVID-19 pandemic: an overview of the telehealth use, advantages, challenges, and opportunities during COVID-19 pandemic. Healthcare (Basel) 2022 Nov 16;10(11):2293. [doi: 10.3390/healthcare10112293] [Medline: 36421617]
15. Fatehi F, Samadbeik M, Kazemi A. What is digital health? review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
16. Iyamu I, Xu AXT, Gómez-Ramírez O, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill 2021 Nov 26;7(11):e30399. [doi: 10.2196/30399] [Medline: 34842555]
17. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414. [doi: 10.2196/36414] [Medline: 35980735]
18. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
19. Lawrence K. Digital health equity. In: Linwood SL, editor. Digital Health: Exon Publications; 2022. [doi: 10.36255/exon-publications-digital-health-health-equity]
20. Eruchalu CN, Pichardo MS, Bharadwaj M, et al. The expanding digital divide: digital health access inequities during the COVID-19 pandemic in New York City. J Urban Health 2021 Apr;98(2):183-186. [doi: 10.1007/s11524-020-00508-9] [Medline: 33471281]
21. Haimi M. The tragic paradoxical effect of telemedicine on healthcare disparities- a time for redemption: a narrative review. BMC Med Inform Decis Mak 2023 May 16;23(1):95. [doi: 10.1186/s12911-023-02194-4] [Medline: 37193960]
22. Jaworski BK, Webb Hooper M, Aklin WM, et al. Advancing digital health equity: directions for behavioral and social science research. Transl Behav Med 2023 Apr 3;13(3):132-139. [doi: 10.1093/tbm/ibac088] [Medline: 36318232]
23. Telehealth competencies. Association of American Medical Colleges. URL: https://www.aamc.org/data-reports/report/telehealth-competencies [accessed 2024-08-21]
24. American Board of Telehealth. Evidence-based competencies to build a telehealth education program [Webinar]. American Telemedicine Association. 2021 Aug 5. URL: https://www.americantelemed.org/resources/abtwebinar [accessed 2024-08-21]
25. Kuhn S, Roda S. Digital competencies. Standing Committee of European Doctors (CPME). 2020. URL: https://www.cpme.eu/policies-and-projects/digital-health/digital-competencies [accessed 2024-08-21]
26. Hsiang EY, Ganeshan S, Patel S, Yurkovic A, Parekh A. Training physicians in the digital health era: how to leverage the residency elective. JMIR Med Educ 2023 Jul 14;9:e46752. [doi: 10.2196/46752] [Medline: 37450323]
27. AMA digital health care 2022 study findings. American Medical Association. 2022 Sep 28. URL: https://www.ama-assn.org/about/research/ama-digital-health-care-2022-study-findings [accessed 2024-08-21]
28. Global digital health strategy. Centers for Disease Control and Prevention. 2024 May 15. URL: https://www.cdc.gov/globalhealth/topics/gdhs/index.

XSL•FO
RenderX

html#:~:text=The%20Centers%20for%20Disease%20Control,%2C%20regional%2C%20and%20global%20levels [accessed 2024-08-21]
29. Competencies for public health professionals. Centers for Disease Control and Prevention. 2024 May 16. URL: https://www.cdc.gov/publichealthgateway/professional/competencies.html [accessed 2024-08-21]
30. 10 Essential public health services. Centers for Disease Control and Prevention. 2024 May 16. URL: https://www.cdc.gov/publichealthgateway/publichealthservices/essentialhealthservices.html [accessed 2024-08-21]
31. Global strategy on digital health 2020-2025. World Health Organization. 2021 Aug 18. URL: https://www.who.int/publications/i/item/9789240020924
32. World Health Organization. 2018 Jan 1. URL: https://www.who.int/publications/i/item/WHO-RHR-18.06 [accessed 2024-08-21]
33. Jimenez G, Spinazze P, Matchar D, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020 Nov;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
34. van Houwelingen CTM, Moerman AH, Ettema RGA, Kort HSM, Ten Cate O. Competencies required for nursing telehealth activities: a Delphi-study. Nurse Educ Today 2016 Apr;39:50-62. [doi: 10.1016/j.nedt.2015.12.025] [Medline: 27006033]
35. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 5;22(11):e22706. [doi: 10.2196/22706] [Medline: 33151152]
36. Hübner U, Shaw T, Thye J, et al. Technology Informatics Guiding Education Reform - TIGER. Methods Inf Med 2018 Jun;57(S 01):e30-e42. [doi: 10.3414/ME17-01-0155] [Medline: 29956297]
37. Tudor Car L, Kyaw BM, Nannan Panday RS, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275. [doi: 10.2196/28275] [Medline: 34287206]
38. Aungst TD, Patel R. Integrating digital health into the curriculum-considerations on the current landscape and future developments. J Med Educ Curric Dev 2020 Jan 20;7:2382120519901275. [doi: 10.1177/2382120519901275] [Medline: 32010795]
39. Boardman D, Wilhite JA, Adams J, et al. Telemedicine training in the COVID era: revamping a routine OSCE to prepare medicine residents for virtual care. J Med Educ Curric Dev 2021 Jun 16;8:23821205211024076. [doi: 10.1177/23821205211024076] [Medline: 34189270]
40. Lawrence K, Hanley K, Adams J, Sartori DJ, Greene R, Zabar S. Building telemedicine capacity for trainees during the novel coronavirus outbreak: a case study and lessons learned. J Gen Intern Med 2020 Sep;35(9):2675-2679. [doi: 10.1007/s11606-020-05979-9] [Medline: 32642929]
41. Open call for experts to serve as members of the Digital Health Competency Framework Committee. World Health Organization. 2023 Jan 16. URL: https://www.who.int/news-room/articles-detail/open-call-for-experts-to-serve-as-members-of-the-digital-health-competency-framework-committee [accessed 2024-08-21]
42. Models and frameworks for the practice of community engagement. Agency for Toxic Substances and Disease Registry. 2015 Jun 25. URL: https://www.atsdr.cdc.gov/communityengagement/pce_models.html [accessed 2024-08-21]
43. NEJM Catalyst. Social determinants of health (SDOH). NEJM Catal Innov Care Deliv 2017 Dec 1;3(6) [FREE Full text]
44. National Academy of Medicine, National Academies of Sciences, Engineering, and Medicine, Committee on the Future of Nursing 2020–2030. In: Wakefield MK, Williams DR, Le Menestrel S, Flaubert JL, editors. The Future of Nursing 2020-2030: Charting a Path to Achieve Health Equity: The National Academies Press; 2021. [doi: 10.17226/25982]
45. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Pub Health Rep 2014;129 Suppl 2(Suppl 2):19-31. [doi: 10.1177/00333549141291S206] [Medline: 24385661]
46. McGinnis JM, Williams-Russo P, Knickman JR. The case for more active policy attention to health promotion. Health Aff (Millwood) 2002;21(2):78-93. [doi: 10.1377/hlthaff.21.2.78] [Medline: 11900188]
47. Campbell M, Liveris M, Caruso Brown AE, et al. Assessment and evaluation in social determinants of health education: a national survey of US medical schools and physician assistant programs. J Gen Intern Med 2022 Jul;37(9):2180-2186. [doi: 10.1007/s11606-022-07498-1] [Medline: 35710668]
48. Lewis JH, Lage OG, Grant BK, et al. Addressing the social determinants of health in undergraduate medical education curricula: a survey report. Adv Med Educ Pract 2020 May 22;11:369-377. [doi: 10.2147/AMEP.S243827] [Medline: 32547288]
49. Kickbusch I, Holly L. Addressing the digital determinants of health: health promotion must lead the charge. Health Promot Int 2023 Jun 1;38(3):daad059. [doi: 10.1093/heapro/daad059] [Medline: 37264549]
50. Doobay-Persaud A, Adler MD, Bartell TR, et al. Teaching the social determinants of health in undergraduate medical education: a scoping review. J Gen Intern Med 2019 May;34(5):720-730. [doi: 10.1007/s11606-019-04876-0] [Medline: 30993619]
51. National Academies of Sciences, Engineering, and Medicine, Institute of Medicine, Board on Global Health, Committee on Educating Health Professionals to Address the Social Determinants of Health. A Framework for Educating Health Professionals to Address the Social Determinants of Health: The National Academies Press; 2016. [doi: 10.17226/21923]

XSL•FO
RenderX

Abbreviations

DDoH: digital determinants of health
EHR: electronic health record
OSCE: Objective Structured Clinical Examination
SDoH: social determinants of health

Edited by B Lesselroth; submitted 31.10.23; peer-reviewed by C Lai, J Sharp, J Wherton, W Evans; revised version received 01.04.24; accepted 27.06.24; published 29.08.24.

Please cite as:
Lawrence K, Levine DL
The Digital Determinants of Health: A Guide for Competency Development in Digital Care Delivery for Health Professions Trainees
JMIR Med Educ 2024;10:e54173
URL: https://mededu.jmir.org/2024/1/e54173
doi:10.2196/54173

© Katharine Lawrence, Defne L Levine. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.8.2024. 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.

XSL•FO
RenderX

Artificial Intelligence in Dental Education: Opportunities and Challenges of Large Language Models and Multimodal Foundation Models

Daniel Claman[1,*], DDS; Emre Sezgin[2,3,*], PhD

1

2

3

*all authors contributed equally

Corresponding Author:
Emre Sezgin, PhD

Abstract

Instructional and clinical technologies have been transforming dental education. With the emergence of artificial intelligence (AI), the opportunities of using AI in education has increased. With the recent advancement of generative AI, large language models (LLMs) and foundation models gained attention with their capabilities in natural language understanding and generation as well as combining multiple types of data, such as text, images, and audio. A common example has been ChatGPT, which is based on a powerful LLM—the GPT model. This paper discusses the potential benefits and challenges of incorporating LLMs in dental education, focusing on periodontal charting with a use case to outline capabilities of LLMs. LLMs can provide personalized feedback, generate case scenarios, and create educational content to contribute to the quality of dental education. However, challenges, limitations, and risks exist, including bias and inaccuracy in the content created, privacy and security concerns, and the risk of overreliance. With guidance and oversight, and by effectively and ethically integrating LLMs, dental education can incorporate engaging and personalized learning experiences for students toward readiness for real-life clinical practice.

(JMIR Med Educ 2024;10:e52346) doi:10.2196/52346

KEYWORDS

artificial intelligence; large language models; dental education; GPT; ChatGPT; periodontal health; AI; LLM; LLMs; chatbot; natural language; generative pretrained transformer; innovation; technology; large language model

Introduction

In recent years, dental education has experienced a significant transformation, driven by the rapid evolution of technology [1-3]. Dentistry faculties and educators have recognized the potential of these advancements to enhance the learning experience and guide patient care and have actively integrated them into their curricula [4]. Therefore, this has led to a change in the state of dental education in dental schools, focusing on the incorporation of technology to foster a more effective, engaging, and innovative learning environment. Specifically, the emergence of artificial intelligence (AI) has created a broader impact [1].

Dentistry has always been a highly specialized field, requiring a combination of theoretical knowledge, practical skills, and clinical acumen. Dental faculty have traditionally employed a combination of lectures, seminars, laboratory work, and supervised clinical practice to deliver a comprehensive educational experience to dental students. However, the advent of cutting-edge technology and AI has created new possibilities for improving the quality of education as well as the practice to better prepare future dental professionals [1,5,6]. By leveraging these technological advancements, dental educators are able to create more interactive and personalized learning experiences. Virtual reality, for instance, allows students to immerse themselves in realistic clinical scenarios, enhancing their understanding of complex dental procedures and techniques [7]. Similarly, haptic devices and 3D printing enable the development of accurate dental models, facilitating hands-on practice and improving students' dexterity and confidence in performing intricate procedures [8]. Finally, the addition of advanced medical charting (eg, integrated electronic medical records or voice-activated periodontal charting) to clinical practice has required the dental school faculty to instruct on how best to use technology to provide safe clinical care when in practice [9-11]. Beyond all, AI has been perceived to improve operations, innovation, and practices in dental education at multiple levels [1].

XSL•FO
RenderX

Transformation in Education and Technology With Large Language Models

As dentistry continues to evolve with the integration of advanced technologies, AI has emerged as a powerful tool with new potentials to improve dental education. One such innovation, which has been highly communicated recently, is large language models (LLMs). An LLM is a type of AI model designed to conceptualize and generate human-like text based on large amounts of data [12]. These models are trained on vast amounts of text from various sources online, enabling them to generate contextually relevant responses, summaries, translations, and more. LLMs have been argued to potentially transform various domains, including education, by providing personalized learning experiences and assisting in content creation [13]. Further ahead, multimodal foundation models (FM) are similar large-scale AI models which are pretrained on extensive data, enabling them to conceptualize and generate image and audio, in addition to the text [14].

Currently well-known LLMs and FMs, such as GPT (OpenAI) [15], LaMDA and PaLM (Google) [16], and LLaMA (Meta) [17], have shown potential in medical education and practice, including problem-solving, question and answering, summarization, and content creation [13,18,19]. Especially in dental education, it may provide innovative methods to enhance the learning experience for dental students. Personalized learning could be one, as these models can be used to create unique experiences for students by generating custom learning materials based on their individual needs, preferences, and learning styles [20,21]. In addition, LLMs and FMs can be used for content creation, where it can create educational content such as quizzes, assessments, and lesson plans which in turn can help educators save time and improve the quality of their teaching materials [20]. Therefore, it is important to explore the application of LLMs and FMs in dental education. In this perspective, to take a glimpse at applications in dental education, we share a use case of periodontal charting, and highlight major opportunities and challenges associated with AI implementation.

Use Case: Periodontal Charting

Overview

Periodontal charting, an important component of dental practice and clinical care, involves the systematic recording of information related to a patient's periodontal health, such as probing depths, gingival recession, clinical attachment levels, and the presence of bleeding or suppuration. Accurate periodontal charting is essential for diagnosis, treatment planning, and monitoring the progress of periodontal therapy. Periodontal health is an important part of the dental school curriculum, and ultimately a significant component of clinical practice. Dental students are asked to complete numerous competency examinations on the assessment and treatment of periodontal disease. Additionally, periodontal assessment and treatment is a critical component of dental licensure examinations. Integrating generative models (eg, LLMs and FMs) into the teaching and learning process of periodontal charting offers several opportunities to improve students' understanding and mastery of this important clinical skill [22].

One major opportunity is to provide personalized feedback and guidance. By inputting the students' charting data and observations, a model can analyze the information, compare it with established guidelines and best practices [23], and generate tailored feedback. This feedback can highlight errors, suggest improvements, and reinforce correct techniques, supporting students in refining their charting skills and enhancing their clinical decision-making capabilities. Another opportunity is the creation of realistic case scenarios and simulations. LLMs and FMs can generate a wide range of patient cases with varying periodontal conditions, enabling students to practice periodontal charting in diverse clinical contexts. This exposure to a multitude of cases can foster deeper comprehension of the underlying principles and techniques, and better prepare students for real-life clinical practice.

However, there are also challenges associated with the integration of these models into dental education. One primary concern is the potential for biases or inaccuracies in the AI-generated content. As these models rely on the data they have been trained on, they may inadvertently reproduce erroneous information or reflect biases present in their training data [18]. To mitigate this issue, it is crucial for educators to carefully review and validate the content generated, ensuring that it aligns with current evidence-based practices and professional standards. Yet, validation might be challenging with tools such as ChatGPT, which do not provide the reference to the source of the information presented. Another challenge is the potential overreliance on AI by students, which may lead to a decreased emphasis on critical thinking and problem-solving skills [24]. To address this concern, educators must keep a balance between leveraging the benefits of AI and fostering independent learning. Encouraging students to critically evaluate the feedback and suggestions provided by AI can promote the development of essential clinical reasoning skills [24,25].

To demonstrate the capability of these generative AI (LLM or FM) models in dental education, we present a series of examples of LLM or FM applications as a periodontal charting educational support mechanism.

Example 1: Periodontal Charting Feedback

In this example, a student enters a prompt to obtain feedback about performed periodontal charting. The prompt and response are presented in Textbox 1. We used the GPT-4 model (via ChatGPT application) [26] to generate this example in March 2023.

XSL•FO
RenderX

Textbox 1. Periodontal charting prompt and response by GPT-4.

Prompt:

I performed periodontal charting on a patient and recorded the following data for tooth #3:

- *Probing depths: 2 mm (mesial), 3 mm (buccal), 2 mm (distal)*
- *Gingival recession: 0 mm (mesial), 1 mm (buccal), 0 mm (distal)*
- *Bleeding on probing: absent (mesial), present (buccal), absent (distal)*
- *Suppuration: absent*

Response:

Based on the information provided, it seems that the periodontal charting for tooth #3 is fairly comprehensive. The probing depths, gingival recession, bleeding on probing, and suppuration have all been recorded, which are important aspects of periodontal charting.

However, there are a few aspects that could be improved or clarified for a more complete assessment:

1. *Six-point probing: To provide a comprehensive picture of the periodontal health around tooth #3, it's important to record probing depths for all six points (mesial-buccal, mid-buccal, distal-buccal, mesial-lingual, mid-lingual, and distal-lingual) rather than just three points (mesial, buccal, and distal). Consider adding the missing data for the lingual side of the tooth.*
2. *Consistency in recording: Ensure that all aspects of the charting (probing depths, gingival recession, bleeding on probing, and suppuration) are documented for each of the six points, if applicable. This will provide a more accurate representation of the periodontal health of tooth #3.*
3. *Mobility and furcation involvement: While not mentioned in the provided data, it is also essential to assess and record tooth mobility and furcation involvement, if present. These factors can have a significant impact on the periodontal prognosis and treatment planning.*
4. *Use of standard notation: Make sure to use standard notation when recording periodontal charting data. This will help to ensure that the information is clear and easily understood by other dental professionals who may be involved in the patient's care.*

Overall, the charting seems to be accurate for the data provided, but additional details are necessary for a complete periodontal assessment of tooth #3.

Example 2: Periodontal Disease Classification

In this example, a student enters a prompt to obtain feedback about periodontal disease assessment. The prompt and response are presented in Textbox 2 (AI-based periodontal health assessment from Figure 1 [27]). The image was captioned as "Upper jaw acute gingivitis in a non-smoker 26-year-old male patient. Abundant plaque deposit is visible on the surfaces of the teeth" [27]. We used the GPT-4 model (via ChatGPT application) [26] to generate this example in January 2024.

Textbox 2. Artifical intelligence–based periodontal health assessment.

Prompt:

What can you say about the periodontal health of these teeth?

Response:

Based on the provided image, there are several clinical signs indicative of periodontal health issues:

Gingival Inflammation: The gums appear to be red and swollen, particularly around the necks of the teeth. This is suggestive of gingival inflammation, which could be due to gingivitis or the initial stages of periodontitis.

Gingival Recession: There seems to be some recession of the gums, especially noticeable around the left central incisor. This could indicate a loss of periodontal attachment.

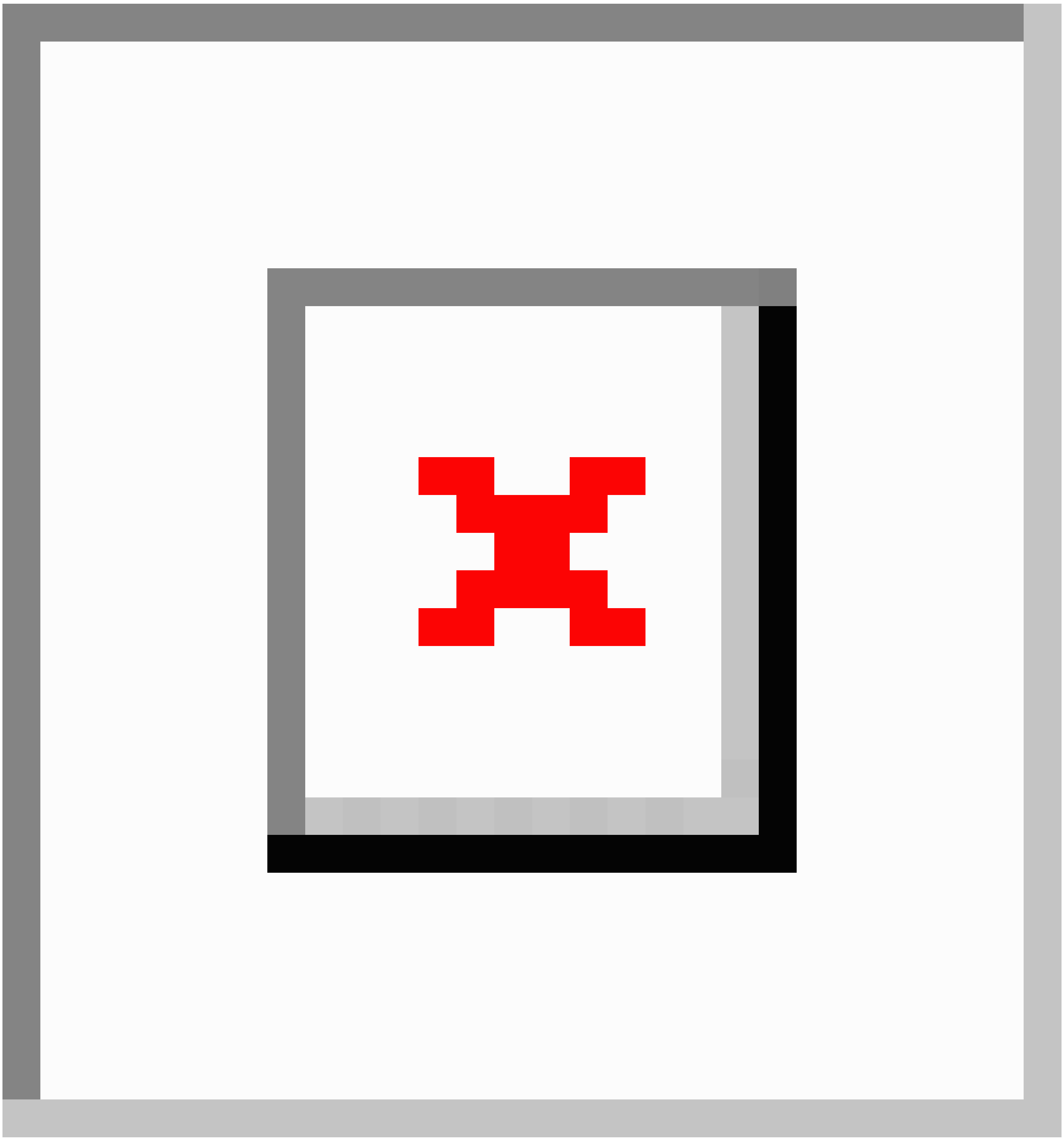
Plaque Accumulation: There are visible yellowish deposits near the gum line of the teeth, which likely represent plaque. Plaque accumulation is a primary factor in the development of both gingivitis and periodontitis.

Tooth Discoloration: There is also evidence of tooth discoloration, which may be related to dietary habits, oral hygiene practices, or possibly the presence of calculus (tartar), which is hardened plaque that can form both above and below the gum line. Gingival Margin Contours: The contour of the gingival margin appears to be irregular, which can be a sign of periodontal disease progression.

It's important to note that a definitive diagnosis would require a full periodontal examination, including periodontal probing to measure pocket depths, assessment of bleeding on probing, mobility testing, and possibly radiographs to evaluate bone levels. This image provides a snapshot that suggests the presence of periodontal issues, but a comprehensive evaluation by a dental professional is necessary for an accurate diagnosis and to formulate an appropriate treatment plan.

XSL•FO
RenderX

Figure 1. Image of upper jaw acute gingivitis (reproduced from Cafiero et al [27], which is published under Creative Commons Attribution 4.0 International License [28])



Example 3: Estimating Periodontal Pocket Depth

In this example, a student enters a prompt to obtain feedback about the periodontal pocket depth (PPD) estimations. The prompt and response are presented in Textbox 3 (AI-based PPD assessment from Figure 2 [27]). The image was captioned as "A periodontal probe is inserted into the sulcus and in a parallel position with respect to the long axis of the tooth. The physiological value of PPD is considered to be ≤3 mm. PPD allows an immediate evaluation of diseased sites" (Textbox 3 and Figure 2) [27]. We used the GPT-4 model (via ChatGPT application) [26] to generate this example in January 2024.

The response provided by GPT-4 (as a multimodal FM) demonstrates a natural language understanding and image recognition for periodontal charting and questions. The readers should note that these examples do not provide the validity or accuracy of the model but rather a demonstration of its capability. However, in terms of the accuracy of exemplified cases, the response by GPT-4 is in line with general dental knowledge and practices.

XSL•FO
RenderX

Textbox 3. Artificial intelligence–based periodontal pocket depth (PPD) assessment.

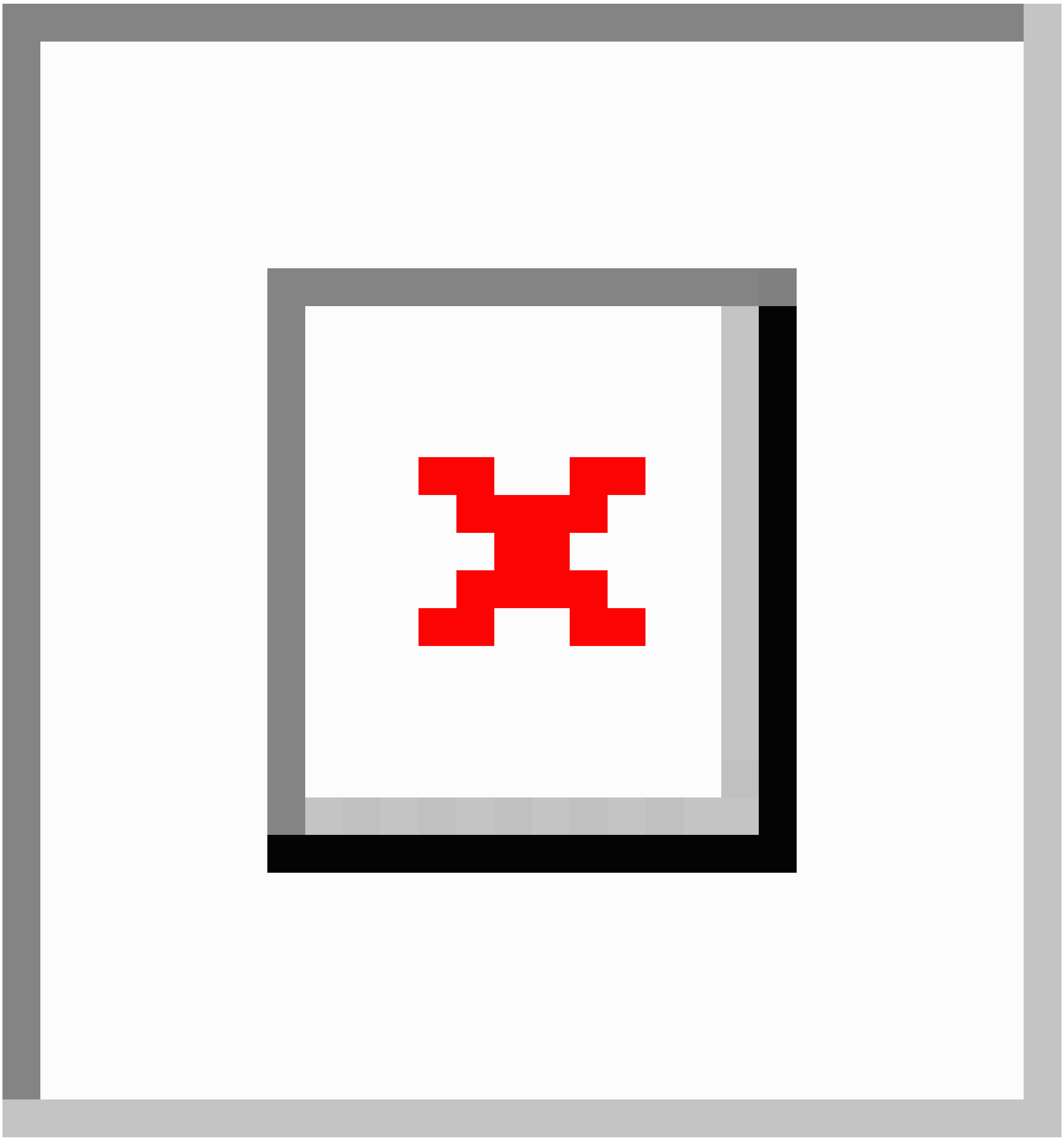
Prompt:

Here is the probe for your reference (Figure 2A). Now estimate the PPD in the next image (Figure 2B).

Response:

Now, with the reference image of the periodontal probe provided, I can give an estimate of the periodontal pocket depth (PPD) in this image. The probe's tip appears to be between the 3 mm and 4 mm marks, as it does not reach the first black band (which I assume, based on common probe designs, indicates 5 mm). It is not clear how much of the probe is within the sulcus due to the angle of the photo and the overlap of the tooth. However, it seems that the probe may be inserted approximately 3 mm into the sulcus. Therefore, the estimated PPD in this image appears to be about 3 mm, which is within the normal physiological range. It's important to note that this is an estimation and that in clinical practice, measurements should be taken from multiple angles and at several points around each tooth for accuracy.

Figure 2. (A) Image of a periodontal probe; (B) image of a periodontal probe inserted into the sulcus (reproduced from Cafiero et al [27] which is published under Creative Commons Attribution 4.0 International License [28]).



XSL•FO
RenderX

In example 1, GPT-4 offers a critique of the data entered by the student. The response identifies areas where the charting is accurate and comprehensive and suggests areas where the charting could be improved or clarified. The feedback highlights the importance of 6-point probing, consistency in recording, mobility and furcation involvement, and the use of standard notation. These aspects are essential for dental students to learn and incorporate into their practice, as they will help to document an accurate and comprehensive periodontal assessment that will be used to enhance patient care. The suggestions for improvement, such as including readings from the lingual aspect of the tooth and ensuring consistency in recording, are valid. From an educational standpoint, the response offers a structured and informative critique that could be beneficial for dental students. The feedback emphasizes the importance of thorough periodontal assessments, which is essential for optimal patient care and treatment planning. Additionally, the response encourages the use of standard notation, which ensures clear communication among dental professionals.

In example 2, GPT-4 offers an evaluation of the given clinical image. The response evaluates the image, providing clinical details present on the image. The feedback highlights the presence of inflammation, plaque accumulation around the gingival margin, and discoloration of the teeth. This evaluation can be helpful in guiding dental students on what to clinically evaluate for periodontal and dental health and can demonstrate examples with analysis of healthy and unhealthy gingiva and dentition. There is a tendency for the model to extrapolate clinical findings from a simple image evaluation; however, the model does caution that a clinical examination should be completed to properly diagnose and evaluate. This response, generally, can be helpful in guiding dental students to properly evaluate and examine patients in the clinical setting.

In example 3, GPT-4 offers an evaluation of the depth indicated on the periodontal probing instrument. The feedback indicates the correct reading, and the appropriate justification for this analysis, while also providing information about common probes and how the probe indicators are arranged. This is helpful for dental students to understand common instrumentation and how to properly read these instruments in the clinical setting. From an educational standpoint, the response offers information on instrumentation from which students can learn. The feedback appropriately evaluates the information, gives context, and indicates the need for a full examination to corroborate findings.

Challenges and Considerations

While the LLM and FM responses provide valuable feedback and highlight areas for improvement within the given use case on periodontal charting, it is essential to consider the potential challenges, risks and limitations associated with this approach on a broader scale.

Publicly available generative AI tools (such as GPT, LaMDA, and LLaMA) are not trained specifically on the dental domain. As probabilistic models, they predict how likely a particular sequence is to occur in the language based on a training data set and they reflect this. Therefore, they may generate different contents for very similar questions depending on the structure of the questions (Textbox 4). This inconsistency could also be the case with repetition of the same prompts with the same model [29]. This may not be ideal for standardization of the educational contents. In some occasions, it may create misleading, biased, or incorrect results, such as responses irrelevant to the question or prompt or responses note based on existing data for the prompt [18]. It is important to be aware that the information generated by the AI may contain such inaccuracies or biases, necessitating a thorough review, experiments, and creation of safeguards or operational guidelines by educators or clinicians to ensure alignment with evidence-based practices and ethical principles [30]. In addition, fine-tuning and advanced prompt engineering (eg, chain-of-thought or in-context learning) could be adopted to mitigate risks [31,32]. Unlike LLMs, FMs are multimodal and more versatile but may require significant fine-tuning with complex domain-specific data to mitigate the risk of generating generalized content that may not align with specialized dental knowledge or practice standards. This adjustment is crucial to tailor the outputs of FMs to reflect up-to-date clinical guidelines and diagnostic criteria specific to dentistry.

Textbox 4. Differences in basic question answering (created by GPT-4 in March 2023). Authors asked GPT-4 to explain periodontal charting in 2 different question forms. The explanations in both responses are slightly different. Whereas response 1 defines periodontal charting as a dental diagnostic and monitoring tool, response 2 defines this as a procedure. Similarly, response 1 addresses the measuring tool as a "specialized probe," whereas response 2 addresses this as a "periodontal probe."

Prompt 1: *What is periodontal charting?*

Response 1: *Periodontal charting is a dental diagnostic and monitoring tool used by dentists and dental hygienists to evaluate the health of a patient's gums and teeth. During periodontal charting, the dental professional measures the depth of the space between the gum line and the tooth root, known as the periodontal pocket, using a specialized probe…*

Prompt 2: *Tell me about periodontal charting.*

Response 2: *Periodontal charting is a procedure performed by dental professionals to evaluate the health of a patient's gums and teeth. It involves measuring the depth of the spaces between the gum line and the tooth root, known as periodontal pockets, using a periodontal probe…*

Regardless of the level of accuracy achieved by technical improvement, while adopting these tools, educators and students must be encouraged to critically assess the feedback provided by the AI to develop their clinical reasoning and problem-solving skills, rather than solely relying on the model's output. Otherwise, there is a risk that students might become overly reliant on AI for decision-making, potentially undermining the development of their independent clinical judgment and manual skills. Similarly, the limitation of the models on detailed tasks, such as critical appraisal of literature,

XSL•FO
RenderX

may further contribute to adverse outcomes, where educators consider adapting teaching and assessment methods to leverage AI's benefits while mitigating risks such as academic dishonesty [21].

It is crucial to design educational programs that balance the use of AI with traditional hands-on and problem-solving training to ensure that students remain adept at both using technology and performing without it. This practice may further necessitate cultural and contextual specific considerations for dental practices in diverse environments, regarding regional differences in dental conditions, treatment preferences, and public health guidelines.

Furthermore, privacy and security of personal health information (PHI) are important to consider. Dental education often involves the use of patient data, including medical histories, diagnostic images, and clinical findings. When using LLMs and FMs in this setting, it is essential to ensure that PHI is not included and that the use of AI has been discussed and approved by the institutions in which they are being used. The entered information (including text and image) should be stripped of any identifying information (and images should be checked for not violating copyright laws) before being input into the AI model to prevent potential privacy breaches, especially with publicly accessible LLM and FMs, which are loosely governed or regulated. These models, as a dental education tool, ideally should be hosted on secure platforms with robust encryption and access controls to prevent unauthorized access and data breaches. Some institutions may provide secure cloud services via compliant service providers (eg, Microsoft Azure, Amazon Web Services, and Google), which may ensure a more private ecosystem for AI use. Various regulations govern the handling of PHI and the use in health care, such as the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in the European Union [33]. These regulations set forth strict requirements for the management of PHI, including data privacy, security, and patient rights. In addition, dental education institutions using such AI models must ensure compliance with the relevant regulations in their jurisdiction to avoid legal repercussions and maintain the trust of patients and the dental community (eg, the recently proposed California AI accountability act necessitates transparency by requiring agencies to disclose interaction with AI and to conduct risk assessment before AI adoption) [34]. Table 1 outlines current challenges and strategies to address them in dental education.

XSL•FO
RenderX

Table . Strategies to address challenges with LLMs[a] and FMs[b] in dental education.

Category and strategy		Details
Bias and inaccuracy mitigation		
	Specialized training data sets and knowledge base	Use data sets compiled from a diverse range of dental texts, research papers, and case studies to train the LLM or FM or for use as part of the knowledge base, ensuring they cover various dental specialties and scenarios.
	Continuous clinical validation	Regularly validate LLM or FM outputs against current dental practices and standards by engaging with dental boards or professional groups.
	Domain-specific fine-tuning and guided prompt engineering	Work with dental faculty and practicing dentists to tailor the LLM or FM outputs to reflect up-to-date clinical guidelines and diagnostic criteria. In addition, use guided prompt engineering and alternative approaches (eg, chain-of-thought) to improve outputs.
Operational guidelines		
	Curriculum integration guidelines	Develop specific guidelines on how LLM or FM integrates into different parts of the dental curriculum, such as diagnostics, treatment planning, and patient communication.
	Professional oversight	Set up a committee of dental professionals to oversee the implementation and use of LLM or FM, ensuring alignment with educational outcomes and clinical accuracy.
Enhancing student interaction		
	Simulation-based learning	Incorporate LLM or FM into simulation settings where students can interact with AI[c] to diagnose and treat virtual patients, enhancing their practical skills without risk.
	Reflective practice sessions	Facilitate sessions where students reflect on the AI’s suggestions compared to standard treatment protocols, promoting critical thinking and decision-making skills.
Privacy and security		
	Scenario-based training	Train students and staff on handling personal health information through scenario-based exercises, ensuring they understand how to manage data securely when using LLM or FM in dental settings.
	Enhanced encryption for dental data	Implement higher levels of encryption and security measures for platforms hosting dental data to ensure compliance and safeguard against breaches.
Regulatory compliance		
	Tailored compliance workshops	Hold workshops focused on the specific legal requirements related to using LLM or FM with PHI in the dental field, such as Health Insurance Portability and Accountability Act in the United States and General Data Protection Regulation in Europe.
	Ethical use guidelines	Develop ethical guidelines that address the nuances of using AI in dental training and practice, including issues of patient consent and AI transparency.
Feedback and continuous improvement		

XSL•FO
RenderX

Category and strategy		Details
	Feedback system for clinical use	Establish a structured feedback system where dental students and professionals can report inaccuracies or ethical concerns with AI outputs, facilitating continuous improvement.

[a]LLM: large language model.

[b]FM: foundation model.

[c]AI: artificial intelligence.

Conclusions

The integration of LLMs and FMs into dental education holds promising opportunities for improving the quality of education and better preparing future dental professionals. By navigating the challenges and leveraging the potential benefits, dental educators can create more interactive, personalized, and innovative learning experiences that effectively prepare students for the complex and evolving world of dental practice. These considerations are also applicable for patient education and self-care practices as well. Considering the accessibility of these models to the public, educational considerations could be further expanded for patient education. Future works are suggested on gathering empirical evidence for the feasibility and utility of LLMs or FMs, including alternative prompt engineering approaches, fine-tuned custom model testing, user testing, cost-benefit analysis, and expanding AI guidelines for including generative AI use in dental education.

Acknowledgments

We used the generative AI tool, ChatGPT by OpenAI [35], to draft example cases (Textbox 1-Textbox 4).

Authors' Contributions

ES conceived the idea. ES and DC contributed to ideation and planning, contributed equally in drafting and finalizing this paper, and reviewed and approved the final paper.

Conflicts of Interest

ES is an associate editor in the editorial board of *Journal of Medical Internet Research* at the time of this publication.

References

1. Islam NM, Laughter L, Sadid-Zadeh R, et al. Adopting artificial intelligence in dental education: a model for academic leadership and innovation. J Dent Educ 2022 Nov;86(11):1545-1551. [doi: 10.1002/jdd.13010] [Medline: 35781809]
2. Arevalo CR, Bayne SC, Beeley JA, et al. Framework for e-learning assessment in dental education: a global model for the future. J Dent Educ 2013 May;77(5):564-575. [Medline: 23658401]
3. Blue C, Henson H. Millennials and dental education: utilizing educational technology for effective teaching. J Dent Hyg 2015 Feb;89 Suppl 1:46-47. [Medline: 25691028]
4. Schwendicke F, Chaurasia A, Wiegand T, et al. Artificial intelligence for oral and dental healthcare: core education curriculum. J Dent 2023 Jan;128:104363. [doi: 10.1016/j.jdent.2022.104363] [Medline: 36410581]
5. Agrawal P, Nikhade P. Artificial intelligence in dentistry: past, present, and future. Cureus 2022 Jul;14(7):e27405. [doi: 10.7759/cureus.27405] [Medline: 36046326]
6. Shan T, Tay FR, Gu L. Application of artificial intelligence in dentistry. J Dent Res 2021 Mar;100(3):232-244. [doi: 10.1177/0022034520969115] [Medline: 33118431]
7. Roy E, Bakr MM, George R. The need for virtual reality simulators in dental education: a review. Saudi Dent J 2017 Apr;29(2):41-47. [doi: 10.1016/j.sdentj.2017.02.001] [Medline: 28490842]
8. Chaudhari PK, Dhillon H, Dhingra K, Alam MK. 3D printing for fostering better dental education. Evid Based Dent 2021 Dec;22(4):154-155. [doi: 10.1038/s41432-021-0217-8] [Medline: 34916647]
9. Nagy M, Hanzlicek P, Zvarova J, et al. Voice-controlled data entry in dental electronic health record. Stud Health Technol Inform 2008;136:529-534. [Medline: 18487785]
10. Virdee J, Thakrar I, Shah R, Koshal S. Going electronic: an epic move. Br Dent J 2022 Jul;233(1):55-58. [doi: 10.1038/s41415-022-4404-6] [Medline: 35804132]
11. Sirrianni J, Sezgin E, Claman D, Linwood SL. Medical text prediction and suggestion using generative pretrained transformer models with dental medical notes. Methods Inf Med 2022 Dec;61(5-06):195-200. [doi: 10.1055/a-1900-7351] [Medline: 35835447]
12. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems 33 (NeurIPS 2020): Neural Information Processing

XSL•FO
RenderX

Systems Foundation, Inc. (NeurIPS); 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html [accessed 2024-08-28]
13. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
14. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv. Preprint posted online on Jul 12, 2022 URL: http://arxiv.org/abs/2108.07258 [accessed 2024-09-17] [doi: 10.48550/arXiv.2108.07258]
15. GPT-4. OpenAI. URL: https://openai.com/product/gpt-4 [accessed 2023-04-25]
16. Collins E, Ghahramani Z. LaMDA: our breakthrough conversation technology. Google AI Blog. 2021. URL: https://blog.google/technology/ai/lamda/ [accessed 2024-08-30]
17. Introducing LLaMA: a foundational, 65-billion-parameter large language model. Meta. 2023 Feb 24. URL: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/ [accessed 2023-04-25]
18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
19. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022 Feb 10;10(2):e32875. [doi: 10.2196/32875] [Medline: 35142635]
20. Thorat VA, Rao P, Joshi N, Talreja P, Shetty A. The role of chatbot GPT technology in undergraduate dental education. Cureus 2024 Feb;16(2):e54193. [doi: 10.7759/cureus.54193] [Medline: 38496058]
21. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-a double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ 2024 Feb;28(1):206-211. [doi: 10.1111/eje.12937] [Medline: 37550893]
22. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2024 Jan 31;10:e51344. [doi: 10.2196/51344] [Medline: 38111256]
23. Versaci MB. ADA releases report on AI in dentistry. ADA News. 2023 Feb 24. URL: https://adanews.ada.org/ada-news/2023/february/ada-releases-report-on-ai-in-dentistry/ [accessed 2024-05-10]
24. Benítez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inform Assoc 2024 Feb 16;31(3):776-783. [doi: 10.1093/jamia/ocad252] [Medline: 38269644]
25. Buldur M, Sezer B. Evaluating the accuracy of Chat Generative Pre-Trained Transformer Version 4 (ChatGPT-4) responses to United States Food and Drug Administration (FDA) frequently asked questions about dental amalgam. BMC Oral Health 2024 May 24;24(1):605. [doi: 10.1186/s12903-024-04358-8] [Medline: 38789962]
26. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-04-26]
27. Cafiero C, Matarasso S. Predictive, preventive, personalised and participatory periodontology: "the 5Ps age" has already started. EPMA J 2013 Jun 14;4(1):16. [doi: 10.1186/1878-5085-4-16] [Medline: 23763842]
28. Attribution 4.0 international (CC BY 4.0). creative commons. Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-09-25]
29. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024 Feb;7(1). [doi: 10.1038/s41746-024-01029-4]
30. Shafique U, Chaudhry US, Towbin AJ. Are the pilots onboard? Equipping radiologists for clinical implementation of AI. J Digit Imaging 2023 Dec;36(6):2329-2334. [doi: 10.1007/s10278-023-00892-z] [Medline: 37556028]
31. Woo B, Huynh T, Tang A, Bui N, Nguyen G, Tam W. Transforming nursing with large language models: from concept to practice. Eur J Cardiovasc Nurs 2024 Jul 19;23(5):549-552. [doi: 10.1093/eurjcn/zvad120] [Medline: 38178303]
32. Schulhoff S, Ilie M, Balepur N, et al. The prompt report: a systematic survey of prompting techniques. arXiv. Preprint posted online on Jul 15, 2024 URL: http://arxiv.org/abs/2406.06608 [accessed 2024-09-17] [doi: 10.48550/arXiv.2406.06608]
33. Tovino SA. The HIPAA Privacy Rule and the EU GDPR: illustrative comparisons. Seton Hall Law Rev 2017;47(4):973-993. [Medline: 28820562]
34. Folks A. Checking in on proposed California privacy and AI legislation. International Association of Privacy Professionals. 2024 Mar 20. URL: https://iapp.org/news/a/checking-in-on-proposed-california-privacy-and-ai-legislation/ [accessed 2024-05-10]
35. ChatGPT. OpenAI. URL: https://openai.com/chatgpt/ [accessed 2024-06-13]

Abbreviations

AI: artificial intelligence
FM: foundation model
LLM: large language model
PHI: personal health information
PPD: periodontal pocket depth

XSL•FO
RenderX

Edited by P Kanzow, TDA Cardoso; submitted 05.09.23; peer-reviewed by D Chrimes, H Feng, JJ Beunza, L Krüdewagen, M Pang, S Markham; revised version received 19.06.24; accepted 19.06.24; published 27.09.24.

Please cite as:
Claman D, Sezgin E
Artificial Intelligence in Dental Education: Opportunities and Challenges of Large Language Models and Multimodal Foundation Models
JMIR Med Educ 2024;10:e52346
URL: https://mededu.jmir.org/2024/1/e52346
doi:10.2196/52346

© Daniel Claman, Emre Sezgin. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.9.2024. 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.

XSL•FO
RenderX

Transforming the Future of Digital Health Education: Redesign of a Graduate Program Using Competency Mapping

Michelle Mun[1,2], DDS; Sonia Chanchlani[1,3], MPH, MD; Kayley Lyons[1], PhD; Kathleen Gray[1], PhD

1

2

3

Corresponding Author:
Michelle Mun, DDS

Abstract

Digital transformation has disrupted many industries but is yet to revolutionize health care. Educational programs must be aligned with the reality that goes beyond developing individuals in their own professions, professionals wishing to make an impact in digital health will need a multidisciplinary understanding of how business models, organizational processes, stakeholder relationships, and workforce dynamics across the health care ecosystem may be disrupted by digital health technology. This paper describes the redesign of an existing postgraduate program, ensuring that core digital health content is relevant, pedagogically sound, and evidence-based, and that the program provides learning and practical application of concepts of the digital transformation of health. Existing subjects were mapped to the American Medical Informatics Association Clinical Informatics Core Competencies, followed by consultation with leadership to further identify gaps or opportunities to revise the course structure. New additions of core and elective subjects were proposed to align with the competencies. Suitable electives were chosen based on stakeholder feedback and a review of subjects in fields relevant to digital transformation of health. The program was revised with a new title, course overview, course intended learning outcomes, reorganizing of core subjects, and approval of new electives, adding to a suite of professional development offerings and forming a structured pathway to further qualification. Programs in digital health must move beyond purely informatics-based competencies toward enabling transformational change. Postgraduate program development in this field is possible within a short time frame with the use of established competency frameworks and expert and student consultation.

(JMIR Med Educ 2024;10:e54112) doi:10.2196/54112

KEYWORDS

digital health; digital transformation; health care; clinical informatics; competencies; graduate education

Introduction

In contrast to simple digitization of processes, digital transformation describes the “comprehensive reorientation of an industry, including its business models, due to the coming of age of digital technologies” [1]. In health care, digital technologies have attracted considerable investment for their potential to reduce costs, improve patient experience, and clinician and system efficiency [2]. However, potential digital health interventions can experience “pilotitis” as innovators and health systems can lack the reciprocal clarity of roles and processes to be able to successfully design, launch, and scale a robust product [3]. This fragmentation explains the observation that while diverse innovative digital health interventions have proliferated in the last 50 years [4], the hope for transformational change and increased value-add of health systems has not yet been delivered [5].

Alongside global recognition of the importance of digital health [6], the domains of digital health and health informatics have become areas of increasing focus for education and workforce development. In Australia, the newly published National Digital Health Capability Action Plan (CAP) and institutional education strategies have a significant role to play in building digital health capability across the health workforce [7-9]. Across a 7-year roadmap, the CAP has outlined priority actions including the development of specialist digital health career pathways, specialist digital health courses, and continuing professional development opportunities for clinicians, informaticians, service management, and related roles in the health sector.

However, the digital transformation of health care cannot be driven by 1 sector alone. For this reason, innovation centers have been established globally in recent years to facilitate collaboration between all stakeholders involved in digital health [3]. In Australia, the University of Melbourne runs multidisciplinary digital health programs through The Center for Digital Transformation of Health, established in 2019 with the vision of “connecting digital innovation to health” [10]. The center sits within the Faculty of Medicine, Dentistry and Health Sciences and operates in conjunction with the School of Computing and Information Systems, Faculty of Engineering

XSL•FO
RenderX

and IT. In 2023, the authors of this paper were commissioned for 3 months to redesign and adapt the existing Graduate Certificate in Health Informatics and Digital Health to meet contemporary national and international digital health standards and align with key center vision and mission objectives.

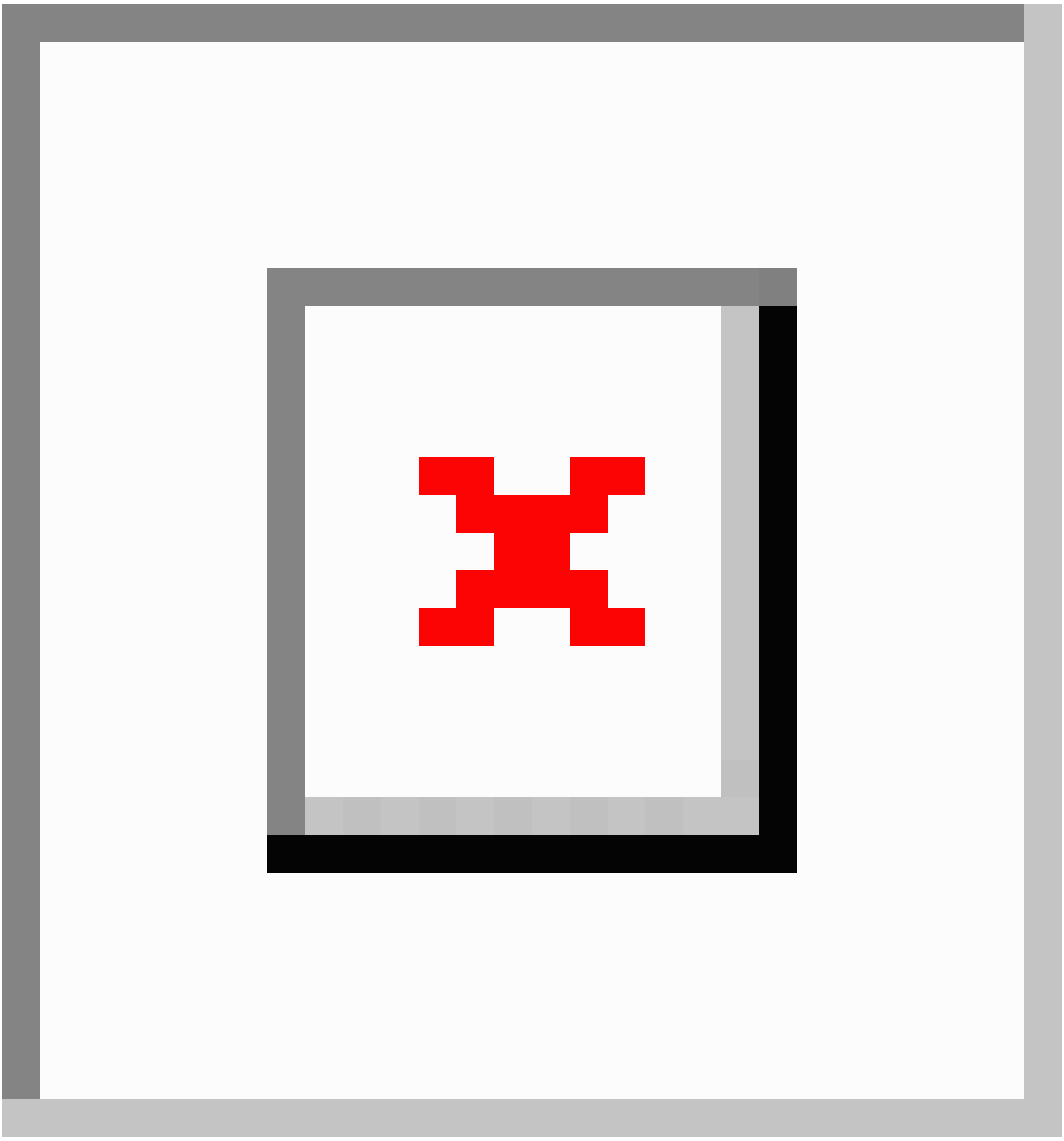
This viewpoint describes the realignment of the existing Graduate Certificate in Health Informatics and Digital Health with the philosophy of "digital transformation," building on the internationally recognized American Medical Informatics Association (AMIA) clinical informatics competency framework [11,12]. The methodology may be used as a blueprint to aid the development of future digital health programs.

Methods

Market Context

The market needs analyses and student feedback informed subsequent stages of competency mapping and course structure review (Figure 1).

Figure 1. Process for evaluation and redesign of the digital health postgraduate program. AMIA: American Medical Informatics Association; HEIMS: Higher Education Information Management System.



A market scan was conducted using data from the Higher Education Information Management System, from the University of Melbourne Department of Education and Training. Over the past 5 years, universities across Australia have seen increased interest in students pursuing higher professional certification in digital health. Many new postgraduate offerings in digital

XSL•FO
RenderX

health, eHealth, health information management, and health and clinical informatics have been created, with 60 - 70 new enrollments in graduate certificates offered in Queensland and Victoria in 2023 alone [13].

Several commissioned needs assessments, market analysis reports, and student reflections within course evaluations were undertaken by the center between 2021 and 2023 to assess future directions for offering digital health professional development, degree, or award programs. The analyses confirmed the challenges of the changing clinical landscape and increased enrollment trends in national graduate certificates in digital health, as well as the presence of an emerging market for people entering or transitioning into the field of digital health. In October 2020, some students (n=14, 20% response rate) enrolled in the University of Melbourne Master of Information Systems Health specialization and the graduate certificate indicated that they were looking for new digital health work opportunities (10 of 14, 71.4% response rate). The survey respondents recommended stronger education in key entry-to-practice degrees and incentives for continued professional development programs.

A recurring theme indicated current clinical career pathways in digital health are not dependent on formal professional certification. Of the 699 advertised digital health–related jobs found over a 3-month period, between October 12, 2020, and January 18, 2021, of which 130 positions were advertised in the state of Victoria, there was significant variation in the range of qualifications, as well as the specialized knowledge and skills relevant to digital health. Short course participant data also highlighted that clinicians are more inclined to consider a graduate certificate if it helped progress their career, with top subjects of interest in domains of data science and artificial intelligence ("machine learning, artificial intelligence and big data" and "data analytics, data linkage, power BI and R"), and the development, implementation, or evaluation of digital health interventions.

Key quotes from participants included "I think you need some ability to put skills learnt into practice in a guided way to really make an impact" and "Theory is just more theory and many of us are way beyond that. Those wanting to do this course would be wanting to make a change, not theorise about it."

In summary, and with consideration to the accelerating pace of digital health technology development, it was anticipated that the observations heralded ongoing and increasing interest in this field.

Competency Mapping

The University of Melbourne's Graduate Certificate in Health Informatics and Digital Health, offered at the postgraduate level, or Australia Qualifications Framework level 8, sits above the professional certificate, and upon completion can be streamlined into a relevant master's degree. The existing certificate consisted of 3 core subjects in biostatistics, health informatics methods, and critical thinking with analytics, alongside limited elective subject options.

To ensure the certificate had core alignment with international informatics standards, SC and MM examined the learning outcomes in subjects in the existing certificate and compared them to the AMIA Core Clinical Informatics Competencies. The Australian Health Informatics Competency Framework (AHICF), developed by the national peak body for informatics, the Australasian Institute of Digital Health (AIDH) [14] was also examined, and syllabi of comparable national and international graduate certificates, sourced from institutional websites, were considered for completeness. The initial mapping phase consisted of a direct comparison of subject-level learning outcomes with statement-level competencies in the AMIA framework. Subsequently, the results were summarized and presented to an expert panel of subject coordinators and center leadership to confirm the accuracy of the mapping and identify further gaps and opportunities.

Results from the competency mapping and the panel interview were used to inform the decision about whether there were existing University of Melbourne subjects that could be used to complement missing competencies, or if there was an opportunity to develop a new subject in digital health, or both. Subjects in the certificate should also be included if they could be accredited toward a relevant master's degree, should the certificate student choose to continue with a higher qualification.

Course Learning Outcomes and Course Structure Redesign

Course-level learning outcomes were developed with consideration to (1) core clinical informatics competencies and (2) desired skills and knowledge beyond clinical informatics, that could equip a graduate to navigate the digital transformation of health care. The latter was derived from course participant feedback, as outlined, as well as key concepts from digital transformation literature:

- "A multi-stakeholders perspective (which) is critical to understanding properly how, in practice, the various players of a (healthcare) ecosystem (patients, pharmaceutical companies, hospitals, public agencies, and many more) exploit (digital transformation) technologies and means to quality of care, value creation, and many more managerial issues" [15]
- Reducing roadblocks that may slow innovation in health systems, including "aligning cross-departmental stakeholders (information technology, security, risk, legal, etc)" [3]
- Introducing the concept of learning health systems, in which "science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the care process, patients and families as active participants in all elements, and new knowledge is captured as an integral by-product of the care experience [16]."

Core subjects were selected from the results of the competency mapping, and their subject-level learning outcomes were reviewed to ensure accordance with the course-level learning outcomes.

Electives for Digital Transformation of Health

The University Handbook, an online catalog of courses and programs, was searched for subjects that could be suitable for

XSL•FO
RenderX

inclusion in the certificate as elective subjects. Inclusion criteria included subjects in domains relevant to data science and statistics, product development, business, leadership and change management, consumer participation in health care, research methods, sustainability, or ethics or legal subjects. A total of 79 subjects were identified. Upon closer consideration, 50 were excluded for the reasons of limited relevance to digital health; required pre-requisite subjects that could not be taken within the constraints of the graduate certificate; discontinued subject; or timing or delivery of the subject not suitable (eg, not semester-long or on-campus only, whereas the certificate required hybrid or online subjects).

As it was intended that the graduate certificate could be used to build a career pathway in digital health, the remaining subjects were assessed for their suitability to scaffold toward a master's pathway. Coordinators of eligible subjects were contacted to determine their availability and interest to be part of the new graduate certificate. After approvals were received, subjects were chosen for their eligibility to form a pathway from professional certificate, graduate certificate, to master's pathways in public health, information systems, and clinical Research.

Results

The mapping and interview phases revealed that many core informatics competencies were already covered within the existing certificate, but there were opportunities to include advanced content about data science, machine learning, and artificial intelligence; the development, implementation, and evaluation of digital health interventions, digital transformation of health care systems, and indigenous data governance. These topics have become increasing areas of interest and debate in recent years, and their need for inclusion in the certificate was evident if the certificate was to both align with recognized standards and be relevant to the modern digital health landscape.

Additionally, the mapping process highlighted the need for a more structured approach to the certificate design, which the center was in a position to provide given the depth and diversity of expertise available.

The new certificate consists of 2 core subjects that were previously electives, that had been identified as achieving the most comprehensive range of AMIA Core Clinical Informatics Competencies (Multimedia Appendix 1). Further mapping to AHICF competencies confirmed alignment with national clinical informatics competencies (Multimedia Appendix 2). The Applied Learning Health Systems subject additionally had a strong focus on practical multidisciplinary learning for the digital transformation of health. The certificate title, course overview, and course-level learning outcomes were updated to align with the skills and competency requirements of the changing market (Textbox 1).

The proposed course-level learning outcomes could now map upstream to the 5 AMIA Clinical Informatics domains (Textbox 2, Table 1) and downstream to subject-level intended learning outcomes of the 2 core subjects (Table 2). Considering the close alignment of the AMIA and AHICF frameworks, for simplicity, the table shows mapping to AMIA competencies only.

A total of 13 electives were identified that would complement the core subjects and allow participants the flexibility to build knowledge and skills in 1 of the self-identified areas of data science and statistics, product development, business, leadership and change management, consumer participation in health care, research methods, sustainability, or ethico-legal contexts in digital health (Table 3). Students would be able to choose 2 elective subjects to complement the core subjects.

To reflect the center's vision of translating digital health innovations into clinical practice, the final metamorphosis of the course included a strategic title change from Graduate Certificate of Health Informatics and Digital Health to Graduate Certificate in Digital Transformation of Health.

Textbox 1. Comparison of previous and newly developed course-intended learning outcomes.

Previous course-level intended learning outcomes

On completion of this course, graduates will be able to:

- Communicate knowledgeably about core health and biomedical informatics concepts, tools and methods, and methods.
- Critically evaluate approaches to information systems and information technology in contemporary health care in Australia and internationally.
- Develop an integrated understanding of how digital data, information, and knowledge are generated and managed for clinical care, biomedical research, public health, and health policy and planning.

New course-level intended learning outcomes

On completion of this course, graduates will be able to:

- Describe how contemporary digital health technologies can be integrated into health care practice in terms of their effect on safety and quality, access and equity, continuity of care, effectiveness, and consumer empowerment.
- Critically evaluate the generation, governance and use of digital data, information and knowledge, including legal and ethical considerations, in the context of electronic health records, clinical decision support systems, virtual care, mobile health, and machine learning and artificial intelligence applications in health.
- Apply the concept of a learning health system and processes of problem assessment, data analysis, design thinking, implementation science, and evaluation frameworks to digital health initiatives in specific contexts.
- Apply principles of governance, leadership, change management, and strategic planning to integrate digital health initiatives and innovation within organizations, across communities and health care systems.

XSL•FO
RenderX

Textbox 2. The American Medical Informatics Association clinical informatics competency domains.

- Fundamentals
- Improving care delivery and outcomes
- Enterprise information systems
- Data governance and data analytics
- Leadership and professionalism

Table . Relationship of clinical informatics competencies and new course-level intended learning outcomes.

Course-level intended learning outcomes (CILO)	Relevance to AMIA[a] domains
CILO1: Describe how contemporary digital health technologies can be integrated into health care practice in terms of their effect on safety and quality, access and equity, continuity of care, effectiveness, and consumer empowerment.	AMIA1, AMIA2, and AMIA3
CILO2: Critically evaluate the generation, governance, and use of digital data, information, and knowledge, including legal and ethical considerations, in the context of electronic health records, clinical decision support systems, virtual care, mobile health, and machine learning and artificial intelligence applications in health.	AMIA2, AMIA3, and AMIA4
CILO3: Apply the concept of a learning health system and processes of problem assessment, data analysis, design thinking, implementation science, and evaluation frameworks to digital health initiatives in specific contexts.	AMIA1 and AMIA2
CILO4: Apply the principles of governance, leadership, change management, and strategic planning to integrate digital health initiatives and innovation within organizations, across communities, and health care systems.	AMIA5

[a]AMIA: American Medical Informatics Association.

XSL•FO
RenderX

Table . Relationship of new course-level and subject-level intended learning outcomes.

Core subject title and subject-level intended learning outcomes (SILO)		Relevance to course-level intended learning outcomes (CILO)
Digital transformation of health		
	SILO1: Explain complex aspects of the structure of health care, including the roles of patients, various professionals, insurance companies and governments.	CILO1
	SILO2: Describe the implications of the generation and the use of biomedical data, information, and knowledge within a variety of relevant systems and settings.	CILO2
	SILO3: Demonstrate the understanding of how core digital health technologies work, through practical activities with simulations of tools such as electronic health records, clinical decision support systems, patient portals, and mobile apps and wearable sensors.	CILO1 and CILO2
	SILO4: Critically analyze how various digital technologies can optimize information use within health care and summarize the potential risks associated with these solutions.	CILO1
	SILO5: Apply ethical frameworks and conceptual models to critique contemporary digital health practices and trends.	CILO1
Applied learning health systems		
	SILO1: Appraise emerging trends and approaches in digital health and informatics.	CILO1 and CILO2
	SILO2: Illustrate how concepts of the learning health system can be applied to your current workplace and role.	CILO3
	SILO3: Outline potential activities in a learning health system project starting with data access and analysis—through designing a virtual care model—and ending with evaluation, implementation, and transformation.	CILO1, CILO3, and CILO4
	SILO4: Create a proposal for a learning health systems (LHS) project that could be implemented at your current or future workplace, which applies digitally enabled LHS concepts.	CILO3 and CILO4

XSL•FO
RenderX

Table . Elective subjects for digital transformation of health.

Elective title	Elective overview
Digital health informatics methods	Overview of major health informatics research areas and methods that contribute to quality improvement, scientific research, and technological innovation in health care and biomedicine.
Biostatistics	Introduction to the fundamental concepts of statistics and the essential methods required to equip students to perform basic statistical analyses and interpret research findings in the public health setting.
Digital health for consumers	Explores wise use of consumer health technologies through dimensions of consumer digital health literacy, global consumer health technology marketplace, lived experiences of active users, and scenarios where consumers are partners in designing and using digitally enabled learning health systems.
Leading health care change for impact	Examines strategies for leading change in clinical settings and health care organizations.
Technology and aging	Examines ways in which recent technological advancements can revolutionize the experience, management, and future of aging.
Health care environment evaluation	Explores the complex, dynamic, interdisciplinary, and multipurpose nature of health care environments focusing on key dimensions of physical workspaces design, virtual work-spaces, and leadership and management practices.
Introduction to programming	Introduction to the fundamental concepts of computer programming and how to solve simple problems using high-level procedural language, with a specific emphasis on data manipulation, transformation, and visualization of data.
Law and emerging health technologies	Examine ways in which law is affecting, and being affected by, the latest advances in medical technology, including genetic, big data analytics, regenerative, therapeutic, artificial intelligence, and reproductive technologies.
Innovation and emerging technologies	Introduction to innovative and contemporary technology that has been recently developed and is currently used in clinical practice and research for the purposes of measurement, diagnosis, and prescription.
Sustainability and health care	Explores the need to urgently formulate adaptation and mitigation strategies, thereby addressing the global climate change emergency, through the lens of sustainable health care.
Natural language processing	Learn computational methods for working with text, in the form of natural language understanding and language generation to develop an understanding of the main algorithms used in natural language processing.
Machine learning applications for health	Introduction to different artificial intelligence applications in health, using different clinical data sources and computational techniques.
Indigenous data governance in health	Provides an overview of the scope of Indigenous data including governance, ethical health research, knowledge translation and evaluation, institutions, and data collections.

Discussion

Principal Results

The redesigned Graduate Certificate in Digital Transformation of Health occurs in the context of increasing awareness of the need to develop a digitally capable health workforce [7-9]. However, broader industry trends show that the digital health sector also contains diverse careers in data analysis, informatics, and application development, with early and mid-career professionals from clinical care, health management, and technology sectors keen for interactive, practical, and interdisciplinary learning [17]. Although the certificate was redesigned with clinical informatics competencies in mind, a high proportion of its students are likely to be from nonclinical backgrounds or nonphysician careers, based on the demographics of previous enrollments.

It was, therefore, the aim of the certificate to align with industry trends and provide flexibility for any professional interested in digital health transformation, not just clinicians, to tailor their learning. This was achieved by mapping the competency frameworks, which revealed 2 subjects that could be used as core subjects, allowing the certificate to be condensed from 3 to 2 cores and creating room for 2 elective subjects. The new “2 core plus 2 elective” structure allows the professional certificate to form a path to a graduate certificate and also allows a range of elective choices that are unique to the certificate. The newly identified electives were not only chosen based on participant feedback and framework gaps but also are aimed to

XSL•FO
RenderX

empower health professionals to lead change within our complex adaptive health system. Newly identified electives with a focus on innovation and emerging technologies, indigenous data governance, sustainable health care, and machine learning applications will facilitate the next transformative phase within the industry.

The new title aligns with existing course offerings including the Professional Certificate of Digital Transformation of Health and organizational strategic vision. The program's content also aligns with new industry trends strengthening management and leadership skills in the core Applied Learning Health Systems subject to allow for a spectrum of digital health pathways. This is critical as although many traditional degrees have set curricula and linear career pathways, learners in digital health come with vast differences in career backgrounds, qualification levels, expertise, amount of work experience, and intrinsic motivations for joining digital health. Consequently, they may end up applying their knowledge and skills to any part of the digital health ecosystem.

Limitations

The multitude of certifications and career pathways in digital health reflects the demand from professionals to enter this pathway. However, the complexity of developing a program in this rapidly progressing field cannot be overstated. This viewpoint describes the practical needs assessment and redevelopment of a graduate program within the time constraints of an institutional schedule. Consultation for this project was informed by reports that included student evaluations but mainly occurred at the faculty educators and executive level. Given the diverse demographics of digital health learners, there is scope to continue co-design the program with past and potential future certificate students, as well as other major stakeholders in digital health such as consumer advocates.

Future work will focus on evaluating the acceptability and effectiveness of the structure and content of the certificate to these stakeholders. A robust evaluation process, modeled on the Kirkpatrick framework, is already in place within 1 core subject [18] and forms the model for evaluation for other subjects in the certificate. In addition, all University subjects undergo continuous evaluation using Student Learning Surveys, in accordance with University standards and processes [19]. As comprehensive program evaluation is critical for its long-term success, assessment of further approaches such as the creation of logic models [20] is underway.

Mapping the previous program to clinical informatics competencies and student feedback was efficient in reaching this point. Challenges ahead lie in maintaining the currency of the learning content and ongoing evaluation and improvement of the effectiveness of its curriculum, learning activities, and assessments. The program will continue to be reviewed against the progress of the CAP, the evolution of the digital health landscape in Australia, and insights from international colleagues and organizations. The next steps will include the development of a decision matrix to aid the prioritization and co-design of new subjects.

Conclusions

A systematic refinement of this postgraduate program has been conducted to align with the center's vision of digital innovation and transformation of health care. Through strategic alignment, competency mapping, and a pedagogical ethos, the transformed graduate certificate aspires to make a substantial impact on the evolving health care ecosystem.

Acknowledgments

The authors sincerely thank Wendy Chapman, Meredith Layton, Daniel Capurro, Brian Chapman, Sathana Dushyanthen, Elizabeth Dent, and Gouri Ligam for their input and assistance in progressing this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Mapping of core subjects to American Medical Informatics Association competencies.

[DOCX File, 36 KB - mededu_v10i1e54112_app1.docx]

Multimedia Appendix 2

Mapping of core subjects to Australian Health Informatics Competency Framework competencies.

[DOCX File, 30 KB - mededu_v10i1e54112_app2.docx]

References

1. Herrmann M, Boehme P, Mondritzki T, Ehlers JP, Kavadias S, Truebel H. Digital transformation and disruption of the health care sector: internet-based observational study. J Med Internet Res 2018 Mar 27;20(3):e104. [doi: 10.2196/jmir.9498] [Medline: 29588274]
2. Ostrovsky A, Barnett M. Accelerating change: fostering innovation in healthcare delivery at academic medical centers. Healthc (Amst) 2014 Mar;2(1):9-13. [doi: 10.1016/j.hjdsi.2013.12.001] [Medline: 26250082]

XSL•FO
RenderX

3. Tseng J, Samagh S, Fraser D, Landman AB. Catalyzing healthcare transformation with digital health: performance indicators and lessons learned from a digital health innovation group. Healthc (Amst) 2018 Jun;6(2):150-155. [doi: 10.1016/j.hjdsi.2017.09.003] [Medline: 28958850]
4. Marques ICP, Ferreira JJM. Digital transformation in the area of health: systematic review of 45 years of evolution. Health Technol 2020 May;10(3):575-586. [doi: 10.1007/s12553-019-00402-8]
5. Perakslis ED. Strategies for delivering value from digital technology transformation. Nat Rev Drug Discov 2017 Feb;16(2):71-72. [doi: 10.1038/nrd.2016.265] [Medline: 28082744]
6. Global strategy on digital health 2020–2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2023-10-16]
7. National digital health capability action plan. Australian Digital Health Agency. 2022. URL: https://www.digitalhealth.gov.au/sites/default/files/documents/national-digital-health-capability-action-plan.pdf [accessed 2023-10-16]
8. Advancing health 2030 strategy. The University of Melbourne. 2022. URL: https://mdhs.unimelb.edu.au/__data/assets/pdf_file/0012/4193868/Advancing-Health-2030-Strategy.pdf [accessed 2023-10-16]
9. Brommeyer M, Liang Z. A systematic approach in developing management workforce readiness for digital health transformation in healthcare. Int J Environ Res Public Health 2022 Oct 25;19(21):13843. [doi: 10.3390/ijerph192113843] [Medline: 36360722]
10. Centre for Digital Transformation of Health. The University of Melbourne. 2023. URL: https://mdhs.unimelb.edu.au/digitalhealth/ [accessed 2023-10-16]
11. Kulikowski CA, Shortliffe EH, Currie LM, et al. AMIA board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. J Am Med Inform Assoc 2012;19(6):931-938. [doi: 10.1136/amiajnl-2012-001053] [Medline: 22683918]
12. Valenta AL, Berner ES, Boren SA, et al. AMIA board white paper: AMIA 2017 core competencies for applied health informatics education at the master's degree level. J Am Med Inform Assoc 2018 Dec 1;25(12):1657-1668. [doi: 10.1093/jamia/ocy132] [Medline: 30371862]
13. The higher education information management system. Australian Government. 2023. URL: https://admin.heims.education.gov.au/Admin/Controller.aspx [accessed 2023-02-20]
14. Australian health informatics competency framework for health informaticians, 2nd ed. AIDH. 2022. URL: https://digitalhealth.org.au/wp-content/uploads/2022/06/AHICFCompetencyFramework.pdf [accessed 2024-02-22]
15. Kraus S, Schiavone F, Pluzhnikova A, Invernizzi AC. Digital transformation in healthcare: analyzing the current state-of-research. J Bus Res 2021 Feb;123:557-567. [doi: 10.1016/j.jbusres.2020.10.030]
16. Friedman CP. What is unique about learning health systems? Learn Health Syst 2022 Jul;6(3):e10328. [doi: 10.1002/lrh2.10328]
17. Dushyanthen S, Perrier M, Chapman W, Layton M, Lyons K. Fostering the use of learning health systems through a fellowship program for interprofessional clinicians. Learn Health Syst 2022 Oct;6(4):e10340. [doi: 10.1002/lrh2.10340] [Medline: 36263261]
18. Dushyanthen S, Choo D, Perrier M, et al. Designing an interprofessional online course to foster learning health systems. Stud Health Technol Inform 2024 Jan 25;310:1241-1245. [doi: 10.3233/SHTI231163] [Medline: 38270013]
19. Student learning surveys: governance and management. The University of Melbourne. 2021. URL: https://www.unimelb.edu.au/sls/governance-and-management/ [accessed 2024-03-13]
20. Hayes H, Parchman ML, Howard R. A logic model framework for evaluation and planning in a primary care practice-based research network (PBRN). J Am Board Fam Med 2011;24(5):576-582. [doi: 10.3122/jabfm.2011.05.110043] [Medline: 21900441]

Abbreviations

AHICF: Australian Health Informatics Competency Framework
AIDH: Australasian Institute of Digital Health
AMIA: American Medical Informatics Association
CAP: National Digital Health Capability Action Plan

Edited by B Lesselroth; submitted 30.10.23; peer-reviewed by M Brommeyer, TH Liu, W LaMendola; revised version received 13.03.24; accepted 20.06.24; published 31.10.24.

Please cite as:
Mun M, Chanchlani S, Lyons K, Gray K
Transforming the Future of Digital Health Education: Redesign of a Graduate Program Using Competency Mapping
JMIR Med Educ 2024;10:e54112
URL: https://mededu.jmir.org/2024/1/e54112
doi:10.2196/54112

XSL•FO
RenderX

© Michelle Mun, Sonia Chanchlani, Kayley Lyons, Kathleen Gray. Originally published in JMIR Medical Education (https://mededu.jmir.org), 31.10.2024. 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.

XSL•FO
RenderX

The Potential of Artificial Intelligence Tools for Reducing Uncertainty in Medicine and Directions for Medical Education

Sauliha Rabia Alli[1], BSc, MD; Soaad Qahhār Hossain[2,3], BSc; Sunit Das[4,5], MD, PhD; Ross Upshur[6], BA, MSc, MA, MD

1

2

3

4

5

6

Corresponding Author:
Soaad Qahhār Hossain, BSc

Abstract

In the field of medicine, uncertainty is inherent. Physicians are asked to make decisions on a daily basis without complete certainty, whether it is in understanding the patient's problem, performing the physical examination, interpreting the findings of diagnostic tests, or proposing a management plan. The reasons for this uncertainty are widespread, including the lack of knowledge about the patient, individual physician limitations, and the limited predictive power of objective diagnostic tools. This uncertainty poses significant problems in providing competent patient care. Research efforts and teaching are attempts to reduce uncertainty that have now become inherent to medicine. Despite this, uncertainty is rampant. Artificial intelligence (AI) tools, which are being rapidly developed and integrated into practice, may change the way we navigate uncertainty. In their strongest forms, AI tools may have the ability to improve data collection on diseases, patient beliefs, values, and preferences, thereby allowing more time for physician-patient communication. By using methods not previously considered, these tools hold the potential to reduce the uncertainty in medicine, such as those arising due to the lack of clinical information and provider skill and bias. Despite this possibility, there has been considerable resistance to the implementation of AI tools in medical practice. In this viewpoint article, we discuss the impact of AI on medical uncertainty and discuss practical approaches to teaching the use of AI tools in medical schools and residency training programs, including AI ethics, real-world skills, and technological aptitude.

(JMIR Med Educ 2024;10:e51446) doi:10.2196/51446

KEYWORDS

artificial intelligence; machine learning; uncertainty; clinical decision-making; medical education; generative AI; generative artificial intelligence

Introduction

In clinical practice, uncertainty refers to a physician's perceived inability to accurately explain or advise on a patient's medical problem [1] and may arise at any stage of the patient encounter, be it assessment, investigation, diagnosis, or treatment [2]. Physicians have both a professional and instinctive desire to be as certain as possible when diagnosing or treating patients [3]. While teaching physicians to tolerate uncertainty is important, there is also a need to overcome problems in medicine that contribute to uncertainty in the physician's mind. Several models of uncertainty have been proposed, but for the ease of our discussion, the distinction between reducible and irreducible uncertainty is most relevant.

Simply put, uncertainties associated with unknowable things are irreducible, and uncertainties associated with knowable things that are currently unknown are reducible [4]. Reducible forms of uncertainty in medicine may stem from the lack of information about the effects of treatment; information overload or complexity; vagueness of terms; or differing beliefs, values, and preferences among providers [5]. Reducible uncertainty can be overcome by obtaining new knowledge. For example, an 85-year-old woman with a headache and a pulsatile temporal artery with a normal erythrocyte sedimentation rate and normal C-reactive protein (CRP) levels may be treated for giant cell arteritis (GCA) with pulse steroids in the emergency department by one provider, but might be deemed unlikely to have the same diagnosis by another. The uncertainty may arise, in this case, from any number of factors, including the lack of a consensus definition of GCA, differing tolerances of risk among providers, and the overall clinical appearance of the patient.

Irreducible forms of uncertainty in medicine, on the other hand, stem from statistical limitations (eg, random error due to natural variation), epistemic problems (such as measurement error,

XSL•FO
RenderX

systematic error, model uncertainty, and uncertainty about inducing case probability from class probability), or numerical vagueness [5]. For example, again considering the diagnosis of GCA, both an elevated erythrocyte sedimentation rate and elevated CRP level (eg, CRP level >10 mg/dL) may help to confirm the diagnosis. However, even a CRP level of 12 mg/dL may not be considered elevated for older age, obesity, and smoking (all of which are factors that may raise the CRP level). The uncertainty in this case is due to the difficulty in applying the broader diagnostic criteria to the specific patient being assessed, and natural variation in numerical criteria used to make the diagnosis. Even with the development of more specific numerical cutoffs for CRP levels, patients with levels slightly above the threshold may not truly have an increased risk of the disease. These are foundational problems of knowledge creation and not specific to medicine. As such, irreducible forms of uncertainty cannot be eliminated by obtaining new knowledge.

Artificial intelligence (AI) tools are being increasingly used as adjuncts to improve diagnosis, medical decision-making, and treatment. Here, the distinction between reducible and irreducible uncertainty becomes important; the forms of uncertainty that can be improved by obtaining knowledge (namely, the reducible forms) may see a benefit with the introduction of AI tools in medical practice. AI tools that have been developed can assist physicians with documenting encounters [6], diagnosing skin cancer [7], providing patient information on medical conditions [8], and teaching surgical skills to medical trainees [9].

Despite the promise AI tools may hold to overcome the sources of uncertainty in medicine, the relationship between AI and medical uncertainty has not been explored in the literature. In this viewpoint article, we consider the potential of AI tools to reduce uncertainty in medical practice, when used as adjuncts to clinical reasoning. In addition, we offer practical approaches to teaching the use of AI in medical schools and residency programs to increase the uptake of these tools in practice.

Impact of AI on Reducible Uncertainty

The potential impact of AI tools on reducible uncertainty in medical practice is vast. We will, however, focus our discussion on three sources of reducible uncertainty: (1) lack of clinical information, (2) provider competence, and (3) provider bias.

Clinical Information

A significant source of uncertainty in medical practice is the lack of availability of relevant information to make a decision. This may include the lack of studies about the disease process or the lack of information about the particular patient. AI tools are currently being employed to gather scientific data, through large database management and integration with biobanks [10], and integrate these biological and clinical variables in prediction outputs [11]. This accelerated pace of data gathering may substantially advance our understanding of disease. In addition to the limited understanding of disease in the scientific community, the lack of information about the particular patient can also create uncertainty. Consider a patient who presents for psychiatric assessment with suicidal ideation. A detailed history is required to arrive at the underlying cause of distress, but it is time-consuming to elicit this information and is affected greatly by patient rapport. AI tools, including scribes [12], can help to address such time constraints, by reducing time spent on documentation and administrative tasks. AI may also provide feedback on a physician's skills in providing patient-centered care, facilitating improvement in this domain [13]. In addition, patients may be more willing to provide socially negative information to AI programs than to physicians, assisting with the collection of data used in clinical decision-making [14].

Provider Competence

Provider skill, knowledge, and experience may also lead to diagnostic uncertainty. AI tools are currently being developed as clinical decision support aids. For example, deep neural networks have been able to classify skin cancer from dermoscopic images at similar levels of accuracy to board-certified dermatologists [15]. Similarly, AI tools have been developed to support the triage process in emergency departments with a 27% greater accuracy than that of the average nurse [16]. These tools have the tremendous potential of reducing human error and contributing to personal learning and process improvement.

Provider Bias

Differing beliefs, values, and preferences among physicians also contribute to reducible uncertainty in diagnosis and treatment. Medical decision-making is ideally a balance of the best available evidence and clinical gestalt, the latter being influenced by unconscious biases. For example, in cases where the incidence of disease is lower in a particular population, a reliance on heuristics may result in underdiagnosis. Take for example, the diagnosis of cutaneous malignant melanoma, which has a lower incidence rate in people with darker skin color compared to non-Hispanic individuals with lighter skin color [17]. Research has shown that patients with darker skin types are more likely to present with later stage cancers [18], resulting in higher mortality rates [19]. AI could assist with addressing disagreements by providing recommendations, irrespective of the decision-making agent's personal perspectives, beliefs, or biases. In fact, a recent study demonstrated that ChatGPT could predict dermatoses in people with lighter and darker skin color with similar levels of accuracy [20], despite established clinical disparities in the diagnosis of skin conditions between these groups [21].

Impact of AI on Irreducible Uncertainty

Despite the potential of AI tools to affect the reducible forms of uncertainty, there are irreducible forms of uncertainty in medicine that these tools will not resolve. Here, we will discuss two irreducible forms of uncertainty, (1) the application of class-to-case probability and (2) model uncertainty, and how AI tools will impact them.

Class-to-Case Probability

The distinction between class probability and case probability as a source of uncertainty was first described by Austrian economist and philosopher, Ludwig von Mises [22]; he described class probability as a general understanding of risk

XSL•FO
RenderX

for a particular group of people, and case probability as the specific understanding of risk for an individual. For example, based on large population studies, we understand that by the age of 80 years, 14% of smokers will develop lung cancer [23]. However, for a particular 65-year-old smoker who comes into the office, whether they will develop lung cancer remains uncertain. Their risk is based on several immeasurable factors, including their comorbidities, environmental exposures, and genetic profile. Regardless, medical decision-making routinely involves an abstraction of class probability to case probability, and we reasonably accept this patient's risk of lung cancer to be 14%.

What is the impact of AI on this problem? AI tools are capable of analyzing large datasets and identifying patterns that may improve the overall accuracy in estimating the risk for groups of patients (class probabilities). Additionally, these tools are being increasingly used to advance the field of precision medicine [24]. For example, genotype-guided treatment is an area of active research in precision medicine. Genomic profiling can be used to provide targeted therapy for patients with lung or breast cancer [25]. By integrating massive amounts of individual data (genetics, lifestyle, and environmental exposure), AI may be able to better predict how a specific patient might respond to treatment, improving our understanding of case probability. Additionally, these tools are also able to learn continuously and may be able to refine their predictions regarding disease risk and prognosis as more information becomes available. Wearable devices, for example, which collect continuous, multidimensional data during daily activities, have captured subtle changes in cognition and functional capacity long before the onset of dementia [26]. Despite these advances, decision-making in medicine will continue to rely on the abstraction of class probability to case probability, as the future outcome of a particular case can never be predicted with complete certainty. This is an epistemological source of uncertainty that AI may be able to mitigate, but never eliminate.

Model Uncertainty

Model uncertainty is another reducible form of uncertainty in medicine. This form of uncertainty arises from the fact that models of disease are approximations of complex systems and involve simplifications of reality and assumptions. As a result, these models may not explain all presentations of a disease. For example, we understand a depletion in serotonin as being a cause of depression [27]; however, this model is imperfect and does not explain why some people who meet the criteria for depression do not respond to selective serotonin reuptake inhibitors. One advantage AI tools offer is that they are data-driven rather than assumption-driven. Deep learning techniques can allow AI tools to learn from raw data rather than predefined model parameters. For example, AI tools used in the COVID-19 pandemic were able to identify unusual cases of pneumonia before public health authorities recognized the threat [28]. Unlike traditional models that are fixed once developed, AI models can also learn and adapt over time as new data becomes available. This means that AI tools can develop models that change over time, at rates much faster than traditional scientific models were developed.

While these tools may reduce model uncertainty, one limitation of AI tools when used to develop models of disease is the lack of explainability or algorithmic transparency. It is not always easy or possible to understand how and why an AI system arrives at its decision [29]. Currently, the tools, methods, policies, and frameworks required for explainable AI have not been well developed [11]. This lack of transparency may increase model uncertainty, due to a lack of physician trust in the model's decision. While explainable AI is foreseeable with time and technological advances, on an epistemological level, AI cannot overcome the fundamental limitation that models are approximations of reality with inherent error. Therefore, model uncertainty will persist as a challenge to medical decision-making despite advances in AI.

Teaching AI to Medical Learners

Overview

Despite the promise AI tools hold in reducing uncertainty, there has been considerable resistance to the implementation of AI tools in medical practice. Several reasons for this reluctance exist, including a lack of transparency, cost, privacy issues, reputation concerns, and legal liability [30]. Some medical professionals also perceive AI systems as threats to medical professional identity (recognition and capability) [31]. In addition, patients worry that these tools may not be able to account for their unique preferences the way a physician might [32]. These concerns are valid and will need to be addressed before AI tools reach widespread implementation. Indeed, our understanding of AI ethics and privacy issues has greatly improved in the last 5 years [33]. Despite these barriers, AI tools have already made their way into medical practice. In fact, learners are increasingly voicing an interest in training on how to use AI technologies in medical practice [34]. Consequently, a shift in perspective is required in medical education to teach learners how to practically use these tools and to understand their benefits and limitations.

Novel teaching approaches are needed to train medical learners to use AI tools practically and responsibly. For educators involved in designing medical curricula, three objectives should be considered: (1) improving students' understanding of capabilities, limitations, and ethics of AI use; (2) increasing practical skills in the use of AI; and (3) increasing technological aptitude needed for producing AI systems.

Teaching the Capabilities, Limitations, and Ethics of AI Use

Students should receive formal education on the capabilities and limitations of AI tools. This includes the scope of technologies presently available in various fields of medicine, in addition to those that are newly emerging, including AI scribes, triage assistants, and patient-facing chatbots. Learners should also be made aware of the limitations of these technologies, including the potential for error due to assumptions of class-to-case probability and model uncertainty. Generative forms of AI, including ChatGPT, experience the lack of context and generalizability and are consequently at risk of spreading misinformation [35]. This phenomenon, known as

XSL•FO
RenderX

"hallucination," refers to the generation of information that appears statically plausible but may not be accurate [36]. Students must be made aware of these limitations. In addition, students should be educated about the ethical and legal risks and issues, misinformation, and hallucinations. This includes social discrimination and racial bias in datasets used to develop these tools, which may be further perpetuated by these tools [37]. The legal implications of AI use should also be discussed. While the legalities around AI use are currently being debated, there is a possibility of medical negligence and liability to both physicians and medical institutions if undue harm to the patient is caused by these tools [38]. Students should also be informed about the privacy and security considerations of sharing personal health information with AI tools, including generative forms of AI, such as ChatGPT [39,40]. The capacity of AI tools to "hallucinate" or provide misinformation, and the efforts needed to improve the technical abilities of present forms of AI should also be taught. Teachers and institutions should find ways to develop and enhance students' critical thinking and analytical skills first, as these technologies are first introduced and refined for practice.

Teaching Practical Skills in AI Use

At present, there is very little teaching on how to practically use AI tools for medical students and residents. This includes the use of these tools for personal, academic, and medical purposes. Perhaps the first step in medical education reform is to normalize the use of AI as a teaching and learning aid. AI has been shown, for example, to ameliorate teaching through helping with teaching concepts; creating and improving scenario modeling, courses, and content; and traditional curriculum and coursework [14]. Technologies that rely on large language models can help with developing curricula and teaching plans [40], generating teaching aids [41], simplifying complex medical concepts [42], and pretesting examination questions [43]. AI can also enhance self-directed learning for medical students. ChatGPT is being used by medical students to practice clinical scenarios, access medical literature, and study for examinations [44]. AI applications are currently being developed, for example, to improve case-based learning and decision-making skills [45,46]. In addition, they can analyze students' responses in real time and provide immediate feedback and insights into students' comprehension and learning progress [47]. As large language models become increasingly integrated, skills in prompt engineering and the development and refinement of appropriate inputs for generative AI are also needed to maximize the efficacy of these tools [48].

In addition to openness around the use of AI as a teaching and learning aid, training is needed on how to practically use AI tools as adjuncts to traditional clinical decision-making tools. The integration of AI into medical practice will require professionals to learn how to adapt workflows and communicate effectively with these tools. For example, AI scribes used to assist with documentation in patient encounters may require providers to learn to explicitly describe physical examination findings or to edit initial documentation effectively [49]. Surgical residents should be taught the use of AI tools used for surgical planning early in training [50]. Additionally, learners should be trained on how to communicate the process, risks, benefits, and alternatives of AI use to patients [34].

While it is encouraged that AI is used in medical education, safeguards will equally be necessary to address issues with academic integrity [14] while using ChatGPT and other such technologies for examinations, assignments, and assessments [44]. Guidelines encompassing accountability systems, ethical considerations, privacy, and moral and integrity issues can be used to help address academic integrity issues [40]. In addition, educating students on how to avoid plagiarism in conjunction with plagiarism detection and language analysis software can promote responsible use of these tools [51].

Increasing Technological Aptitude

In addition to training medical learners on the use of existing AI technologies, a possible long-term goal to improve medical education on AI is to develop the technological aptitude. This includes skills in coding, Python language, mechanisms of data leakage, and an overview of how AI tools are developed [34]. The combination of clinical aptitude and understanding of the practical problems these tools need to solve make physicians uniquely positioned to assist with the production of these technologies. At present, however, many physicians lack the technical skills to help with AI development. Down the line, medical schools will need to consider how they can train physicians to do both.

Conclusion

As AI tools become increasingly integrated into medical practice, they will offer powerful solutions to problems of uncertainty in medicine. These tools have the potential to address reducible forms of uncertainty, including the lack of available clinical information and scientific studies, limits to physician ability, and provider personal bias in decision-making. These tools may also improve the irreducible forms of uncertainty to some extent, increasing our ability to make case predictions from class probabilities and develop models of disease. Despite these capabilities, these tools will never be able to overcome foundational knowledge problems in medicine and pose ethical concerns that must be addressed. Nonetheless, AI tools are being used in practice, and trainees must learn the scope of these technologies, the ethical and legal challenges they pose, and how to use them practically. In the future, trainees should also be taught technical skills of how to develop these technologies. AI has reached medicine, and the medical profession is being asked time and time again to adapt; medical education reform is crucial in this transformation.

Authors' Contributions

SRA and SQH contributed equally to the preparation of this manuscript. All authors were involved in the conception and design of the work. SRA and SQH performed the literature search, conducted the analysis, and created the initial draft. SD and RU

XSL•FO
RenderX

substantively revised the work. All authors approve the final manuscript and assume accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

SD is a member of the advisory board of the Subcortical Surgery Group. He is a member of the speakers bureau for the Congress of Neurological Surgeons and American Association of Neurological Surgeons. He receives research funding from Synaptive and VPIX. He receives royalty payments from Oxford University Press. He serves as the provincial lead for CNS Cancers, Ontario Health (Cancer Care Ontario). RU has received research funding from the Canadian Institutes of Health Research, Health Canada and Wellcome Trust. He serves on advisory boards for the World Health Organization, Doctors Without Borders, the College of Family Physicians of Canada, the Royal College of Physicians and Surgeons of Canada, and the Canadian Medical Association.

References

1. Hillen MA, Gutheil CM, Strout TD, Smets EMA, Han PKJ. Tolerance of uncertainty: conceptual analysis, integrative model, and implications for healthcare. Soc Sci Med 2017 May;180:62-75. [doi: 10.1016/j.socscimed.2017.03.024] [Medline: 28324792]
2. Albrecht O. In: Gary LA, Ray F, Susan CS, editors. Handbook of Social Studies in Health and Medicine: Sage Publications; 2000. [doi: 10.4135/9781848608412]
3. Simpkin AL, Schwartzstein RM. Tolerating uncertainty - the next medical revolution? N Engl J Med 2016 Nov 3;375(18):1713-1715. [doi: 10.1056/NEJMp1606402] [Medline: 27806221]
4. Kiureghian AD, Ditlevsen O. Aleatory or epistemic? Does it matter? Struct Saf 2009 Mar;31(2):105-112. [doi: 10.1016/j.strusafe.2008.06.020]
5. Djulbegovic B, Hozo I, Greenland S. Uncertainty in clinical medicine. In: Philosophy of Medicine: Elsevier; 2011:299-356. [doi: 10.1016/B978-0-444-51787-6.50011-8]
6. Bundy H, Gerhart J, Baek S, et al. Can the administrative loads of physicians be alleviated by AI-facilitated clinical documentation? J Gen Intern Med 2024 Jun 27. [doi: 10.1007/s11606-024-08870-z] [Medline: 38937369]
7. Melarkode N, Srinivasan K, Qaisar SM, Plawiak P. AI-powered diagnosis of skin cancer: a contemporary review, open challenges and future research directions. Cancers (Basel) 2023 Feb 13;15(4):1183. [doi: 10.3390/cancers15041183] [Medline: 36831525]
8. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol 2023 Nov;55(11):2717-2732. [doi: 10.1007/s11255-023-03729-4] [Medline: 37528247]
9. Satapathy P, Hermis AH, Rustagi S, Pradhan KB, Padhi BK, Sah R. Artificial intelligence in surgical education and training: opportunities, challenges, and ethical considerations - correspondence. Int J Surg 2023 May 1;109(5):1543-1544. [doi: 10.1097/JS9.0000000000000387] [Medline: 37037597]
10. Frascarelli C, Bonizzi G, Musico CR, et al. Revolutionizing cancer research: the impact of artificial intelligence in digital biobanking. J Pers Med 2023 Sep 16;13(9):1390. [doi: 10.3390/jpm13091390] [Medline: 37763157]
11. Kim K, Lee YM. Understanding uncertainty in medicine: concepts and implications in medical education. Korean J Med Educ 2018 Sep;30(3):181-188. [doi: 10.3946/kjme.2018.92] [Medline: 30180505]
12. Cao DY, Silkey JR, Decker MC, Wanat KA. Artificial intelligence-driven digital scribes in clinical documentation: pilot study assessing the impact on dermatologist workflow and patient encounters. JAAD Int 2024 Jun;15:149-151. [doi: 10.1016/j.jdin.2024.02.009] [Medline: 38571698]
13. Ryan P, Luz S, Albert P, Vogel C, Normand C, Elwyn G. Using artificial intelligence to assess clinicians' communication skills. BMJ 2019 Jan 18;364:l161. [doi: 10.1136/bmj.l161] [Medline: 30659013]
14. Lucas GM, Gratch J, King A, Morency LP. It's only a computer: virtual humans increase willingness to disclose. Comput Human Behav 2014 Aug;37:94-100. [doi: 10.1016/j.chb.2014.04.043]
15. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nat New Biol 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
16. Ivanov O, Wolf L, Brecher D, et al. Improving ED emergency severity index acuity assignment using machine learning and clinical natural language processing. J Emerg Nurs 2021 Mar;47(2):265-278. [doi: 10.1016/j.jen.2020.11.001] [Medline: 33358394]
17. Brunsgaard EK, Wu YP, Grossman D. Melanoma in skin of color: part I. Epidemiology and clinical presentation. J Am Acad Dermatol 2023 Sep;89(3):445-456. [doi: 10.1016/j.jaad.2022.04.056] [Medline: 35533771]
18. Kabigting FD, Nelson FP, Kauffman CL, Popoveniuc G, Dasanu CA, Alexandrescu DT. Malignant melanoma in African-Americans. Dermatol Online J 2009;15(2):3. [doi: 10.5070/D33K77P755]
19. Cormier JN, Xing Y, Ding M, et al. Ethnic differences among patients with cutaneous melanoma. Arch Intern Med 2006 Sep 25;166(17):1907-1914. [doi: 10.1001/archinte.166.17.1907] [Medline: 17000949]
20. Qureshi S, Alli SR, Ogunyemi B. Accuracy of ChatGPT-3.5 and GPT-4 in diagnosing clinical scenarios in dermatology involving skin of color. Int J Dermatol 2024 Aug 9. [doi: 10.1111/ijd.17425] [Medline: 39123282]

XSL•FO
RenderX

21. Fenton A, Elliott E, Shahbandi A, et al. Medical students' ability to diagnose common dermatologic conditions in skin of color. J Am Acad Dermatol 2020 Sep;83(3):957-958. [doi: 10.1016/j.jaad.2019.12.078] [Medline: 32017947]
22. Mises LV. Human Action: Ludwig von Mises Institute; 1949. URL: https://cdn.mises.org/Human%20Action_3.pdf [accessed 2024-09-26]
23. Weber MF, Sarich PEA, Vaneckova P, et al. Cancer incidence and cancer death in relation to tobacco smoking in a population-based Australian cohort study. Int J Cancer 2021 Sep 1;149(5):1076-1088. [doi: 10.1002/ijc.33685] [Medline: 34015143]
24. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
25. Hartmaier RJ, Albacker LA, Chmielecki J, et al. High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis. Cancer Res 2017 May 1;77(9):2464-2475. [doi: 10.1158/0008-5472.CAN-16-2479] [Medline: 28235761]
26. Gold M, Amatniek J, Carrillo MC, et al. Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer's disease clinical trials. Alzheimers Dement (N Y) 2018;4(1):234-242. [doi: 10.1016/j.trci.2018.04.003] [Medline: 29955666]
27. Moncrieff J, Cooper RE, Stockmann T, et al. The serotonin theory of depression: a systematic umbrella review of the evidence. Mol Psychiatry 2023;28:3243-3256. [doi: 10.1038/s41380-022-01661-0]
28. Hasan MM, Islam MU, Sadeq MJ, Fung WK, Uddin J. Review on the evaluation and development of artificial intelligence for COVID-19 containment. Sensors (Basel) 2023 Jan 3;23(1):527. [doi: 10.3390/s23010527] [Medline: 36617124]
29. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(4):e1312. [doi: 10.1002/widm.1312] [Medline: 32089788]
30. Yang Y, Ngai EWT, Wang L. Resistance to artificial intelligence in health care: literature review, conceptual framework, and research agenda. Inf Manag 2024 Jun;61(4):103961. [doi: 10.1016/j.im.2024.103961]
31. Jussupow E, Spohrer K, Heinzl A. Identity threats as a reason for resistance to artificial intelligence: survey study with medical students and professionals. JMIR Form Res 2022 Mar 23;6(3):e28750. [doi: 10.2196/28750] [Medline: 35319465]
32. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019 Dec 1;46(4):629-650. [doi: 10.1093/jcr/ucz013]
33. Stahl BC. Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies: Springer Nature; 2021. [doi: 10.1007/978-3-030-69978-9]
34. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
35. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
36. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Teach 2024 May;46(5):657-664. [doi: 10.1080/0142159X.2023.2271159] [Medline: 37862566]
37. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C. AI in medical education: global situation, effects and challenges. Educ Inf Technol 2024 Mar;29(4):4611-4633. [doi: 10.1007/s10639-023-12009-8]
38. Mehta D. The role of artificial intelligence in healthcare and medical negligence. Liverp Law Rev 2024 Apr;45(1):125-142. [doi: 10.1007/s10991-023-09340-y]
39. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
40. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024;21:6. [doi: 10.3352/jeehp.2024.21.6] [Medline: 38486402]
41. Grassini S. Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. Educ Sci 2023;13(7):692. [doi: 10.3390/educsci13070692]
42. Benítez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inform Assoc 2024 Feb 16;31(3):776-783. [doi: 10.1093/jamia/ocad252] [Medline: 38269644]
43. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
44. Lakshan MTD, Chandratilake M, Drahaman AMP, Perera MB. Exploring the pros and cons of integrating artificial intelligence and ChatGPT in medical education: a comprehensive analysis. Ceylon J Otolaryngology 2024;13(1):39-45. [doi: 10.4038/cjo.v13i1.5380]
45. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024 Apr;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
46. Ossa LA, Rost M, Lorenzini G, Shaw DM, Elger BS. A smarter perspective: learning with and from AI-cases. Artif Intell Med 2023 Jan;135:102458. [doi: 10.1016/j.artmed.2022.102458] [Medline: 36628794]

XSL•FO
RenderX

47. Onesi-Ozigagun O, Ololade YJ, Eyo-Udo NL, Ogundipe DO. Revolutionizing education through AI: a comprehensive review of enhancing learning experiences. Int J Appl Res Soc Sci 2024;6(4):589-607. [doi: 10.51594/ijarss.v6i4.1011]
48. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25:e50638. [doi: 10.2196/50638] [Medline: 37792434]
49. van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a digital scribe system on clinical documentation time and quality: usability study. JMIR AI 2024 Sep 23;3:e60020. [doi: 10.2196/60020] [Medline: 39312397]
50. Varghese C, Harrison EM, O'Grady G, Topol EJ. Artificial intelligence in surgery. N Med 2024 May;30(5):1257-1268. [doi: 10.1038/s41591-024-02970-3] [Medline: 38740998]
51. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int 2024 Mar 3;61(2):228-239. [doi: 10.1080/14703297.2023.2190148]

Abbreviations

AI: artificial intelligence
CRP: C-reactive protein
GCA: giant cell arteritis

Edited by TDA Cardoso; submitted 01.08.23; peer-reviewed by B Meskó, R Onaisi; revised version received 26.09.24; accepted 27.09.24; published 04.11.24.

Please cite as:
Alli SR, Hossain SQ, Das S, Upshur R
The Potential of Artificial Intelligence Tools for Reducing Uncertainty in Medicine and Directions for Medical Education
JMIR Med Educ 2024;10:e51446
URL: https://mededu.jmir.org/2024/1/e51446
doi:10.2196/51446

© Sauliha Rabia Alli, Soaad Qahhār Hossain, Sunit Das, Ross Upshur. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.11.2024. 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.

XSL•FO
RenderX

Viewpoint

Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education

Lukas Weidener[1], Dr Med; Michael Fischer[1], PhD

UMIT TIROL – Private University for Health Sciences and Health Technology, Hall in Tirol, Austria

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

Abstract

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

(JMIR Med Educ 2024;10:e55368) doi:10.2196/55368

KEYWORDS

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

Introduction

Background

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

XSL•FO
RenderX

use of AI in medicine can also lead to erroneous diagnoses such as unnecessary treatment, which violates the basic principles of medical ethics [5].

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

Definitions of AI

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

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

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

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

AI Ethics

The field of AI ethics was an area of interest for both scientific and governmental communities, even before the emergence of

XSL•FO
RenderX

AI applications such as ChatGPT, which has gained widespread public attention [29,30]. However, there remains a lack of consensus on the definition of AI ethics, which can be attributed to several factors, including the novelty and interdisciplinary nature of the field as well as the absence of a widely accepted definition of AI [31].

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

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

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

AI Ethics in Medical Education

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

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

A recurrent theme related to the teaching of AI ethics as part of medical education focuses on the principles of medical ethics according to Beauchamp and Childress (autonomy, beneficence, nonmaleficence, and justice) [10]. This emphasis is also echoed by existing guidelines and frameworks regarding AI ethics [30]. Additional recommendations on AI ethics teaching content include "explainability," "liability," and "accountability," which are also considered important by available guidelines [30,40]. On the basis of the analysis of existing publications on teaching AI ethics in medical education, 12 potential subjects were considered for teaching AI ethics. The 12 identified potential teaching subjects for AI ethics in medical education are listed in Textbox 1.

XSL•FO
RenderX

Textbox 1. Recommended artificial intelligence (AI) ethics teaching content with specific descriptions.

Informed consent

Informed consent in the context of AI in medicine requires that patients be fully informed about treatment options and risks, necessitating a comprehensive understanding and explanation of AI technologies by physicians.

Bias

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

Safety

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

Transparency

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

Privacy

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

Allocation

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

Fairness

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

Responsibility

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

Empathy

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

Explainability

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

Liability

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

Accountability

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

Objective

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

Theoretical Analysis of Recommended AI Ethics Teaching Subjects in Medical Education

Overview

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

While these 4 principles have been an integral part of current scientific publications on AI ethics in medical education, the

XSL•FO
RenderX

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

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

Informed Consent

Overview

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

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

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

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

Informed Consent in the Context of AI in Medicine

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

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

Intersections of Informed Consent With Key AI Ethics Teaching Subjects

Overview

This section aims to underscore interdisciplinarity and intersectionality among the recommended teaching subjects in AI ethics, as outlined in the Informed Consent section, with informed consent serving as a representative example. Focusing on these intersections, this section highlights the importance of an integrated educational approach in the context of medical AI ethics. Such an approach acknowledges that topics such as

XSL•FO
RenderX

bias, privacy, and transparency, among others, are not merely isolated subjects but instead require a comprehensive, holistic evaluation. Embracing this integrated perspective is important for a comprehensive understanding of AI ethics in medical practice and education, underscoring the need to re-evaluate and potentially refine current teaching recommendations. To effectively illustrate the interdisciplinarity and interconnectedness of frequently recommended teaching subjects for AI ethics in medical education, "informed consent" should be discussed in the context of 5 frequently proposed teaching subjects: bias, safety, transparency, privacy, and liability.

Bias

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

Safety

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

Transparency

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

Privacy

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

Liability

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

Medical AI Ethics

Overview

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

In the context of traditional medical ethics education, the emphasis on foundational principles provides a broad and adaptable framework that is essential for understanding and addressing complex ethical dilemmas. This approach facilitates the holistic comprehension of ethical issues, offering the flexibility to accommodate the diverse and evolving nature of medical scenarios. Similarly, when considering AI in medicine, a focus on core ethical principles rather than solely on specific challenges lays the groundwork for a robust and comprehensive educational strategy. Future medical professionals should be equipped with a deeper and more nuanced understanding of ethical decision-making by emphasizing ethical principles in the context of implementing and using AI in medicine. This principle-based approach ensures that medical ethics education

XSL•FO
RenderX

remains relevant and responsive to the dynamic landscape of AI integration in medicine. The goal is for medical students to be able to effectively navigate the ethical complexities associated with AI technologies in medicine, not just focusing on potential challenges but also emphasizing the ethical values that are essential to medical practice.

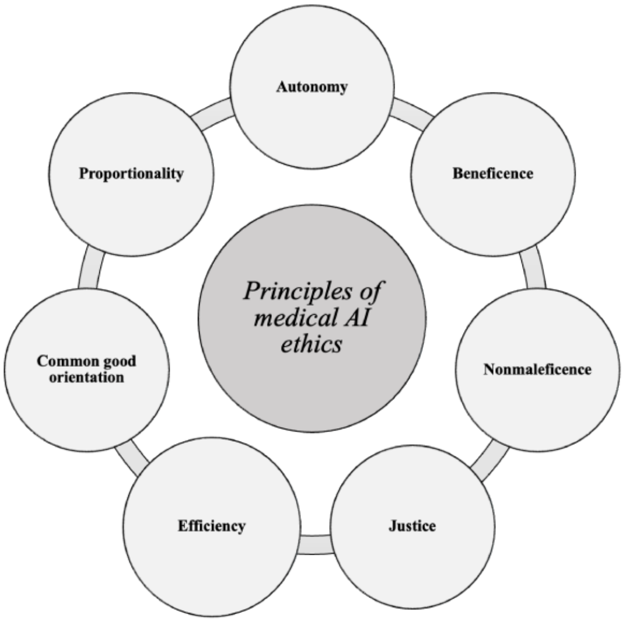
Owing to the paramount importance and relevance of the 4 principles of medical ethics formulated by Beauchamp and Childress [34], the principles of autonomy, beneficence, nonmaleficence, and justice should provide the essential foundation for medical AI ethics. These 4 principles are subsequently introduced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

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

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

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

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



Autonomy

The principle of autonomy in medical ethics emphasizes the right to make independent decisions regarding health care [34]. This principle recognizes an individual's capacity for self-determination and personal choice, affirming that patients have the authority to provide or withhold consent for medical treatment. Respecting autonomy in medical practice involves providing patients with sufficient information, ensuring comprehension, and facilitating independent decision-making [59]. This respect for autonomy is closely tied to the principle of informed consent, which ensures that patients actively participate in decisions regarding their care and treatment.

XSL•FO
RenderX

In the context of using AI in medicine, particularly in diagnostics and treatment recommendations, technology introduces new challenges and opportunities to maintain patient autonomy [60]. For example, when using AI-based diagnostic applications, it is crucial to inform patients about how these tools impact their health care decisions, ensuring that informed consent is comprehensive. Equally important is equipping physicians with the knowledge to balance AI-generated insights with their clinical expertise, thus upholding both patient and physician autonomy in decision-making processes. The incorporation of AI into health care decision-making can affect the presentation and comprehension of options by patients. Ensuring that patients retain their autonomous decision-making power in an AI-driven environment requires the careful consideration of how the information is communicated and understood [60]. Autonomy in this context extends to ensuring that patients have a clear understanding of AI interventions and their capabilities, limitations, and impact on personal health decisions. Moreover, the principle of autonomy extends to physicians. If AI increasingly assists in medical decision-making, it is imperative that physicians remain empowered to make independent professional judgments, balancing AI insights with their clinical expertise and ethical considerations.

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

Beneficence

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

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

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

Nonmaleficence

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

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

To raise the awareness of potential conflicts with the principle of nonmaleficence regarding the use of AI in medicine, medical education should focus on the ethical design, development, and

XSL•FO
RenderX

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

Justice

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

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

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

Efficiency

Efficiency within public health ethics underscores the strategic use of resources to maximize health benefits for the population [57]. This principle is not only solely an economic concern but also a moral imperative to ensure the equitable and judicious use of medical technologies and services. Ethical considerations regarding the principle of efficiency are especially relevant in health care settings where resources are limited and demand is high, as exemplified in the context of the COVID-19 pandemic.

Owing to the capabilities of AI in medicine with the potential to enhance the efficiency of medical services through faster and more accurate diagnostics, it is crucial to consider the ethical implications of these developments [19]. The ability of AI to rapidly analyze large data sets can greatly enhance the speed of diagnostic procedures, which could result in more timely patient care and improved treatment choices that are more precise. However, this benefit is contingent on the data quality. Poor-quality data can result in AI models that incorrectly predict outcomes based on artifacts in the data rather than actual clinical results [21]. Therefore, the ethical use of AI in health care must include rigorous validation of the data quality to ensure accurate and reliable outcomes. For example, physicians must balance the efficiency gains offered by AI with the need for clinical judgment and personalized patient care and upholding and maintaining the quality of physician-patient relationships [63].

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

Common Good Orientation

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

The principle of common good orientation in the context of AI, crucial in guiding the integration of technology into medical practice, calls for a delicate balance between individual patient benefits and the collective well-being of the community. It is essential to recognize how AI in medicine can address or potentially exacerbate health disparities [64]. The ability of AI to process and analyze data can be harnessed to identify and address gaps in health care delivery, offering insights into underserved populations and tailoring interventions to meet their specific needs. Conversely, if not carefully managed, AI could unintentionally increase these disparities by favoring populations with better access to the technology. This duality underscores the need for AI advancements in health care to contribute positively and equitably to public health, promoting fairness in health care access and outcomes. It is important to note that the selective application of AI not only undermines the principle of common good orientation but also risks creating

XSL•FO
RenderX

a perception of elitism in the medical profession. Such a scenario could harm the reputation of the medical field, rendering it as unevenly benefiting certain populations. Furthermore, using AI in medical practice could potentially lead to events where patients are harmed, for example, through biased decision-making or errors made by users. This could potentially lead to a negative perception of AI within the broader population, which in turn may result in a general unwillingness or resistance to adopting AI technologies. This hesitance could directly conflict with the principle of common good orientation, as it hinders the widespread and equitable implementation of AI that could benefit the entire community [25].

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

Proportionality

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

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

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

Discussion

Overview

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

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

The high level of intersectionality and interdisciplinarity inherited by the implementation and use of AI in medicine highlights the importance of a principle-based approach rather than solely focusing on anticipated challenges. While the proposed ethical principles also show a high level of interconnectivity, the chosen educational approach aims to encourage a more nuanced understanding, not limited to specific anticipated challenges but rather to enable future physicians to adapt to the changing landscape associated with the use of AI in medicine, facilitating the consideration of multiple ethical dimensions simultaneously. In addition to the proposed principles, medical education should incorporate practical case studies and simulations to reflect real-world scenarios. For example, applying AI to patient triage during health emergencies such as the COVID-19 pandemic can offer practical contexts for students. This approach would not only enhance their understanding of ethical principles but also prepare them for decision-making in complex, real-life medical situations influenced by AI. It is important for future physicians to

XSL•FO
RenderX

understand the balance between the potential benefits of AI and the ethical implications of its use, particularly in scenarios in which biased algorithms could lead to unequal treatment of diverse patient groups. Therefore, a comprehensive curriculum that includes both theoretical knowledge and practical applications is essential to cultivate ethically informed medical professionals.

An in-depth and interdisciplinary understanding of ethics is important in the dynamic field of medical AI. This importance is underscored by the fact that the integration of AI into medical education may not always keep pace with rapid advancements in medical practice. A focus on ethical principles rather than solely on specific challenges of AI use in medicine aims to prepare medical students for various scenarios in the medical context. This approach maintains relevance even if the AI applications used in education are not representative of the latest state-of-the-art developments in medical AI. The principle-based approach to AI ethics offers broader applicability and reduces dependence on the most recent AI technologies, potentially benefiting medical schools with limited financial resources. In addition, AI products for teaching, often sourced from third parties and guided by cost considerations, may pose unique challenges such as the risk of bias or rapid obsolescence [66,67]. This necessitates awareness, among medical students, of the potential ethical issues associated with these tools. By emphasizing a principle-based approach to AI ethics, educators can equip students with the necessary understanding to navigate the evolving landscape of AI in medicine, fostering adaptability and ethical sensitivity in future medical professionals. This adaptability is crucial to ensure that future physicians are prepared for the ethical dilemmas they may encounter in a rapidly evolving AI landscape.

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

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

Limitations

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

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

Moreover, integrating new teaching content into medical curricula is challenging due to the need for time-intensive accreditation processes and extensive teaching content. The integration of new teaching content such as medical AI ethics education requires careful planning to ensure that future

XSL•FO
RenderX

physicians are adequately prepared and not overwhelmed by information. In addition, limited access to instructors knowledgeable in ethics, medicine, and AI may pose a challenge to implementing the proposed teaching of medical AI ethics, as these experts may not be available in most institutions.

Conclusions

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

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

Conflicts of Interest

None declared.

References

1. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
2. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
3. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
4. Hedlund M, Persson E. Expert responsibility in AI development. AI Soc 2022 Jun 13:1-12 [FREE Full text] [doi: 10.1007/s00146-022-01498-9]
5. Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc 2020 Jun 09;19(1):61-86 [FREE Full text] [doi: 10.1108/jices-12-2019-0138]
6. Wartman SA, Combs CD. Reimagining medical education in the Age of AI. AMA J Ethics 2019 Feb 01;21(2):E146-E152 [FREE Full text] [doi: 10.1001/amajethics.2019.146] [Medline: 30794124]
7. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. JMIR Med Educ 2023 Apr 24;9:e46428 [FREE Full text] [doi: 10.2196/46428] [Medline: 36946094]
8. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018 Sep 27;1(1):54 [FREE Full text] [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
9. Katznelson G, Gerke S. The need for health AI ethics in medical school education. Adv Health Sci Educ Theory Pract 2021 Oct 03;26(4):1447-1458. [doi: 10.1007/s10459-021-10040-3] [Medline: 33655433]
10. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023;12(1):399-410 [FREE Full text] [doi: 10.5334/pme.954] [Medline: 37868075]
11. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020 Jun 19;3(1):86 [FREE Full text] [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
12. Pinto Dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr 6;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
13. Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. MedEdPublish 2021;10(1):75. [doi: 10.15694/mep.2021.000075.1]
14. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ 2021 Feb 18;21(1):112 [FREE Full text] [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
15. Haenlein M, Kaplan A. A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. Calif Manag Rev 2019 Jul 17;61(4):5-14. [doi: 10.1177/0008125619864925]
16. Russell S, Norvig P. Artificial Intelligence: A Modern Approach (Prentice Hall Series in Artificial Intelligence). London, UK: Pearson; 2010.

XSL•FO
RenderX

17. Dick S. Artificial Intelligence. Harvard Data Sci Rev 2019 Jun 23;1(1) [FREE Full text] [doi: 10.1162/99608f92.92fe150c]
18. Wang P, Monett D, Lewis CW, Thórisson KR. On defining artificial intelligence. J Artif Gen Intell 2019;10(2):1-37. [doi: 10.2478/jagi-2019-0002]
19. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York, NY: Basic Books; 2019.
20. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Cham, Switzerland: Springer; 2016:795-840.
21. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar 12;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
22. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
23. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
24. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019 Dec 24;322(24):2377-2378 [FREE Full text] [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
25. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. Intell Based Med 2020 Nov;1-2:100001. [doi: 10.1016/j.ibmed.2020.100001]
26. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
27. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
28. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023 Dec 21;28(1):2181052 [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
29. Ethics guidelines for trustworthy AI. European Commission's High-Level Expert Group on Artificial Intelligence. 2019. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 [accessed 2023-10-20]
30. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. Minds Mach 2020 Feb 01;30(1):99-120. [doi: 10.1007/s11023-020-09517-8]
31. Christoforaki M, Beyan O. AI ethics—a bird's eye view. Appl Sci 2022 Apr 20;12(9):4130. [doi: 10.3390/app12094130]
32. Whittlestone J, Alexandrova A, Nyrup R. The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019 Presented at: AIES '19; January 27-28, 2019; Honolulu, HI p. 195-200 URL: https://dl.acm.org/doi/10.1145/3306618.3314289 [doi: 10.1145/3306618.3314289]
33. Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf [accessed 2023-10-20]
34. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. Oxford, UK: Oxford University Press; 2001.
35. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019 Sep 02;1(9):389-399. [doi: 10.1038/s42256-019-0088-2]
36. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. J Med Educ Curric Dev 2021 Sep 06;8:23821205211036836 [FREE Full text] [doi: 10.1177/23821205211036836] [Medline: 34778562]
37. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 01;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
38. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 09;22(1):772 [FREE Full text] [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
39. Paranjape K, Schinkel M, Nannan Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
40. Quinn TP, Coghlan S. Readying medical students for medical AI: the need to embed AI ethics education. arXiv. Preprint posted online 7 September, 2021 [FREE Full text] [doi: 10.48550/arXiv.2109.02866]
41. Giubilini A, Milnes S, Savulescu J. The medical ethics curriculum in medical schools: present and future. J Clin Ethics 2016 Jun 01;27(2):129-145. [doi: 10.1086/jce2016272129]
42. Spatz ES, Krumholz HM, Moulton BW. The new era of informed consent: getting to a reasonable-patient standard through shared decision making. JAMA 2016 May 17;315(19):2063-2064 [FREE Full text] [doi: 10.1001/jama.2016.3070] [Medline: 27099970]

XSL•FO
RenderX

43. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367 [FREE Full text] [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
44. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? SSRN J 2020;108:1425-1469 [FREE Full text] [doi: 10.2139/ssrn.3529576]
45. Kotsenas AL, Balthazar P, Andrews D, Geis JR, Cook TS. Rethinking patient consent in the era of artificial intelligence and big data. J Am Coll Radiol 2021 Jan;18(1 Pt B):180-184. [doi: 10.1016/j.jacr.2020.09.022] [Medline: 33413897]
46. Ursin F, Timmermann C, Orzechowski M, Steger F. Diagnosing diabetic retinopathy with artificial intelligence: what information should be included to ensure ethical informed consent? Front Med (Lausanne) 2021 Jul 21;8:695217 [FREE Full text] [doi: 10.3389/fmed.2021.695217] [Medline: 34368192]
47. Leslie D. Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. SSRN J 2019 [FREE Full text] [doi: 10.2139/ssrn.3403301]
48. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
49. Barkal JL, Stockert JW, Ehrenfeld JM, Cohen LK. AI and the evolution of the patient – physician relationship. In: Byrne MF, Parsa N, Greenhill AT, Chahal D, Ahmad O, Bagci U, editors. AI in Clinical Medicine: A Practical Guide for Healthcare Professionals. Hoboken, NJ: John Wiley & Sons; 2023:478-487.
50. Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. Electron Mark 2020 Oct 01;31(2):447-464. [doi: 10.1007/s12525-020-00441-4]
51. Price 2nd WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
52. Molnár-Gábor F. Artificial intelligence in healthcare: doctors, patients and liabilities. In: Wischmeyer T, Rademacher T, editors. Regulating Artificial Intelligence. Cham, Switzerland: Springer; 2019:337-360.
53. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q 2021 Sep 06;99(3):629-647 [FREE Full text] [doi: 10.1111/1468-0009.12504] [Medline: 33822422]
54. Sabetkish N, Rahmani A. The overall impact of COVID-19 on healthcare during the pandemic: a multidisciplinary point of view. Health Sci Rep 2021 Dec;4(4):e386 [FREE Full text] [doi: 10.1002/hsr2.386] [Medline: 34622020]
55. Mastroianni AC, Khan PJ, Kass NE. The Oxford Handbook of Public Health Ethics. Oxford, UK: Oxford Academic; 2019.
56. Faden R, Bernstein J, Shebaya S. Public health ethics. Stanford Encyclopedia of Philosophy Archive. 2022. URL: https://plato.stanford.edu/archives/spr2022/entries/publichealth-ethics [accessed 2024-01-29]
57. Schro¨der-Ba¨ck P. Ethische Prinzipien für die Public-Health-Praxis. New York, NY: Campus Verlag GmbH; 2014.
58. Kahrass H, Mertz M. Ethik in der Public Health. Bremen, Germany: Apollon University Press; 2021.
59. Varelius J. The value of autonomy in medical ethics. Med Health Care Philos 2006 Oct 11;9(3):377-388 [FREE Full text] [doi: 10.1007/s11019-006-9000-z] [Medline: 17033883]
60. Bitterman DS, Aerts HJ, Mak RH. Approaching autonomy in medical artificial intelligence. Lancet Digit Health 2020 Sep;2(9):e447-e449 [FREE Full text] [doi: 10.1016/S2589-7500(20)30187-4] [Medline: 33328110]
61. Bester JC. Beneficence, interests, and wellbeing in medicine: what it means to provide benefit to patients. Am J Bioeth 2020 Mar 27;20(3):53-62. [doi: 10.1080/15265161.2020.1714793] [Medline: 32105204]
62. Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV. Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Med Exp 2019 Dec 10;7(1):70 [FREE Full text] [doi: 10.1186/s40635-019-0286-6] [Medline: 31823128]
63. Blasimme A, Vayena E. The ethics of AI in biomedical research, patient care and public health. SSRN Journal. Preprint posted online April 9, 2019 [FREE Full text] [doi: 10.2139/ssrn.3368756]
64. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 2020 Jun 01;3(1):81 [FREE Full text] [doi: 10.1038/s41746-020-0288-5] [Medline: 32529043]
65. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
66. Akgun S, Greenhow C. Artificial intelligence in education: addressing ethical challenges in K-12 settings. AI Ethics 2022 Sep 22;2(3):431-440 [FREE Full text] [doi: 10.1007/s43681-021-00096-7] [Medline: 34790956]
67. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019 Oct 28;16(1):1-27. [doi: 10.1186/s41239-019-0171-0]
68. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 06;9:e48163 [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
69. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9(6):e45312-e45350 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]

XSL•FO
RenderX

70. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
71. Khlaif ZN, Mousa A, Hattab MK, Itmazi J, Hassan AA, Sanmugam M, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR Med Educ 2023 Sep 14;9:e47049 [FREE Full text] [doi: 10.2196/47049] [Medline: 37707884]
72. Lee LM, Zarowsky C. Foundational values for public health. Public Health Rev 2015 May 29;36(1):2 [FREE Full text] [doi: 10.1186/s40985-015-0004-1] [Medline: 29450030]

Abbreviations

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

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

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

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

XSL•FO
RenderX

Viewpoint

Rolling the DICE (Design, Interpret, Compute, Estimate): Interactive Learning of Biostatistics With Simulations

Robert Thiesmeier[1*], MMSc; Nicola Orsini[1*], PhD

Department of Global Public Health, Karolinska Institutet, Solna, Sweden

*all authors contributed equally

Corresponding Author:
Robert Thiesmeier, MMSc
Department of Global Public Health
Karolinska Institutet
Tomtebodavägen 18
Solna, 171 65
Sweden
Phone: 46 735779719
Email: robert.thiesmeier@ki.se

Abstract

Despite the increasing relevance of statistics in health sciences, teaching styles in higher education are remarkably similar across disciplines: lectures covering the theory and methods, followed by application and computer exercises in given data sets. This often leads to challenges for students in comprehending fundamental statistical concepts essential for medical research. To address these challenges, we propose an engaging learning approach—DICE (design, interpret, compute, estimate)—aimed at enhancing the learning experience of statistics in public health and epidemiology. In introducing DICE, we guide readers through a practical example. Students will work in small groups to plan, generate, analyze, interpret, and communicate their own scientific investigation with simulations. With a focus on fundamental statistical concepts such as sampling variability, error probabilities, and the construction of statistical models, DICE offers a promising approach to learning how to combine substantive medical knowledge and statistical concepts. The materials in this paper, including the computer code, can be readily used as a hands-on tool for both teachers and students.

(JMIR Med Educ 2024;10:e52679) doi:10.2196/52679

KEYWORDS

learning statistics; Monte Carlo simulation; simulation-based learning; survival analysis; Weibull

Introduction

The correct use and application of statistics plays a fundamental role in the health sciences, in turn providing objective and quantitative evidence to support decision-making in public health [1]. Despite the increasing relevance of statistics in health research, it is often taught in isolation, usually through standard lectures covering the theory and methods followed by computer exercises with given data sets. This can lead to a disconnect between statistical and epidemiological methods such as study design, as well as insufficient awareness of important statistical concepts such as sampling variability [2]. Therefore, teaching methods that deliver statistical concepts in conjunction with epidemiology for students in the health sciences are crucial for educational development [3].

Simulation-based learning has previously been proposed as a tool to support engaging learning [4] and has been shown to be an effective learning method to develop critical thinking and reflective skills [5-7]. In the context of public health and epidemiology, 2 articles in particular highlight Monte Carlo simulations [8] (hereafter simulations) as a method to illustrate, learn, and understand statistical and epidemiological concepts. First, Rudolph et al [3] demonstrate how to use simulations to teach and learn nondifferential misclassification and understand the concept of the *P* value. Second, Fox et al [9] illustrate how to design simple simulations from directed acyclic graphs and use them to explain epidemiological concepts. Both papers provide helpful resources for students to familiarize themselves with the basics of setting up a simulation.

However, despite a broad acceptance of simulations as a helpful tool to learn statistical and epidemiological concepts [5], in our experience, they are rarely implemented as the main teaching and learning method for students in health sciences. Rather than using simulations to learn a stand-alone element of statistics,

XSL•FO
RenderX

we propose a learning method that uses simulations to explore and understand the major steps involved in conducting a scientific investigation. In expanding upon the current foundations of simulation-based learning in health sciences, we introduce DICE (design, interpret, compute, estimate), an engaging, problem- and simulation-based learning method. The overall aim is to promote statistical reasoning in the health sciences by combining medical and statistical knowledge in designing epidemiological studies. The purpose of this viewpoint paper is therefore to describe the concept of DICE and discuss its potential strengths and limitations in learning statistics in the health sciences. The statements expressed in this paper are based on the experiences and opinions of the authors.

The remaining part is structured as follows: we will first describe the proposed method—DICE—and explain the intended learning objectives and outcomes. We will then illustrate the use of DICE with an example of a time-to-event outcome. Finally, we will discuss some potential strengths and limitations of applying the method in a classroom setting.

The DICE Approach

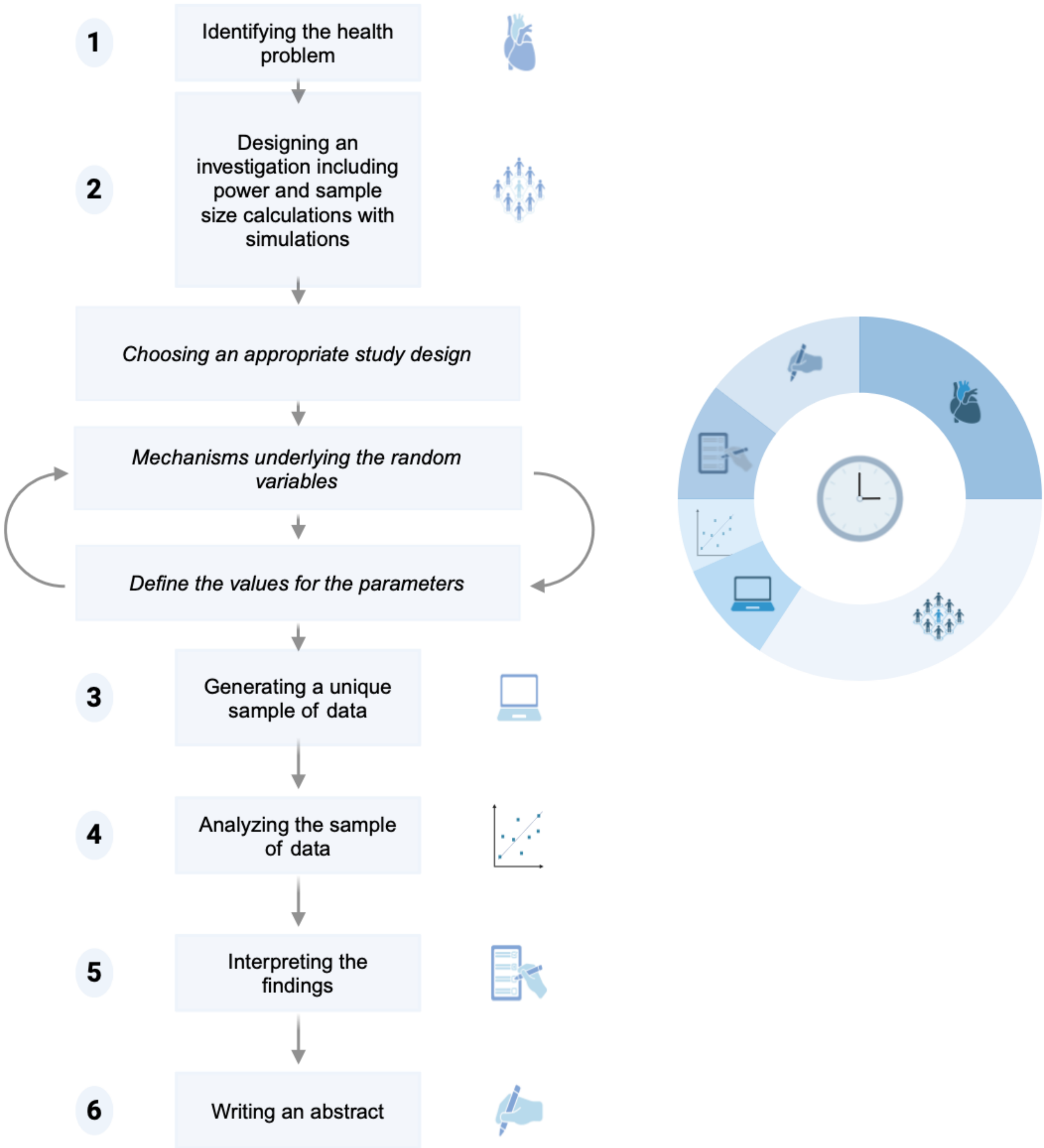
DICE is an engaging learning method that enables students to use simple simulations to design, analyze, and interpret a realistic epidemiological study (note that the acronym DICE represents the learning steps involved, but not in order). The use of DICE as a learning tool combines problem-oriented learning [10,11] with simulations [12]. A detailed description of Monte Carlo methods can be found elsewhere [13]. While there are numerous ways to simulate artificial data, we focus on the approach presented by Fox et al [9] due to its simplicity to implement in statistical software and its easy-to-follow translation from a causal framework. In brief, simulations enable us to study a mechanism empirically by sampling from a statistical model that governs the mechanism. Data are then sampled from a predefined probability distribution (eg, Bernoulli, normal, or Weibull) that defines the mechanism, commonly referred to as the inverse transformation sampling method [14]. Further, the data are analyzed with an appropriate statistical model (preferably the same model that generated the data). These steps can be repeated a large number of times to empirically observe the variability of the sampling process [13].

Steps and Learning Objectives

DICE includes 6 major steps that cover the major stages of a scientific investigation. Figure 1 visualizes the chronological order of these steps and highlights the approximate amount of time that one step requires. The second step—designing an investigation including power and sample size calculations with simulations—is further divided into 3 parts, which can be repeated to calibrate the power and sample size of a study before moving on to step 3.

XSL•FO
RenderX

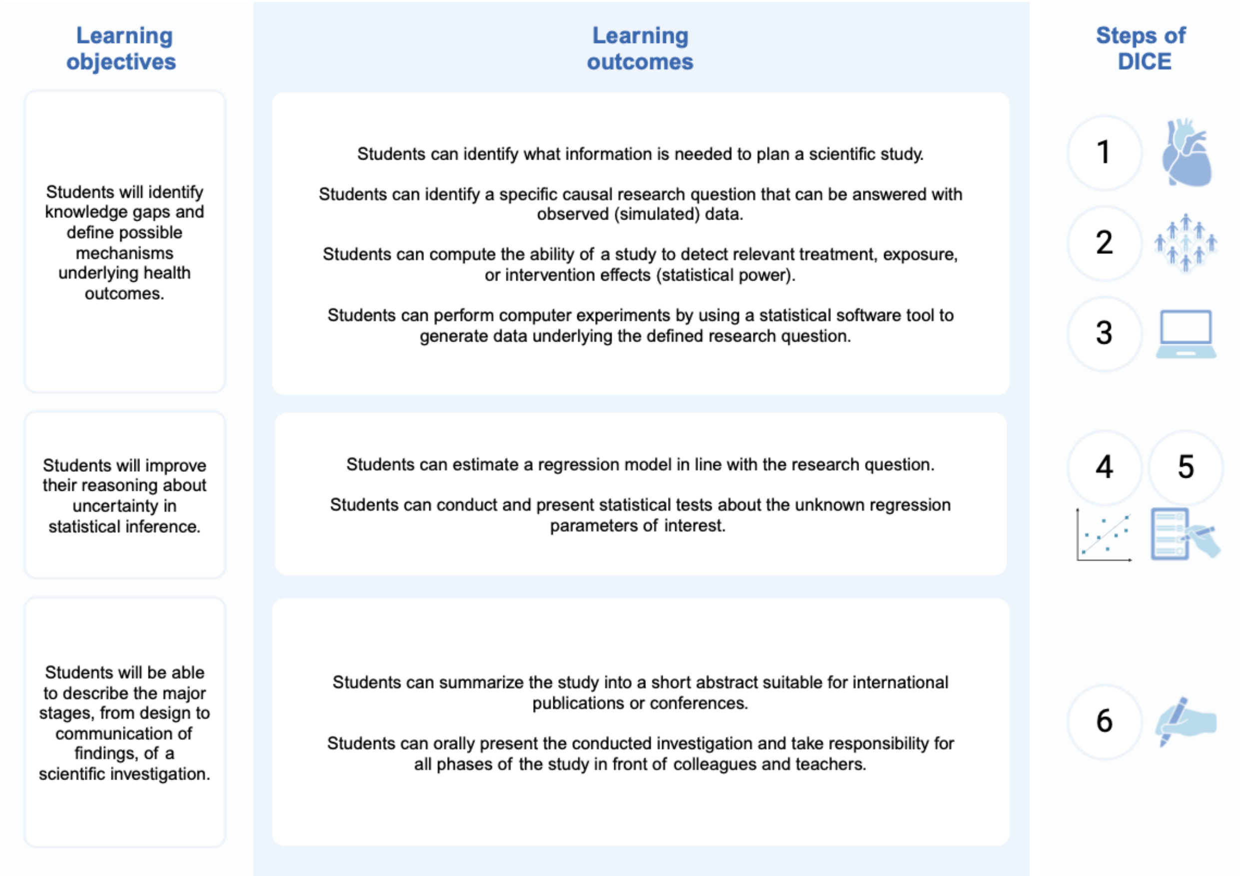
Figure 1. Flowchart of the 6 steps involved in DICE (design, interpret, compute, estimate) with an illustrative pie chart showing the approximate time dedicated to each step. The curved arrows in step 2 indicate that these activities can be completed multiple times to calibrate the sample size and power of a study before moving on to the next step.



The key learning objectives and outcomes of DICE, as highlighted in Figure 2, target experiential learning [15] and active learning styles [16] according to Bloom's taxonomy of educational objectives, including applying recently learned concepts and theories, making informed judgements and evaluations, and generating new knowledge [17]. DICE is a flexible method that accommodates different learning styles that have been shown to play an important role in medical education [18]. As such, each student can work according to their strengths (eg, taking a leading role in the group to cover a specific aspect of the design of a simulated study, like computer coding or result interpretation). Due to the heterogeneity in the working groups, it can be expected that students will use their own learning styles and strengths to learn from other students with different skills [18].

XSL•FO
RenderX

Figure 2. The main learning objectives and outcomes across the steps of DICE (design, interpret, compute, estimate). The steps of DICE include (1) identifying the health problem, (2) designing an investigation including power and sample size calculations with simulations, (3) generating a unique sample of data, (4) analyzing the sample according to the plan, (5) interpreting the findings carefully, and (6) writing a short abstract to be presented in class.



A Guide Through an Example

Each step is now practically explained with an example. The following example is inspired by 2 recent epidemiological studies [19,20]. All information and data are simulated and only serve educational purposes. The computer code in Stata (StataCorp) and R (R Core Team) can be readily used to replicate the example (the code is provided in Multimedia Appendix 1).

Step 1: Identifying the Health Problem

During the first step of DICE, students should think about a particular problem, population, and area that they would like to investigate. This can be somewhat time-consuming and requires a decision about the nature of the research question (ie, causal, descriptive, or predictive) [21]. As we focus on simulating data according to a causal framework explained by Fox et al [9], the research questions are intended to answer a causal question. Other forms of research questions can, of course, be incorporated and simulated; however, they are not the focus of this example. In our example, the aim is to examine the effect of physical activity on the 10-year mortality rate in a large cohort of older people.

Step 2: Designing an Investigation Including Power and Sample Size Calculations With Simulations

The second step addresses the overall design of the study, including the assessment methods for the specified variables. The step is further divided into three specific parts: (1) students should reflect on the appropriate study design (eg, experimental or observational), (2) put forward the possible mechanisms (confounding, interaction, etc) underlying all the random variables involved in the study, and (3) discuss plausible values for all of the parameters. These are discussed in more detail below.

Part 1: Choosing an Appropriate Study Design

Designing an investigation that includes power and sample size calculations with simulations requires careful consideration of available literature and substantive knowledge about the underlying health problem. We recommend allocating sufficient time for this step of planning a realistic simulation study.

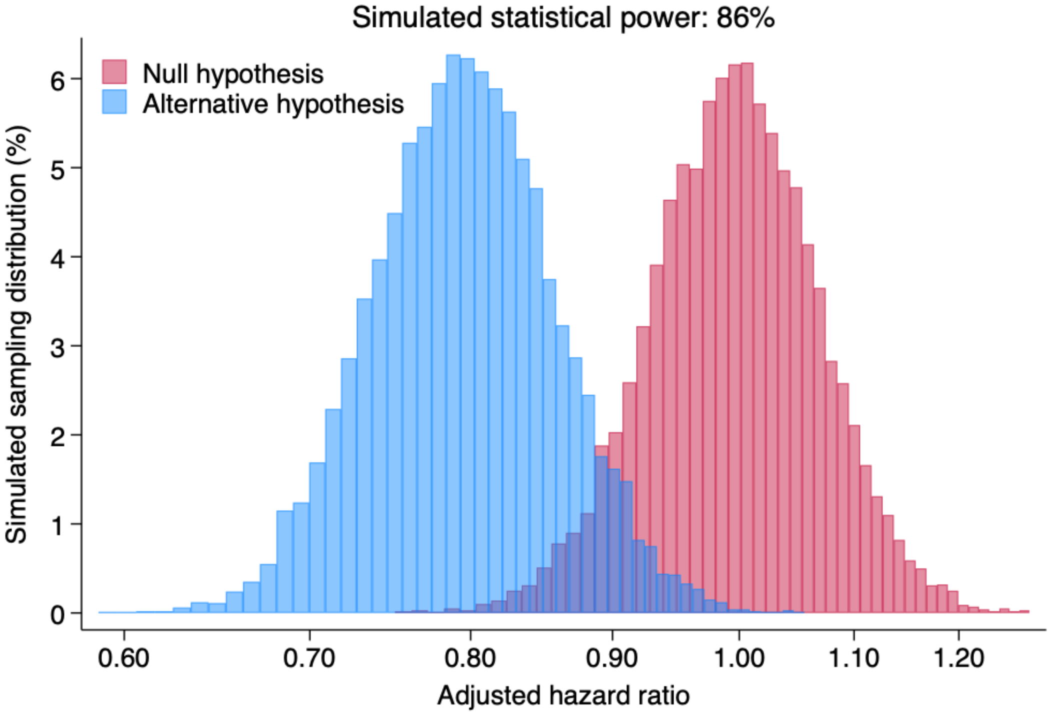
In our example, we design a large, observational cohort study with a confounding effect by age. Information on physical activity (3.5 hours per week of moderate to vigorous physical activity [MVPA] vs less), together with age (≥80 years vs <80 years), is assessed at baseline in a short questionnaire. The mortality rate in a cohort of older people is likely to increase over time due to aging, among both physically active and inactive populations. Assuming a baseline mortality rate in the

XSL•FO
RenderX

younger and physically inactive population of 7 deaths per 1000 person-years, we determined that 5000 individuals (about 1005 deaths during 10-year follow-up) would provide a statistical power of about 86% to detect at least a 20% lower mortality rate (age-adjusted hazard ratio 0.8) in the physically active population relative to the inactive population. A 2-sided Wald-type test for the age-adjusted hazard ratio conferred by physical activity equal to 1 with a type II error of 5% is conducted based on a multivariable Weibull survival model including physical activity and age as covariates.

Figure 3 shows the sampling distribution of the age-adjusted hazard ratio comparing physically active versus inactive individuals under the null and alternative hypotheses.

Figure 3. Simulated sample distribution of the age-adjusted mortality hazard ratio comparing active versus inactive individuals under the null and alternative hypotheses (hazard ratio 0.8). The simulated statistical power was obtained by counting the number of studies that correctly rejected the null hypothesis with a 2-sided Wald-type test at a significance level of 5% based on a multivariable Weibull survival model. The number of simulations is 10,000, the sample size of each study is 5000, and the average number of deaths within each study is 1005.



Part 2: Mechanisms Underlying the Random Variables

Parameters and their distributions can be inspired by previous studies, textbooks, or substantive knowledge from group members. For example, if the exposure is defined as systolic blood pressure (mmHg), students can assume an approximately symmetric and bell-shaped distribution with a given mean and SD and derive the parameter from a normal distribution function. For this study, we need the following variables: (1) *z*, an indicator variable for the older population (1 ">80" vs 0 "≤80 years"); (2) *x*, an indicator variable for the physically active population (1 ">3.5 h/w of MVPA" vs 0 "≤3.5 h/w MVPA"); and (3) *t*, the time from baseline to death (in years) or the end of follow-up (10 years), whichever came first.

Part 3: Define the Values for the Parameters

During this step, students should write a few lines of code or a function capable of generating data according to the desired study and mechanism. Simulations can be used to calibrate the sample size and statistical power of the study. To achieve the desired statistical power (eg, 80%), the sample size can be changed accordingly during this step. This requires some time, and we recommend students try to adapt certain values for the parameters or underlying mechanisms from the previous step (Figure 1). This process is commonly referred to as the data generating mechanism (DGM). We understand DGM as the mechanism underlying the causal structure, including the uncertainty governing the observed data. The simulated power of the statistical test to detect an effect is simply given by the sum of studies that reject the null hypothesis of no effect divided by the total number of simulated studies.

In our example, the first variable to be generated is baseline age (about 60% are older than 80 years of age), which is a confounding variable in the relationship between physical activity and mortality:

For the exposure model, the second variable to be generated is baseline physical activity as a function of age. People aged ≤80 years have a probability of being physically active of 50%, whereas the odds of being physically active among older people are 1/3 (67% lower odds) relative to younger people:

XSL•FO
RenderX

Of note, Bernoulli is a statistical function, whereas logit and ln are a mathematical function.

For the outcome model, individual time-to-death (in years) conditional on the variables physical activity and age is obtained under the Weibull survival model, as follows:

where γ is the parameter defining the departure from a simpler exponential (constant mortality rate) survival model. The value of γ is set to 1.1, indicating a slight increase in the mortality rate over the follow-up period in all the covariate patterns.

The natural logarithm of the baseline mortality rate (per 1 year) among young and inactive people is assumed to be 7 deaths per 1000 person-years, so the intercept is β_0=ln(7/1000)=–4.962

The age-adjusted mortality hazard ratio comparing physically active versus inactive people is set to 0.80, as determined by β_1=ln(0.80)=–0.2231.

The physical activity-adjusted mortality hazard ratio comparing older versus younger people is set to 4, computed as β_2=ln(4)=1.386.

Given a random value for the survival probability *S* ranging over the 0 to 1 interval, a random value of individual time-to-death (in years) conditionally on the variables physical activity and age is obtained as follows:

In addition, any randomly generated time-to-death beyond the follow-up time of 10 years is set to 10 and considered censored (ie, still alive at the end of the follow-up). An indicator variable for death or censored status is also created to inform any survival analysis.

Step 3: Generating a Unique Sample of Data

Once the study has been designed with sufficient statistical power to detect the relevant effect, the next step is to draw one unique sample. Students will analyze and present only this sample in class. The uniqueness and reproducibility of the simulated data are guaranteed by setting a numerical sequence, called a seed, before obtaining realizations of the random variables. This is important for the exact replication of the study. Every group of students is asked to use a common seed in generating the analytical sample of data so that all groups replicate the study under the same conditions. Each group will have a different research question and an underlying health problem with varying parameters. The reason for choosing a seed in the beginning is to highlight the uniqueness of a single study generated under a known DGM. The easiest choice is to specify the seed according to the date of the DICE activity. In our example, we use the seed 20230413 (based on the year, month, and date: "YYYYMMDD"). However, for specific tasks such as power calculations or simulating a distribution of effects, the seed must be deleted to ensure variability in the simulations.

Step 4: Analyzing the Sample of Data

The outcome model is specified according to the process underlying the data, and it is estimated based on the only sample available. Students estimate the statistical model whose performance was evaluated in the initial step of the study design. In our example, we estimate a multivariable Weibull regression model including physical activity and age as covariates.

Step 5: Interpreting the Findings

Students carefully interpret the estimated model and write about the inferential results. In our example, during the 10-year follow-up period, a total of 974 people died out of 5000. Compared with inactive people, the age-adjusted hazard ratio for active people was 11% lower (hazard ratio=0.89; 95% CI 0.78-1.03). A Wald-type 2-sided test indicates some compatibility between this sample of data and the hypothesis of a null age-adjusted mortality hazard ratio for physical activity (z=–1.52; P=.13). This unique sample of data is an example of type II error (failing to reject the null hypothesis, which is indeed incorrect). Nonetheless, the magnitude and direction of the hazard ratio indicate a beneficial effect of physical activity on the 10-year mortality rate. This provides an example of correctly differentiating statistical and scientific inference.

Step 6: Writing an Abstract

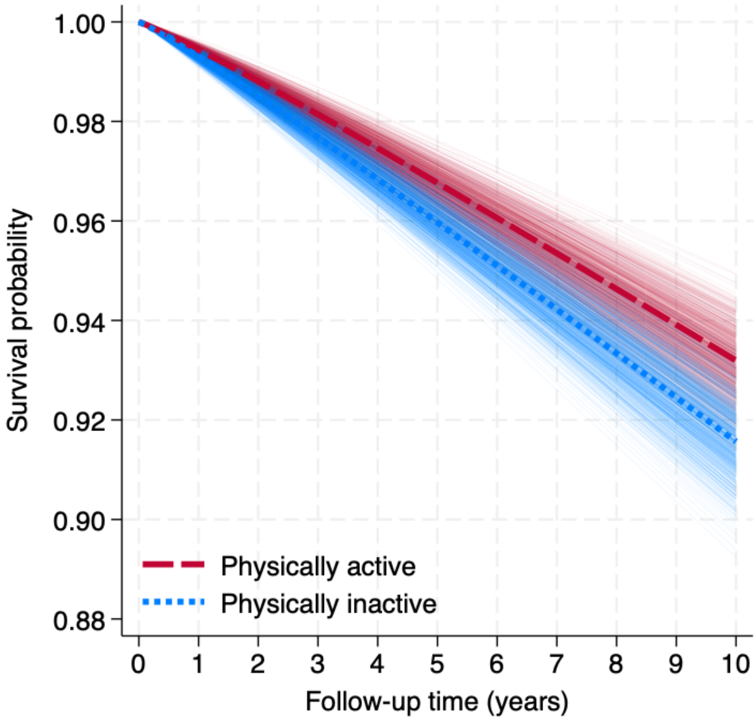
Each group of students should then write a structured scientific abstract (200-250 words) summarizing all the previous steps suitable for an epidemiological conference. The findings and interpretation are then presented in class. Each group of students briefly presents their findings and reasoning behind the study design. Teachers and peers have the possibility of asking questions. The presentations of each group should not exceed 10 minutes per group.

What Have we Learned?

Based on our experiences teaching with DICE and to conclude the steps of DICE shown in the practical example, we hope the key learning lessons for students will include the following: First, students should realize that the most challenging and time-consuming step is the design of the study and identifying a plausible distribution of the random variables involved, the mechanisms underlying the data, and all the parameters included. Second, students should understand that error probabilities (type I, type II, and power) in conducting a test of hypothesis can be easily evaluated by replicating the study many times under similar conditions (Figure 3) using a simulation. Third, students should appreciate the fundamental distinction between the analysis of a single study and the analysis of a collection of estimates obtained from its replication (Figure 4). Fourth, students should understand that the ability of a study to find a relevant exposure or intervention effect (statistical power) can be achieved only with respect to one parameter of interest. Fifth, students should learn that the correct use of statistics plays a key role in all stages of a scientific investigation.

XSL•FO
RenderX

Figure 4. Sampling variability of the estimated survival probability comparing physically active and physically inactive young participants based on 900 simulated studies. The thick dashed lines for both physically active and inactive groups show the functions that were set under the original data-generating mechanism.



Strengths and Limitations of DICE

We proposed an engaging learning method, DICE, to stimulate experimental, active, and enjoyable learning of statistical concepts, fostering key scientific skills in designing and conducting experiments. While the main strengths of this approach lie in its interactivity and group-based nature, we acknowledge several limitations.

First, the proposed simulation method is practically limited to only a few numbers of parameters that can be included in the design of a study. Each additional variable increases the complexity of the DGM exponentially. Thus, this approach is best suited for illustrative, simplified examples of realistic health problems. More sophisticated data derived from multivariate distributions would exceed the simplicity of the method but can, of course, be considered for more advanced classes.

Second, implementing DICE is resource-intensive and should not be done in a short time frame (eg, less than 1 hour). Although this is not a direct limitation of the method, it might be a limitation of its implementation in a classroom.

Third, the effectiveness of DICE in conveying statistical concepts in epidemiology has not been formally evaluated yet. This paper is a description and discussion of the method as implemented in class at a medical university. A formal evaluation of its effectiveness in learning statistics is being devised.

Implementing DICE in the Classroom

Based on the experiences of the authors in using DICE, we summarize the following points for its implementation in the classroom for graduate students in medical sciences, including public health and epidemiology.

First, to implement DICE in a classroom, we recommend a classroom size of approximately 20-40 students, with small groups of 3-5 students from heterogeneous scientific backgrounds. Each group should consist of students who have different strengths and learning styles. We experienced that this could improve interaction between students and increase the joy of learning statistics.

Second, throughout the group work, students are encouraged to discuss and reflect upon the study design, practice the generation and simulation of data under a certain mechanism, and communicate their findings and interpretation of the study. We experienced that some students require more support to understand and use the provided computer code, particularly in settings with fewer students experienced in coding. It can help to go through an example of a simulated study with Stata or R code in front of the class.

Third, DICE can be implemented within a full day of teaching or over several days. For a 1-day implementation, the morning can be used for students to frame their research question and develop the study using simulations (steps 1-3). The afternoon can then be reserved for steps 4-6, ending with the presentation of the abstracts. It is important to keep in mind that the first 2 steps require most of the time (Figure 1). Students should not be rushed through these steps and should be provided with sufficient guidance and support to find an adequate research question, study design, and set up the simulations. Alternatively, DICE can be implemented over several days. An introduction to DICE is given in class, and students can work over several

XSL•FO
RenderX

days in their respective groups. The final day can be used for presenting and discussing the studies and outcomes of each group.

Conclusion

This paper introduces an engaging simulation-based method, DICE, to learn statistics in the health sciences. We argue that DICE can boost statistical reasoning and bridge the gap between substantive knowledge and statistics for all major steps of a scientific investigation. Students can learn fundamental statistical and epidemiological concepts with simulations and combine learning of technical aspects such as coding with theoretical concepts such as error probabilities. The materials in this paper can be readily used by teachers and students.

Acknowledgments

We would like to thank all students of the 2023 master's in public health sciences at Karolinska Institutet for participating in the weekly DICE activities and giving us valuable feedback to improve this teaching method. In addition, the authors would like to thank Stephanie Pitt (Karolinska Institutet) for valuable insights into the draft of the manuscript. Figures 1 and 2 were created on the website of BioRender.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

NO supervised the project. RT wrote the original draft of the manuscript. Both authors contributed equally to the conceptualization, methodology, and software. The authors approved of the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Computer code in Stata and R to replicate the example used in the paper.
[DOCX File , 19 KB - mededu_v10i1e52679_app1.docx]

References

1. Ashby D. Pigeonholes and mustard seeds: growing capacity to use data for society. J R Stat Soc Ser A Stat Soc 2019;182(4):1121-1137 [FREE Full text] [doi: 10.1111/rssa.12483]
2. Tintle N, Chance B, Cobb G, Roy S, Swanson T, VanderStoep J. Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. Am Stat 2015;69(4):362-370. [doi: 10.1080/00031305.2015.1081619]
3. Rudolph JE, Fox MP, Naimi AI. Simulation as a tool for teaching and learning epidemiologic methods. Am J Epidemiol 2021;190(5):900-907 [FREE Full text] [doi: 10.1093/aje/kwaa232] [Medline: 33083814]
4. Hodgson T, Burke M. On simulation and the teaching of statistics. Teach Stat 2000;22(3):91-96. [doi: 10.1111/1467-9639.00033]
5. Novak E. Effects of simulation‐based learning on students' statistical factual, conceptual and application knowledge. J Comput Assist Learn 2014;30(2):148-158. [doi: 10.1111/jcal.12027]
6. Tintle N, Clark J, Fischer K, Chance B, Cobb G, Roy S, et al. Assessing the association between precourse metrics of student preparation and student performance in introductory statistics: results from early data on simulation-based inference vs. nonsimulation-based inference. J Stat Educ 2018;26(2):103-109. [doi: 10.1080/10691898.2018.1473061]
7. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res 2020;90(4):499-541 [FREE Full text] [doi: 10.3102/0034654320933544]
8. Metropolis N, Ulam S. The Monte Carlo method. J Am Stat Assoc 1949;44(247):335-341. [doi: 10.2307/2280232] [Medline: 18139350]
9. Fox MP, Nianogo R, Rudolph JE, Howe CJ. Illustrating how to simulate data from directed acyclic graphs to understand epidemiologic concepts. Am J Epidemiol 2022;191(7):1300-1306 [FREE Full text] [doi: 10.1093/aje/kwac041] [Medline: 35259232]
10. Wood EJ. Problem-based learning: exploiting knowledge of how people learn to promote effective learning. Biosci Educ 2004;3(1):1-12 [FREE Full text] [doi: 10.3108/beej.2004.03000006]
11. Wood DF. Problem based learning. BMJ 2003;326(7384):328-330 [FREE Full text] [doi: 10.1136/bmj.326.7384.328] [Medline: 12574050]
12. Tintle N, Chance BL, Cobb GW, Rossman AJ, Roy S, Swanson T, et al. Introduction to Statistical Investigations, 2nd Edition. Hoboken, NJ: John Wiley & Sons; 2021.

XSL•FO
RenderX

13. Rotondi A, Pedroni P, Pievatolo A. Monte Carlo methods. In: Probability, Statistics and Simulation: With Application Programs Written in R. Cham, Switzerland: Springer; 2022:319-367.
14. Tannenbaum SJ, Holford NHG, Lee H, Peck CC, Mould DR. Simulation of correlated continuous and categorical variables using a single multivariate distribution. J Pharmacokinet Pharmacodyn 2006;33(6):773-794. [doi: 10.1007/s10928-006-9033-1] [Medline: 17053984]
15. Wurdinger SD, Carlson J. Teaching for Experiential Learning: Five Approaches That Work. Lanham, MD: Rowman & Littlefield; 2010.
16. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, et al. Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci USA 2014;111(23):8410-8415 [FREE Full text] [doi: 10.1073/pnas.1319030111] [Medline: 24821756]
17. Anderson LW, Krathwohl DR. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York, NY: Addison Wesley Longman; 2001.
18. Gayef A, Çaylan A, Temiz SA. Learning styles of medical students and related factors. BMC Med Educ 2023 Apr 25;23(1):282 [FREE Full text] [doi: 10.1186/s12909-023-04267-4] [Medline: 37098595]
19. Hassan L, Huhndorf P, Mikolajczyk R, Kluttig A. Physical activity trajectories at older age and all-cause mortality: a cohort study. PLoS One 2023;18(1):e0280878 [FREE Full text] [doi: 10.1371/journal.pone.0280878] [Medline: 36701298]
20. Watts EL, Matthews CE, Freeman JR, Gorzelitz JS, Hong HG, Liao LM, et al. Association of leisure time physical activity types and risks of all-cause, cardiovascular, and cancer mortality among older adults. JAMA Netw Open 2022;5(8):e2228510 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.28510] [Medline: 36001316]
21. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. Chance 2019;32(1):42-49 [FREE Full text] [doi: 10.1080/09332480.2019.1579578]

Abbreviations

DGM: data generating mechanism
DICE: design, interpret, compute, estimate

Edited by T de Azevedo Cardoso; submitted 12.09.23; peer-reviewed by M Waleed, E Ogut, M Marques da Cruz; comments to author 06.11.23; revised version received 17.11.23; accepted 14.02.24; published 15.04.24.

Please cite as:
Thiesmeier R, Orsini N
Rolling the DICE (Design, Interpret, Compute, Estimate): Interactive Learning of Biostatistics With Simulations
JMIR Med Educ 2024;10:e52679
URL: https://mededu.jmir.org/2024/1/e52679
doi:10.2196/52679
PMID:38619866

©Robert Thiesmeier, Nicola Orsini. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.04.2024. 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.

XSL•FO
RenderX

Tutorial

Sharing Digital Health Educational Resources in a One-Stop Shop Portal: Tutorial on the Catalog and Index of Digital Health Teaching Resources (CIDHR) Semantic Search Engine

Julien Grosjean[1,2*], PhD; Arriel Benis[3,4*], PhD; Jean-Charles Dufour[5], MD, PhD; Émeline Lejeune[1], BSc; Flavien Disson[1], MSc; Badisse Dahamna[1,2], MSc; Hélène Cieslik[1], MSc; Romain Léguillon[1,2,6], MSc, PharmD; Matthieu Faure[7], PhD; Frank Dufour[8], PhD; Pascal Staccini[8], MD, PhD; Stéfan Jacques Darmoni[1,2,4], MD, PhD

[1]Department of Digital Health, Rouen University Hospital, Rouen, France

[2]LIMICS, INSERM U1142, Sorbonne Université, Paris, France

[3]Department of Digital Medical Technologies, Holon Institute of Technology, Holon, Israel

[4]European Federation for Medical Informatics, Le Mont-sur-Lausanne, Switzerland

[5]SESSTIM, Aix Marseille Univ, APHM, INSERM, IRD, Hop Timone, BioSTIC, Marseille, France

[6]Department of Pharmacy, Rouen University Hospital, Rouen, France

[7]Délégation du Numérique en Santé, Paris, France

[8]RETINES, Université de Nice Côté d'Azur, Nice, France

[*]these authors contributed equally

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

Abstract

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

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

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

XSL•FO
RenderX

Results: Resource identification and evaluation processes were executed by a dedicated team of medical librarians, aiming to collect and curate an extensive collection of digital health teaching and learning resources. The resources that successfully passed the evaluation process were promptly included in CIDHR. These resources were diligently indexed (with MIMO and FNRDH) and tagged for the study field and degree level. By October 2023, a total of 371 indexed resources were available on a dedicated portal.

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

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

KEYWORDS

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

Introduction

Background

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

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

In 2022, the Delegation of Digital Health of the French Ministry of Health and the French National Research Agency published an open call for projects to support the development of digital health teaching and learning technologies, in French, and dedicated to the community of French health–related professions students and practitioners [6]. These include medicine; dental medicine; pharmacy; midwifery; nursing; physiotherapy; ergotherapy; and, more broadly, any related field such as social work, health administration, and biomedical engineering. By 2027, this heterogeneous community, which includes postgraduates and continuous learners, will reach 210,000 members trained simultaneously in France.

Thus, the association of the departments of digital health (DDHs) of the University of Rouen Normandy (URN) and Côte d'Azur University (CAU) is developing and implementing the SaNuRN (*Santé Numérique Rouen Nice*) [7], a 5-year project started in September 2022 and granted with €3,951,200 (US $4,163,775) for a total cost of €6,891,923 (US $7,262,708), in the context of the said open call (grant #ANR_22-CMAS-0014) having an overall budget of €71 million (US $77.6 million) dedicated by the government to digital health education.

From an educational perspective, SaNuRN is currently based on existing pedagogical resources developed by the DDHs of URN and CAU. In addition, a large part of these resources follows the concepts structuring the French National Referential on Digital Health (FNRDH) [8] that provides French higher education institutions educating health-related professionals with a guideline to support teaching in digital health. Thus, students and lecturers from URN, CAU, and other higher education institutions and professionals have free and unrestricted access to the *Catalog and Index of Digital Health Teaching Resources* (CIDHR) as a platform providing structured and validated information contributing to the body of knowledge necessary to master the field [9].

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

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

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

One of the primary reasons for emphasizing the need for a French-speaking knowledge catalog in the digital health domain, such as CIDHR, is to bridge the language gap. Although English

XSL•FO
RenderX

is a dominant language in scientific literature and teaching platforms, it excludes a substantial portion of the global population, particularly those more comfortable with other languages and, more particularly, French in this specific case. Thus, this language barrier can hinder the dissemination of critical information and knowledge transfer in digital health education and the development of a dedicated platform in French (which can comprise resources in other languages) [16-19].

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

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

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

Aim, Objective, and Hypothesis

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

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

Methods

Highlights

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

Resources Identification

To identify new or updated digital health teaching and learning resources, a group of 3 librarians from URN DDH is working on a continuous information watch, according to an internally developed and validated process comprising the steps and actions.

XSL•FO
RenderX

Thus, the librarians search proprietarily on a predefined list of academic websites of Schools of Health Sciences (eg, Medicine, Dental Medicine, Pharmacy, Nursing, Rehabilitation), National Agencies (eg, the French Ministry of Health [26]; the French National Authority for Health—*La Haute Autorité de Santé* [27]; the French national agency for medicines and health products safety—*Agence Nationale de sécurité du médicament*; and the French Agency for Food, Environmental and Occupational Health & Safety—*Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail*); and other organizations involved in digital health education such as universities in France and around the world. They are also using search engine alerts, allowing reception of emails with potentially interesting content detected by their algorithms.

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

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

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

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

Librarians evaluate each potential resource against the following three criteria:

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

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

Resources Indexation

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

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

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

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

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

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

For example, a user (eg, a medical student) interested in learning about the use of artificial intelligence in digital health can use CIDHR to find learning resources that are indexed with the following MIMO concepts: "artificial intelligence," "digital health," "machine learning," and "data mining"; or the same user can find resources indexed with the following FNRDH skill: "use of artificial intelligence in digital health." Accordingly, CIDHR provides a list of relevant educational resources.

XSL•FO
RenderX

Using KOSs and metadata sets is a common practice in digital learning to organize and represent digital learning resources in a flexible, comprehensive, and user-friendly manner.

Resources Tagging and Integration to the Curricula

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

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

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

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

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

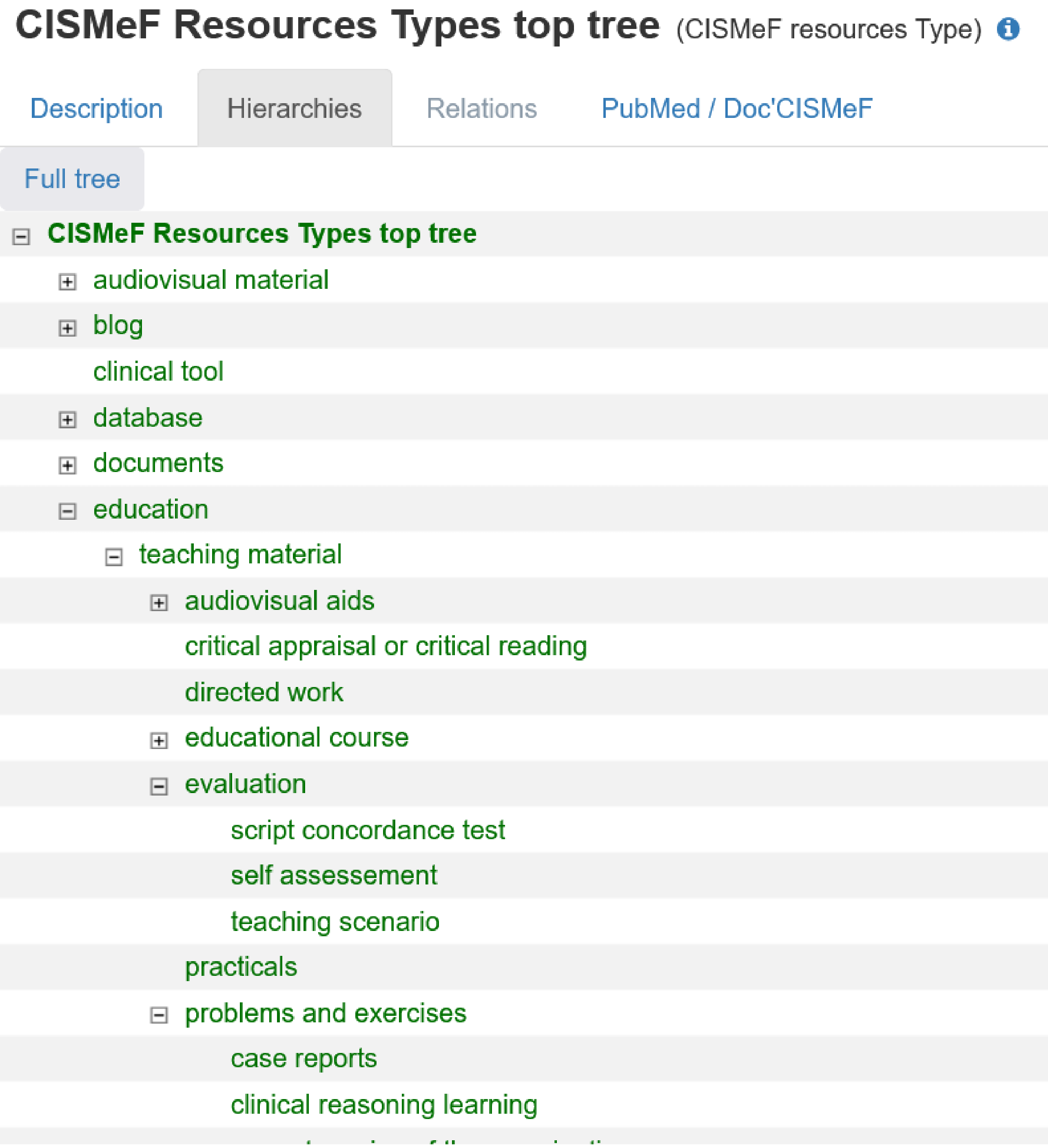
Moreover, any kind of teaching resource is cataloged in CIDHR, thanks to an extensive resources type hierarchy created for CISMeF based on a conceptual extension of the MeSH publication type [20,34]. This resource-type hierarchy has been used fruitfully for more than 20 years by users (health students,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

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

These combinations of the metadata tags "field of study" and "degree level" with the "resource type" tag as filters allow delivery to the user more or fewer indexed resources relevant to the knowledge fields submitted in the query to CIDHR depending on the filters selection submitted with the query.

XSL•FO
RenderX

Figure 2. List (sample) of resource types for teaching resources in the Health Terminology/Ontology Portal. CISMeF: Catalogue et Index des Sites Médicaux en langue Française (Catalog and Index of Medical Sites in French Language).



User Experience Assessment

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

Ethical Considerations

This research is dispensed of the ethical committee's approval, the User Feedback for Continuous Improvement being a normal educational practice and classroom management method

XSL•FO
RenderX

conducted in educational settings. Specifically, as non-interventional research dealing with practical habits analysis the Rouen University Hospital ethical committee does not ask for submitting such kind of research to the ethical committee. Moreover, the whole project SaNuRN that comprises CIDHR has been approved as a whole by the Delegation of Digital Health of the French Ministry of Health and the French National Research Agency [38].

Results

Resource Discovery and Indexation in CIDHR

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

In the last year, we identified approximately 500 valuable resources. It is noteworthy that the identified resources reflect a multilingual character (in particular, English). However, to align with the SaNuRN grant requirements, a substantial proportion (>90%) of the resources is in French. However, we ensured a representation of diverse languages to accommodate a wide-ranging audience interested in digital health education. In addition, we supported the ongoing internationalization and English-drafted teaching and self-learning introduced in the French higher education curricula.

Resources Evaluation

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

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

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

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

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

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

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

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

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

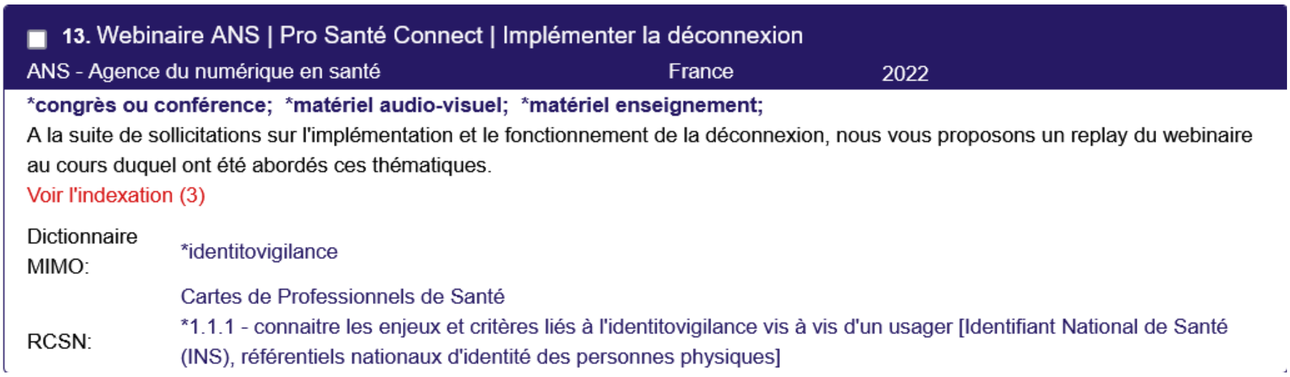
Figure 3 shows an example of the results for the query "dossiers médicaux électroniques" (in English, "electronic health records" or EHRs).

XSL•FO
RenderX

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

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

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



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

To facilitate the indexing process with FNRDH, which presents considerable complexity for medical librarians, the SaNuRN pedagogical team has established manual associations between

XSL•FO
RenderX

MIMO and FNRDH concepts. For instance, this involves manually linking the MIMO concept with the FNRDH competency. It is essential to clarify that this mapping relation does not constitute a strict "exact match"; instead, it means that when a librarian indexes a teaching resource using a MIMO concept (eg, "electronic medical records") associated with an FNRDH ability (eg, "Interact appropriately between professionals, with the healthcare customer, caregivers and companions and with institutions and administrations"), the educational resource is also indexed with this corresponding FNRDH competency.

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

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

User Feedback for Continuous Improvement

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

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

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

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

Discussion

Overview

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

Strengths and Limitations

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

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

XSL•FO
RenderX

diverse linguistic needs and are available to all individuals seeking to enhance their digital health knowledge. With CIDHR's vast collection of indexed educational resources, individuals can access high-quality information and support regardless of their native or daily spoken language.

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

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

However, some limitations have been identified. First, although CIDHR supports mainly French resources, it would benefit from expanding its multilingual and international support to make it more accessible to a global audience of the digital health education community. Second, it is necessary to expand CIDHR resource collection to incorporate more digital health resources from diverse sources allowing providing them to the educational community and industry insights. Third, although SaNuRN plans to provide personalized learning paths to users, via CIDHR, it is crucial to ensure that these paths are effective and tailored to the individual needs of each user, which requires further research and development [45,46]. Fourth, integrating CIDHR with the learning management systems used by educational institutions would streamline access to digital health resources for students and educators. However, it is crucial to ensure that the integration is smooth and that CIDHR is easy to use within these systems. Finally, developing a feedback and rating system for resources would be helpful in enabling users to identify the most valuable and reliable materials within the platform. However, it is vital to design the system carefully to ensure that it is fair and unbiased. Moreover, it is important to note that CIDHR is under development, and there may be some bugs or glitches in the system. In addition, some features may not be fully implemented.

Future Perspectives

Handling the current limitations of CIDHR opens a wide range of perspectives.

To improve the accessibility and user-friendliness of CIDHR, the SaNuRN team will look at different paths. First, expanding multilingual support to cater to a wider global audience (over the French-speaking community) by indexing (based on MIMO as a multilingual ontology dedicated to digital health) more resources in more languages MIMO on the platform. In addition, CIDHR enrichment will benefit from the SaNuRN team's international partnerships and collaborations to expand CIDHR resource collection and promote knowledge exchange to enrich the user experience [23,47]. Moreover, an additional enhancement is planned to provide personalized learning paths to users based on their profiles, such as their field of study, degree level, CIDHR personal and similar user use, to enable tailored educational experiences and effectiveness. Furthermore, CIDHR will be integrated with the learning management systems used by educational institutions to streamline access to digital health resources for students and educators (eg, Moodle [48]). Finally, CIDHR will benefit from the development of a feedback and rating system for resources not only to help users identify the most valuable and reliable materials within the platform but also to allow the SaNuRN team project to get feedback on the resource collection, indexing, and tagging processes from mass users' practice. All these measures will augment CIDHR utility and enrich the user experience.

Conclusions

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

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

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

A European project called the HosmartAI (Hospital Smart development based on AI) project deals with the digital transformation of the European health care sector to make the European health care system more strong, efficient, sustainable, and resilient. CIDHR can play an important role in acquisition of literacy in digital health for professionals [49]. The European

XSL•FO
RenderX

Federation for Medical Informatics is taking part in different projects such as HosmartAI and as a collaboration and cooperation-oriented scientific and academic international organization, it can help disseminate information about CIDHR to promote its use by an increasing number of members of the digital health educational community worldwide.

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

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

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

Acknowledgments

This work was partially supported by the SaNuRN project (ANR_22-CMAS-0014 3.951.200) granted by the Delegation of Digital Health of the French Ministry of Health and the French National Research Agency.

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

Data Availability

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

Authors' Contributions

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

Conflicts of Interest

None declared.

References

1. Steinhubl SR, Topol EJ. Digital medicine, on its way to being just plain medicine. NPJ Digit Med 2018 Jan 15;1(1):20175 [FREE Full text] [doi: 10.1038/s41746-017-0005-1] [Medline: 31304349]
2. Ma M, Li Y, Gao L, Xie Y, Zhang Y, Wang Y, et al. Correction: the need for digital health education among next-generation health workers in China: a cross-sectional survey on digital health education. BMC Med Educ 2023 Sep 22;23(1):688 [FREE Full text] [doi: 10.1186/s12909-023-04613-6] [Medline: 37737169]
3. Global strategy on digital health 2020-2025. World Health Organization. URL: https://tinyurl.com/58d488cu [accessed 2024-01-03]
4. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://apps.who.int/iris/handle/10665/260480 [accessed 2022-08-13]
5. Tudor Car L, Poon S, Kyaw BM, Cook DA, Ward V, Atun R, et al. Digital education for health professionals: an evidence map, conceptual framework, and research agenda. J Med Internet Res 2022 Mar 17;24(3):e31977 [FREE Full text] [doi: 10.2196/31977] [Medline: 35297767]

XSL•FO
RenderX

6. Appel à manifestation d'intérêt (AMI) « compétences et métiers d'avenir ». Ministère du Travail, du Plein Emploi et de L'insertion. 2022 Feb 11. URL: https://tinyurl.com/59uxf7fk [accessed 2024-01-04]
7. Digital health training program - ANR_22-CMAS-0014 3.951.200. Sante Numerique Rouen Nice. URL: https://sanurn.eu/ [accessed 2024-01-04]
8. Arrêté du 10 novembre 2022 relatif à la formation socle au numérique en santé des étudiants en santé. Légifrance. 2022 Nov 11. URL: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000046548689 [accessed 2023-09-29]
9. CIDHR: catalog and index of digital health teaching resources on the internet. CHU Rouen Normandie. URL: https://doccismef.chu-rouen.fr/dc/#env=cidhr [accessed 2023-09-23]
10. Beux P, Duff F, Fresnel A, Berland Y, Beuscart R, Burgun A, et al. The French virtual medical university. Stud Health Technol Inform 2000;77:554-562. [Medline: 11187614]
11. Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J, Thirion B. CISMeF: a structured health resource guide. Methods Inf Med 2000 Mar;39(1):30-35. [Medline: 10786067]
12. Cours de santé numérique (digital health). CHU Rouen Normandie. URL: https://www.cismef.org/cismef/d2im/cours/ [accessed 2023-09-15]
13. Pioneering digital healthcare education platform celebrates 5-year anniversary. Healthcare Leadership Academy. URL: https://tinyurl.com/epkhae9y [accessed 2023-09-29]
14. Platforms and content. National Health Service England. URL: https://tinyurl.com/mread9msn [accessed 2024-01-03]
15. IMD Health home page. IMD Health. URL: https://www.imdhealth.com/ [accessed 2024-01-04]
16. McCall M, Spencer E, Owen H, Roberts N, Heneghan C. Characteristics and efficacy of digital health education: an overview of systematic reviews. Health Educ J 2018;77(5):497-514. [doi: 10.1177/0017896918762013]
17. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
18. Cresswell K, Sheikh A, Franklin BD, Krasuska M, The Nguyen H, Hinder S, et al. Interorganizational knowledge sharing to establish digital health learning ecosystems: qualitative evaluation of a national digital health transformation program in England. J Med Internet Res 2021 Aug 19;23(8):e23372 [FREE Full text] [doi: 10.2196/23372] [Medline: 34420927]
19. Carlson ES, Barriga TM, Lobo D, Garcia G, Sanchez D, Fitz M. Overcoming the language barrier: a novel curriculum for training medical students as volunteer medical interpreters. BMC Med Educ 2022 Jan 10;22(1):27 [FREE Full text] [doi: 10.1186/s12909-021-03081-0] [Medline: 35012526]
20. Darmoni SJ, Thirion B, Leroy JP, Douyère M, Lacoste B, Godard C, et al. Doc'CISMEF: a search tool based on "encapsulated" MeSH thesaurus. Stud Health Technol Inform 2001;84(Pt 1):314-318. [Medline: 11604754]
21. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, et al. Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform 2011;166:129-138. [Medline: 21685618]
22. Grosjean J, Merabti T, Griffon N, Dahamna B, Darmoni S. Multiterminology cross-lingual model to create the European health terminology/ontology portal. In: Proceedings of the 9th International Conference on Terminology and Artificial Intelligence. 2011 Presented at: 9th International Conference on Terminology and Artificial Intelligence; November 8-10, 2011; Paris, France.
23. Benis A, Crisan-Vida M, Stoicu-Tivadar L, Darmoni S. A multi-lingual dictionary for health informatics as an international cooperation pillar. Stud Health Technol Inform 2019 Jul 04;262:31-34. [doi: 10.3233/SHTI190009] [Medline: 31349258]
24. Benis A, Grosjean J, Billey K, Montanha G, Dornauer V, Cri an-Vida M, et al. Medical informatics and digital health multilingual ontology (MIMO): a tool to improve international collaborations. Int J Med Inform 2022 Nov;167:104860. [doi: 10.1016/j.ijmedinf.2022.104860] [Medline: 36084537]
25. Darmoni S, Benis A, Lejeune E, Disson F, Dahamna B, Weber P, et al. Digital health multilingual ontology to index teaching resources. Stud Health Technol Inform 2022 Aug 31;298:19-23. [doi: 10.3233/SHTI220900] [Medline: 36073449]
26. Actualités. Ministère de la Santé et de la Prévention. URL: https://sante.gouv.fr/ [accessed 2024-01-04]
27. Haute Autorité de Santé home page. Haute Autorité de Santé. URL: https://www.has-sante.fr/ [accessed 2024-01-04]
28. HeTOP (Health Terminology/Ontology Portal) home page. HeTOP. URL: https://www.hetop.eu/hetop/en/ [accessed 2022-04-29]
29. MIMO dictionary: MIMO medical informatics thesaurus. HeTOP. URL: https://www.hetop.eu/hetop/rep/en/EFMIMIMO/ [accessed 2024-01-04]
30. RCSN: référentiel socle et transversal de compétences en santé numérique. HeTOP. URL: https://www.hetop.eu/hetop/rep/fr/TER_RCSN/ [accessed 2024-01-04]
31. Standard for learning metadata. IEEE Standards Association. URL: https://standards.ieee.org/ieee/2881/10248/ [accessed 2024-01-04]
32. DCMI metadata terms. DublinCore. URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ [accessed 2024-01-04]
33. Accueil LOMFR. LOM-fr. URL: http://www.lom-fr.fr/ [accessed 2024-01-04]
34. Publication characteristics (publication types) with scope notes. National Institutes of Health National Library of Medicine. URL: https://www.nlm.nih.gov/mesh/pubtypes.html [accessed 2024-01-04]

XSL•FO
RenderX

35. Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP, et al. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 2004 Dec;21(4):253-261 [FREE Full text] [doi: 10.1111/j.1471-1842.2004.00526.x] [Medline: 15606883]
36. Darmoni SJ, Pereira S, Sakji S, Merabti T, Prieur É, Joubert M, et al. Multiple terminologies in a health portal: automatic indexing and information retrieval. In: Proceedings of the 12th Conference on Artificial Intelligence in Medicine in Europe. 2009 Presented at: 12th Conference on Artificial Intelligence in Medicine in Europe; July 18-22, 2009; Verona, Italy. [doi: 10.1007/978-3-642-02976-9_37]
37. Kleib M, Arnaert A, Nagle LM, Ali S, Idrees S, Kennedy M, et al. Digital health education and training for undergraduate and graduate nursing students: a scoping review protocol. JBI Evid Synth 2023 Jul 01;21(7):1469-1476. [doi: 10.11124/JBIES-22-00266] [Medline: 36728743]
38. Regulatory Approaches In Research. University of Rouen. URL: https://dumg-rouen.fr/p/ethique-et-protection-des-donnees [accessed 2024-01-04]
39. Google alerts. Google. URL: https://www.google.com/alerts [accessed 2024-01-04]
40. Creating alerts: PubMed. National Institutes of Health. URL: https://tinyurl.com/cwzyev47 [accessed 2024-01-04]
41. LinkedIn log in. LinkedIn. URL: https://www.linkedin.com/ [accessed 2024-01-04]
42. Burgstahler S, Corrigan B, McCarter J. Making distance learning courses accessible to students and instructors with disabilities: a case study. Internet High Educ 2004;7(3):233-246. [doi: 10.1016/j.iheduc.2004.06.004]
43. Zhang X, Tlili A, Nascimbeni F, Burgos D, Huang R, Chang TW, et al. Accessibility within open educational resources and practices for disabled learners: a systematic literature review. Smart Learn Environ 2020 Jan 03;7:1. [doi: 10.1186/s40561-019-0113-2]
44. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010 Dec 04;376(9756):1923-1958. [doi: 10.1016/S0140-6736(10)61854-5] [Medline: 21112623]
45. Valenta AL, Berner ES, Boren SA, Deckard GJ, Eldredge C, Fridsma DB, et al. AMIA board white paper: AMIA 2017 core competencies for applied health informatics education at the master's degree level. J Am Med Inform Assoc 2018 Dec 01;25(12):1657-1668 [FREE Full text] [doi: 10.1093/jamia/ocy132] [Medline: 30371862]
46. Benis A, Tamburis O, Chronaki C, Moen A. One digital health: a unified framework for future health ecosystems. J Med Internet Res 2021 Feb 05;23(2):e22189 [FREE Full text] [doi: 10.2196/22189] [Medline: 33492240]
47. Benis A, Crisan-Vida M, Stoicu-Tivadar L. The EFMI working group "healthcare informatics for interregional cooperation": an evolving strategy for building cooperation bridges. Stud Health Technol Inform 2019 Aug 21;264:1907-1908. [doi: 10.3233/SHTI190707] [Medline: 31438401]
48. Moodle home page. Moodle. URL: https://moodle.org/ [accessed 2024-01-04]
49. HosmartAI home page. HosmartAI. URL: https://www.hosmartai.eu/ [accessed 2022-03-08]
50. Mantas J, Hasman A. IMIA educational recommendations and nursing informatics. Stud Health Technol Inform 2017;232:20-30. [Medline: 28106578]
51. Chen D, Gorla J. The need to develop digital health competencies for medical learners. Med Teach 2023 Jul;45(7):790-791. [doi: 10.1080/0142159X.2023.2178886] [Medline: 36787406]
52. Seemann RJ, Mielke AM, Glauert DL, Gehlen T, Poncette AS, Mosch LK, et al. Implementation of a digital health module for undergraduate medical students: a comparative study on knowledge and attitudes. Technol Health Care 2023;31(1):157-164 [FREE Full text] [doi: 10.3233/THC-220138] [Medline: 35754241]
53. Benis A, Haghi M, Deserno TM, Tamburis O. One digital health intervention for monitoring human and animal welfare in smart cities: viewpoint and use case. JMIR Med Inform 2023 May 19;11:e43871 [FREE Full text] [doi: 10.2196/43871] [Medline: 36305540]
54. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
55. Wilkinson MD, Sansone SA, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. Sci Data 2018 Jun 26;5(1):180118 [FREE Full text] [doi: 10.1038/sdata.2018.118] [Medline: 29944145]
56. Tamburis O, Benis A. One digital health for more FAIRness. Methods Inf Med 2022 Dec;61(S 02):e116-e124 [FREE Full text] [doi: 10.1055/a-1938-0533] [Medline: 36070786]

Abbreviations

CAU: Côte d'Azur University
CIDHR: Catalog and Index of Digital Health Teaching Resources
CISMeF: Catalogue et Index des Sites Médicaux en langue Française (Catalog and Index of Medical Sites in French Language)
DCMI-MT: Dublin Core Metadata Terms

XSL•FO
RenderX

DDH: department of digital health
EHR: electronic health record
FNRDH: French National Referential on Digital Health
HosmartAI: Hospital Smart development based on AI
KOS: knowledge organization system
LOM: Learning Object Metadata
MeSH: Medical Subject Headings
MIMO: Medical Informatics Multilingual Ontology
SaNuRN: Santé Numérique Rouen Nice
SoM: sets of metadata
URN: University of Rouen Normandy

Edited by T Leung, T de Azevedo Cardoso; submitted 21.04.23; peer-reviewed by M Wolfien, HY Yoon; comments to author 09.06.23; revised version received 13.10.23; accepted 18.12.23; published 04.03.24.

Please cite as:
Grosjean J, Benis A, Dufour JC, Lejeune É, Disson F, Dahamna B, Cieslik H, Léguillon R, Faure M, Dufour F, Staccini P, Darmoni SJ
Sharing Digital Health Educational Resources in a One-Stop Shop Portal: Tutorial on the Catalog and Index of Digital Health Teaching Resources (CIDHR) Semantic Search Engine
JMIR Med Educ 2024;10:e48393
URL: https://mededu.jmir.org/2024/1/e48393
doi:10.2196/48393
PMID:38437007

©Julien Grosjean, Arriel Benis, Jean-Charles Dufour, Émeline Lejeune, Flavien Disson, Badisse Dahamna, Hélène Cieslik, Romain Léguillon, Matthieu Faure, Frank Dufour, Pascal Staccini, Stéfan Jacques Darmoni. Originally published in JMIR Medical Education (https://mededu.jmir.org), 04.03.2024. 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.

XSL•FO
RenderX

Tutorial

How to Develop an Online Video for Teaching Health Procedural Skills: Tutorial for Health Educators New to Video Production

Komal Srinivasa[1], MBChB, PGDip (ClinEd); Amanda Charlton[2,3], MBChB, PGDip (ClinEd); Fiona Moir[1], MBChB, PhD; Felicity Goodyear-Smith[1], MBChB, MD

[1]Department of General Practice & Primary Health Care, The University of Auckland, Auckland, New Zealand

[2]Department of Histopathology, Auckland City Hospital, Auckland, New Zealand

[3]Department of Molecular Medicine and Pathology, The University of Auckland, Auckland, New Zealand

Corresponding Author:
Komal Srinivasa, MBChB, PGDip (ClinEd)
Department of General Practice & Primary Health Care
The University of Auckland
Private Bag 92019
Auckland, 1010
New Zealand
Phone: 64 9 923 1975
Email: komal.srinivasa@auckland.ac.nz

Abstract

Background: Clinician educators are experts in procedural skills that students need to learn. Some clinician educators are interested in creating their own procedural videos but are typically not experts in video production, and there is limited information on this topic in the clinical education literature. Therefore, we present a tutorial for clinician educators to develop a procedural video.

Objective: We describe the steps needed to develop a medical procedural video from the perspective of a clinician educator new to creating videos, informed by best practices as evidenced by the literature. We also produce a checklist of elements that ensure a quality video. Finally, we identify the barriers and facilitators to making such a video.

Methods: We used the example of processing a piece of skeletal muscle in a pathology laboratory to make a video. We developed the video by dividing it into 3 phases: preproduction, production, and postproduction. After writing the learning outcomes, we created a storyboard and script, which were validated by subject matter and audiovisual experts. Photos and videos were captured on a digital camera mounted on a monopod. Video editing software was used to sequence the video clips and photos, insert text and audio narration, and generate closed captions. The finished video was uploaded to YouTube (Google) and then inserted into open-source authoring software to enable an interactive quiz.

Results: The final video was 4 minutes and 4 seconds long and took 70 hours to create. The final video included audio narration, closed captioning, bookmarks, and an interactive quiz. We identified that an effective video has six key factors: (1) clear learning outcomes, (2) being engaging, (3) being learner-centric, (4) incorporating principles of multimedia learning, (5) incorporating adult learning theories, and (6) being of high audiovisual quality. To ensure educational quality, we developed a checklist of elements that educators can use to develop a video. One of the barriers to creating procedural videos for a clinician educator who is new to making videos is the significant time commitment to build videography and editing skills. The facilitators for developing an online video include creating a community of practice and repeated skill-building rehearsals using simulations.

Conclusions: We outlined the steps in procedural video production and developed a checklist of quality elements. These steps and the checklist can guide a clinician educator in creating a quality video while recognizing the time, technical, and cognitive requirements.

(JMIR Med Educ 2024;10:e51740) doi:10.2196/51740

KEYWORDS

online video; developing video; procedural video; medical education; clinician educator; health education

XSL•FO
RenderX

Introduction

Up to 87% of surgical trainees routinely watch online laparoscopic videos as a part of multimedia learning, meeting a "user demand" [1]. Research has demonstrated that users consider online videos significantly more helpful than other resources due to improved self-confidence and navigability of the resource [2]. Furthermore, videos enhance active learning components for health professionals, foster a community of inquiry, and facilitate social interactions in online courses [3].

From an educator's perspective, there are pedagogical reasons for using videos, such as creating a student-centered learning environment by enabling students to take an active role in their learning, increasing student engagement, and demonstrating procedures in a standardized and stepwise fashion [4]. While videos can be educational and interactive, these qualities depend on how they are designed and used in a learning context.

There are many options for health educators wanting to create a procedural video, ranging from outsourcing the entire process to doing it all themselves. Educators can engage professionals from the media production department in their institutions or private companies. They can also collaborate with students, residents, and colleagues with experience and interest in creating videos. The choice depends on the time, skills, funding, and, most importantly, the level of involvement the educator desires.

While the reasons for using an online educational video are numerous and well-researched, there is limited literature on developing a high-quality medical educational online video for teaching procedural skills. Several studies have shown that 10%-40% of medical videos on YouTube lack essential safety information [5,6]. Several studies have assessed the quality and content of medical procedural videos on YouTube and found them to be of variable educational value [7,8]. Health educators aiming to create high-quality videos will benefit from a clear understanding of the video development process and the required production skills. Therefore, we provide stepwise guidance and a quality checklist and identify barriers and facilitators to make a quality procedural video.

In this tutorial, we present the adaption of a previously described structure of dividing the development of a video into (1) preproduction, (2) production, and (3) postproduction phases to create a medical procedural educational online video. These phases are further divided into background factors and practical applications.

Methods

Preproduction Phase

Background Factors in the Preproduction Phase

A scoping review before starting the video recording process and previous attendance at courses on clinical education theories informed the steps of the process. In the preproduction phase, certain background factors need to be considered which are listed below in detail.

Needs Analysis

A video's purpose, format, and content can be determined by performing a needs assessment before incorporating the video into a lesson plan [9].

Learning Outcomes

The learning outcomes should be clearly defined and aligned with an assessment taxonomy for clinical skills, such as Miller's pyramid [10], as a video relates to procedural skills acquisition. Miller's pyramid is a clinical assessment framework that defines a learner's ability into the knows, knows how, shows how, and does categories that test progress from basic knowledge to practical application [10].

Learner-Centered

Allowing for different learning preferences will ensure maximal student engagement and active learning and contribute to a learner-centered environment [4].

Mayer's Principles of Multimedia Learning

Different parts of the cerebral cortex process audio and visual content, and only a certain amount of information is held in working memory at any given time (cognitive load). Applying Mayer's principles of multimedia while developing e-learning resources can reduce cognitive overload [11] as these systems can become overwhelmed [12]. These include 12 principles, such as signaling (visual cues are added to multimedia to signal the main concepts), coherence (unnecessary content is removed), and segmentation (an online video is broken into small units or pieces to enable the learner to process this information before moving to the next stage) [11,12]. While narration (using the educator's voice) is essential [12], adding other audio media, such as music or background noise, adds to the cognitive load [3]. Multimedia principles should be considered when weighing up the optimal combination of text, images, and narration to communicate information.

Guidelines for the Design of Instructional Videos

In 2022, Meij and Hopfner suggested guidelines for the design of instructional videos for software design, which can also provide guidance for medical instructional videos [13]. These include (1) keeping instructional videos short, (2) supporting users in finding a suitable video by including the video purpose in the title, (3) previewing the task, (4) using a screencast with narration, (5) supporting an action-oriented approach, (6) consider key components of a well-designed procedure, (7) make task demonstration easy to follow and mimic, (8) support users in handling the transitory nature of the video, (9) review the task, (10) strengthen demonstration with practice, and (11) occasionally include music.

In addition to these, a clinical educator should be aware that the current video size guidelines for full high-definition videos are 1920×1080 pixels and 149 MB per minute [14].

Storyboard, Script, and Shot List

Preparing a storyboard, script, and shot list outlining the exact steps and scenes of the video optimizes the chance of smooth recording on the day [15,16]. If the content is unfamiliar then this should be peer-reviewed by a subject matter expert (who

XSL•FO
RenderX

confirms the subject content is accurate and in line with the learning outcomes). For advice on audiovisual matters, checking the storyboard and shot list with an audiovisual expert would be useful [9,15]. The audiovisual expert is a person trained in media production who has expertise in producing videos. They may be from the audiovisual department of the organization or an outside company. Following this process ensures optimal video quality [15]. Validating the video script may address 6 questions related to the video's objective, content, relevance, environment, verbal language, and inclusion of topics [15]. A knowledge of common video recording language and techniques will help inform the writing of the storyboard and shot list [12]. These include (1) framing a shot, (2) camera placement, (3) camera angle, (4) zooming in or out, (5) panning, and (6) cutting with purpose. These are explained further in Table 1.

Table 1. Common video recording language and techniques.

Technical word	Definition
Framing a shot	A shot is what is presented on screen. So, aim for the smallest frame necessary for the shot to balance detail and context.
Camera placement	The placement of the camera is essential. Examples include the 180-degree rule and the 30-degree rule. The 180-degree rule states that when recording 2 objects, the camera placement should not exceed 180 degrees from each other to provide consistency. The 30-degree rule states that the camera angles should be at least 30 degrees apart when changing angles.
Camera angle	Camera angles refer to the placement of the camera (ie, what the viewer sees). Examples include eye level, low angle, high angle, and bird's eye, each with a different purpose. Tilting up or down is changing camera angles while the video is still rolling.
Zooming	It adjusts the focal length of the lens (in or out).
Panning	It is moving the camera from left to right.
Cutting with purpose	Is moving between scenes or camera angles on purpose.

Video Duration

The duration of an online video is another factor to be considered. There is no consensus on the optimal length of a procedural video, as this depends on several contextual factors. Videos of shorter duration have been shown to have better viewer engagement, and some authors state that videos should be limited to 6 minutes or less [4,17]. Others suggest they should be 10-15 minutes long [3]. Ideally, the length should range from 6 to 15 minutes, and videos longer than this should be subdivided into shorter parts or chapters labeled with time-stamped content [3,4].

Title

A video title should make the learner aware of the video's goal and the session's intended learning outcomes [13]. A good title increases the ease with which an educator or learner can search for the online video [13].

Educator Presence and Narration

Having the educator's face or head and shoulders in the video and narrating with the educator's voice [18] is also suggested to increase the learner's connectedness with the content and educator [3]. This is especially true in asynchronous and remote learning situations where this "human touch" improves a learner's sense of teacher presence. This improved sense of connectivity, either in a planned lesson or during self-directed learning, can lead to a more learner-centered environment, promote active learning, and create a sense of community [18]. Audio narration in a conversational and enthusiastic tone (Mayer's principle of personalization) and at a reasonably fast pace (185-254 words per minute) is preferable [17].

Table of Contents and Time Stamping

A table of contents with links to specific time points (time stamping, chapters, or bookmarks) within a video also increases user navigation and control. These features save time as well as improve the ease of access and functional interactivity [17]. In instructional videos, interactive features such as pauses and quizzes can test the learners' knowledge, enable reflection, and improve engagement [17].

Health Information Governance

Health information governance issues are essential in the preproduction phase and can be divided into ethical or professional considerations. Ethical approval might be required from specific jurisdictions and institutional groups before recording. Any patient images (still or video form) must be collected after their consent (in some jurisdictions), used respectfully, and stored to ensure patient privacy and confidentiality. Professional considerations relate to copyright, data protection, and indigenous populations' sovereignty issues. As using content created by someone else may have copyright issues, specific permission may be required, especially if the video content is modified. Live streaming procedures in surgical broadcasting or coaching require knowledge of confidentiality and health information laws [19]. The final video and raw data must be stored with data protection considerations. There may be Indigenous Populations' data sovereignty issues to consider, such as Māori data sovereignty in New Zealand (Te Mana Raraunga Maori Data Sovereignty Network), the US Indigenous Data Sovereignty Network, and the International Indigenous Data Sovereignty Interest Group. Specific bodies within or outside a university or health care organization can provide guidelines on appropriate processes to follow for data sharing for teaching.

Time, Cost, Feasibility, and Permissions

Finally, before recording can commence, the project's time, costs, and feasibility must be considered. Permission might be required before recording at sites such as laboratories, hospitals, and university campuses. The cost and feasibility include a list

XSL•FO
RenderX

of the necessary equipment for recording, the availability and the ability to use the equipment, and the cost of the entire project with a budget. The educator must consider the feasibility of recording time-dependent or rare procedures.

Making Our Video: Practical Considerations in the Preproduction Phase

If doing the recording yourself, this stage should also involve attending courses on videography and editing skills and audiovisual skill building by deliberate practice with feedback, using simulations and rehearsals. Our steps for making an online video are shown in Figure 1.

We identified the target audience for this educational video (postgraduate pathology trainees and histology scientists), conducted a needs assessment, and listed the learning objectives from an educator's perspective. We then converted these to learning outcomes from the learner's perspective. We used the recommendations by Fleming et al [16] to guide our video's script and storyboard, which also involved deciding the sequence of what, how, and where the recording would occur. The latter was structured in a table with 4-column labeled steps of the procedure, image or scene, audio (script), and photos (Multimedia Appendix 1).

A subject matter expert (AC), an anatomical pathologist with expertise in skeletal muscle pathology, assessed and validated the educational content of the storyboard and script. The technical content was evaluated by a media production contact from the Media Productions Department at the University of Auckland, who provided input on the technical aspects of the storyboard (Multimedia Appendix 2). We amended the storyboard and script to incorporate our experts' suggestions. The final script and storyboard used for recording are presented in Multimedia Appendix 3. We created a checklist of the technical audiovisual equipment required on the day, listed under the production phase. We also needed to co-ordinate the participants' schedules involved in the recording. Throughout this process, KS kept a journal to document the time each phase took. The repetitive cycle of feedback from the team and personal reflection led to the documentation of factors that had progressed well or not during each stage.

Figure 1. Steps in developing an online instructional video.

Production Phase

Background Factors in the Production Phase

The background factors to consider in the production phase include ensuring all the equipment on the list is available, the videographer is familiar with using them, the audiovisual technical aspects of the recording process, and the planned recording schedule.

Ensuring that all equipment on the list is available and that the videographer is familiar with using them. The list consists of (1) adequate lighting, (2) a digital camera with a microphone, (3) a digital camera mount such as a monopod or tripod, (4) fully charged batteries, (5) a secure digital (SD) card with speed and capacity to record video, and (6) a method to transfer the video files from the camera to the computer, such as a Wi-Fi enabled connection, an SD to USB card reader, or a camera to computer cable.

An option is to record the images and videos on hand-held devices, such as personal mobile phones. This can be a cost-effective and convenient option, especially since current devices have sufficient camera resolution to produce high-quality images and increased memory storage options. The user is also familiar with using this device. However, there are patient confidentiality and data protection issues to consider before choosing to use a personal mobile phone. The use of personal mobile devices and personal cloud storage can potentially lead to breaches of data protection and patient privacy [20,21]. The shared video content may be widely accessible, and sometimes not in the way the authors of the video intended, such as by family members on a shared cloud storage. Currently, there is no health information governance legislation that applies worldwide. However, institutions and countries have specific guidance or legislation that must be adhered to (such as HIPAA [Health Insurance Portability Accountability Act] in the United States or the GDPR [General Data Protection Regulation] from the European Union).

Also, the audiovisual technical aspects of the recording process and the planned recording schedule must be considered. Choosing the appropriate demonstrator(s) for the video and

XSL•FO
RenderX

rehearsing the recording process will streamline recording [12]. Tasks and responsibilities can be clarified before the day of recording. Recording more takes and angles than required may create a more time-effective and smoother video [12]. Ideally, the video should have different shots with judicious camera movements.

Making Our Video: Practical Considerations in the Production Phase

Setting

The video was recorded in a histopathology laboratory that routinely processes human skeletal muscle samples, as the video content was on macroscopically handling a fresh skeletal muscle biopsy. The videographer (KS) recorded a pathologist (AC) handling and preparing a fresh skeletal muscle biopsy in the laboratory. FM, FG-S, and AC provided feedback on the video. FM and FG-S are not experts in this content, and their feedback was from a naïve learner perspective informed by the literature. AC, a subject matter expert, also reviewed the initial video and provided feedback on content and editing.

Data Collection

The video was recorded immediately upon arrival of the specimen in the laboratory. KS recorded still images and short video clips using a digital camera mounted on a monopod. The process of recording the video clips and still photos took 4 hours. The equipment checklist, script, and storyboard created in the preproduction phase guided the recording. KS ensured all aspects of recording a particular step were completed before moving to the next step. Some steps needed rerecording as the initial images or shots were technically suboptimal. The suboptimal time-sensitive shots (as a muscle biopsy cannot be delayed in the fixation process upon receipt into the laboratory) were rerecorded immediately. In contrast, images of equipment or steps that were not time-sensitive were rerecorded at the end. FM and FG-S reviewed the initial video and provided feedback. AC provided feedback on the modified version of the video. KS iteratively modified the video in response to this feedback.

Postproduction Phase

Background Factors in the Postproduction Phase

The background factors to consider in the postproduction phase include video editing [9,15,16], video hosting platforms, video quality, and health information governance issues. Various video editing software (such as Microsoft Video Editor, Camtasia [TechSmith], Adobe Premier Pro, iMovie [Apple Inc], Wondershare Filmora [Wondershare Technology] and many others) range in price, complexity, and ease of use. Technical considerations include internet speeds and the video file type (.mp4 file type is preferred). The video editing should conform to the principles of multimedia learning [11] and add audio narration and optional subtitles (or close captioning) to improve accessibility and inclusivity. Educators can also add active learning to a video by embedding the video in software, such as H5P (H5P group), to create interactive videos with quizzes.

The video should be published on an online platform that provides easy access, for example, YouTube or embedded in the learner's learning management system. Many online platforms include videos for both patient and medical personnel use, and the aim and content of these videos will differ as they target different audiences [22]. YouTube is open access and popular with users; however, in a study, only 12% of medical videos were from university channels and professional organizations [22]. Alternatively, professional organizations like The Royal College of Pathologists of Australasia can host videos on their website such as the open-access macroscopic cut-up manual videos. Novice learners are more likely to access online videos on YouTube.

On the other hand, faculty members are more likely to access videos on professional organization web pages or YouTube channels specific to the organization [23]. However, YouTube hosts videos of variable quality, and novice users may not be able to identify poor-quality ones [23]. This clearly could have safety implications, depending on the procedure being learned.

Finally, the video author must ensure that health information governance issues described in the preproduction phase are all fulfilled before the video is published online.

Making Our Video: Practical Considerations in the Postproduction Phase: Analysis

The analysis phase of making a video involves postproduction editing, refinement, and publishing on a platform. The recorded images were transferred from the SD card into a computer with sufficient hard drive capacity. Various software, each chosen for a different purpose, were used to edit the sequence of scenes into an online video. KS used Wondershare Filmora 11 software to edit the video clips due to familiarity with the software. The final .mp4 file was transferred to Panopto, so that close captions in English could be enabled. The authoring software H5P was used to add a quiz to the video (drag and drop boxes) to provide formative feedback to the students and make the video interactive. H5P use also allowed bookmarking within the video to improve user navigation. The factors that made an effective online video from an educational and technical point of view, derived from the scoping review and broader literature, were incorporated into the video during the editing stage. These factors are elaborated in the Discussion section.

Ethical Considerations

We were granted ethical approval by the Auckland Health Research Ethics Committee (AHREC) on May 6, 2022 (ref AH23813). We obtained appropriate written patient consent before the recording. The patient data is anonymized, and no compensation was provided to the patient. The consent form and the online video will be stored according to the institutional ethics requirements.

Results

General Points About our Video

Our final video is 4 minutes and 4 seconds long. Its creation took 70 hours (20 hours of preproduction, 4 hours of recording, and 46 hours of editing and revising), which KS recorded manually in a journal. A link to the video is available at the following reference [24].

XSL•FO
RenderX

We identified that a clinician educator creating an online educational video needs to align the video with clear learning outcomes and provide an engaging, learner-centric video that promotes active learning. Other aspects involve incorporating the principles of multimedia learning and adult learning theories and producing a video of high audiovisual technical quality. This is shown in diagrammatic form in Figure 2.

We also created a checklist of elements that ensures a high-educational quality video is produced based on previous literature on this topic and the process of developing a new video. These include a needs assessment, a validated storyboard and script, clear learning outcomes, a title that reflects video content, a video duration of less than 15 minutes, the face of the educator, narration by the educator, close captioning, and factors to enable interactivity with the video. The specific characteristics of our video that align with these elements are summarized in Table 2.

Figure 2. Diagrammatic depiction of components of an effective online procedural video.

Table 2. Checklist of suggested quality elements in an online video.

Quality elements	Present in our video	Further details
Needs assessment	Not applicable	Needs assessment with learners will follow
Script and storyboard validated by an educational expert	Yes	—[a]
Script and storyboard validated by an audiovisual expert	Yes	—
Easy for a learner to access video	Yes	—
The title reflects the video content	Yes	—
Learning outcomes	Yes	—
Video duration <15 minutes	Yes	—
Time-stamping or bookmarks	Yes	—
Image of the instructor	Yes	—
Audio narration	Yes	—
Principals of multimedia followed:	Yes	—
Signaling	—	Highlighting with arrows
Coherence	—	Removal of extraneous material
Personalization	—	The instructor’s voice was used for narration.
Segmentation	—	Video is broken into small units or segments.
Functional interactivity	Yes	Interactive quiz inserted using H5P software
Close caption option	Yes	—

[a]Not applicable.

XSL•FO
RenderX

Barriers and Facilitators to Creating Our Online Video

We documented the barriers and facilitators to creating an instructional video from the perspective of a clinician educator who was a novice at making videos of any type. One of the barriers to developing this video was the significant time required for creation. According to the literature, it takes about an hour to make 1 minute of online video content [25]. Still, with 70 hours of production time for a 4-minute video, the novice video producer (KS) vastly exceeded this. This time estimate is for experienced video creators or professionals. Due to inexperience, all phases of recording a video took a notable amount of time. The preproduction phase involved reading educational resources about theories of adult learning in the clinical setting and principles of multimedia and reading about the technical aspects of audiovisual recording [4,11-13,16,17]. The 46 hours of postproduction time included accessing and learning how to use the editing and interactive software, as well as revising the video based on iterative feedback.

The recording took 4 hours. Performing the videography also proved difficult, especially without previous rehearsal. For example, selecting camera settings, the way to frame a shot, and the lighting and camera placement to get an optimal picture had been learned in theory but proved challenging in practice. This was mainly due to KS having unfamiliarity with the equipment and a lack of technical videography skills. Several retakes were needed. It was also challenging to record shots to avoid confidential patient details. As this was not a simulated procedure, shots could not be repeated, and because the tissue was fresh, the procedure was time-critical.

A facilitator of video creation is the use of deliberate practice with feedback using simulation and rehearsals. The planned rehearsal using a simulated specimen could not be done, therefore technical skills on the day were minimal [26]. Rehearsals would have also identified the extra equipment and the storyboard or shot board adjustments. Therefore, the recording session took longer than scheduled due to necessary repetitions. The videographer (KS) could have completed online video editing courses, visited some video recording sessions with the university's audiovisual department before recording, and practiced video recording and editing using nonmedical subjects, as these practical skills are more important than reading theory around this topic. These steps would have reduced the time spent recording and editing the video as more suitable images would have been captured on the day of recording.

Another facilitator was creating a community of practice (CoP) through discussions with colleagues with various skills and expertise. These discussions included specific technical equipment, audiovisual techniques such as camera angles, and the software to edit the raw recording. This dynamic group helped troubleshoot solutions to practical problems, which was especially important as KS chose software that was not industry standard for editing.

Other facilitators are organized and flexible, resulting in less stress during the recording process. Having a storyboard or script and an equipment list is essential. We found that it was not easy to stick to the storyboard, but the task would have been chaotic without it. Throughout this process, being cognitively and psychologically flexible to change makes the job easier and less frustrating.

Discussion

Principal Findings

While online educational videos are widely used in medical education, there is a lack of literature on creating a high-quality video. Our tutorial provides a step-by-step method for developing a quality medical video for clinician educators new to video creation.

Numerous studies have shown that the quality of videos on online platforms is heterogeneous [6-8,27]. Therefore, we propose a checklist to ensure that the educational factors in a video are optimized (Table 2). This checklist is a synthesis of the literature on this topic, so using it may ensure a high-quality video that is interactive and creates an engaging and learner-centric environment.

The time required to gain proficiency in making a procedural online video efficiently is a potential barrier. An educator new to making videos requires considerable time to produce new medical video content [25]. The professional body that makes online videos for the University of California San Francisco medical school estimates that 9-30 hours of an educator's time is needed when developing such content, in addition to professional videography personnel time [28]. This time range includes both the time required to learn to create video content and use the technology. Often, for the educator, this is a time in addition to usual work hours. However, this can be a rewarding process, and the time requirements may be lessened by preparing and validating a script and storyboard, repeated rehearsals, especially using simulations to improve skills, creating a CoP, and intentionally developing and transferring video production skills gained in nonmedical settings. The time requirements will also lessen over time as a clinician educator becomes more experienced in video production. As an alternative to the clinician educator learning video production skills, they can engage and work with others with audiovisual expertise to do the production, reducing the time requirements.

The amount of knowledge and skills required to create a quality online video means that a novice may find this task daunting. However, KS found this process was aided by developing a CoP—a supportive and educational system known to be valued by teaching staff [29]. A CoP allows people to connect and learn about a particular topic, such as creating online videos. Our CoP group has variable amounts of experience in developing online medical videos of various types, creating a technically and emotionally supportive environment. Such groups work on the background theory of constructivism as they allow for social learning and mentorship [30]. The iterative process of seeking and providing feedback also improved video quality.

While several quality assessment tools have been developed in the last 4 years [31-34], an easy-to-use tool is desirable to assess the quality of a procedural educational video. Berrocal et al [31] have a one-page rubric for peer-reviewing microlectures. The instructional video quality checklist is another quality assessment tool, that uses a 26-item checklist assessing aspects

XSL•FO
RenderX

of educational design, source reliability, multimedia principle adherence, and accessibility [32]. The LAP-VEGaS (LAParoscopic surgery Video Educational GuidelineS) video assessment tool examines 9 elements of a video [33], but it only applies to laparoscopic videos.

The authors of the LAP-VEGaS guidelines and LAP-VEGaS video assessment tool suggest using these during the preproduction and production phases to ensure a video has high educational value [33,34]. Several studies have subsequently used either the LAP-VEGaS guidelines or assessment tools to assess the quality of online medical videos. When de'Angelis et al [7] used the guidelines to review the quality of videos on appendectomy available on YouTube, 36% of the videos showed poor image quality, audio and written commentary were rarely present, and the overall conformity to the LAP-VEGaS guidelines was low. A similar result was found when videos on robotic-assisted laparoscopic pyeloplasty available on YouTube were assessed using the LAP-VEGaS guidelines [35]. This highlights the use of quality assessment tools or checklists for medical procedural videos.

Using online videos to teach procedural skills is no substitute for hands-on teaching. Karadas and O'Brien have demonstrated that repeatedly watching a video without physical practice can lead to an illusion of skill acquisition, where the learner assumes to have a greater skill level than they have [36]. Instead, online videos ought to be a part of the learning cycle for practical skills, perhaps based on an experiential learning cycle, so that the online video is followed by repeated practice exercises, such as simulation, adequate supervision, and a feedback cycle [37].

Strengths and Limitations

A strength of this tutorial is the use of theoretical considerations in the preproduction, production, and postproduction phases of making a high-quality video to demonstrate a procedural skill. Furthermore, we illustrate this with our own experience of making such a video from the perspective of a clinician-educator new to making videos, therefore the relevance and usefulness of this tutorial to an experienced video maker will be limited. Furthermore, our checklist of elements for a procedural video that aligns with educational and audiovisual quality factors needs further validation before implementing its use.

Conclusion

The process of creating procedural online videos is rewarding; however, it takes significant time and cognitive requirements, especially for clinician educators new to the process. Barriers include the time required for deliberate practice to gain competence in video production. Facilitators include technical skill building by deliberate practice with feedback, using simulations and rehearsals, and intentionally developing and transferring video production skills gained in nonmedical settings. For some, creating a CoP is supportive.

Acknowledgments

We thank Dr Rachelle Singleton, University of Auckland, for advice on using H5P and Panopto and Folko Boermans, Media Productions, University of Auckland. No funding was received.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Subject matter expert's input storyboard and script creating a video.
[DOCX File , 3019 KB - mededu_v10i1e51740_app1.docx]

Multimedia Appendix 2
Technical expert input storyboard and script creating a video.
[DOCX File , 3019 KB - mededu_v10i1e51740_app2.docx]

Multimedia Appendix 3
Final script and storyboard for an online medical video (modified after both experts input).
[DOCX File , 2801 KB - mededu_v10i1e51740_app3.docx]

References

1. Celentano V, Smart N, Cahill RA, McGrath JS, Gupta S, Griffith JP, et al. Use of laparoscopic videos amongst surgical trainees in the United Kingdom. Surgeon 2019;17(6):334-339. [doi: 10.1016/j.surge.2018.10.004] [Medline: 30420320]
2. Poon C, Stevens SM, Golub JS, Pensak ML, Samy RN. Pilot study evaluating the impact of otology surgery videos on otolaryngology resident education. Otol Neurotol 2017;38(3):423-428. [doi: 10.1097/MAO.0000000000001303] [Medline: 28192383]
3. Di Paolo T, Wakefield JS, Mills LA, Baker L. Lights, camera, action: facilitating the design and production of effective instructional videos. TechTrends 2017;61:452-460. [doi: 10.1007/s11528-017-0206-0]

XSL•FO
RenderX

4. Dong C, Goh PS. Twelve tips for the effective use of videos in medical education. Med Teach 2015;37(2):140-145. [doi: 10.3109/0142159X.2014.943709] [Medline: 25110154]
5. Murugiah K, Vallakati A, Rajput K, Sood A, Challa NR. YouTube as a source of information on cardiopulmonary resuscitation. Resuscitation 2011;82(3):332-334. [doi: 10.1016/j.resuscitation.2010.11.015] [Medline: 21185643]
6. Rössler B, Lahner D, Schebesta K, Chiari A, Plöchl W. Medical information on the internet: quality assessment of lumbar puncture and neuroaxial block techniques on YouTube. Clin Neurol Neurosurg 2012;114(6):655-658. [doi: 10.1016/j.clineuro.2011.12.048] [Medline: 22310998]
7. de'Angelis N, Gavriilidis P, Martínez-Pérez A, Genova P, Notarnicola M, Reitano E, et al. Educational value of surgical videos on YouTube: quality assessment of laparoscopic appendectomy videos by senior surgeons vs. novice trainees. World J Emerg Surg 2019;14:22 [FREE Full text] [doi: 10.1186/s13017-019-0241-6] [Medline: 31086560]
8. Rouhi AD, Roberson JL, Kindall E, Ghanem YK, Ndong A, Yi WS, et al. What are trainees watching? Assessing the educational quality of online laparoscopic cholecystectomy training videos using the LAP-VEGaS guidelines. Surgery 2023;174(3):524-528. [doi: 10.1016/j.surg.2023.05.021] [Medline: 37357097]
9. Hayden EL, Seagull FJ, Reddy RM. Developing an educational video on lung lobectomy for the general surgery resident. J Surg Res 2015;196(2):216-220. [doi: 10.1016/j.jss.2015.02.020] [Medline: 25828933]
10. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
11. Mayer RE. Using multimedia for e‐learning. JCAL 2017;33(5):403-423 [FREE Full text] [doi: 10.1111/jcal.12197]
12. Shi G, Lee S, Yuen Y, Liu J, Rothman Z, Milaire P, et al. 12 Tips for creating high impact clinical encounter videos - with technical pointers. MedEdPublish 2019;8:92 [FREE Full text] [doi: 10.15694/mep.2019.000092.1] [Medline: 38089391]
13. van der Meij H, Hopfner C. Eleven guidelines for the design of instructional videos for software training. Tech Commun 2022;69(3):5-23 [FREE Full text] [doi: 10.55177/tc786532]
14. Catoy K. The basics of video resolution.: Video4Change; 2020. URL: https://video4change.org/the-basics-of-video-resolution/ [accessed 2024-07-04]
15. Alves MG, Batista DFG, Cordeiro ALPDC, Silva MD, Canova JDCM, Dalri MCB. Production and validation of a video lesson on cardiopulmonary resuscitation. Rev Gaucha Enferm 2019;40:e20190012 [FREE Full text] [doi: 10.1590/1983-1447.2019.20190012] [Medline: 31389480]
16. Fleming SE, Reynolds J, Wallace B. Lights... camera... action! a guide for creating a DVD/video. Nurse Educ 2009;34(3):118-121. [doi: 10.1097/NNE.0b013e3181a0270e] [Medline: 19412052]
17. Brame CJ. Effective educational videos: principles and guidelines for maximizing student learning from video content. CBE Life Sci Educ 2016;15(4):es6 [FREE Full text] [doi: 10.1187/cbe.16-03-0125] [Medline: 27789532]
18. Hehir E, Zeller M, Luckhurst J, Chandler T. Developing student connectedness under remote learning using digital resources: a systematic review. Educ Inf Technol (Dordr) 2021;26(5):6531-6548 [FREE Full text] [doi: 10.1007/s10639-021-10577-1] [Medline: 34220282]
19. Srinivasa K, Moir F, Goodyear-Smith F. The role of online videos in teaching procedural skills in postgraduate medical education: a scoping review. J Surg Educ 2022;79(5):1295-1307. [doi: 10.1016/j.jsurg.2022.05.009] [Medline: 35725724]
20. Butler DJ. A review of published guidance for video recording in medical education. Fam Syst Health 2018;36(1):4-16. [doi: 10.1037/fsh0000328] [Medline: 29369649]
21. Srinivasa K, Chen Y, Henning MA. The role of online videos in teaching procedural skills to post-graduate medical learners: a systematic narrative review. Med Teach 2020;42(6):689-697. [doi: 10.1080/0142159X.2020.1733507] [Medline: 32174211]
22. Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis--a wakeup call? J Rheumatol 2012;39(5):899-903. [doi: 10.3899/jrheum.111114] [Medline: 22467934]
23. Rapp AK, Healy MG, Charlton ME, Keith JN, Rosenbaum ME, Kapadia MR. YouTube is the most frequently used educational video source for surgical preparation. J Surg Educ 2016;73(6):1072-1076 [FREE Full text] [doi: 10.1016/j.jsurg.2016.04.024] [Medline: 27316383]
24. Macroscopic processing of fresh skeletal muscle biopsy in a Pathology Laboratory with CC. URL: https://auckland.h5p.com/content/1291809298360807019 [accessed 2024-07-25]
25. Douglas SS, Aiken JM, Greco E, Schatz M, Lin SY. Do-it-yourself whiteboard-style physics video lectures. Phys Teach 2017;55(1):22-24 [FREE Full text] [doi: 10.1119/1.4972492]
26. Ericsson KA. The influence of experience and deliberate practice on the development of superior expert performance. In: The Cambridge Handbook of Expertise and Expert Performance. Cambridge, UK: Cambridge University Press; 2006:683-704.
27. Nason GJ, Kelly P, Kelly ME, Burke MJ, Aslam A, Giri SK, et al. YouTube as an educational tool regarding male urethral catheterization. Scand J Urol 2015;49(2):189-192. [doi: 10.3109/21681805.2014.975837] [Medline: 25363608]
28. Creating Video Based Education.: UCSF medical education school of medicine; 2022. URL: https://meded.ucsf.edu/tee/vided#Instructor-Considerations [accessed 2022-12-07]
29. Clune M, Charlton A, Kam M, Jowsey T, Ruiz CD, Singleton R. Strengthening online teaching capability: medical and health sciences faculty development. : ASCILITE Publications; 2022 Presented at: ASCILITE 2022 Conference Companion Materials; 2022 Nov 18; Australia p. e22029 URL: https://doi.org/10.14742/apubs.2022.29 [doi: 10.14742/apubs.2022.29]

XSL•FO
RenderX

30. de Carvalho-Filho MA, Tio RA, Steinert Y. Twelve tips for implementing a community of practice for faculty development. Med Teach 2020;42(2):143-149. [doi: 10.1080/0142159X.2018.1552782] [Medline: 30707855]
31. Berrocal Y, Regan J, Fisher J, Darr A, Hammersmith L, Aiyer M. Implementing rubric-based peer review for video microlecture design in health professions education. Med Sci Educ 2021;31(6):1761-1765 [FREE Full text] [doi: 10.1007/s40670-021-01437-1] [Medline: 34956695]
32. Schooley SP, Tackett S, Peraza LR, Shehadeh LA. Development and piloting of an instructional video quality checklist (IVQC). Med Teach 2022;44(3):287-293. [doi: 10.1080/0142159X.2021.1985099] [Medline: 34666585]
33. Celentano V, Smart N, Cahill RA, Spinelli A, Giglio MC, McGrath J, et al. Development and validation of a recommended checklist for assessment of surgical videos quality: the LA paroscopic surgery video educational guidelines (LAP-VEGaS) video assessment tool. Surg Endosc 2021;35(3):1362-1369 [FREE Full text] [doi: 10.1007/s00464-020-07517-4] [Medline: 32253556]
34. Celentano V, Smart N, McGrath J, Cahill RA, Spinelli A, Obermair A, et al. LAP-VEGaS practice guidelines for reporting of educational videos in laparoscopic surgery: a joint trainers and trainees consensus statement. Ann Surg 2018;268(6):920-926. [doi: 10.1097/SLA.0000000000002725] [Medline: 29509586]
35. Haslam RE, Seideman CA. Educational value of YouTube surgical videos of pediatric robot-assisted laparoscopic pyeloplasty: a qualitative assessment. J Endourol 2020;34(11):1129-1133. [doi: 10.1089/end.2020.0102] [Medline: 32709213]
36. Kardas M, O'Brien E. Easier seen than done: merely watching others perform can foster an illusion of skill acquisition. Psychol Sci 2018;29(4):521-536. [doi: 10.1177/0956797617740646] [Medline: 29451427]
37. Taylor DCM, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE Guide No. 83. Med Teach 2013;35(11):e1561-e1572. [doi: 10.3109/0142159X.2013.828153] [Medline: 24004029]

Abbreviations

AHREC: Auckland Health Research Ethics Committee
CoP: community of practice
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability Accountability Act
LAP-VEGaS: LAParoscopic surgery Video Educational GuidelineS
SD card: secure digital card

Edited by B Lesselroth; submitted 10.08.23; peer-reviewed by Y Berrocal, L Shehadeh, K Dumon; comments to author 21.02.24; revised version received 06.04.24; accepted 27.06.24; published 07.08.24.

Please cite as:
Srinivasa K, Charlton A, Moir F, Goodyear-Smith F
How to Develop an Online Video for Teaching Health Procedural Skills: Tutorial for Health Educators New to Video Production
JMIR Med Educ 2024;10:e51740
URL: https://mededu.jmir.org/2024/1/e51740
doi:10.2196/51740
PMID:39110488

©Komal Srinivasa, Amanda Charlton, Fiona Moir, Felicity Goodyear-Smith. Originally published in JMIR Medical Education (https://mededu.jmir.org), 07.08.2024. 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.

XSL•FO
RenderX

Evaluating the Effectiveness of an Online Course on Pediatric Malnutrition for Syrian Health Professionals: Qualitative Delphi Study

Amal Sahyouni[1,2,*], MSc, MD; Imad Zoukar[1,3,*], MSc, MD; Mayssoon Dashash[1,4,*], MSc, PhD, DDS

1

2

3

4

*all authors contributed equally

Corresponding Author:
Amal Sahyouni, MSc, MD

Abstract

Background: There is a shortage of competent health professionals in managing malnutrition. Online education may be a practical and flexible approach to address this gap.

Objective: This study aimed to identify essential competencies and assess the effectiveness of an online course on pediatric malnutrition in improving the knowledge of pediatricians and health professionals.

Methods: A focus group (n=5) and Delphi technique (n=21 health professionals) were used to identify 68 essential competencies. An online course consisting of 4 educational modules in Microsoft PowerPoint (Microsoft Corp) slide form with visual aids (photos and videos) was designed and published on the Syrian Virtual University platform website using an asynchronous e-learning system. The course covered definition, classification, epidemiology, anthropometrics, treatment, and consequences. Participants (n=10) completed a pretest of 40 multiple-choice questions, accessed the course, completed a posttest after a specified period, and filled out a questionnaire to measure their attitude and assess their satisfaction.

Results: A total of 68 essential competencies were identified, categorized into 3 domains: knowledge (24 competencies), skills (29 competencies), and attitudes (15 competencies). These competencies were further classified based on their focus area: etiology (10 competencies), assessment and diagnosis (21 competencies), and management (37 competencies). Further, 10 volunteers, consisting of 5 pediatricians and 5 health professionals, participated in this study over a 2-week period. A statistically significant increase in knowledge was observed among participants following completion of the online course (pretest mean 24.2, SD 6.1, and posttest mean 35.2, SD 3.3; $P<.001$). Pediatricians demonstrated higher pre- and posttest scores compared to other health care professionals (all P values were <.05). Prior malnutrition training within the past year positively impacted pretest scores ($P=.03$). Participants highly rated the course (mean satisfaction score >3.0 on a 5-point Likert scale), with 60% (6/10) favoring a blended learning approach.

Conclusions: In total, 68 essential competencies are required for pediatricians to manage children who are malnourished. The online course effectively improved knowledge acquisition among health care professionals, with high participant satisfaction and approval of the e-learning environment.

(JMIR Med Educ 2024;10:e53151) doi:10.2196/53151

KEYWORDS

effectiveness; online course; pediatric; malnutrition; essential competencies; e-learning; health professional; Syria; pilot study; acquisition knowledge

Introduction

Severe acute malnutrition (SAM) increases the risk of mortality among children aged younger than 5 years, affecting an estimated 17 million children worldwide, particularly in low- and middle-income countries [1,2]. The Syrian conflict has exacerbated this crisis, with half a million children enduring chronic malnutrition and 137,000 aged younger then 5 years experiencing acute malnutrition, increasing their susceptibility to preventable diseases [3]. Scaling up the management of SAM is crucial for reducing child mortality [4], but training and resourcing health care professionals to effectively identify and treat children who are malnourished remain major challenges, especially in conflict zones such as Syria [5,6].

XSL•FO
RenderX

While a wealth of knowledge exists about pediatric malnutrition, a crucial gap remains. This gap lies in the delivery of practical, accessible, and contextually relevant training for health care professionals in conflict zones [7]. This study addresses this gap by focusing on the development and evaluation of a self-directed online course for Syrian pediatricians and other health care professionals in the management of pediatric malnutrition.

e-Learning offers a promising solution for addressing this training gap, providing a flexible, scalable, and cost-effective method to deliver high-quality instruction [8]. By using e-learning, we can empower health care professionals with the necessary knowledge and skills to combat pediatric malnutrition; bridge access barriers, particularly in conflict zones; and tailor training to meet specific needs [7,9].

However, existing e-learning platforms may not adequately address the unique challenges and context of conflict zones such as Syria, where resources are limited [10]. Local training initiatives are crucial to ensure contextual relevance and maximize impact. By identifying specific knowledge gaps and skill deficits, developing culturally sensitive materials, and integrating with local resources, tailored e-learning solutions can foster a sense of ownership and engagement, leading to more effective knowledge transfer and application [11].

This study investigates the efficacy of e-learning as a training solution for pediatric malnutrition management in Syria, a context marked by conflict and limited access to traditional training opportunities. Specifically, this study aims to identify essential competencies needed by pediatricians and health care professionals for effective management of pediatric malnutrition, evaluate the efficacy of an online course on pediatric malnutrition in improving their knowledge and skills, and explore the potential of e-learning as a scalable solution for training in challenging contexts. By addressing this crucial training gap, the research seeks to contribute to improved patient care; enhanced staff retention; and, ultimately, a reduction in pediatric malnutrition in Syria.

Methods

Ethical Considerations

This study was approved by the Ethical Committee at the Syrian Virtual University (SVU; #2154/0, November 25, 2021). An electronic consent form was obtained from all participants, ensuring their understanding that their data would be kept confidential and used solely for the purposes of this research study.

Competency Development

Overview

The Delphi technique was used to develop a consensus regarding essential competencies for managing children who are malnourished, as it is a helpful strategy for identifying medical education competencies [12]. Further, 2 qualitative methods were applied sequentially—focus groups and the Delphi technique.

Focus Groups

The online focus group consisted of 5 participants: 3 postgraduate students working in an inpatient stabilization center in Lattakia and 2 pediatricians serving as pediatric malnutrition therapists on projects funded by UNICEF (United Nations Children's Fund). The participants were recruited via telephone and social media to participate in the virtual discussion.

The research approach was fully explained to the focus group participants, who were asked to provide a brief report on the competencies needed by pediatricians to manage pediatric malnutrition. Consequently, a preliminary list of essential competencies was formulated during a 3-hour meeting.

Delphi Rounds

All pediatricians and health professionals enrolled in the Medical Education Master Program at the SVU were invited to participate in this study (n=21). Ultimately, 18 of them took part. Participants were instructed, via the Virtual University Management System, to review relevant protocols and guidelines on pediatric malnutrition published by recognized health organizations, such as the World Health Organization (WHO) and UNICEF [13,14]. A methodology for writing competencies and vocabulary for job descriptions [15] was established to prepare the initial competency list.

Participants were randomly assigned to 1 of 3 groups, each focused on knowledge, skills, or attitudes, for the identification of essential competencies.

Through 3 virtual meetings conducted over a month, all groups agreed on a classification of competencies based on etiology, assessment, diagnosis, and managing pediatric malnutrition. The team leader reviewed all responses, removed repetitive items, reformulated inappropriate ones, and discussed transferring fields between specialties. Participants then independently reviewed the competencies over a 1-week period, and the revised list was distributed to all participants for rating on a 5-point Likert scale within a week.

Competencies endorsed by at least 80% (15/18, 83%) of participants were subsequently combined and merged. The weighted response for each competency was obtained by calculating the responses at each level and the mean score for each competency, ranging from 0.0 to 3.0 [16]. All competencies were ranked, and the relative importance of each competency was identified.

Training Development

Course Design

Drawing upon a variety of sources and expertise, an online pediatric malnutrition course was developed, informed by the competencies identified in the first phase of this study. The course was designed to meet the specific needs and context of Syrian health professionals, taking into account the unique challenges and priorities of the Syrian health care system.

The development of the online course and its accompanying multiple-choice questions involved a collaborative effort among pediatricians possessing years of experience in managing both

XSL•FO
RenderX

inpatients and outpatients with acute malnutrition, as well as in training health professionals.

Available evidence-based guidelines and protocols from reputable organizations, such as the WHO and UNICEF, were carefully reviewed to ensure the content validity and accuracy of the course materials. Visual aids, such as images and videos, were also incorporated to enhance engagement and learning.

Course Content

The e-learning content was categorized into 4 modules covering all identified competencies. Module 1 provided information on definition, classification, prevalence, differentiation between types, and associated pathophysiological changes in pediatric malnutrition. Module 2 focused on the diagnosis and anthropometric measurement techniques for accurate assessment. Module 3 presented treatment modalities according to WHO guidelines and the 10-steps approach. Module 4 addressed the prognosis and medical complications associated with pediatric malnutrition. Each module contained interlinked subunits for easy offline access, designed to be completed in less than 3 hours.

The principal investigator, AS, developed online content, incorporating visual aids such as images and videos, and created 40 multiple-choice pre- and posttests. The course was delivered in Arabic.

Participants

Recruitment

This study's participants comprised 10 individuals: 5 pediatricians and 5 other health care professionals, including medical doctors (general practitioners and specialists), dentists, and pharmacists. These participants, 8 females and 2 males, were recruited via social media to evaluate the online course.

Recruitment for this study was announced via the researcher's email and on social networking sites within official groups for doctors and resident students. Participation was optional, and the purpose of the research was clearly communicated as solely for scientific purposes. Participant grades would not be used outside the research and would remain confidential. Participants provided explicit consent via email, with a "yes" or "no" response.

Training Delivery

The SVU's learning management system was used for the online course presentation, with access restricted to participants via individual usernames and passwords. Data security was maintained through the platform's security protocols.

Prior to accessing the course, participants were surveyed regarding their prior attendance in malnutrition-related training courses. A designated date for the pretest was communicated via email to each participant. Upon completion of the pretest, participants were granted access to the online course, with a 1-month time frame allotted for completion. Following course completion, a second email was sent to each participant, scheduling a specific date for the online, synchronous posttest.

Evaluation

Assessment Instruments

Both the pretest and posttest, administered electronically via a Google Forms link within the SVU platform, consisted of 40 identical multiple-choice questions, ensuring a consistent measure of knowledge acquisition. The assessment instrument was distributed across the 4 modules, with a weighting reflecting the number of subunits within each module: modules 1 and 2, covering foundational knowledge of pediatric malnutrition, each contained 10 questions; module 3, focusing on assessment and diagnosis, included 5 questions; and module 4, on management, contained 15 questions.

Participant Satisfaction

Participant satisfaction was assessed using a 30-item questionnaire encompassing 3 domains: content presentation style, knowledge gained, and the e-learning environment (10 items per domain). Participants rated each item using a 5-point Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree). Cronbach α, a measure of internal consistency [17], was calculated. The mean and SD were then used to analyze the distribution of responses for each item, providing insights into participant satisfaction levels across the different domains.

The qualitative interpretation of Likert scale measurements is shown in Table S1 in Multimedia Appendix 1.

Data Analysis

To evaluate knowledge gains, paired *t* tests (2-tailed) were used to compare pre- and posttest scores. Subgroup analyses were conducted to examine potential differences in knowledge acquisition based on participant characteristics, including gender, specialty (pediatricians or other health professionals), and prior training experience. Comparisons between groups (eg, male vs female) were facilitated using the Mann-Whitney test, while comparisons within groups (eg, pretest vs posttest scores for pediatricians) were performed using the Wilcoxon signed rank test [18,19].

These nonparametric tests were chosen due to the relatively small sample size (n=10) [20], followed by parametric tests, and the results were compared. Statistical significance was set at $P<.05$.

IBM SPSS Statistics for Windows (version 25.0; IBM Corp) was used to perform all the analyses.

Results

The Delphi technique yielded an 85% (18/21) response rate. At least 80% (15/18, 83%) of participants suggested 68 essential competencies for managing children who are malnourished, organized into knowledge, skills, and attitude. As outlined in Table 1, competencies also fall under 3 subheadings: etiology, assessment or diagnosis, and management.

XSL•FO
RenderX

Table . Essential competencies for managing children who are malnourished.

Domain		Competency
Knowledge		
	Etiology	1. Recognize malnutrition terminology. 2. Recognize the epidemiology of malnutrition in children. 3. Differentiate between types of malnutrition. 4. Identify the causes and prevalence of pediatric malnutrition worldwide.
	Assessment and diagnosis	1. Identify the clinical signs and symptoms of acute malnutrition in children. 2. Recognize different methods for assessing children who are malnourished. 3. Describe admission and discharge criteria for managing SAM[a] with medical complications under inpatient care. 4. Identify high-risk groups for SAM in children. 5. Discuss strategies to detect cases of pediatric malnutrition. 6. Recognize complications and prognosis of pediatric malnutrition. 7. Identify target age groups to screen for malnutrition. 8. Explain methods for diagnosing malnutrition in children.
	Management	1. Describe outpatient care for managing SAM without medical complications. 2. Describe admission criteria for outpatient care (infants under 6 mo and children 6 - 59 mo) 3. Describe the outpatient care and follow-up process for children 6 - 59 mo. 4. Explain medical treatment for SAM without complications under outpatient care. 5. Explain nutrition rehabilitation for SAM without complications under outpatient care for children 6 - 59 mo. 6. Describe the key messages for mothers or caregivers of children 6 - 59 mo in outpatient care. 7. Explain managing at-risk mothers and infants aged younger than 6 mo without complications in outpatient care. 8. Explain discharge criteria and procedures for at-risk mothers, infants under 6 mo, and children 6 - 59 mo. 9. Outline management of SAM with medical complications under inpatient care. 10. Review medical and dietary treatment in inpatient care. 11. Describe programs to manage MAM[b]. 12. Describe the admission and discharge process for MAM management.
Skills		

XSL•FO
RenderX

Domain		Competency
	Etiology	1. Classify nutritional vulnerability in at-risk mothers and infants aged younger than 6 mo. 2. Educate parents of children who are malnourished to understand the risks of pediatric malnutrition. 3. Develop a plan to monitor cases of pediatric malnutrition. 4. Monitor and respond to barriers to access.
	Assessment and diagnosis	1. Take accurate clinical history. 2. Measure the weight, length, and mid-upper arm circumference of children. 3. Calculate the child's age. 4. Classify acute malnutrition. 5. Perform appropriate medical examination. 6. Provide correct diagnosis in each pediatric malnutrition case. 7. Assess and admit a child to outpatient care. 8. Assess and manage at-risk mothers and infants under 6 mo without medical complications in outpatient setting. 9. Conduct field visits for children who are malnourished in supplementary feeding programs.
	Management	1. Provide outpatient care for SAM without medical complications. 2. Identify when further action is required, such as referral to inpatient care and follow-up home visits. 3. Treat a child during outpatient care follow-on. 4. Practice referral between inpatient care and outpatient care. 5. Make referrals from supplementary feeding to outpatient or inpatient care. 6. Complete patient records and interpret findings. 7. Calculate and review service or program performance. 8. Calculate therapeutic doses accurately. 9. Correctly apply treatment in terms of timing and adjustments for each case. 10. Accurate diagnosis of pediatric malnutrition. 11. Manage clinical cases based on stage, development, and complications. 12. Discuss medical and nutritional treatment for MAM management. 13. Discuss treatment protocols according to malnutrition severity in a child. 14. Apply therapeutic protocols. 15. Administer therapeutic foods according to the malnutrition severity. 16. List indications and contraindications of medications and procedures.
Attitude		

XSL•FO
RenderX

Domain		Competency
	Etiology	1. Explain all malnutrition information to parents. 2. Promote health education about malnutrition and when to take action.
	Assessment and diagnosis	1. Keep children who are malnourished safe and protected from harm. 2. Demonstrate investigative and analytical thinking to meet the needs of children who are malnourished. 3. Provide the best possible health care to children who are malnourished regardless of age, gender, culture, and economic status. 4. Communicate effectively with children who are malnourished and families to explain case progression.
	Management	1. Demonstrate professionalism with peers, staff, patients, and families. 2. Collaborate with health care professionals. 3. Respect patient privacy and confidentiality. 4. Show sympathy and compassion for children who are malnourished. 5. Provide spiritual support to children who are malnourished and parents. 6. Develop strategies for consultation, collaboration, and referral. 7. Exhibit leadership, initiative, and optimism to manage cases effectively. 8. Work flexibly under stress and changing conditions while remaining calm. 9. Apply WHO[c] general principles for routine care (10-steps).

[a]SAM: severe acute malnutrition.

[b]MAM: moderate acute malnutrition.

[c]WHO: World Health Organization

The final competencies comprised 24 knowledge, 29 skills, and 15 attitudes items. By competencies classification, 10 addressed etiology, 21 assessment or diagnosis, and 37 management.

In total, 10 participants were recruited between July 1, 2021, and July 15, 2021. Following the administration of the pretest, the online course commenced on July 21 and continued for 1 month. The posttest was administered on August 28, 2021, upon completion of the course. The cohort comprised 5 (50%) pediatricians and 5 (50%) other health professionals, with 8 (80%) females and 2 (20%) males. All pediatricians had prior experience in managing children who are malnourished and had received training on pediatric malnutrition, with 3 having been trained over a year prior and 2 within the past year. None of the other health professionals in the cohort had previous training.

Table 2 displays knowledge gained across participants, assessed before and after the course. Comparisons were made between groups based on gender, specialty, and prior training. All participants achieved higher posttest than pretest scores. The overall mean increase in knowledge scores from pretest to posttest was 11.0 points, representing a 45% relative gain. This significant difference (P<.001) indicates that the online course was effective in enhancing knowledge about pediatric malnutrition for the entire cohort.

XSL•FO
RenderX

Table . Knowledge gains between pre- and posttests based on participant demographics.

Variable	n	Pretest mean[a](SD)	Posttest mean (SD)[a]	Mann-Whitney (*P* value)		Paired *t* test (*df*; *P* value)	
				Pretest	Posttest	Pretest	Posttest
Female	8	23.6 (6.7)	35.4 (3.7)	−0.26 (.79)	−0.53 (.60)	−0.58 (8;.58)	0.32 (8;.76)
Male	2	26.5 (2.1)	34.5 (0.7)	−0.26 (.79)	−0.53 (.60)	−0.58 (8;.58)	0.32 (8;.76)
Health professionals	5	20.2 (5.9)	33 (2.5)	−2 (.046)	−2.11 (.04)	−2.73 (8;.03)	−2.81 (8;.02)
Pediatricians	5	28.2 (2.9)	37.4 (2.4)	−2 (.046)	−2.11 (.04)	−2.73 (8;.03)	−2.81 (8;.02)
Nonprior attendance	5	20.2 (5.9)	—[b]	−2 (.046)	—	−2.73 (8;.03)	—
Prior attendance	5	28.2 (2.9)	—	−2 (.046)	—	−2.73 (8;.03)	—
More than 1 year	3	26.3 (1.5)	—	—	—	−3.43 (8;.04)	—
Last year	2	31 (1.4)	—	—	—	−3.43 (8;.04)	—
Total	10	24.2 (6.1)	35.2 (3.3)	−2.81 (.01)	−2.81 (.01)	−7.79 (9;<.001)	−7.79 (9;<.001)

[a]Maximum score=40.

[b]Not applicable.

Notably, the paired *t* test revealed a statistically significant improvement in posttest scores for all participants (5/5) with prior malnutrition training (*P*=.03). This finding suggests that the course not only builds upon existing knowledge but also serves as a valuable refresher for those previously trained in this area (Table 2).

Female participants had average pretest scores of 23.6 (SD 6.7) and posttest scores of 35.3 (SD 3.7), while males averaged 26.5 (SD 2.1) and 34.5 (SD 0.7), respectively. Gender did not significantly impact knowledge gains, with *P*=.79 for the pretest and *P*=.60 for posttest scores. Pediatricians had higher mean pretest scores 28.2 (SD 2.9) and posttest scores 37.4 (SD 2.4) than other health professionals 20.2 (SD 5.9) and 33 (SD 2.5), respectively. Specialty significantly affected pre- and posttest scores (*P*=.03 and *P*=.02, respectively). Health professionals with prior training attendance demonstrated a significant difference in pretest scores compared to those without prior attendance (*P*=.046).

Even though gender had little significant impact on test scores, knowledge improved across all educational modules.

When test scores were compared by specialty, pediatricians outperformed health care professionals in module 1 (defining, classifying, and determining malnutrition prevalence) and module 4 (managing pediatric malnutrition; *P*=.02 and *P*=.02, respectively). Details are presented in Table 3.

XSL•FO
RenderX

Table . Comparison of educational modules results by specialty.

Module and specialty		n	Pretest mean (SD)	Posttest mean (SD)	Paired *t* test (*df*; *P* value)		Mann-Whitney (*P* value)	
					Pretest	Posttest	Pretest	Posttest
One					−2.15 (8;.06)	−2.89 (8;.02)	−1.89 (.06)	−2.15 (.03)
	Health professionals	5	4.8 (1.6)	8.4 (0.5)				
	Pediatricians	5	6.6 (0.9)	9.4 (0.5)				
Two					−2.12 (8;.07)	−1.37 (8;.21)	−1.61 (.11)	−1.21 (.23)
	Health professionals	5	5.2 (2.8)	8.2 (0.8)				
	Pediatricians	5	8 (1)	9 (1)				
Three					−0.45 (8;.67)	−1.27 (8;.24)	−0.35 (.73)	−1.23 (.22)
	Health professionals	5	2.2 (0.8)	4.2 (0.4)				
	Pediatricians	5	2.4 (0.5)	4.6 (0.5)				
Four					−2.76 (8;.03)	−2.84 (8;.02)	−2.02 (.04)	−2.24 (.03)
	Health professionals	5	8 (1.9)	12.2 (1.5)				
	Pediatricians	5	11.2 (1.8)	14.4 (0.9)				

The effect size, calculated using Cohen *d*, was 2.25, indicating a very large effect, suggesting a substantial improvement in knowledge among participants from the pretest to the posttest. This finding highlights the significant impact of the online course on participants' understanding of pediatric malnutrition.

The analysis of 30-item participant satisfaction questionnaire revealed a Cronbach α of .74, indicating a satisfactory level of internal consistency and reliability for the instrument.

Participants reported high satisfaction with the content presentation style (mean 3.16, SD 0.24), very high satisfaction with the scientific knowledge gained (mean 3.26, SD 0.30), and high satisfaction with the e-learning environment (mean 3.06, SD 0.4). The details can be accessed on Table S2 in Multimedia Appendix 1.

Most participants (6/10) preferred a combination of traditional and e-learning, while 3 preferred e-learning only, and 1 preferred traditional education.

Discussion

Principal Findings

This study investigates the efficacy of online learning in addressing pediatric malnutrition, a significant global health concern. The research identified 68 essential competencies for effective pediatric malnutrition management, encompassing knowledge, skills, and attitudes. These competencies were developed through a collaborative process with Syrian health care professionals, offering a valuable resource for future training initiatives. This study demonstrated that a self-directed online course significantly enhanced participants' knowledge acquisition, highlighting the potential of e-learning as a scalable solution for addressing training needs in resource-constrained environments.

This study's findings underscore the importance of comprehensive competency frameworks for addressing pediatric malnutrition, especially in challenging contexts such as Syria. The participants highly rated the online course, suggesting its effectiveness in bridging training gaps in resource-constrained settings. This study provides valuable insights for developing and implementing effective training initiatives to improve the management of pediatric malnutrition in resource-limited and conflict-affected settings.

This study expands upon existing research on essential competencies in medical education and the feasibility of e-learning in Syria [21-23]. While international organizations have focused on pediatric malnutrition in low-income countries [1-3], research specifically addressing the competencies required by pediatricians and other health care professionals to manage these cases has been limited. This study aimed to identify essential competencies for managing pediatric malnutrition and evaluating the effectiveness of e-learning modules in enhancing knowledge among health care professionals.

This study identified essential competencies by drawing upon existing training guides and WHO guidelines on SAM [24,25]. Although this study was conducted before the release of updated WHO guidelines in 2023 [26], future revisions of the identified competencies and online course content are recommended to align with these new recommendations.

This study builds upon previous work by Meeker et al [27] on a technical competency framework for nutrition in emergencies, specifically focusing on pediatric malnutrition and the needs of Syrian health professionals. While prior studies [11,27,28] have highlighted the importance of managing pediatric malnutrition,

XSL•FO
RenderX

this research fills a gap by identifying the essential competencies pediatricians require to effectively manage these cases.

This study demonstrates the potential of e-learning to effectively scale up malnutrition management knowledge, a recognized challenge in most low- and middle-income countries [29,30]. Consistent with Annan et al's [8] findings, knowledge acquisition was enhanced when the course was linked to career and academic progress. Participants reported increased learning and understanding of effective pediatric malnutrition management. These findings align with Choi et al's [7] demonstration across 4 countries that e-learning enhanced health care practices and reduced malnutrition mortality through increased facility-based management of SAM. While this study's time frame precluded measuring clinical outcomes, the significant knowledge gains could be attributed to the course's practical relevance and scientifically sound content, as reflected in high participant approval [31].

Pediatricians' pretest performance was positively influenced by previous malnutrition training, indicating knowledge retention [32]. This highlights the need for posttests and follow-up evaluations to compare the effectiveness of e-learning with traditional training, especially given its lower implementation costs per participant [8].

Developing the educational course in Arabic was a key challenge due to the limited availability of Arabic language online resources [33]. However, offering the course in the local language facilitated better knowledge comprehension and potentially increased long-term retention and practical application [34].

Strength

This study highlights the potential of online learning to improve pediatric malnutrition management, especially in challenging environments such as Syria. By identifying essential competencies and demonstrating the effectiveness of a self-directed online course, it offers a model for developing similar training programs in other resource-constrained and conflict-affected areas. The course's accessibility, adaptability, and positive feedback from participants suggest a promising way to address the shortage of skilled health care professionals in these regions.

Limitation

This pilot study, conducted with a small sample size, limits the generalizability of the findings to a broader population of health care professionals. While this study demonstrates positive knowledge acquisition, it does not assess the impact on clinical practice or patient outcomes. Additionally, this study does not delve deeply into the specific challenges and facilitators of implementing the online course in the Syrian context, nor does it include long-term follow-up to assess the persistence of knowledge gains and their impact on clinical practice. To mitigate these limitations, we strived to recruit a diverse sample within our limited scope and used rigorous data collection and analysis methods.

Future Research

Future research should expand on this study by (1) conducting larger studies with diverse participants to ensure findings are broadly applicable; (2) assessing the online course's impact on real-world clinical practice and patient outcomes, not just knowledge acquisition; (3) understanding practical considerations and implementation strategies for e-learning in challenging environments such as Syria; (4) translating identified skills and attitudes into practical training methods; and (5) conducting long-term follow-up to track knowledge retention and its impact on clinical practice.

Conclusion

This study demonstrates the effectiveness of self-directed online learning in improving knowledge and skills related to pediatric malnutrition management among Syrian health care professionals. This study identified 68 essential competencies across various domains, highlighting the breadth of knowledge needed for effective pediatric malnutrition management. The findings suggest e-learning as a powerful tool for scaling up training in challenging contexts such as Syria, while acknowledging the need for careful consideration of contextual factors.

Acknowledgments

The authors would like to thank all the participants in this study for their time and willingness to share their experiences. All authors declared that they had insufficient or no funding to support open access publication of this paper, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The file contains three explanatory tables about qualitative interpretation of 5-point Likert scale measurements, questionnaire assessment results, and CHERRIES. CHERRIES: Checklist for Reporting Results of Internet E-Surveys.

[PDF File, 154 KB - mededu_v10i1e53151_app1.pdf]

References

XSL•FO
RenderX

1. Anato A. Severe acute malnutrition and associated factors among children under-five years: a community based-cross sectional study in Ethiopia. Heliyon 2022 Oct;8(10):e10791. [doi: 10.1016/j.heliyon.2022.e10791] [Medline: 36203897]
2. UNICEF, WHO, World Bank Group. Levels and trends in child malnutrition. UNICEF. 2019. URL: https://www.unicef.org/media/60626/file/Joint-malnutrition-estimates-2019.pdf [accessed 2023-09-03]
3. UNICEF. UNICEF whole of Syria humanitarian situation report. ReliefWeb. 2021. URL: https://reliefweb.int/report/syrian-arab-republic/unicef-whole-syria-humanitarian-situation-report-august-2021 [accessed 2023-09-03]
4. Osendarp S, Akuoku JK, Black RE, et al. The COVID-19 crisis will exacerbate maternal and child undernutrition and child mortality in low- and middle-income countries. Nat Food 2021 Jul;2(7):476-484. [doi: 10.1038/s43016-021-00319-4] [Medline: 37117686]
5. Gillespie S, Haddad L, Mannar V, Menon P, Nisbett N. The politics of reducing malnutrition: building commitment and accelerating progress. Lancet 2013 Aug;382(9891):552-569. [doi: 10.1016/S0140-6736(13)60842-9]
6. Jackson A, Ashworth A, Annan RA. The International Malnutrition Task Force: a model for the future? Trends Food Sci Technol 2022 Dec;130:11-19. [doi: 10.1016/j.tifs.2022.09.002]
7. Choi S, Yuen HM, Annan R, et al. Effectiveness of the malnutrition elearning course for global capacity building in the management of malnutrition: cross-country interrupted time-series study. J Med Internet Res 2018 Oct 3;20(10):e10396. [doi: 10.2196/10396] [Medline: 30282620]
8. Annan RA, Aduku LNE, Kyei-Boateng S, et al. Implementing effective eLearning for scaling up global capacity building: findings from the malnutrition elearning course evaluation in Ghana. Glob Health Action 2020 Dec 31;13(1):1831794. [doi: 10.1080/16549716.2020.1831794] [Medline: 33086945]
9. Choi S, Annon R. eLearning: a means to widen the opportunity for malnutrition education. In: Barton S, Hedberg J, Suzuki K, editors. Proceedings of Global Learn Asia Pacific 2011--Global Conference on Learning and Technology: Association for the Advancement of Computing in Education (AACE); 2011:1751 URL: https://www.learntechlib.org/p/37396 [accessed 2023-09-03]
10. Al-Shorbaji N, Atun R, Car J, Majeed A, Wheeler E. eLearning for undergraduate health professional education: a systematic review informing a radical transformation of health workforce development. World Health Organization. 2015. URL: https://apps.who.int/iris/handle/10665/330089 [accessed 2023-09-03]
11. Annan RA, Choi S, Turyashemererwa F, Pickup T, Jackson AA. Building core competencies for the prevention and treatment of severe malnutrition in infants and children: the role of elearning. Presented at: 2011 FANUS Meeting; Sep 11-15, 2011; Abuja, Nigeria p. 12-16.
12. de Villiers MR, de Villiers PJT, Kent AP. The Delphi technique in health sciences education research. Med Teach 2005 Nov;27(7):639-643. [doi: 10.1080/13611260500069947] [Medline: 16332558]
13. World Health Organization. Guideline: Updates on the Management of Severe Acute Malnutrition in Infants and Children: World Health Organization; 2013. URL: https://www.who.int/publications/i/item/9789241506328 [accessed 2024-10-19]
14. World Health Organization. Management of the Child with a Serious Infection or Severe Malnutrition: Guidelines for Care at the First-Referral Level in Developing Countries: World Health Organization Publications; 2000.
15. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010 Aug;32(8):638-645. [doi: 10.3109/0142159X.2010.501190]
16. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000 Oct;32(4):1008-1015. [doi: 10.1046/j.1365-2648.2000.t01-1-01567.x] [Medline: 11095242]
17. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334. [doi: 10.1007/BF02310555]
18. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 1947 Mar;18(1):50-60 [FREE Full text]
19. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull 1945 Dec;1(6):80 [FREE Full text]
20. Frey J, Ozturk O, Deshpande JV. Nonparametric tests for perfect judgment rankings. J Am Stat Assoc 2007 Jun;102(478):708-717. [doi: 10.1198/016214506000001248]
21. Khiami A, Dashash M. Identification of the role of oral health educators in elementary schools during COVID-19 pandemic: a competency framework. BMC Res Notes 2022 Dec 10;15(1):6. [doi: 10.1186/s13104-021-05887-z] [Medline: 35012621]
22. Kenjrawi Y, Dashash M. The first asynchronous online evidence-based medicine course for Syrian health workforce: effectiveness and feasibility pilot study. JMIR Form Res 2022 Oct 25;6(10):e36782. [doi: 10.2196/36782] [Medline: 36282556]
23. Zoukar I, Dashash M. Using a modified Delphi method for identifying competencies in a Syrian undergraduate neonatology curriculum. Matern Child Health J 2023 Jun;27(11):1921-1929. [doi: 10.1007/s10995-023-03719-z] [Medline: 37289293]
24. Management of severe malnutrition: a manual for physicians and other senior health workers. World Health Organization. 1999. URL: https://www.who.int/publications/i/item/9241545119 [accessed 2023-09-03]
25. Tickell KD, Denno DM. Inpatient management of children with severe acute malnutrition: a review of WHO guidelines. Bull World Health Organ 2016 Sep 1;94(9):642-651. [doi: 10.2471/BLT.15.162867] [Medline: 27708469]

XSL•FO
RenderX

26. World Health Organization. WHO guideline on the prevention and management of wasting and nutritional oedema (acute malnutrition) in infants and children under 5 years. MAGICapp. 2023. URL: https://app.magicapp.org/#/guideline/noPQkE [accessed 2024-07-01]
27. Meeker J, Perry A, Dolan C, et al. Development of a competency framework for the nutrition in emergencies sector. Public Health Nutr 2014 Mar;17(3):689-699. [doi: 10.1017/S1368980013002607] [Medline: 24103388]
28. Schofield C, Ashworth A, Annan R, Jackson AA. Malnutrition treatment to become a core competency. Arch Dis Child 2012 May;97(5):468-469. [doi: 10.1136/adc.2010.209015] [Medline: 21427122]
29. Daniel T, Mekkawi T, Garelnabi H, Sorkti S, Mutunga M. Scaling up CMAM in protracted emergencies and low resource settings: experiences from Sudan. Field Exch 2016 Aug 15(55):74.
30. Frehywot S, Vovides Y, Talib Z, et al. e-Learning in medical education in resource constrained low- and middle-income countries. Hum Resour Health 2013 Dec 4;11(1):4. [doi: 10.1186/1478-4491-11-4] [Medline: 23379467]
31. Gerdeman R, Russell A, Worden K. Web-based student writing and reviewing in a large biology lecture course. J Coll Sci Teach 2007 Mar;36(5):46-52.
32. Gowda RS, Suma V. A comparative analysis of traditional education system vs. e-learning. Presented at: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA); Feb 21-23, 2017; Bengaluru, India p. 567-571. [doi: 10.1109/ICIMIA.2017.7975524]
33. Alkoudmani R, Elkalmi R. Challenges to web-based learning in pharmacy education in Arabic language speaking countries. Arch Pharma Pract 2015;6(3):41. [doi: 10.4103/2045-080X.160989]
34. Ali S, Amaad Uppal M, Basir M, Zahid Z. An empirical investigation of digital learning via mobile phones in higher education institutes. Webol 2021;18(2) [FREE Full text]

Abbreviations

SAM: severe acute malnutrition
SVU: Syrian Virtual University
UNICEF: United Nations Children's Fund
WHO: World Health Organization

Edited by A Bahattab; submitted 27.09.23; peer-reviewed by AD Pucchio, M Kerac, M Nojomi, MS Shafi; revised version received 11.08.24; accepted 01.09.24; published 28.10.24.

Please cite as:
Sahyouni A, Zoukar I, Dashash M
Evaluating the Effectiveness of an Online Course on Pediatric Malnutrition for Syrian Health Professionals: Qualitative Delphi Study
JMIR Med Educ 2024;10:e53151
URL: https://mededu.jmir.org/2024/1/e53151
doi:10.2196/53151

© Amal Sahyouni, Imad Zoukar, Mayssoon Dashash. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.10.2024. 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.

XSL•FO
RenderX

Original Paper

Unpacking the Experiences of Health Care Professionals About the Web-Based Building Resilience At Work Program During the COVID-19 Pandemic: Framework Analysis

Wei How Darryl Ang[1], RN, BSN (Hons), PhD; Zhi Qi Grace Lim[1], BA; Siew Tiang Lau[1], RN, BHS, MHS, PhD; Jie Dong[1], RN; Ying Lau[2], RN, RM, BSc, BN (Hons), MN, PhD

[1]Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[2]The Nethersole School of Nursing, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

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

Abstract

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

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

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

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

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

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

(JMIR Med Educ 2024;10:e49551) doi:10.2196/49551

KEYWORDS

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

XSL•FO
RenderX

Introduction

Background

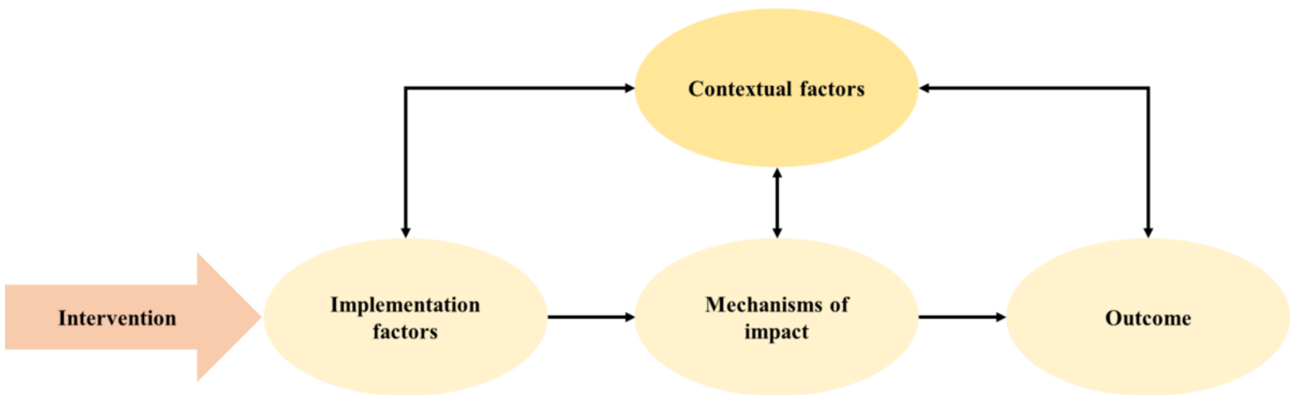
The emergence of the COVID-19 pandemic has led to extensive changes in the health care landscape. Globally, the repeated waves of COVID-19 infections have led to health care professionals (HCPs) grappling with occupational health hazards and overstretched assignments [1,2]. These constant stressors have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

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

Resilient individuals are adept at using personal, relational, and environmental resources to overcome adversity [11,12]. At the personal level, individuals with certain personality traits such as a positive outlook can appraise stressful situations from an optimistic point of view [15]. Based on the transactional model of stress and coping [16], positive emotions may reduce the negative effect that arises when one experiences adversities. Furthermore, individuals with collegial relationships with colleagues and peers can rely on social support resources to overcome adversities [11]. Finally, environmental protective factors in the form of workplace culture can influence an individual's resilience [11,12]. For instance, an organization that focuses on building a collegial and harmonious workplace culture can in turn facilitate one's access to social support resources and thus develop resilience [17,18].

Existing resilience interventions have focused on modifiable personal and relational factors such as the use of cognitive behavioral techniques [19], mindfulness training [20,21], and social competency skills [22,23]. However, most existing literature focused on evaluating the effects of resilience training using quantitative approaches [13]. In line with the development and evaluation of complex interventions [24], using qualitative approaches will be useful in gathering in-depth information about the various contextual and implementational factors that can alter the intended outcomes of the intervention. Particularly, the process evaluation framework [25] proposes that an intervention should be further examined by identifying the contextual factors, implementation processes, mechanisms of impact, and outcomes of the intervention (Figure 1).

Figure 1. Process evaluation framework.



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

Objectives

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

XSL•FO
RenderX

Methods

Ethical Considerations

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

Research Design

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

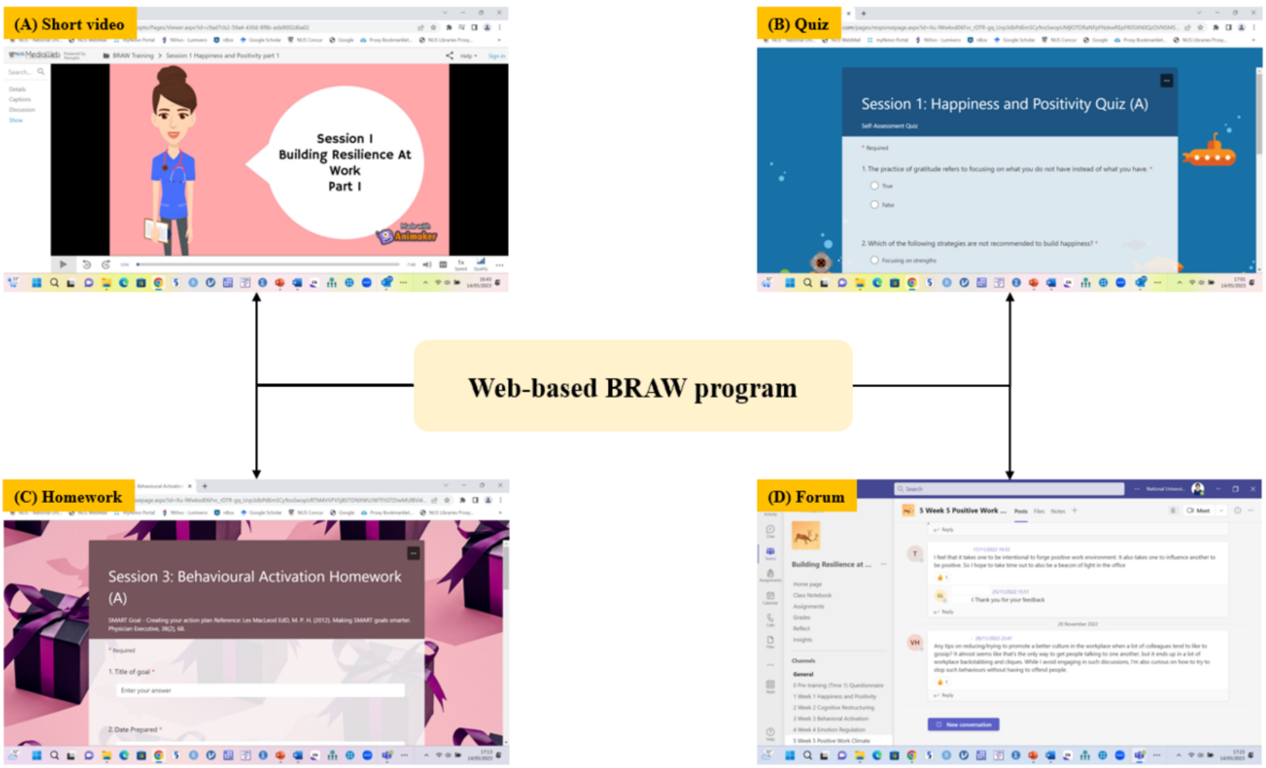
Setting and Participants

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

Web-Based BRAW Program

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

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



XSL•FO
RenderX

Table 1. Overview of the Building Resilience At Work program.

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

Data Collection

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

XSL•FO
RenderX

Textbox 1. Semistructured interview guide.

Questions

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

Data Analysis

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

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

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

Rigor

The principles of credibility, transferability, dependability, and conformability were used to demonstrate rigor [38]. First, a reflexivity journal was maintained by all members of the research team to improve their self-awareness and reduce any potential personal influences on the data. Second, the data

XSL•FO
RenderX

analyses were conducted by 2 independent researchers (WHDA and YL). Third, participants were invited to review their transcripts to clarify the context of the statements and ensure that the final themes and subthemes were representative of their experiences [39]. Subsequently, an audit trail detailing the recruitment, data collection, and analysis process was conducted to ensure ease of replication, transparency, and dependability [38]. Finally, a thick description of the context and the intervention was provided, this facilitates the transferability of the findings of this study [38].

Results

Overview

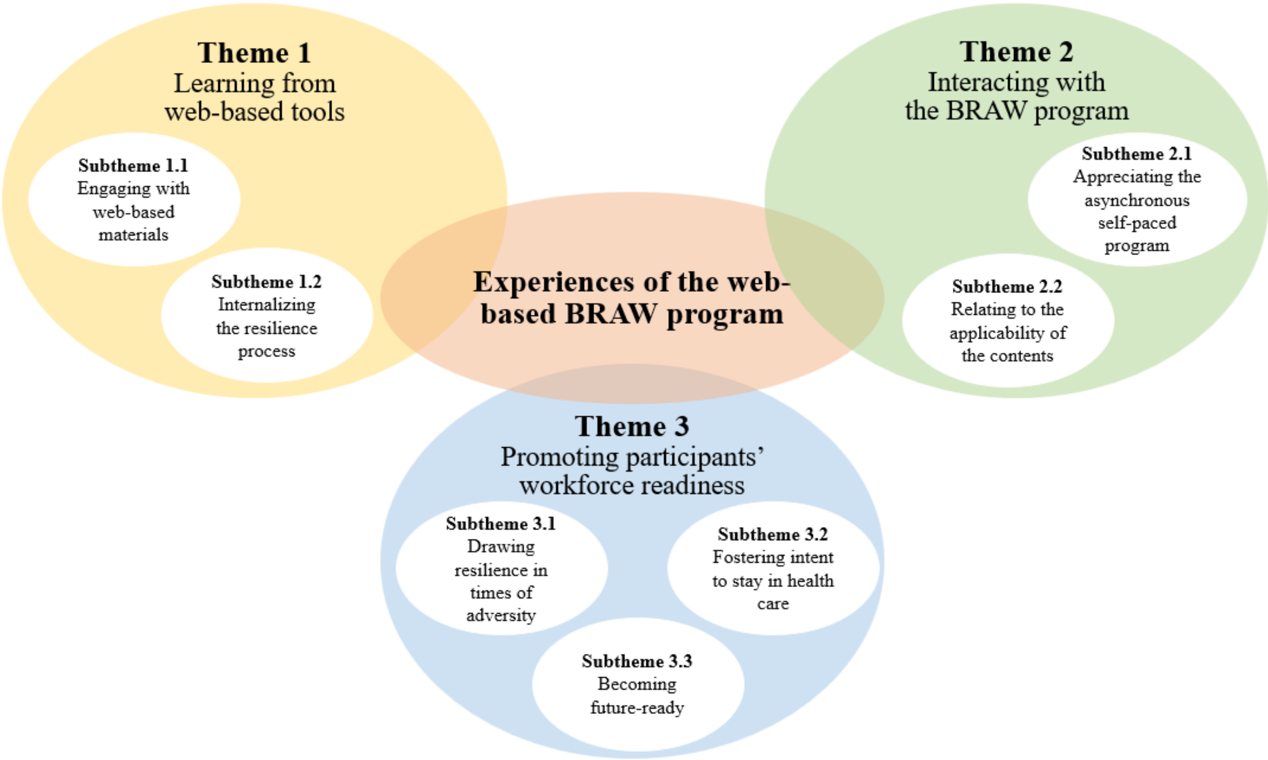
A total of 33 HCPs participated in this qualitative study. The sociodemographic variables are presented in Table 2. Most of the participants were between the ages of 31-40 years (n=11, 34%), female (n=24, 73%), ethnic Chinese (n=25, 76%), and nurses (n=15, 46%). The findings from the framework analysis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

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

Variables	Values
Age group (years), n (%)	
21-25	5 (15)
26-30	9 (27)
31-40	11 (34)
41-50	6 (18)
51-60	2 (6)
Sex, n (%)	
Male	9 (27)
Female	24 (73)
Ethnicity, n (%)	
Chinese	25 (76)
Malay	7 (21)
Indian	1 (3)
Profession, n (%)	
Allied health worker	12 (36)
Clinical administrator	1 (3)
Clinical researcher	4 (12)
Nurse (registered and enrolled)	15 (46)
Physician	1 (3)
Duration of interviews (minutes)	
Mean (SD)	35.48 (7.83)
Range	20-54

XSL•FO
RenderX

Figure 3. Participants' experiences of the web-based BRAW program. BRAW: Building Resilience At Work.



Theme 1: Learning From Web-Based Tools

Overview

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

Engaging With Web Materials

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

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

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

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

Internalizing the Resilience Process

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

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

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

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

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

> *Especially some of the terms, erm, maybe a bit technical? I'm not that acquainted. So, it [referring to the quizzes] allows me to clarify, review and*

XSL•FO
RenderX

> *understand and get it correct.* [Participant 8, female, Chinese, clinical administrator]

Theme 2: Interacting With the BRAW Program

Overview

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

Appreciating the Asynchronous Self-Paced Program

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

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

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

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

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

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

Relating to the Applicability of the Contents

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

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

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

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

Theme 3: Promoting Participants' Workforce Readiness

Overview

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

Drawing Resilience in Times of Adversity

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

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

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

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

Promoting Intent to Stay in Health Care

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

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

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

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

XSL•FO
RenderX

Becoming Future-Ready

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

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

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

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

Discussion

Principal Findings

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

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

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

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

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

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

XSL•FO
RenderX

emotion regulation. More literature has highlighted the pivotal role of noncognitive skills on employment outcomes [56,57].

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

Limitations

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

Conclusions

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

Conflicts of Interest

None declared.

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

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

References

1. Sethi BA, Sethi A, Ali S, Aamir HS. Impact of coronavirus disease (COVID-19) pandemic on health professionals. Pak J Med Sci 2020;36(COVID19-S4):S6-S11 [FREE Full text] [doi: 10.12669/pjms.36.COVID19-S4.2779] [Medline: 32582306]
2. Razu SR, Yasmin T, Arif TB, Islam MS, Islam SMS, Gesesew HA, et al. Challenges faced by healthcare professionals during the COVID-19 pandemic: a qualitative inquiry from Bangladesh. Front Public Health 2021;9:647315 [FREE Full text] [doi: 10.3389/fpubh.2021.647315] [Medline: 34447734]
3. Prasad K, McLoughlin C, Stillman M, Poplau S, Goelz E, Taylor S, et al. Prevalence and correlates of stress and burnout among U.S. healthcare workers during the COVID-19 pandemic: a national cross-sectional survey study. EClinicalMedicine 2021;35:100879 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100879] [Medline: 34041456]
4. Jalili M, Niroomand M, Hadavand F, Zeinali K, Fotouhi A. Burnout among healthcare professionals during COVID-19 pandemic: a cross-sectional study. Int Arch Occup Environ Health 2021;94(6):1345-1352 [FREE Full text] [doi: 10.1007/s00420-021-01695-x] [Medline: 33864490]
5. Pappa S, Ntella V, Giannakas T, Giannakoulis VG, Papoutsi E, Katsaounou P. Prevalence of depression, anxiety, and insomnia among healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. Brain Behav Immun 2020;88:901-907 [FREE Full text] [doi: 10.1016/j.bbi.2020.05.026] [Medline: 32437915]

XSL•FO
RenderX

6. Poon YSR, Lin YP, Griffiths P, Yong KK, Seah B, Liaw SY. A global overview of healthcare workers' turnover intention amid COVID-19 pandemic: a systematic review with future directions. Hum Resour Health 2022;20(1):70 [FREE Full text] [doi: 10.1186/s12960-022-00764-7] [Medline: 36153534]
7. Albougami AS, Almazan JU, Cruz JP, Alquwez N, Alamri MS, Adolfo CA, et al. Factors affecting nurses' intention to leave their current jobs in Saudi Arabia. Int J Health Sci (Qassim) 2020;14(3):33-40 [FREE Full text] [Medline: 32536847]
8. Chen YC, Wu HC, Kuo FT, Koh D, Guo YLL, Shiao JSC. Hospital factors that predict intention of health care workers to leave their job during the COVID-19 pandemic. J Nurs Scholarsh 2022;54(5):607-612 [FREE Full text] [doi: 10.1111/jnu.12771] [Medline: 35187777]
9. Ang WHD, Chew HSJ, Rusli KDB, Ng WHD, Zheng ZJ, Liaw SY, et al. Spotlight on noncognitive skills: views from nursing students and educators. Nurse Educ Today 2022;117:105486. [doi: 10.1016/j.nedt.2022.105486] [Medline: 35917708]
10. Smithers LG, Sawyer ACP, Chittleborough CR, Davies NM, Smith GD, Lynch JW. A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. Nat Hum Behav 2018;2(11):867-880 [FREE Full text] [doi: 10.1038/s41562-018-0461-x] [Medline: 30525112]
11. Szanton SL, Gill JM. Facilitating resilience using a society-to-cells framework: a theory of nursing essentials applied to research and practice. ANS Adv Nurs Sci 2010;33(4):329-343. [doi: 10.1097/ANS.0b013e3181fb2ea2] [Medline: 21068554]
12. Van Breda A. A critical review of resilience theory and its relevance for social work. Soc Work 2018;54(1):1-18 [FREE Full text] [doi: 10.15270/54-1-611]
13. Ang WHD, Chew HSJ, Dong J, Yi H, Mahendren R, Lau Y. Digital training for building resilience: systematic review, meta-analysis, and meta-regression. Stress Health 2022;38(5):848-869 [FREE Full text] [doi: 10.1002/smi.3154] [Medline: 35460533]
14. Chmitorz A, Kunzler A, Helmreich I, Tüscher O, Kalisch R, Kubiak T, et al. Intervention studies to foster resilience—a systematic review and proposal for a resilience framework in future intervention studies. Clin Psychol Rev 2018;59:78-100 [FREE Full text] [doi: 10.1016/j.cpr.2017.11.002] [Medline: 29167029]
15. Gómez-Molinero R, Zayas A, Ruíz-González P, Guil R. Optimism and resilience among university students. Int J Dev Educ Psychol 2018;1(1):147 [FREE Full text] [doi: 10.17060/ijodaep.2018.n1.v1.1179]
16. Lazarus RS. Toward better research on stress and coping. Am Psychol 2000;55(6):665-673. [doi: 10.1037//0003-066x.55.6.665] [Medline: 10892209]
17. Cao X, Li J, Gong S. Effects of resilience, social support, and work environment on turnover intention in newly graduated nurses: the mediating role of transition shock. J Nurs Manag 2021;29(8):2585-2593. [doi: 10.1111/jonm.13418] [Medline: 34252240]
18. Cusack L, Smith M, Hegney D, Rees CS, Breen LJ, Witt RR, et al. Exploring environmental factors in nursing workplaces that promote psychological resilience: constructing a unified theoretical model. Front Psychol 2016;7:600 [FREE Full text] [doi: 10.3389/fpsyg.2016.00600] [Medline: 27242567]
19. Yi-Frazier JP, O'Donnell MB, Adhikari EA, Zhou C, Bradford MC, Garcia-Perez S, et al. Assessment of resilience training for hospital employees in the era of COVID-19. JAMA Netw Open 2022;5(7):e2220677 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.20677] [Medline: 35796151]
20. Mistretta EG, Davis MC, Temkit M, Lorenz C, Darby B, Stonnington CM. Resilience training for work-related stress among health care workers: results of a randomized clinical trial comparing in-person and smartphone-delivered interventions. J Occup Environ Med 2018;60(6):559-568. [doi: 10.1097/JOM.0000000000001285] [Medline: 29370014]
21. DeTore NR, Sylvia L, Park ER, Burke A, Levison JH, Shannon A, et al. Promoting resilience in healthcare workers during the COVID-19 pandemic with a brief online intervention. J Psychiatr Res 2022;146:228-233 [FREE Full text] [doi: 10.1016/j.jpsychires.2021.11.011] [Medline: 34857369]
22. Ang WHD, Chew HSJ, Ong YHN, Zheng ZJ, Shorey S, Lau Y. Becoming more resilient during COVID-19: insights from a process evaluation of digital resilience training. Int J Environ Res Public Health 2022;19(19):12899 [FREE Full text] [doi: 10.3390/ijerph191912899] [Medline: 36232196]
23. Heath C, Sommerfield A, von Ungern-Sternberg BS. Resilience strategies to manage psychological distress among healthcare workers during the COVID-19 pandemic: a narrative review. Anaesthesia 2020;75(10):1364-1371 [FREE Full text] [doi: 10.1111/anae.15180] [Medline: 32534465]
24. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
25. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
26. Sanetti LMH, Cook BG, Cook L. Treatment fidelity: what it is and why it matters. Learn Disabil Res Pract 2021;36(1):5-11 [FREE Full text] [doi: 10.1111/ldrp.12238]
27. Cheshire A, Hughes J, Lewith G, Panagioti M, Peters D, Simon C, et al. GPs' perceptions of resilience training: a qualitative study. Br J Gen Pract 2017;67(663):e709-e715 [FREE Full text] [doi: 10.3399/bjgp17X692561] [Medline: 28893767]

XSL•FO
RenderX

28. Nissim R, Malfitano C, Coleman M, Rodin G, Elliott M. A qualitative study of a compassion, presence, and resilience training for oncology interprofessional teams. J Holist Nurs 2019;37(1):30-44 [FREE Full text] [doi: 10.1177/0898010118765016] [Medline: 29598225]
29. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357 [FREE Full text] [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
30. Healthcare workforce statistics. Health Hub Singapore. 2022. URL: https://www.healthhub.sg/a-z/health-statistics/12/health-manpower [accessed 2024-01-06]
31. Beck JS. Cognitive-behavioral therapy. In: Mack AH, Frances RJ, Miller SI, editors. Clinical Textbook of Addictive Disorders, 3rd Edition. New York: Guilford Publications; 2011:474-501.
32. Hayes SC, Luoma JB, Bond FW, Masuda A, Lillis J. Acceptance and commitment therapy: model, processes and outcomes. Behav Res Ther 2006;44(1):1-25. [doi: 10.1016/j.brat.2005.06.006] [Medline: 16300724]
33. Nezu AM, Nezu CM, D'Zurilla TJ. Problem-Solving Therapy: A Treatment Manual. New York: Springer Publishing Company; 2012.
34. Fusch PI, Ness LR. Are we there yet? Data saturation in qualitative research. Qual Rep 2015;20(9):1408-1416 [FREE Full text] [doi: 10.46743/2160-3715/2015.2281]
35. Ritchie J, Spencer L, O'Connor W. Carrying out qualitative analysis. In: Lewis J, Ritchie J, editors. Qualitative Research Practice: A Guide for Social Science Students and Researchers. London: Sage Publications; 2003:219-262.
36. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
37. Parkinson S, Eatough V, Holmes J, Stapley E, Midgley N. Framework analysis: a worked example of a study exploring young people’s experiences of depression. Qual Res Psychol 2016;13(2):109-129. [doi: 10.1080/14780887.2015.1119228]
38. Krefting L. Rigor in qualitative research: the assessment of trustworthiness. Am J Occup Ther 1991;45(3):214-222. [doi: 10.5014/ajot.45.3.214] [Medline: 2031523]
39. Lincoln YS. Emerging criteria for quality in qualitative and interpretive research. Qual Inq 2016;1(3):275-289 [FREE Full text] [doi: 10.1177/107780049500100301]
40. Luo Y, Li HCW, Cheung AT, Ho LLK, Xia W, Zhang J. Evaluating the experiences of parents of children with cancer engaging in a mobile device-based resilience training programme: a qualitative study. Support Care Cancer 2022;30(7):6205-6214 [FREE Full text] [doi: 10.1007/s00520-022-07066-7] [Medline: 35441899]
41. Ang WHD, Shorey S, Lopez V, Chew HSJ, Lau Y. Generation Z undergraduate students' resilience during the COVID-19 pandemic: a qualitative study. Curr Psychol 2022;41(11):8132-8146 [FREE Full text] [doi: 10.1007/s12144-021-01830-4] [Medline: 34253948]
42. Jones JA. Scaffolding self-regulated learning through student-generated quizzes. Act Learn High Educ 2017;20(2):115-126 [FREE Full text] [doi: 10.1177/1469787417735610]
43. Zainuddin Z, Shujahat M, Haruna H, Chu SKW. The role of gamified e-quizzes on student learning and engagement: an interactive gamification solution for a formative assessment system. Comput Educ 2020;145:103729. [doi: 10.1016/j.compedu.2019.103729]
44. Coutts PM. Meanings of homework and implications for practice. Theory Pract 2004;43(3):182-188. [doi: 10.1207/s15430421tip4303_3]
45. Schatt MD. High school instrumental music students‘ attitudes and beliefs regarding practice: an application of attribution theory. Update Appl Res Music Educ 2011;29(2):29-40 [FREE Full text] [doi: 10.1177/8755123310396981]
46. Giovannetti AM, Quintas R, Tramacere I, Giordano A, Confalonieri P, Uccelli MM, et al. A resilience group training program for people with multiple sclerosis: results of a pilot single-blind randomized controlled trial and nested qualitative study. PLoS One 2020;15(4):e0231380 [FREE Full text] [doi: 10.1371/journal.pone.0231380] [Medline: 32271833]
47. Agarwal B, Brooks SK, Greenberg N. The role of peer support in managing occupational stress: a qualitative study of the sustaining resilience at work intervention. Workplace Health Saf 2020;68(2):57-64 [FREE Full text] [doi: 10.1177/2165079919873934] [Medline: 31538851]
48. Smith B, Shatté A, Perlman A, Siers M, Lynch WD. Improvements in resilience, stress, and somatic symptoms following online resilience training: a dose-response effect. J Occup Environ Med 2018;60(1):1-5 [FREE Full text] [doi: 10.1097/JOM.0000000000001142] [Medline: 28820863]
49. Muthuprasad T, Aiswarya S, Aditya KS, Jha GK. Students' perception and preference for online education in India during COVID-19 pandemic. Soc Sci Humanit Open 2021;3(1):100101 [FREE Full text] [doi: 10.1016/j.ssaho.2020.100101] [Medline: 34173507]
50. Wanner T, Palmer E. Personalising learning: exploring student and teacher perceptions about flexible learning and assessment in a flipped university course. Comput Educ 2015;88:354-369. [doi: 10.1016/j.compedu.2015.07.008]
51. Lieberman JT, Lobban K, Flores Z, Giordano K, Nolasco-Barrientos E, Yamasaki Y, et al. "We all have strengths": a retrospective qualitative evaluation of a resilience training for Latino immigrants in Philadelphia, PA. Health Equity 2019;3(1):548-556 [FREE Full text] [doi: 10.1089/heq.2019.0070] [Medline: 31681906]

XSL•FO
RenderX

52. Mappamiring M, Putra AHPK. Understanding career optimism on employee engagement: broaden-built and organizational theory perspective. J Asian Finance Econ Bus 2021;8(2):605-616 [FREE Full text] [doi: 10.13106/jafeb.2021.vol8.no2.0605]
53. Hampton D, Rayens MK. Impact of psychological empowerment on workplace bullying and intent to leave. J Nurs Adm 2019;49(4):179-185. [doi: 10.1097/NNA.0000000000000735] [Medline: 30829723]
54. Robson A, Robson F. Investigation of nurses' intention to leave: a study of a sample of UK nurses. J Health Organ Manag 2016;30(1):154-173 [FREE Full text] [doi: 10.1108/JHOM-05-2013-0100] [Medline: 26964855]
55. Warshawsky NE, Wiggins AT, Rayens MK. The influence of the practice environment on nurse managers' job satisfaction and intent to leave. J Nurs Adm 2016;46(10):501-507. [doi: 10.1097/NNA.0000000000000393] [Medline: 27571545]
56. Gutman LM, Schoon I. The impact of non-cognitive skills on outcomes for young people. A literature review. Education Endowment Foundation. 2013. URL: https://discovery.ucl.ac.uk/id/eprint/10125763/ [accessed 2024-01-06]
57. Kautz T, Heckman JJ, Diris R, Weel BT, Borghans L. Fostering and measuring skills: improving cognitive and non-cognitive skills to promote lifetime success. National Bureau of Economic Research. 2014. URL: https://www.nber.org/system/files/working_papers/w20749/w20749.pdf [accessed 2024-01-06]

Abbreviations

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

Edited by T Leung, T de Azevedo Cardoso; submitted 01.06.23; peer-reviewed by T Mu; comments to author 12.11.23; revised version received 21.11.23; accepted 28.12.23; published 31.01.24.

Please cite as:
Ang WHD, Lim ZQG, Lau ST, Dong J, Lau Y
Unpacking the Experiences of Health Care Professionals About the Web-Based Building Resilience At Work Program During the COVID-19 Pandemic: Framework Analysis
JMIR Med Educ 2024;10:e49551
URL: https://mededu.jmir.org/2024/1/e49551
doi:10.2196/49551
PMID:38294866

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

XSL•FO
RenderX

A Pilot Project to Promote Research Competency in Medical Students Through Journal Clubs: Mixed Methods Study

Mert Karabacak[1], MD; Zeynep Ozcan[2]; Burak Berksu Ozkara[3], MD; Zeynep Sude Furkan[2]; Sotirios Bisdas[4], MSc, MD, PhD

1

2

3

4

Corresponding Author:
Sotirios Bisdas, MSc, MD, PhD

Abstract

Background: Undergraduate medical students often lack hands-on research experience and fundamental scientific research skills, limiting their exposure to the practical aspects of scientific investigation. The Cerrahpasa Neuroscience Society introduced a program to address this deficiency and facilitate student-led research.

Objective: The primary goal of this initiative was to enhance medical students' research output by enabling them to generate and publish peer-reviewed papers within the framework of this pilot project. The project aimed to provide an accessible, global model for research training through structured journal clubs, mentorship from experienced peers, and resource access.

Methods: In January 2022, a total of 30 volunteer students from various Turkish medical schools participated in this course-based undergraduate research experience program. Students self-organized into 2 groups according to their preferred study type: original research or systematic review. Two final-year students with prior research experience led the project, developing training modules using selected materials. The project was implemented entirely online, with participants completing training modules before using their newly acquired theoretical knowledge to perform assigned tasks.

Results: Based on student feedback, the project timeline was adjusted to allow for greater flexibility in meeting deadlines. Despite these adjustments, participants successfully completed their tasks, applying the theoretical knowledge they had gained to their respective assignments. As of April 2024, the initiative has culminated in 3 published papers and 3 more under peer review. The project has also seen an increase in student interest in further involvement and self-paced learning.

Conclusions: This initiative leverages globally accessible resources for research training, effectively fostering research competency among participants. It has successfully demonstrated the potential for undergraduates to contribute to medical research output and paved the way for a self-sustaining, student-led research program. Despite some logistical challenges, the project provided valuable insights for future implementations, showcasing the potential for students to engage in meaningful, publishable research.

(JMIR Med Educ 2024;10:e51173) doi:10.2196/51173

KEYWORDS

medical student; research; peer education; student society; journal club; skills; scientific investigation; undergraduate; student-led; initiative; resources; research training; competency; continuing education; research improvement; motivation; mentor; mentorship; medical education

Introduction

Undergraduate medical students frequently face limited opportunities for hands-on research experience [1,2]. Current medical school curricula often fail to equip students adequately with fundamental scientific research skills. Despite a high proportion of students expressing interest in research, only a small fraction possesses a thorough understanding of the medical research process [3]. In addition, empirical evidence underlines the contribution of undergraduate research engagement to career progression in medicine [4]. Consequently, medical students, cognizant of research's significance, are increasingly seeking opportunities to augment their involvement [5]. To facilitate this quest, various course-based undergraduate research experience (CURE) programs have emerged, albeit with room for further refinement [6].

The emergence of remote learning, coupled with the proliferation of web-based platforms and open-access journals, has amplified data accessibility and the availability of research tools. This shift has catalyzed scientific literacy development and enabled self-paced learning among students across diverse disciplines. Students can now opt for extracurricular web-based

XSL•FO
RenderX

courses or access materials of varying media to deepen their understanding of selected topics, acquire new skills, and enhance their overall capabilities. Beyond an individualistic approach, web-based platforms have simplified the process for students to find groups for information exchange, thereby bolstering their scientific understanding. The Cerrahpasa Neuroscience Society, a student-led organization, hosts 4 journal clubs wherein students gather online to discuss neurosurgery, neurology, psychiatry, and neuroscience through selected papers and subsequent discussions [7]. The primary aim of this research initiative is to stimulate and mentor students within these journal clubs to undertake their own research projects, leveraging a structured program replete with experienced near-peer guidance and comprehensive information access. The program's ultimate objective is the inception and publication of fully student-run studies and papers in peer-reviewed journals, marrying theoretical knowledge with research fundamentals in a hands-on setting.

Our organizing team, composed of 1 second-year and 2 final-year medical students (MK, ZO, and BBO), sought to exploit the omnipresence of information and the scientific curiosity of journal club participants. We embarked on a pilot project with the clear ambition of significantly enhancing the research output of undergraduate medical students.

Methods

Ethical Considerations

Ethical approval was deemed unnecessary by the Istanbul University-Cerrahpasa — Cerrahapasa Faculty of Medicine Institutional Review Board as the survey responses were anonymous and participants consented to their data being used for research purposes. Participants' data were anonymized and no compensation was provided for the participants. In addition, the data originated from the activities of the Cerrahpasa Neuroscience Society, which were conducted remotely and independently of the university.

Planning

The project involved participants exclusively from the journal clubs, encompassing undergraduate medical students from various Turkish medical faculties and academic levels. We presented the project idea to all club members in December 2021, with 30 of the 40 members volunteering to partake in the project.

As the organizing committee, we compiled a series of introductory papers and vetted web-based courses centered around research fundamentals, which we arranged into scheduled training modules (Multimedia Appendices 1 and 2). We slated monthly briefings to guide and track participants' progress, consistent with the project timeline. The final-year students in the organizing team (hereafter referred to as "the tutors"), who had accrued prior research experience, pinpointed research topics appropriate for undergraduate projects (Table 1).

Table . Study types and topics selected for research project implementation.

Journal club subject and study type		Research topic
Neurosurgery		
	Systematic Review	Medulloblastoma subgroup classification with radiomics
	Systematic Review	Radiosensitizing agents in medulloblastoma
Neurology		
	Original Study	Differentiation of SPMS[a] formation using machine learning
	Systematic Review	The concurrence of multiple sclerosis and glioblastoma
Psychiatry		
	Original Study	Medical student stress, burnout, and depression in Turkey
	Original Study	Substance use and mental health among medical students in Turkey
Neuroscience		
	Original Study	Adolescents' sleep and academic standing
	Systematic Review	Neuropsychological outcomes following radiation therapy of pediatric posterior fossa tumors

[a]SPMS: secondary progressive multiple sclerosis.

Study type selection was based on practical considerations and journal club subjects. Two systematic reviews were assigned to the neurosurgery journal club, as conducting original studies in this field would pose challenges for undergraduate students.

XSL•FO
RenderX

The psychiatry journal club was tasked with creating 2 original studies using survey methods, whereas the neurology and neuroscience journal clubs were each assigned 1 systematic review and 1 original study. The project was planned to be entirely online, spanning 6 months from January 2022 to June 2022.

Implementation

In January 2022, participants from the 4 journal clubs were segregated into 2 groups based on their chosen study type: original papers or systematic reviews, resulting in 8 project groups. Each group, composed of 3‐5 students, incorporated journal club participants and the clubs' tutors. The project was guided by academic supervisors, who ensured methodological rigor and provided expert advice on the research topics. Tutors BBO and MK, both final-year medical students with prior research experience, actively identified suitable research questions, led regular web-based meetings, provided ongoing feedback, responded to participants' queries, and facilitated navigation through the various stages of the research projects. Communication within these groups was facilitated through web-based chat platforms, with the organizing team included.

The educational materials for the project were meticulously selected based on the tutors' personal experiences and an extensive review of available web-based resources. This process ensured that the materials were both relevant and of high educational quality. Following their initial briefing on project fundamentals and expectations, participants embarked on the first training module intended to acquaint them with research basics (Multimedia Appendix 1). This module contained a video series on using PubMed and Zotero (Center for History and New Media at George Mason University), along with 6 papers detailing the general steps of a research project. Participants were encouraged to complete these materials at their own pace within a 1-month time frame.

In February 2022, we classified tasks into 3 categories: "literature review and data extraction," "statistical analysis," and "manuscript writing." Participants were allocated these tasks primarily based on their skills and interests. The second training module offered specific web-based courses for these tasks (Multimedia Appendix 2). Unlike the first module, the deadline for the second module was tailored to each participant's task timeline. During monthly web-based briefings, the tutors provided assistance and feedback while illustrating task examples. Furthermore, tutors guided participants in ancillary tasks such as database access and ethics committee application form preparation.

Unexpected constraints led us to revise some study designs and the schedule. Two projects—a systematic review and an original study requiring database access—were discontinued and substituted with bibliometric analysis projects, supplemented by additional peer training. Heeding participant feedback, we also extended the original deadlines.

Bibliometric studies necessitated unique procedures, executed throughout March and April 2022 with regular briefings. In May 2022, a tutor conducted an auxiliary academic writing workshop. Although not every participant was tasked with study documentation, we believed all could glean valuable insights from this near-peer workshop within the project's ambit. This workshop was made available to all participants, with a recorded version disseminated for those unable to attend the live session.

Results

The project began with 30 undergraduate medical students, some contributing to multiple projects, and concluded with 25 participants successfully adhering to the full schedule. Those who withdrew from the project did so during the implementation of the second module, necessitating adjustments in task assignments and study configurations. The 25 students successfully adhering to the full schedule were from 5 universities, with a significant concentration (18 students) at Istanbul University—Cerrahpasa. The remainder was distributed to 4 other universities. Ten of these participants were enrolled in English language medical programs. The cohort consisted of 19 preclinical students, who were primarily enrolled in basic medical sciences courses, and 6 clinical students, who were completing clerkships and internships. The group included 14 female and 11 male students. None of the students had previous research experience.

The completion of the first module was gauged through participants' feedback on the materials and their demonstrated proficiency in operating the platforms integrated into the module. Given that the second module encompassed verified web-based courses, completion was monitored via certifications from the respective platforms. As this module required the practical application of learned theoretical skills, successful task execution within the research study denoted each participant's successful project completion.

Participants tasked with "literature review and data collection" and "statistical analysis" adeptly applied their theoretical knowledge acquired from the courses and briefings, creating necessary data tables and thus fulfilling their tasks. Owing to requisite timeline adjustments, those delegated to academic writing courses completed their tasks at disparate times relative to the original schedule. Nevertheless, manuscript creation for all studies was achieved, signifying that all participants made their respective contributions to the project.

During the implementation phase, we made some timeline alterations in response to student feedback, with participants requesting more accommodating deadlines. The tutors, who had previously conducted independent research projects, provided substantial support throughout the project's execution. They shared their experiences and offered guidance, aiding participants in gaining a deeper understanding of their tasks. Upon the conclusion of second module training, participants shared feedback on the project's implementation. As of April 2024, 3 papers have been published in peer-reviewed journals [8-10], and 4 papers have been submitted for peer review.

Discussion

Principal Findings

A multi-institutional study revealed that although 83% of the students surveyed believed that participating in research was

XSL•FO
RenderX

educationally beneficial, only 31% thought that there was enough time allocated for it. In addition, just 15% felt that they received adequate training in research methodology, and only 25% considered the training in critical appraisal to be sufficient [11]. CUREs may help alleviate some of these issues, potentially improving access to both time and quality instruction in research skills. In parallel, CUREs have been adopted in universities to inspire students to pursue research, although these programs predominantly involve student participation in laboratory settings for data collection [6]. Efforts to enhance student contributions to research include a microbiology research initiative at a Canadian university aimed at publishing student-authored papers [12] and remote CUREs focusing on ecology at a US university [13]. However, our program offers a distinct approach.

The pilot project's foremost insight is the successful operation of a fully student-run research program, from the training process to paper publication, using a method that is universally applicable to students interested in research. Instead of creating new lectures, we leveraged preexisting, verified web-based resources on research basics for our training modules. This strategy emphasizes the global accessibility of research training for students, as these resources are publicly available and facilitate self-paced learning. Notably, our research initiative aimed to exploit information accessibility not just only for skill acquisition but also for data collection to complete a research project. As undergraduate medical students have limited access to hands-on research, effectively using available information is crucial. To underscore global applicability and facilitate research involvement, the data used were sourced from 2 web-based platforms: academic literature and web-based surveys. Our project enabled the conduct of both systematic reviews and original studies by students lacking prior hands-on research experience, thereby enhancing student research output through remote involvement.

Regarding skill development, the project achieved the anticipated results. All participants completed their respective tasks, culminating in the production of completed manuscripts. This outcome demonstrates a tangible enhancement in participants' research skills, particularly considering their nonexistent prior research experience. It is critical to note that participants did not receive identical training; instead, a division of labor was used. Participants learned about and practiced various aspects of research study design, with the project's methodology allowing them to hone their skills in their chosen task within a study. As an immediate benefit, some participants expressed interest in completing the remaining web-based courses to further their skills, suggesting potential for project continuation. Feedback indicated a keen interest in furthering the concept. For instance, students initially assigned to work with databases expressed a desire to design and conduct a study upon completing their current work. Previous research suggests that students gain an improved understanding of the benefits of research following participation, and our project participants' enthusiasm supports this finding [14].

A key success factor was the experienced near-peer tutoring that accompanied the project's full duration. The final-year students on the organizing team designed the project outline, selected suitable research questions, monitored progress, and provided guidance as needed. Monthly briefings fostered an environment where participants reported progress and posed questions. Within these meetings, tutors also demonstrated tasks to facilitate student understanding. Participants gained comprehensive insight into the research process by first completing a course, then practicing an example task, and finally executing their respective tasks independently. Participant feedback suggested that near-peer tutoring facilitated question asking, contributing to a comfortable learning environment. Overall, including student guidance in the initiative increased the efficiency of the modules, as also affirmed by participants.

A significant advantage of the project's design is its potential for self-sustainability. The students trained during this pilot project now have the experience to guide subsequent student cohorts looking to enhance their research skills. They can also offer fresh ideas for improving the training modules based on their experiences. Through this cyclical process, we aim to establish a fully student-led research group that cultivates student training, ultimately enhancing medical students' research skills, experience, and productivity.

Research interest among medical students has been found to diminish as they advance through their academic years [15]. Our study supports these findings somewhat, as the majority of our cohort consisted of 19 preclinical students, compared with just 6 clinical students who were involved in clerkships and internships. To counteract this decline, efforts to promote research could be strengthened throughout their university education, potentially through the integration of CUREs. In addition, research indicates a decrease in the number of clinician-scientists in the United Kingdom, attributed partly to an inadequate influx of individuals into the "clinician-scientist pipeline" to replenish an aging workforce [16]. CUREs could potentially boost enthusiasm and familiarity with research, encouraging more individuals to pursue these career paths. Finally, although empirical evidence on the impact of student-led initiatives in academic medicine is limited, their widespread acceptance and popularity may suggest that students recognize a need for these programs and gain some value from participating in them [17]. Our study addresses this gap and serves as a call to action for policy makers.

Limitations

As a pilot project, this research initiative revealed several areas needing revision and adjustment alongside the desired outcomes. First, the journal club's membership was self-selected, which is likely to have influenced the project's results. Participants were predisposed to be more motivated and interested in research, which may have increased engagement and contribution quality. This self-selection helped the project succeed by ensuring that participants were highly committed. However, it introduced a potential selection bias, reducing the generalizability of our findings. The predominance of motivated individuals may not accurately reflect the larger medical student population, particularly those who are less inclined or confident in conducting research. Moreover, while the publication of research papers by participants is an objective indicator of the project's success, it is acknowledged that publication does not

XSL•FO
RenderX

fully capture the breadth of research competencies sought by this program. Focusing solely on publication outcomes has the potential to overlook broader research skills such as ethical considerations, data management, and long-term research planning, all of which are critical to the sustainability of research practices. While we emphasized the importance of ethical considerations during our sessions, it is important to remember that the primary goal of our research was to help students take their first steps into the world of research. As a result, while we acknowledge the project's scope and depth of limitations, it effectively served as an entry point for participants, many of whom had no prior research experience, to begin engaging with the research process.

Furthermore, various challenges surfaced during the project's execution, leading to alterations in specific project elements and deadline extensions. One of these challenges was the inability to gain access to databases initially planned for original research study designs. Our requests were not met with a positive response, necessitating a change in the content and methodology of the projects requiring such access. For future iterations, we intend to include study proposals that leverage access to these databases. Another unexpected obstacle was the delay in obtaining ethical committee approvals for survey studies. This issue was not factored into the original timeline and, in light of this experience, we will allow for greater flexibility in project schedules moving forward. The project implementation process highlighted key factors that require consideration to ensure the project's sustainability and ease of execution. One significant challenge was coordinating teamwork among participants with varied schedules. Sticking to the initial timeline was difficult for all participants, leading us to recommend gathering schedules beforehand and grouping students with similar availability together for future implementations. Decreased commitment from some participants was another issue. Over the course of the lengthy project, some students withdrew, primarily due to time constraints. This situation required the reassignment of certain tasks and additional courses for some students. Also, the project targeted a small, specific group of students and lacked a selection process. To address this, we propose the inclusion of an application process for more efficient training in future iterations. Another limitation was the lack of active engagement from participants in the question design. The tutors provided guidance and designed the research questions themselves as a starting point. Adding a course on this topic in future iterations could foster more active involvement from participants, thereby potentially improving project outcomes. Finally, dividing tasks among participants posed challenges in ensuring full research competency for all, as each participant focused on specific aspects of the project.

Conclusions

This project effectively capitalized on the widespread accessibility of information to educate and enable students to partake in medical research, irrespective of their lack of direct hands-on experience. This approach carries significant weight as it equips students with the skills to draw data from preexisting studies, thereby exploiting the incremental nature of science. In addition, this method provides students from campuses with limited access to research facilities the opportunity to acquire experience in conducting a scientific project. Despite the encountered challenges, the project was successfuly implemented, resulting in a notable advancement of the research skill set among previously inexperienced students. This translated into a demonstrable increase in undergraduate research output. The limitations identified during the project's course provide a crucial understanding for improving future iterations of this initiative. Our goal is to perpetually refine and use this project as a supplement to traditional medical training, thus providing students with a keen interest in research the opportunity for self-paced learning and research training.

Acknowledgments

We would like to thank all Cerrahpasa Neuroscience Society members who contributed to the journal club meetings and this project. This project was funded by the Association for Medical Education in Europe (AMEE), as a part of the Student Initiative Grant Awards (SIGA) 2022.

Conflicts of Interest

None declared.

Multimedia Appendix 1

First module materials.

[DOCX File, 16 KB - mededu_v10i1e51173_app1.docx]

Multimedia Appendix 2

Second module materials.

[DOCX File, 15 KB - mededu_v10i1e51173_app2.docx]

References

1. Mass-Hernández LM, Acevedo-Aguilar LM, Lozada-Martínez ID, et al. Undergraduate research in medicine: a summary of the evidence on problems, solutions and outcomes. Ann Med Surg 2022;74:103280. [doi: 10.1016/j.amsu.2022.103280]

XSL•FO
RenderX

2. Mabvuure NT. Twelve tips for introducing students to research and publishing: a medical student's perspective. Med Teach 2012;34(9):705-709. [doi: 10.3109/0142159X.2012.684915] [Medline: 22905656]
3. Burgoyne LN, O'Flynn S, Boylan GB. Undergraduate medical research: the student perspective. Med Educ Online 2010 Sep 10;15(1). [doi: 10.3402/meo.v15i0.5212] [Medline: 20844608]
4. Sorial AK, Harrison-Holland M, Young HS. The impact of research intercalation during medical school on post-graduate career progression. BMC Med Educ 2021 Jan 8;21(1):39. [doi: 10.1186/s12909-020-02478-7] [Medline: 33419435]
5. El Achi D, Al Hakim L, Makki M, et al. Perception, attitude, practice and barriers towards medical research among undergraduate students. BMC Med Educ 2020 Jun 17;20(1):195. [doi: 10.1186/s12909-020-02104-6] [Medline: 32552801]
6. Bangera G, Brownell SE. Course-based undergraduate research experiences can make scientific research more inclusive. CBE Life Sci Educ 2014;13(4):602-606. [doi: 10.1187/cbe.14-06-0099] [Medline: 25452483]
7. Ozkara BB, Karabacak M, Alpaydin DD. Student-run online journal club initiative during a time of crisis: survey study. JMIR Med Educ 2022;8(1):e33612. [doi: 10.2196/33612]
8. Karabacak M, Hakkoymaz M, Ukus B, et al. Final-year medical student mental wellness during preparation for the examination for specialty in Turkey: a cross-sectional survey study. BMC Med Educ 2023 Feb 1;23(1):79. [doi: 10.1186/s12909-023-04063-0] [Medline: 36726114]
9. Karabacak M, Ozkara BB, Ozturk A, et al. Radiomics-based machine learning models for prediction of medulloblastoma subgroups: a systematic review and meta-analysis of the diagnostic test performance. Acta Radiol 2023 May;64(5):1994-2003. [doi: 10.1177/02841851221143496]
10. Karabacak M, Kose EB, Bahadir Z, et al. Factors associated with substance use among preclinical medical students in Turkey: a cross-sectional study. Can Med Educ J 2024 Jul;15(3):37-44. [doi: 10.36834/cmej.77088] [Medline: 39114776]
11. Siemens DR, Punnen S, Wong J, Kanji N. A survey on the attitudes towards research in medical school. BMC Med Educ 2010 Jan 22;10:4. [doi: 10.1186/1472-6920-10-4] [Medline: 20096112]
12. Sun E, Graves ML, Oliver DC. Propelling a course-based undergraduate research experience using an open-access online undergraduate research journal. Front Microbiol 2020;11:589025. [doi: 10.3389/fmicb.2020.589025] [Medline: 33329466]
13. Fey SB, Theus ME, Ramirez AR. Course-based undergraduate research experiences in a remote setting: two case studies documenting implementation and student perceptions. Ecol Evol 2020 Nov;10(22):12528-12541. [doi: 10.1002/ece3.6916] [Medline: 33250991]
14. Imafuku R, Saiki T, Kawakami C, Suzuki Y. How do students' perceptions of research and approaches to learning change in undergraduate research? Int J Med Educ 2015 Apr 12;6:47-55. [doi: 10.5116/ijme.5523.2b9e] [Medline: 25863495]
15. Sanabria-de la Torre R, Quiñones-Vico MI, Ubago-Rodríguez A, Buendía-Eisman A, Montero-Vílchez T, Arias-Santiago S. Medical students' interest in research: changing trends during university training. Front Med (Lausanne) 2023;10:1257574. [doi: 10.3389/fmed.2023.1257574] [Medline: 37928463]
16. Parameswaran G, Bowman A, Swales C, et al. Cross-sectional survey of medical student perceptions of and desires for research and training pathways (SMART): an analysis of prospective cohort study of UK medical students. BMC Med Educ 2023 Dec 15;23(1):964. [doi: 10.1186/s12909-023-04881-2] [Medline: 38102619]
17. Funston G. The promotion of academic medicine through student-led initiatives. Int J Med Educ 2015 Nov 21;6:155-157. [doi: 10.5116/ijme.563a.5e29] [Medline: 26590359]

Abbreviations

CURE: course-based undergraduate research experience

Edited by B Lesselroth; submitted 23.07.23; peer-reviewed by J Abbas, M Akyol, P Jagtiani, R Jenkin; revised version received 17.04.24; accepted 13.07.24; published 31.10.24.

Please cite as:
Karabacak M, Ozcan Z, Ozkara BB, Furkan ZS, Bisdas S
A Pilot Project to Promote Research Competency in Medical Students Through Journal Clubs: Mixed Methods Study
JMIR Med Educ 2024;10:e51173
URL: https://mededu.jmir.org/2024/1/e51173
doi:10.2196/51173

© Mert Karabacak, Zeynep Ozcan, Burak Berksu Ozkara, Zeynep Sude Furkan, Sotirios Bisdas. Originally published in JMIR Medical Education (https://mededu.jmir.org), 31.10.2024. 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

XSL•FO
RenderX

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.

XSL•FO
RenderX

Original Paper

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

Susanne G Johnson[1*], MSc; Birgitte Espehaug[1*], Prof Dr; Lillebeth Larun[2*], PhD; Donna Ciliska[3*], Prof Dr; Nina Rydland Olsen[1*], PhD

[1]Department of Health and Functioning, Western Norway University of Applied Sciences, Bergen, Norway

[2]Division of Health Services, Norwegian Institute of Public Health, Oslo, Norway

[3]Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

[*]all authors contributed equally

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

Abstract

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

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

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

Results: Of 4 cohorts with 150 students, 119 (79.3%) students used the app and produced 240 critically appraised topics (CATs) in the app. The EBP steps "ask," "assess," and "appraise" were often correctly performed. The clinical question was formulated correctly in 53.3% (128/240) of the CATs, and students identified research evidence in 81.2% (195/240) of the CATs. Critical appraisal checklists were used in 81.2% (195/240) of the CATs, and most of these checklists were assessed as relevant for the type of research evidence identified (165/195, 84.6%). The least frequently correctly reported steps were "apply" and "audit." In 39.6% (95/240) of the CATs, it was reported that research evidence was applied. Only 61% (58/95) of these CATs described how the research was applied to clinical practice. Evaluation of practice changes was reported in 38.8% (93/240) of the CATs. However, details about practice changes were lacking in all these CATs. A positive association was found between correctly reporting the "population" and "interventions/interest" elements of the PICO/PICo and identifying research evidence ($P<.001$).

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

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

KEYWORDS

active learning strategies; application; cross-sectional study; development; education; higher education; interactive; mobile application; mobile app; occupational therapy students; occupational therapy; students; usability; use

XSL•FO
RenderX

Introduction

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

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

Methods

Design

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

Mobile App

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

Textbox 1. The EBPsteps app content.

Ask

- Reflect on information needs
- Formulate the clinical question
- Identify the type of clinical question (drop-down menu)
- Identify the Population, Intervention, Comparison, and Outcome/Population, Interest, and Context (PICO/PICo) elements

Assess

- Report information source used to identify research evidence
- Report links to research evidence identified

Appraise

- Choose a relevant critical appraisal checklist
- Complete the critical appraisal using the integrated checklist

Apply

- Report how research evidence was applied in practice (drop-down menu)

Audit

- Report if changes in practice were completed and evaluated
- Describe changes if changes were implemented
- Evaluate the EBP process (ask, assess, appraise, apply, and audit)

XSL•FO
RenderX

By documenting the EBP process in the app, students produced critically appraised topics (CATs). A CAT can be explained as a summary of research evidence on a clinical question [18]. The CATs completed in the EBPsteps app included information on all EBP steps, and the CATs could be sent through email and shared as a PDF document. The CATs produced in the app were stored on the HVL research server and were accessible to the researchers in this project.

Participants

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

Setting

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

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

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

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

Year 1

- Standalone sessions about "ask" (2 hours) and "assess" (2 hours). Total duration is 4 hours.

Year 2

- Standalone sessions about "ask" (1 hour), "assess" (1 hour), "appraise" (3 hours), and "apply" (2 hours). Total duration is 7 hours.

Year 3

- Standalone sessions about "ask" (2 hours), "assess" (2 hours), "appraise" (8 hours), "apply" (3 hours), and "audit" (1 hour). Total duration is 16 hours.

Data Collection

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

XSL•FO
RenderX

Textbox 3. Overview of the scoring plan. Includes the EBP steps and what was assessed. EBP: evidence-based practice.

Ask

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

Assess

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

Appraise

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

Apply

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

Audit

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

Analysis

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

Ethical Considerations

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

Results

Participants

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

Step 1: Ask

A need for more knowledge on a clinical problem was reported in 94.6% (227/240) CATs. In 80% (192/240) of the CATs, the type of clinical question was identified using a drop-down menu. A clinical question was formulated in 53.3% (128/240) of the CATs. The "effect of therapy" was the most prevalent clinical question reported (100/240, 41.7%) (Table 1).

XSL•FO
RenderX

All PICO/PICo elements were reported correctly in 10.4% (25/240) of the CATs. Assessing the different PICO/PICo elements separately, the "population" and "intervention/interest" elements were more often correctly reported (187/240, 77.9% and 189/240, 78.8%) than the "comparison" and "outcome/context" elements (44/240, 18.3% and 103/240, 42.9%). This applied to all question types, including when the question had been formulated as a background question (Table 1). In CATs without a clinical question identified, most PICO/PICo elements were incorrectly reported.

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

	Population, n (%)	Intervention/interest, n (%)	Comparison, n (%)	Outcome/context, n (%)
Effect of therapy (n=100)	90 (90)	96 (96)	30 (30)	53 (53)
Qualitative (n=27)	25 (93)	25 (93)	N/R[a]	13 (48)
Background (n=64)	55 (86)	52 (81)	11 (17)	32 (50)
Other (n=1) or missing (n=48)	17 (35)	16 (33)	3 (6)	5 (10)

[a]Not relevant.

Step 2: Assess

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

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

Step 3: Appraise

A checklist was used in 81.3% (195/240) of the CATs. Of these, the correct checklist was used in 84.6% (165/195) of the CATs; that is, there was agreement between the type of checklist and the research evidence identified (Table 2).

In 98.2% (162/165) of the CATs with a correct checklist, more than 75% of the checklist questions had been answered. Effect estimates from identified research evidence were documented in 27% (21/77) of the checklists for systematic reviews and 36% (15/42) of the checklists for RCTs.

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

Type of research evidence	The agreement between research evidence and checklist, n (%)
Systematic reviews (n=85)	77 (89)
Randomized controlled trials (n=51)	42 (82)
Qualitative research (n=44)	42 (95)
Guidelines (n=4)	2 (50)
Observational studies[a] (n=11)	2 (18)
The total number of research evidence identified (n=195)	165 (84.6)

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

Step 4: Apply

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

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

Step 5: Audit

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

XSL•FO
RenderX

Discussion

Principal Findings

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

Comparison to Previous Work

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

Ask and Assess

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

Appraise

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

Apply and Audit

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

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

Future Directions

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

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

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

Methodological Considerations

The main limitation of this study was that we included students from only one profession and from the same educational institution, and thus the generalizability of the results to other institutions and to other health and social care students is reduced. However, the sample consisted of 4 student cohorts from different academic years (from 2018 to 2021; n=119), including 240 CATs. Accordingly, we believe the results from this study can be recognizable and relevant across other populations.

XSL•FO
RenderX

A strength of this study was that the EBPsteps app allowed us to objectively measure the performance of the EBP process using an app that includes all 5 EBP steps. It is recommended that educators select instruments that objectively measure EBP performance [11]. Shaneyfelt et al [6] emphasized the use of online documentation of the EBP steps as a promising approach.

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

Conclusions

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

Acknowledgments

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

Data Availability

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

Authors' Contributions

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

Conflicts of Interest

None declared.

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

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

References

1. Dawes M, Summerskill W, Glasziou P, Cartabellotta A, Martin J, Hopayian K, et al. Sicily statement on evidence-based practice. BMC Med Educ 2005;5(1):1 [FREE Full text] [doi: 10.1186/1472-6920-5-1] [Medline: 15634359]
2. Hoffmann T, Bennett S, Del MC. Evidence-Based Practice Across the Health Professions, 3rd Edition. Chatswood, Australia: Elsevier; 2017.
3. Albarqouni L, Hoffmann T, Straus S, Olsen NR, Young T, Ilic D, et al. Core competencies in evidence-based practice for health professionals: consensus statement based on a systematic review and Delphi survey. JAMA Netw Open 2018;1(2):e180281 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.0281] [Medline: 30646073]
4. Tilson JK, Kaplan SL, Harris JL, Hutchinson A, Ilic D, Niederman R, et al. Sicily statement on classification and development of evidence-based practice learning assessment tools. BMC Med Educ 2011;11:78 [FREE Full text] [doi: 10.1186/1472-6920-11-78] [Medline: 21970731]
5. Bala MM, Poklepović Peričić T, Zajac J, Rohwer A, Klugarova J, Välimäki M, et al. What are the effects of teaching Evidence-Based Health Care (EBHC) at different levels of health professions education? An updated overview of systematic reviews. PLoS One 2021;16(7):e0254191 [FREE Full text] [doi: 10.1371/journal.pone.0254191] [Medline: 34292986]
6. Shaneyfelt T, Baum KD, Bell D, Feldstein D, Houston TK, Kaatz S, et al. Instruments for evaluating education in evidence-based practice: a systematic review. JAMA 2006;296(9):1116-1127. [doi: 10.1001/jama.296.9.1116] [Medline: 16954491]

XSL•FO
RenderX

7. Thomas A, Saroyan A, Dauphinee WD. Evidence-based practice: a review of theoretical assumptions and effectiveness of teaching and assessment interventions in health professions. Adv Health Sci Educ Theory Pract 2011;16(2):253-276. [doi: 10.1007/s10459-010-9251-6] [Medline: 20922477]
8. Kumaravel B, Hearn JH, Jahangiri L, Pollard R, Stocker CJ, Nunan D. A systematic review and taxonomy of tools for evaluating evidence-based medicine teaching in medical education. Syst Rev 2020;9(1):91 [FREE Full text] [doi: 10.1186/s13643-020-01311-y] [Medline: 32331530]
9. Roberge-Dao J, Maggio L, Zaccagnini M, Rochette A, Shikako-Thomas K, Boruff J, et al. Quality, methods, and recommendations of systematic reviews on measures of evidence-based practice: an umbrella review. JBI Evid Synth 2022;20(4):1004-1073. [doi: 10.11124/JBIES-21-00118] [Medline: 35220381]
10. van de Mortel TF. Faking it: social desirability response bias in self-report research. Aust J Adv Nurs 2008;25(4):40-48 [FREE Full text]
11. Buchanan H, Siegfried N, Jelsma J. Survey instruments for knowledge, skills, attitudes and behaviour related to evidence-based practice in occupational therapy: a systematic review. Occup Ther Int 2016;23(2):59-90 [FREE Full text] [doi: 10.1002/oti.1398] [Medline: 26148335]
12. Albarqouni L, Hoffmann T, Glasziou P. Evidence-based practice educational intervention studies: a systematic review of what is taught and how it is measured. BMC Med Educ 2018;18(1):1-8 [FREE Full text] [doi: 10.1186/s12909-018-1284-1] [Medline: 30068343]
13. Sophonhiranrak S, Sakonnakron SPN. Limitations of mobile learning: a systematic review. 2017 Presented at: E-Learn: World Conference on e-learning in Corporate, Government, Healthcare, and Higher Education; October 17-20, 2017; Vancouver, BC p. 965-971 URL: https://www.learntechlib.org/p/181279
14. Lall P, Rees R, Law GCY, Dunleavy G, Cotič Ž, Car J. Influences on the implementation of mobile learning for medical and nursing education: qualitative systematic review by the digital health education collaboration. J Med Internet Res 2019;21(2):1-15 [FREE Full text] [doi: 10.2196/12895] [Medline: 30816847]
15. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology 2007;18(6):805-835 [FREE Full text] [doi: 10.1097/EDE.0b013e3181577511] [Medline: 18049195]
16. Johnson SG, Titlestad KB, Larun L, Ciliska D, Olsen NR. Experiences with using a mobile application for learning evidence-based practice in health and social care education: An interpretive descriptive study. PLoS One 2021;16(7):1-16 [FREE Full text] [doi: 10.1371/journal.pone.0254272] [Medline: 34252136]
17. EBPsteps. Western Norway University of Applied Sciences. 2015. URL: https://www.ebpsteps.no [accessed 2023-08-29]
18. Callander J, Anstey AV, Ingram JR, Limpens J, Flohr C, Spuls PI. How to write a Critically Appraised Topic: evidence to underpin routine clinical practice. Br J Dermatol 2017;177(4):1007-1013. [doi: 10.1111/bjd.15873] [Medline: 28967117]
19. Forskrift om felles rammeplan for helse- og sosialfagutdanninger [Regulations on a common framework for health- and social care training programmes]. Kunnskapsdepartementet [Ministry of Education]. 2019. URL: https://lovdata.no/dokument/SF/forskrift/2017-09-06-1353 [accessed 2023-09-01]
20. Studieplan - bachelor i ergoterapi [Study plan—bachelor in occupational therapy]. Høgskulen på Vestlandet. 2022. URL: https://www.hvl.no/studier/studieprogram/ergoterapi/2022h/studieplan/ [accessed 2023-09-01]
21. McEvoy MP, Williams MT, Olds TS. Evidence based practice profiles: differences among allied health professions. BMC Med Educ 2010;10:1-8 [FREE Full text] [doi: 10.1186/1472-6920-10-69] [Medline: 20937140]
22. Microsoft Excel. Microsoft Corporation. 2023. URL: https://microsoft.com [accessed 2023-05-01]
23. Johnson SG, Olsen NR. Replication data for: occupational therapy students' evidence-based practice skills as reported in a mobile app: a cross-sectional study. DataverseNO. 2023. URL: https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/ETCEOE [accessed 2023-09-01]
24. Gregg M, Datta S, Lorenz D. R package version 0.2.2. htestClust: Reweighted Marginal Hypothesis Tests for Clustered Data. URL: https://cran.r-project.org/web/packages/htestClust/index.html [accessed 2023-05-01]
25. IBM SPSS Statistics for Windows. 2021. URL: https://www.ibm.com/products/spss-statistics [accessed 2023-05-01]
26. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2018. URL: https://www.R-project.org/ [accessed 2023-05-01]
27. Davies KS. Formulating the evidence based practice question: A review of the frameworks. Evid Based Libr Inf Pract 2011;6(2):75-80. [doi: 10.18438/B8WS5N]
28. Speckman RA, Friedly JL. Asking structured, answerable clinical questions using the Population, Intervention/Comparator, Outcome (PICO) framework. PM R 2019 May;11(5):548-553. [doi: 10.1002/pmrj.12116] [Medline: 30729707]
29. DeCleene Huber K, Nichols A, Bowman K, Hershberger J, Marquis J, Murphy T, et al. The correlation between confidence and knowledge of Evidence-Based Practice among occupational therapy students. Open J Occup Ther 2015;3(1):1-19. [doi: 10.15453/2168-6408.1142]
30. Olsen NR, Bradley P, Lomborg K, Nortvedt MW. Evidence based practice in clinical physiotherapy education: a qualitative interpretive description. BMC Med Educ 2013;13:52 [FREE Full text] [doi: 10.1186/1472-6920-13-52] [Medline: 23578211]

XSL•FO
RenderX

31. Olsen NR, Lygren H, Espehaug B, Nortvedt MW, Bradley P, Bjordal JM. Evidence-based practice exposure and physiotherapy students' behaviour during clinical placements: a survey. Physiother Res Int 2014;19(4):238-247. [doi: 10.1002/pri.1590] [Medline: 24664886]
32. Crabtree JL, Justiss M, Swinehart S. Occupational therapy master-level students' evidence-based practice knowledge and skills before and after fieldwork. Occup Ther Health Care 2012;26(2-3):138-149. [doi: 10.3109/07380577.2012.694584] [Medline: 23899138]
33. Hitch D, Nicola-Richmond K. Instructional practices for evidence-based practice with pre-registration allied health students: a review of recent research and developments. Adv Health Sci Educ Theory Pract 2017;22(4):1031-1045. [doi: 10.1007/s10459-016-9702-9] [Medline: 27469244]
34. Lehane E, Leahy-Warren P, O'Riordan C, Savage E, Drennan J, O'Tuathaigh C, et al. Evidence-based practice education for healthcare professions: an expert view. BMJ Evid Based Med 2019;24(3):103-108 [FREE Full text] [doi: 10.1136/bmjebm-2018-111019] [Medline: 30442711]
35. Kumar BA, Mohite P. Usability of mobile learning applications: a systematic literature review. J Comput Educ 2017;5(1):1-17. [doi: 10.1007/s40692-017-0093-6]

Abbreviations

CAT: critically appraised topic
EBP: evidence-based practice
ECTS: European Credit Transfer System
HVL: Western Norway University of Applied Sciences
OT: occupational therapy
PBL: problem-based learning
PICO/PICo: Population, Intervention, Comparison, and Outcome/Population, Interest, and Context
RCT: randomized controlled trial
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by T de Azevedo Cardoso; submitted 26.04.23; peer-reviewed by G Kian Liang, M Johnson, M Stein, M Mostafa, M Gasmi ; comments to author 09.08.23; revised version received 18.09.23; accepted 29.01.24; published 21.02.24.

Please cite as:
Johnson SG, Espehaug B, Larun L, Ciliska D, Olsen NR
Occupational Therapy Students' Evidence-Based Practice Skills as Reported in a Mobile App: Cross-Sectional Study
JMIR Med Educ 2024;10:e48507
URL: https://mededu.jmir.org/2024/1/e48507
doi:10.2196/48507
PMID:38381475

©Susanne G Johnson, Birgitte Espehaug, Lillebeth Larun, Donna Ciliska, Nina Rydland Olsen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Students' Perceived Knowledge, Skills, Attitudes, and Experience

Lena Rettinger[1,2], BSc, MSc; Peter Putz[3], Mag, Dr Rer Nat; Lea Aichinger[1], BSc, MSc; Susanne Maria Javorszky[4], BSc, MSc; Klaus Widhalm[5], MSc; Veronika Ertelt-Bach[6], Mag, MAS; Andreas Huber[7], MSc; Sevan Sargis[8], BSc, MSc; Lukas Maul[1], BSc, MSc; Oliver Radinger[9], BA, Dr; Franz Werner[1], Mag, Dr Tech; Sebastian Kuhn[2], MME, Prof Dr

[1]Health Assisting Engineering, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[2]Institute of Digital Medicine, Philipps-University & University Hospital of Giessen and Marburg, Marburg, Germany
[3]Competence Center INDICATION, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[4]Logopedics – Phoniatrics - Audiology, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[5]Physiotherapy, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[6]Occupational Therapy, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[7]Orthoptics, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[8]Midwifery, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[9]Competence Center Nursing Sciences, FH Campus Wien, University of Applied Sciences, Vienna, Austria

Corresponding Author:
Lena Rettinger, BSc, MSc
Health Assisting Engineering
FH Campus Wien
University of Applied Sciences
Favoritenstrasse 226
Vienna, 1100
Austria
Phone: 43 1 606 68 77 ext 4382
Email: lena.rettinger@fh-campuswien.ac.at

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2024/1/e59919

Abstract

Background: The COVID-19 pandemic has highlighted the growing relevance of telehealth in health care. Assessing health care and nursing students' telehealth competencies is crucial for its successful integration into education and practice.

Objective: We aimed to assess students' perceived telehealth knowledge, skills, attitudes, and experiences. In addition, we aimed to examine students' preferences for telehealth content and teaching methods within their curricula.

Methods: We conducted a cross-sectional web-based study in May 2022. A project-specific questionnaire, developed and refined through iterative feedback and face-validity testing, addressed topics such as demographics, personal perceptions, and professional experience with telehealth and solicited input on potential telehealth course content. Statistical analyses were conducted on surveys with at least a 50% completion rate, including descriptive statistics of categorical variables, graphical representation of results, and Kruskal Wallis tests for central tendencies in subgroup analyses.

Results: A total of 261 students from 7 bachelor's and 4 master's health care and nursing programs participated in the study. Most students expressed interest in telehealth (180/261, 69% very or rather interested) and recognized its importance in their education (215/261, 82.4% very or rather important). However, most participants reported limited knowledge of telehealth applications concerning their profession (only 7/261, 2.7% stated profound knowledge) and limited active telehealth experience with various telehealth applications (between 18/261, 6.9% and 63/261, 24.1%). Statistically significant differences were found

XSL•FO
RenderX

between study programs regarding telehealth interest (*P*=.005), knowledge (*P*<.001), perceived importance in education (*P*<.001), and perceived relevance after the pandemic (*P*=.004). Practical training with devices, software, and apps and telehealth case examples with various patient groups were perceived as most important for integration in future curricula. Most students preferred both interdisciplinary and program-specific courses.

Conclusions: This study emphasizes the need to integrate telehealth into health care education curricula, as students state positive telehealth attitudes but seem to be not adequately prepared for its implementation. To optimally prepare future health professionals for the increasing role of telehealth in practice, the results of this study can be considered when designing telehealth curricula.

(JMIR Med Educ 2024;10:e51112) doi:10.2196/51112

KEYWORDS

telehealth; health care education; student perspectives; curriculum; interdisciplinary education

Introduction

Background

Telehealth has become increasingly important in recent years, particularly considering technological and societal developments. Telehealth is the use of information and communications technologies to deliver health services where there is a physical separation between care providers or recipients over both long and short distances [1]. It has the potential to help overcome barriers to accessing care, particularly in remote or underserved areas [2], and can be particularly beneficial for patients with chronic diseases [3] to improve long-term adherence [4], as well as addressing shortages in the health care workforce [5]. Owing to the COVID-19 pandemic, the integration of telehealth into the services of health care providers was further increased to prevent further infections and to serve patients in isolation [6-10].

However, the growing use has further highlighted the need for telehealth education for health care providers [11-16]. To successfully and sustainably implement telehealth and subsequently reap the benefits, it is necessary to integrate telehealth into the curricula of future health care providers [5]. A lack of knowledge and experience, as well as a lack of appropriate telehealth training, have been identified as major barriers to telehealth implementation among health care providers [17]. Conversely, telehealth education and training can increase the willingness to adopt telehealth, the perceived readiness, and confidence [5,13,18-20].

Providing telehealth services not only requires a basic understanding of telehealth and its applications but also an assortment of competencies spanning from theoretical knowledge to practical skills, closely mirroring the concepts of Miller pyramid of clinical competence [21] or its adapted version, the Miller prism [22]. As they outline, there are different levels of competence, such as knowledge, skills, and attitudes. In terms of telehealth competencies, *knowledge* involves the basic understanding of telehealth, its tools, and its applications. This also includes knowledge on how to ensure privacy and confidentiality [11,23-27]. The second competence level *skills* refers to the know-how. In telehealth, it requires health care professionals to organize and apply their knowledge to conduct physical assessments via telehealth, make perceptive observation-based examinations, and communicate effectively in a nontraditional clinical setting [5,10,11,15,16,19,23,24,26-29]. In the *performance* or *show* level, professionals demonstrate their ability to select, implement, and use appropriate telehealth tools in a simulated or controlled environment. This is where technological skills become crucial [11,23-27]. Finally, at the *action* or *does* level, health care professionals are expected to perform these skills in real-life situations, providing high-quality and safe telehealth services, and effectively incorporating ethical considerations into their practice [23-25]. *Attitude* is considered a vital component, along with knowledge, skills, and performance, that contributes to actual work competency. It refers to the behavioral and emotional aspects that influence how knowledge and skills are applied in practice [30]. *Attitude* can encompass elements such as motivation, ethical considerations, professionalism, and openness to learning, which seem to be important in the telehealth context.

In accordance with the principles of competency-based frameworks, curricula of health care study programs need to be adapted to qualify health care professionals at all levels of competency, increasing the probability that telehealth is effectively implemented in daily practice [11]. Two reviews [26,31] conducted in 2021 highlighted significant shortcomings in the training and curricula in allied health and nursing. They showed that there was a lack of consistency and absence of a systematic approach in integrating telehealth into these curricula [26,31]. Thus, it is crucial to design telehealth curricula with competency-based frameworks in mind to meet the diverse needs of students and ensure they are equipped with the necessary knowledge, performance skills, and attitude to effectively use telehealth technologies in their future health care practices. An increasing number of standards and guidelines are becoming available to guide the development of individual telehealth courses. They focus on various aspects such as administrative [32,33], ethical [32,34], clinical [32], technical [32,35], or soft skills [36]. However, they often do not address the specialized needs of allied health professionals [37]. Therefore, identifying the specific interests and learning needs of students can help educators to plan their teaching methods and provide tailored curricula or courses in individual study programs. This can further help to promote student engagement and motivation, ensure that the education is relevant and meaningful to their future professional practice, and ultimately improve learning outcomes.

XSL•FO
RenderX

Aim

The primary objective of this study was to assess the perceived telehealth knowledge, skills, attitude, and experience among health care professionals and nursing students to understand students' current self-assessed telehealth competencies and identify their learning needs. Our secondary objective was to evaluate students' preferences for telehealth content and teaching methods withing their respective curricula. This dual focus is intended to provide a rounded perspective of the students' perceived readiness for telehealth practice and to inform effective educational strategies.

Methods

Study Design

We conducted an anonymous cross-sectional web-based survey among the total population of selected health care profession students at *FH Campus Wien* (University of Applied Sciences). Reporting followed the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [38].

Sample Characteristics

Given the exploratory nature of this study, the sample size was not predetermined but was derived from the number of students enrolled in the targeted health care and nursing programs who were available and consented to participate during the survey period. At the time of the survey's release, 2273 students in the following selected academic health care professions were actively studying at the "FH Campus Wien" (University of Applied Sciences, Vienna, Austria) and were thus eligible to participate: BSc dietetics (DIE), BSc occupational therapy (OT), BSc health care and nursing (NUR), BSc midwifery (MID), BSc speech and language therapy (SLT), BSc orthoptics (ORT), BSc physiotherapy (PT), MSc health assisting engineering (HAE), MSc advanced nursing counseling (ANC), MSc advanced nursing education (ANE), and MSc advanced nursing practice (ANP). Participants who answered <50% of the questions were excluded.

Survey Administration

Students of all semesters were contacted directly with an email invitation. The survey was not listed publicly, no advertisement or incentive offers were put in place, and survey participation was voluntary. The survey was created using the web-based platform LimeSurvey (version 5.3.12 [39]), and it was open for participation between May 2 and May 30, 2022. An invitation email was sent on May 2, 2022. As a measure to improve the response rate, a first reminder was sent on May 9, 2022, and a second reminder was sent on May 25, 2022. Data were stored in a password-secured folder to which only selected study team members had access. Cookies were used to prevent users from accessing the survey twice, and IP addresses were not stored. No other measures to identify multiple entries were used. To ensure anonymous participation no registration process was put in place.

Ethical Considerations

Anonymous surveys currently do not require a formal review by a research ethics committee under Austrian research governance, in which the Declaration of Helsinki defines applicability to research on identifiable human data [40]. Exemption from ethical review has been formally confirmed by the Ethics Committee of the FH Campus Wien University of Applied Sciences (waiver no. W02/24). The survey followed ethical research practices (ie, voluntary participation; reassurance of anonymity, data protection, and confidentiality; advance information on purpose and content; provision of contact details of the research team; and full disclosure of involved organizations). This information was summarized on the first page of the web-based survey. Anonymous electronic consent to voluntary participation was required to begin the survey, but no signatures were obtained. All data processing procedures have been discussed in detail with the data protection officer of *FH Campus Wien* (University of Applied Sciences, Vienna). All data obtained in this survey will be stored for 10 years in compliance with national research legislation and the funding body.

Data Collection Methods

We used a newly developed, project-specific questionnaire (Multimedia Appendices 1 and 2). Feedback from project members on topics, constructs, and scales was iteratively incorporated into a first complete survey draft. Subsequent face-validity testing for usability and technical functionality was performed by 3 persons, not involved in the project, requiring minor usability and wording revisions. The survey consisted of 5 pages with 20 questions, of which 16 questions were mandatory: 6 demographical questions (including a question on the self-assessed information and communications technology competence to further describe the technology skills of the sample), 5 questions about personal perceptions of telehealth, 1 question on professional experience with telehealth, and 4 questions on potential content for telehealth courses or curriculum. The 4 optional questions were included to facilitate additional input or clarification.

Eligibility criteria were queried at the beginning of the survey: "Do you study at FH Campus Wien?" and "Which study program do you attend?" Respondents who clicked the survey link but were not eligible were taken directly to the end of the survey. Telehealth interest and perceived importance of telehealth in education were rated on a 4-point Likert scale (1=not interested/important, 2=less interested/important, 3=rather interested/important, and 4=very interested/important), and perceived relevance of telehealth after the pandemic was also rated on a 4-point Likert scale (1=for sure not, 2=rather not, 3=probably, and 4=for sure). Telehealth knowledge was rated by selecting 1 of 5 statements (1=I have never heard of telehealth, 2=I know the term but not more about it, 3=I know telehealth in medical services but not so much about it in my own profession, 4=I know some telehealth applications in my own profession, and 5=I know a lot of telehealth applications in my own profession). Experience with telehealth was rated among the options "performed," "observed," and "neither nor" for given examples. The perceived relevance of types of telehealth for the profession was assessed with multiple selections of given examples. Participants rated their interest in telehealth content on a 4-point Likert scale (1=for sure, 2=rather yes, 3=rather not, and 4=for sure not) for given

XSL•FO
RenderX

examples. The preferred setting for learning about telehealth was assessed using single choice selection. The option “Don’t know” was implemented, where applicable. Items were not randomized and were always presented in the same order to maintain the survey structure.

Statistical Analysis

Questionnaires with a completion rate of at least 50% were analyzed. Predefined subgroup analysis to compare for study programs, age, gender, and study year was undertaken. Descriptive statistics of categorical variables were reported as absolute and relative frequencies, and ordinal variables were reported with median. Histograms, heat maps, and boxplots were deployed for graphical illustration of the results. Boxplots display the first and third quartiles as a rectangular box, with whiskers extending from the box to indicate the minimum and maximum values, except for outliers. The median is depicted by a horizontal line. Outliers are represented by individual dots, whereas the mean is denoted using an “x” symbol. Stacked bar charts represent the frequencies of positive (right to 0) and negative responses (left to 0) for categorical variables with higher values in the middle. Kruskal Wallis tests were conducted to test for central tendencies in the subgroup analyses. The α value was set at .05, and exact *P* values were reported. The following mergers were made to achieve a minimum of 5 participants in each subgroup for Kruskal Wallis tests: semesters were merged into study years, for example, first and second bachelor’s semesters combined; age groups were assessed in 8 age group categories (<20, 21-25, 26-30,..., and >50 years) but combined for subgroup analysis into 3 generation groups. In the literature, generational affiliations vary among different publications [41,42]. In this study, Generation Z was defined as students aged up to 25 years, Generation Y encompassed students aged between 26 and 40 years, and students aged ≥41 years were grouped into Generation X and baby boomers. Furthermore, pairwise comparisons were conducted. Test statistic H, SE, standardized test statistic, unadjusted *P* values, and Bonferroni-adjusted *P* values were reported. The Bonferroni-adjusted statistical significance was summarized graphically using spider web figures. Pairwise comparisons were not conducted if the alternative hypothesis was rejected by the overall Kruskal Wallis test.

Results

Overview

A total of 2273 students of the selected academic health care professions were potentially eligible to participate. The link to the web-based survey was accessed by 281 students, of whom 261 (92.9%) completed the questionnaire (ie, answered at least 50% of the questions) and were therefore included in the analysis, resulting in a completion rate of 93%. Overall, 206 students were attending a bachelor’s degree program and 55 students were attending a master’s degree program (Table 1). The demographic characteristics of the survey participants are presented in Table 2.

Table 1. Participation across the selected bachelor’s and master’s programs (N=261).

Programs	Values, n (%)	Response rate (%)
Bachelor’s programs	206 (79)	9
Dietetics	20 (9.7)	36
Occupational therapy	24 (11.7)	24
Nursing	32 (15.5)	2
Midwifery	35 (17)	31
Speech and language therapy	25 (12.1)	37
Orthoptics	23 (11.2)	51
Physiotherapy	47 (22.8)	13
Master’s programs	55 (21)	32
Advanced nursing counseling	7 (13)	35
Advanced nursing education	16 (29)	27
Advanced nursing practice	13 (24)	28
Health assisting engineering	19 (34)	40

XSL•FO
RenderX

Table 2. Demographic characteristics of survey participants (N=261).

Characteristics	Total, n (%)	Bachelor’s (n=206), n (%)	Master’s (n=55), n (%)
Generation Z (years)	157 (60.15)	151 (73.3)	6 (10.9)
<20	33 (12.64)	33 (16.02)	0 (0)
21-25	124 (47.51)	118 (57.28)	6 (10.9)
Generation Y (years)	82 (31.42)	51 (24.76)	31 (56.36)
26-30	48 (18.39)	33 (16.02)	15 (27.27)
31-35	27 (10.34)	14 (6.8)	13 (23.64)
36-40	7 (2.68)	4 (1.94)	3 (5.45)
Generation X, baby boomers (years)	22 (8.43)	4 (1.94)	18 (32.73)
41-45	13 (4.98)	4 (1.94)	9 (16.36)
46-50	3 (1.15)	0 (0)	3 (5.45)
>50	6 (2.3)	0 (0)	6 (10.9)
Gender			
Man	228 (87.36)	184 (89.32)	44 (80)
Woman	30 (11.49)	19 (9.22)	11 (20)
Nonbinary	3 (1.15)	3 (1.46)	0 (0)
Semester			
BSc 1-2	103 (39.46)	103 (50)	N/A[a]
BSc 3-4	61 (23.37)	61 (29.62)	N/A
BSc 5-6	42 (16.09)	42 (20.39)	N/A
MSc 1-2	33 (12.64)	N/A	33 (60)
MSc 3-4	22 (8.43)	N/A	22 (40)
Self-assessed ICT[b] competence[c]			
1=very good	80 (30.65)	65 (31.55)	15 (27.27)
2=good	129 (49.43)	101 (49.03)	28 (50.91)
3=medium	50 (19.16)	38 (18.45)	12 (21.82)
4=sufficient	1 (0.38)	1 (0.49)	0 (0)
5=not sufficient	1 (0.38)	1 (0.49)	0 (0)

[a]N/A: not applicable.

[b]ICT: information and communications technology.

[c]Corresponding to the Austrian school grading system.

Subgroup Differences

A Kruskal Wallis H test (Table 3) showed that there was a statistically significant difference between the study programs in telehealth interest (*P*=.005), telehealth knowledge (*P*<.001), perceived importance of telehealth in education (*P*<.001), and perceived relevance of telehealth after the pandemic (*P*=.004). Corresponding box plots are shown in Figures 1-3. There were no significant differences between genders in telehealth interest (*P*=.63), telehealth knowledge (*P*=.19), perceived importance of telehealth in education (*P*=.73), and perceived relevance of telehealth after the pandemic (*P*=.55). On the basis of age and generation, there were significant differences in the perceived importance of telehealth education (*P*=.01) but no significant differences in telehealth interest (*P*=.14), telehealth knowledge (*P*=.19), and perceived relevance of telehealth after the pandemic (*P*=.06). There was a significant difference between students of different semesters in telehealth knowledge (*P*<.001) and perceived relevance of telehealth after the pandemic (*P*=.008) but not in telehealth interest (*P*=.09) and perceived importance of telehealth in education (*P*=.09). Details on pairwise comparisons between the different subgroups are described in Multimedia Appendix 3. For each item, smaller values indicate better (more positive) agreement. In summary, significant pairwise differences were observed mainly for the study programs, specifically when comparing the ratings regarding telehealth knowledge (HAE<ORT, HAE<DIE, HAE<ANE, HAE<MID, HAE<NUR, SLT<MID, SLT<NUR, OT<MID, OT<NUR, PT<MID, and PT<NUR), telehealth importance (HAE<PT, HAE<MID, and SLT<MID), and the postpandemic role of telehealth (ANP<PT, ANP<MID, and

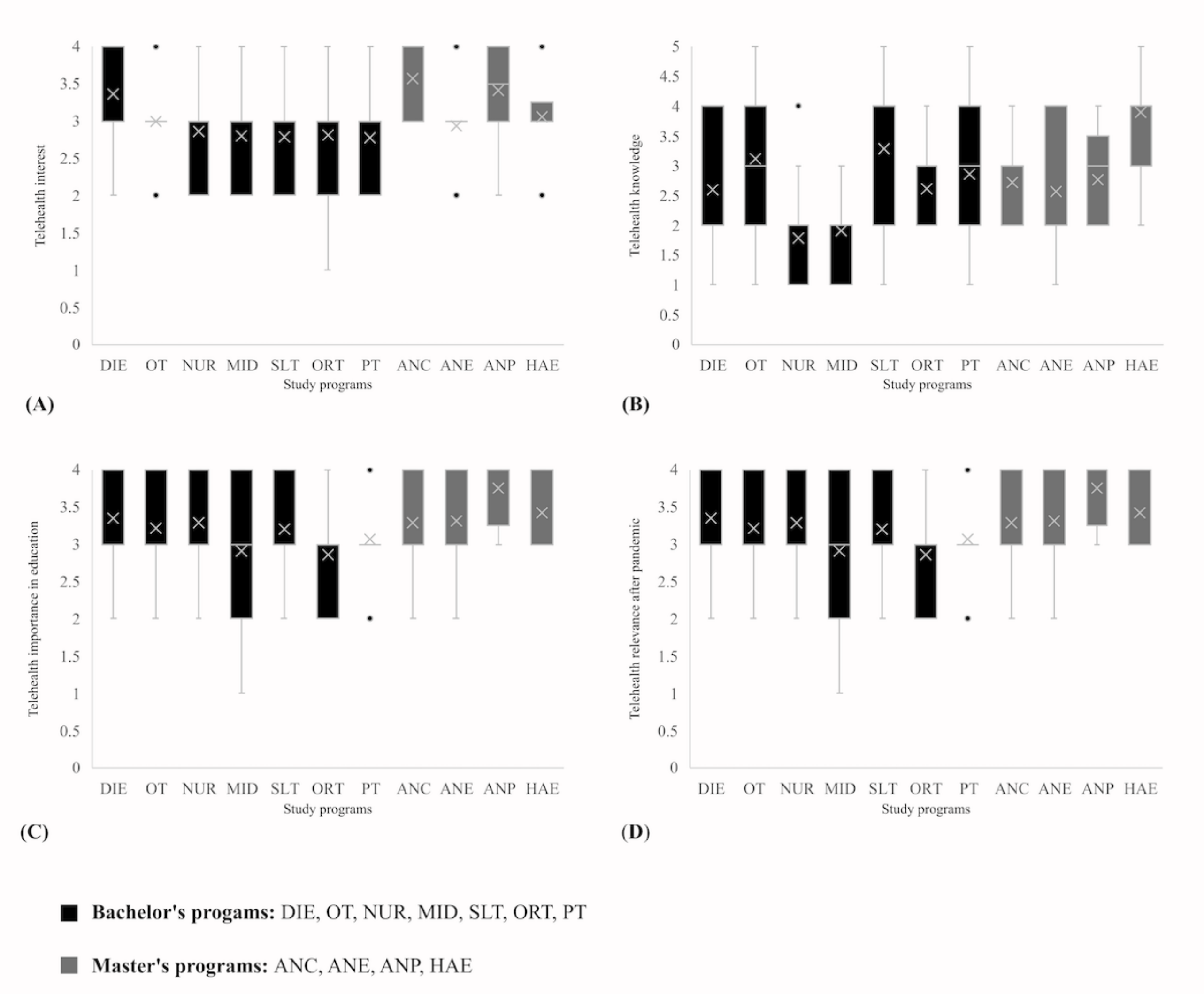
XSL•FO
RenderX

ANP<ORT). For gender, the null hypotheses were rejected by the overall Kruskal Wallis tests for all 4 domains, and thus, no subsequent pairwise comparisons were conducted. For generations, the only significant pairwise comparison was for the role after the pandemic, where Generation Z had more positive ratings than Generation Y. For study progress, the only significant pairwise comparison was for the role after the pandemic, where the first 2 master's semesters had more positive ratings than the fifth to sixth bachelor's semesters.

Table 3. Results of the Kruskal Wallis H test for each subgroup test.

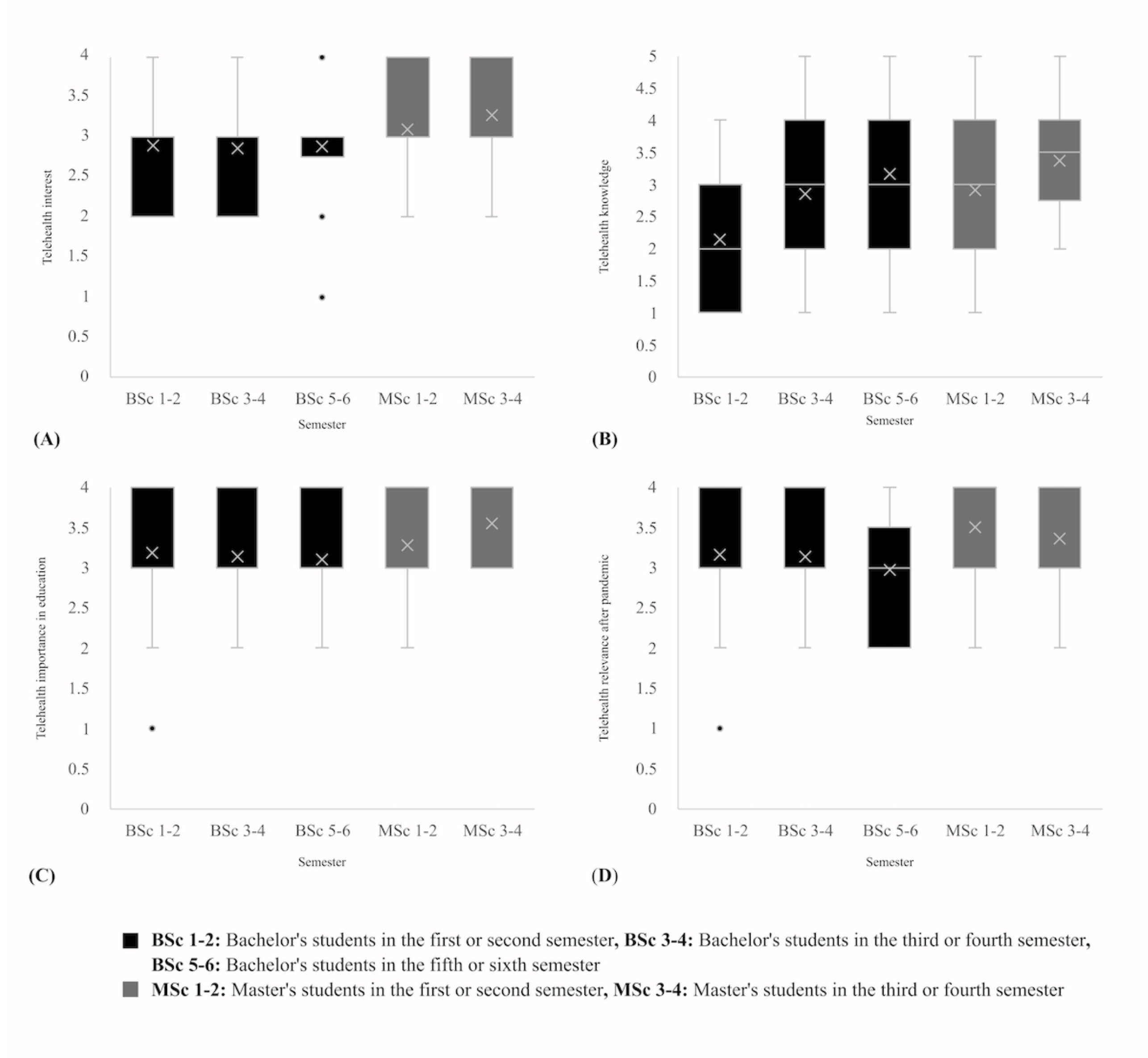
	Telehealth interest		Telehealth knowledge		Telehealth importance in education		Telehealth relevance after pandemic	
	Kruskal Wallis H test (*df*)	*P* value	Kruskal Wallis H test (*df*)	*P* value	Kruskal Wallis H test (*df*)	*P* value	Kruskal Wallis H test (*df*)	*P* value
Study programs	25.3 (10)	.005	70.6 (10)	<.001	33.0 (10)	<.001	25.8 (10)	.004
Genders	0.9 (2)	.63	3.3 (2)	.19	0.6 (2)	.73	1.2 (2)	.55
Age and generation	4.0 (2)	.14	3.3 (2)	.19	8.8 (2)	.01	5.5 (2)	.06
Semester	8.2 (4)	.09	43.1 (4)	<.001	8.0 (4)	.09	13.7 (4)	.008

Figure 1. Box plots of (A) telehealth interest, (B) telehealth knowledge, (C) perceived telehealth importance in education, and (D) perceived telehealth relevance after pandemic for each study program. Higher values represent higher interest, knowledge, importance, and perceived relevance. ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.



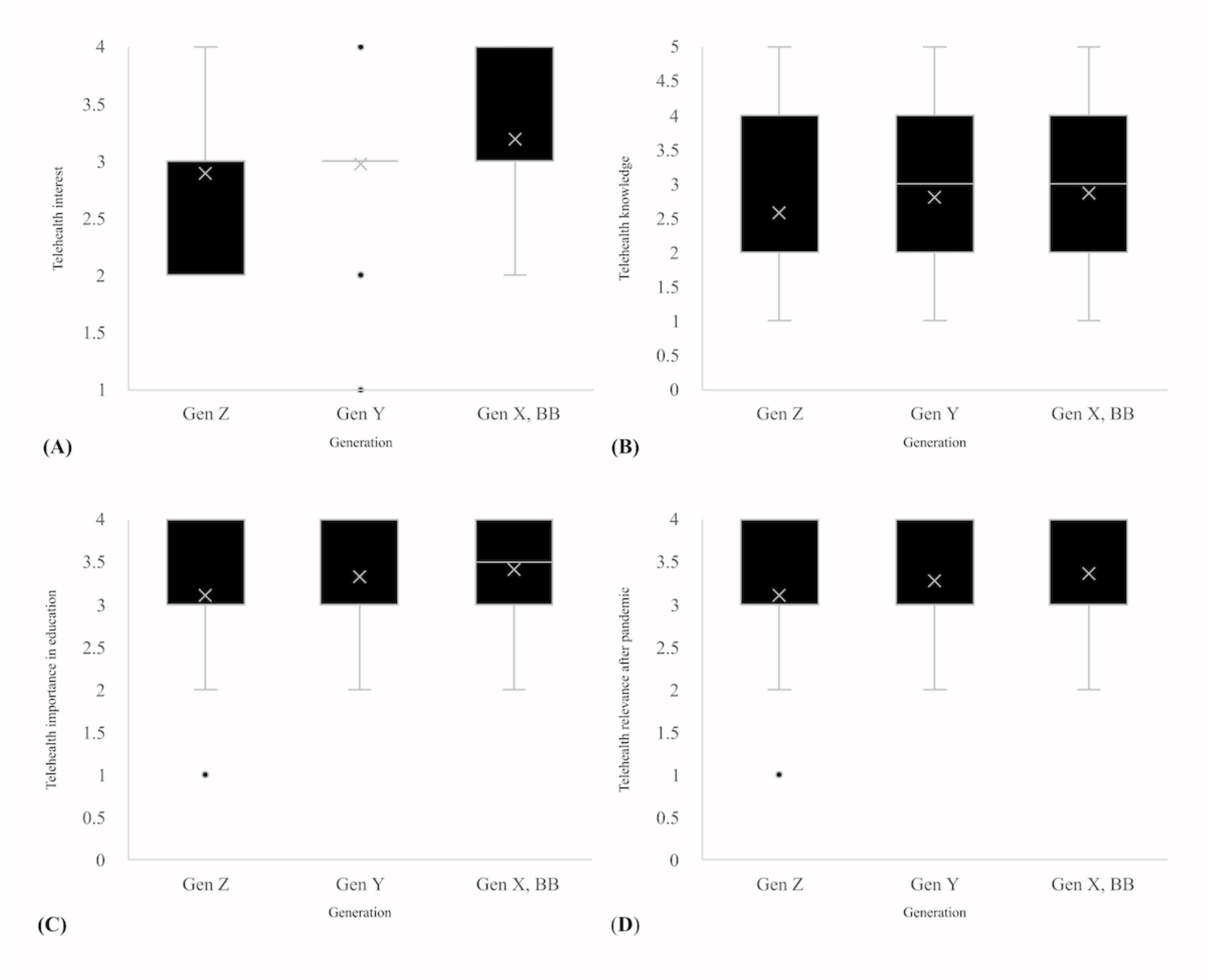
XSL•FO
RenderX

Figure 2. Box plots of (A) telehealth interest, (B) telehealth knowledge, (C) perceived telehealth importance in education, and (D) perceived telehealth relevance after the pandemic, based on semester. Higher values represent higher interest, knowledge, importance, and perceived relevance.



XSL•FO
RenderX

Figure 3. Box plots of (A) telehealth interest, (B) telehealth knowledge, (C) perceived telehealth importance in education, and (D) perceived telehealth relevance after the pandemic for different generations. Higher values represent higher interest, knowledge, importance, and perceived relevance. Gen Z (Generation Z): up to 25 years; Gen Y (Generation Y): 26-40 years; Gen X, BB (Generation X, baby boomer): ≥41 years.

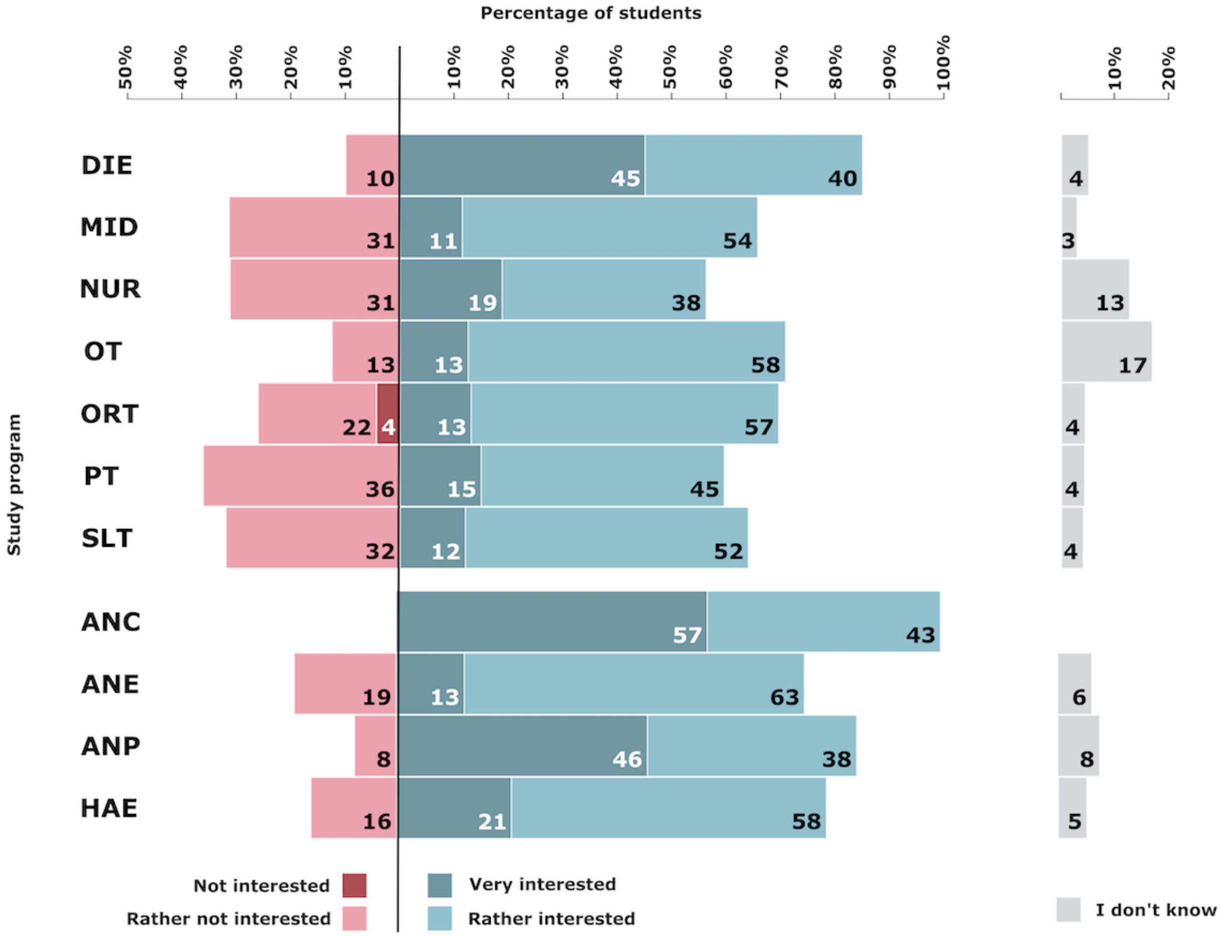


Telehealth Interest

Overall, 19.5% (51/261) of the students were very interested and 49.4% (129/261) of the students were rather interested in telehealth. Study programs with the highest interest ratings (very or rather interested) were ANC (7/7, 100%), DIE (17/20, 85%), and ANP (11/13, 84%). Moreover, 24.1% (63/261) of students were less interested and 0.4% (1/261) were not interested in telehealth. The study programs with the most uninterested students (rather not or not interested) were PT (17/47, 36%), SLT (8/25, 32%), MID (11/35, 31%), and NUR (10/32, 31%). The percentages by study program are presented in Figure 4.

XSL•FO
RenderX

Figure 4. Students' interest in telehealth based on the study program. ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.

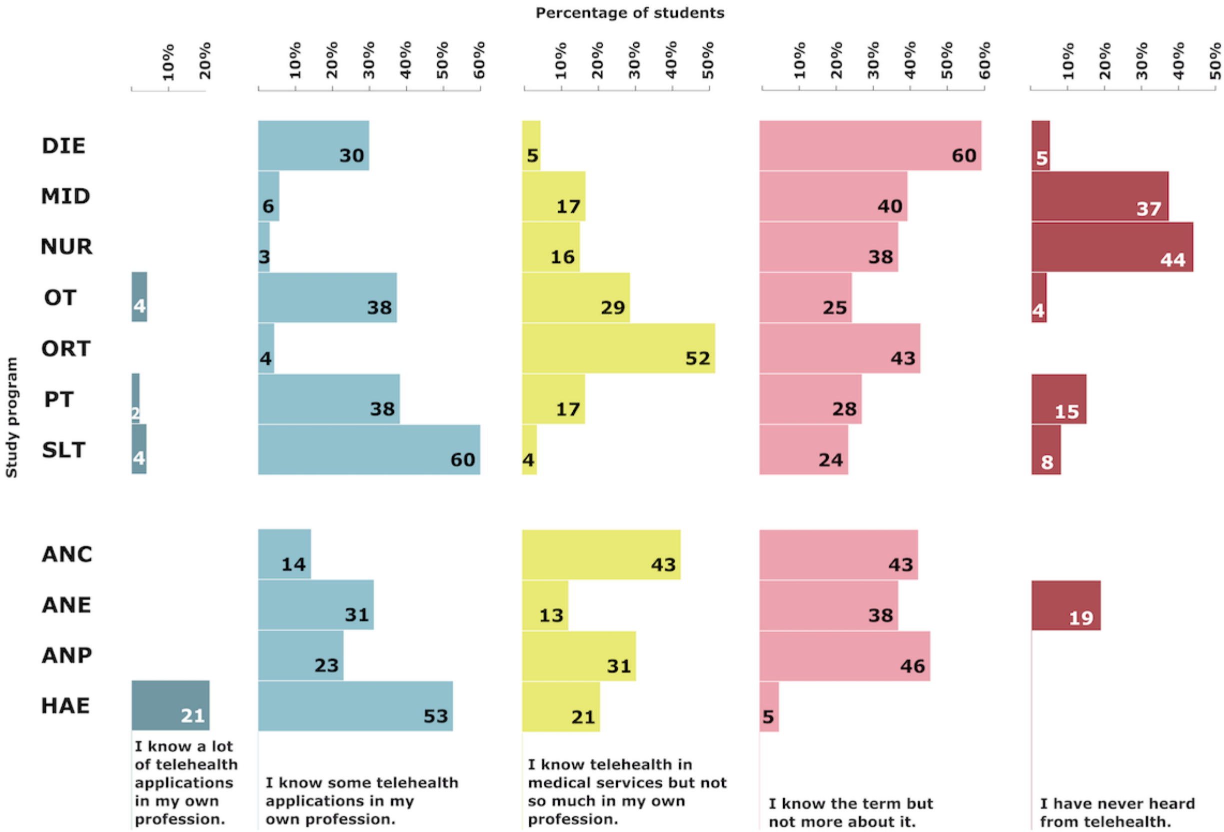


Telehealth Knowledge

Only 2.7% (7/261) of the students stated that they have already dealt intensively with telehealth in their own profession and that they knew a lot of applications, 27.2% (71/261) stated that they knew some telehealth applications in their own profession, 20.3% (53/261) stated that they knew telehealth in medical services but not in their own profession, 34.1% (89/261) stated that they knew the term but nothing more about it, and 15.7% (41/261) had never heard of telehealth. The percentages by study program are presented in Figure 5.

XSL•FO
RenderX

Figure 5. Students' self-assessed knowledge about telehealth based on the study program. ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.

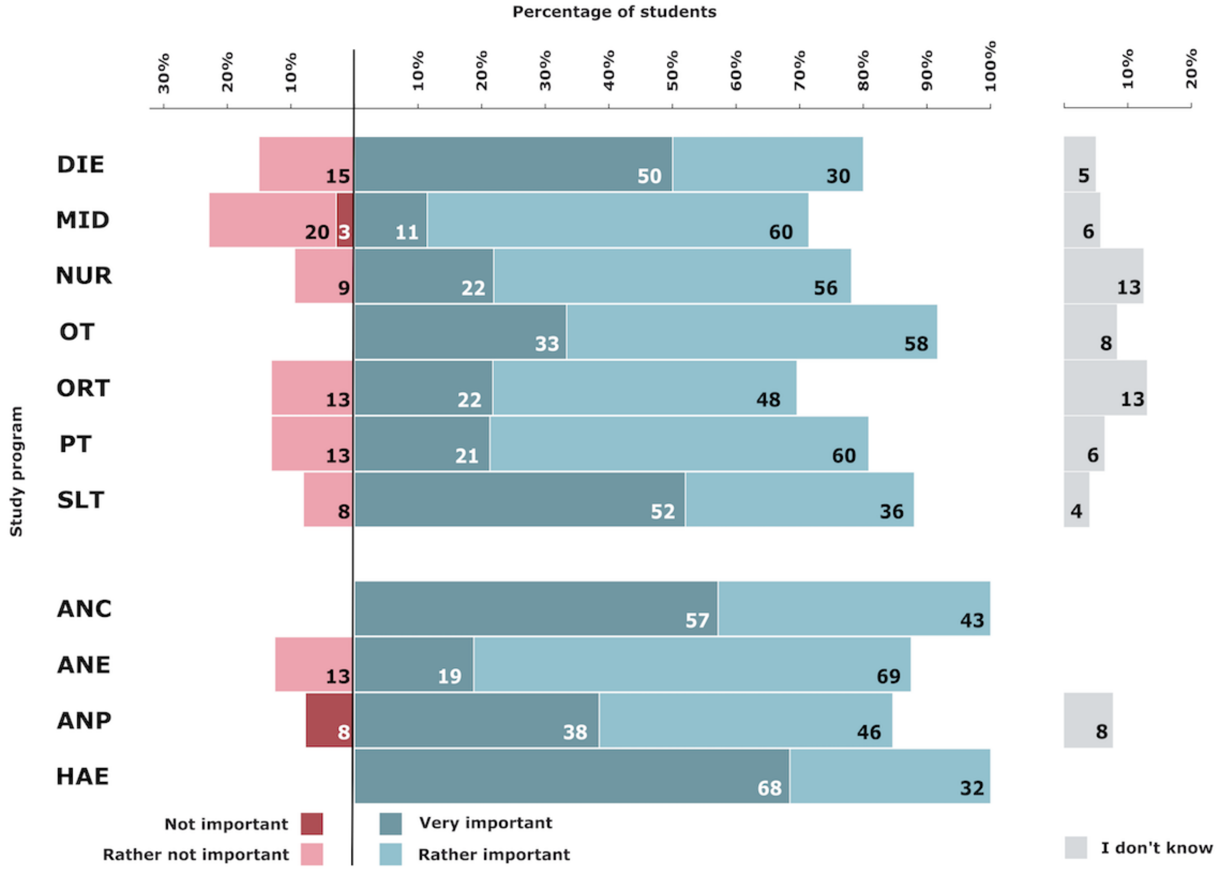


Telehealth Importance in Education

Overall, 31.4% (82/261) of the students thought telehealth was very important for their education, and 50.9% (133/261) of the students rated it as rather important. The study programs with the highest importance ratings (very or rather important combined) were ANC (7/7, 100%), HAE (19/19, 100%), and OT (22/24, 91%). Moreover, 10% (26/261) of the students thought it was rather not important, and 1.1% (3/261) thought it was not important. The highest percentages of unimportance ratings (not or rather not important) were in MID (8/35, 23%), DIE (3/20, 15%), and ORT (4/23, 13%), PT (3/47, 13%), and ANE (1/13, 13%). The percentages by study program are depicted in Figure 6.

XSL•FO
RenderX

Figure 6. Students' perceived importance of telehealth in education. ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.

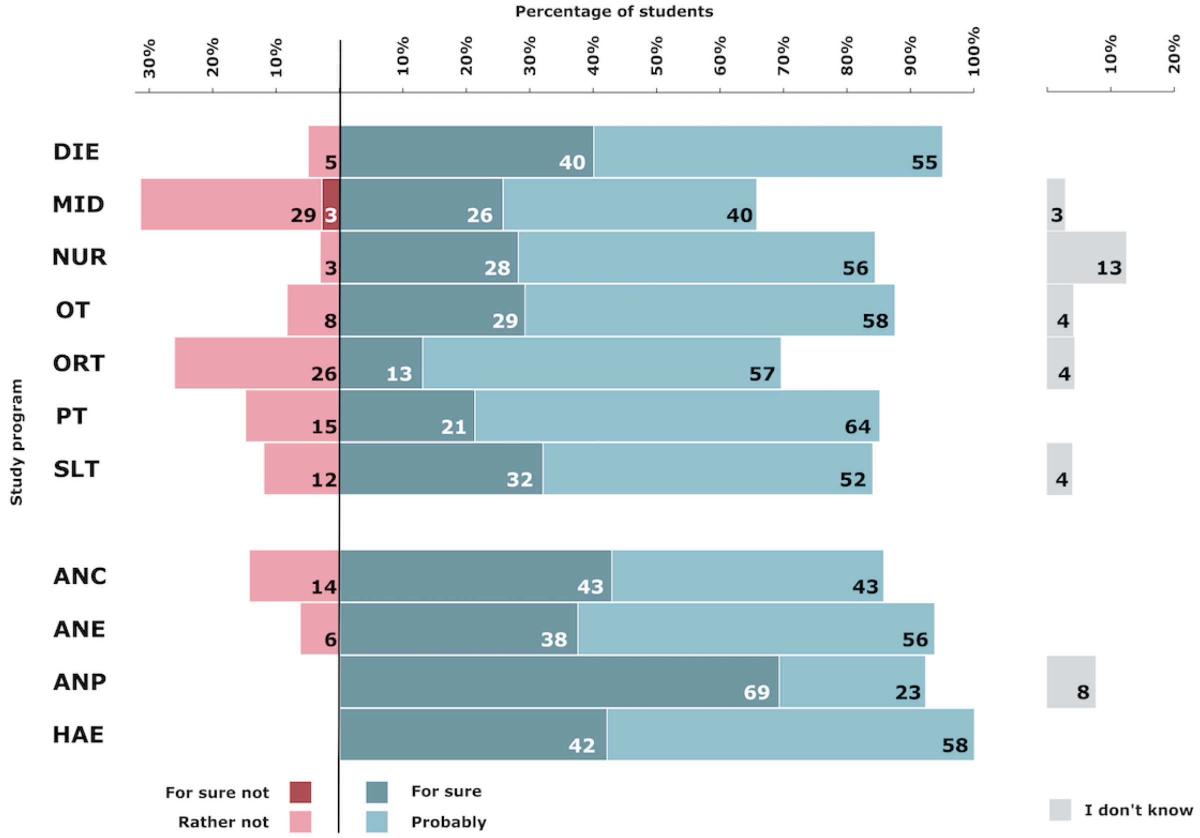


Telehealth Relevance After the Pandemic

Overall, 30.7% (80/261) of the students thought that telehealth will, for sure, be relevant in their profession after the pandemic, and 53.3% (139/261) of the students thought that telehealth would probably be relevant in their profession. The study programs that rated the future relevance of telehealth as highest were HAE (19/19, 100%), DIE (19/20, 95%), and ANE (15/16, 94%). Furthermore, 12.3% (32/261) of the students stated that it will rather not be relevant and 3.4% (9/261) stated that it will for sure not be relevant. The study programs that least anticipated a future relevance of telehealth were MID (11/35, 32%), ORT (6/23, 26%), and PT (7/47, 15%). The percentages by study program are depicted in Figure 7.

XSL•FO
RenderX

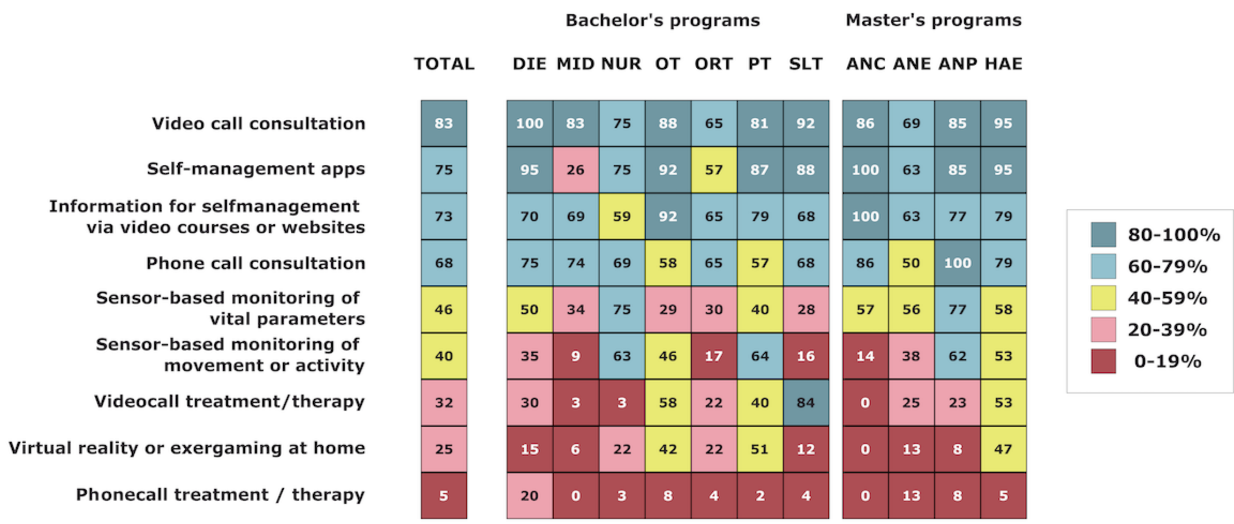
Figure 7. Students' perceived relevance of telehealth after the pandemic. ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.



Relevance of Different Forms of Telehealth Provision

The given relevance of different forms of telehealth provision in their own profession was confirmed as follows: video call consultation, 82.8% (216/261); apps for self-management, 75.1% (196/261); information for self-management via video courses or websites, 72.8% (190/261); phone call consultation, 68.2% (178/261); sensor-based monitoring of vital parameters, 46% (120/261); sensor-based monitoring of movement or activity, 39.8% (104/261); video call treatment or therapy, 32.2% (84/261); virtual reality or exergaming at home, 25.3% (66/261); and phone call treatment or therapy, 5.4% (14/261). The details of the study program are shown in Figure 8.

Figure 8. Students' perception of the relevance of different forms of telehealth concerning their own profession (the percentage of students that believes this telehealth form is relevant in their profession). ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.

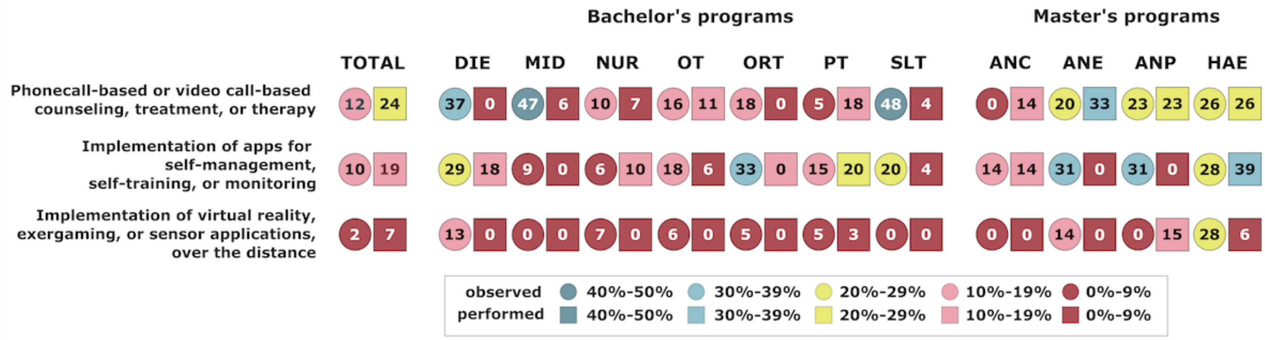


XSL•FO
RenderX

Telehealth Experience

Overall, 45.6% (119/261) of the students already had telehealth experience. Furthermore, 22.2% (58/261) of the students had observed and 10.7% (28/261) of the students had performed phone call–based or video call–based counseling, treatment, or therapy; 17.2% (45/261) of the students had observed and 9.2% (24/261) of the students had performed the implementation of apps for self-management and self-training of monitoring; and 5.7% (15/261) of the students had observed and 1.5% (4/261) of the students had performed the implementation of virtual reality, exergaming, or sensors over the distance. The details of the study program are shown in Figure 9.

Figure 9. Students' telehealth experience. The circles represent the percentage of students that have observed this form of telehealth, the squares represent the percentage of students that have performed this form of telehealth. ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.



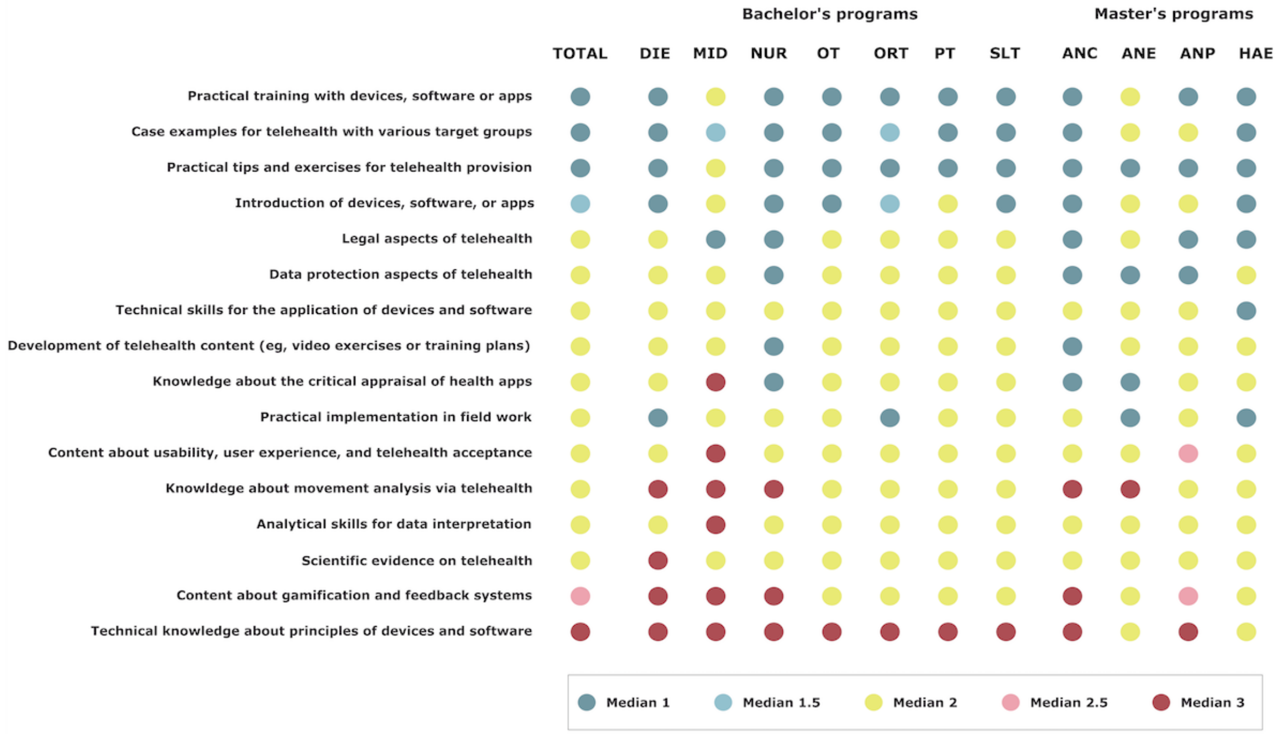
Telehealth Content Within the Curriculum

Students' preferences for telehealth content within their curriculum from highest to lowest ranking were practical training with devices, software, or apps (median 1), case examples for telehealth with various target groups (median 1), practical tips and exercises for telehealth provision (median 1), introduction of devices, software or apps (median 1.5), legal aspects of telehealth (median 2), data protection aspects of telehealth (median 2), technical skills for the application of devices and software (median 2), development of telehealth content (eg, video exercises or training plans; median 2), knowledge about the critical appraisal of health apps (median 2), practical implementation in field work (median 2), content about usability, user experience, and telehealth acceptance (median 2), knowledge about movement analysis via telehealth (median 2), analytical skills for data interpretation (median 2), scientific evidence on telehealth (median 2), content about gamification and feedback systems (median 2.5), and technical knowledge about principles of devices and software (median 3). Details by the study program are shown in Figure 10.

Overall, 21.8% (55/252) of the students preferred to learn about telehealth with students in their study program, 10% (25/252) preferred interdisciplinary courses, 60.7% (153/252) preferred both of them, 3.2% (8/252) did not want to learn about telehealth at all, and 4.4% (11/252) did not know. Furthermore, 30.6% (77/252) of the students wanted telehealth to be taught within required subjects, 62.3% (157/252) wanted telehealth to be taught within elective subjects, 1.6% (4/252) thought it should not be incorporated into the curriculum, and 5.6% (14/252) did not know. In bachelor's programs, 10.5% (20/190) of the students preferred the first or second semester, 60.5% (115/190) preferred the third or fourth semester, 26.8% (51/190) preferred the fifth or sixth semester, and 2.1% (4/190) of participants preferred none of them. Overall, 30% (13/44) of the master's students thought that the first or second semester and 68% (30/44) thought that the third or fourth semester would be most appropriate, and 1 (N=1, 2%) student felt that none was appropriate.

XSL•FO
RenderX

Figure 10. Students' preferences for telehealth content based on the study program. The circles represent the median of all answers with 1=for sure, 2=rather yes, 3=rather not, or 4=for sure not. ANC: advanced nursing counseling; ANE: advanced nursing education; ANP: advanced nursing practice; DIE: dietetics; HAE: health assisting engineering; MID: midwifery; NUR: health care and nursing; ORT: orthoptics; OT: occupational therapy; PT: physiotherapy; SLT: speech and language therapy.



Discussion

Overview

Our study provides novel insight into the telehealth knowledge, skills, attitudes, and experience of health care and nursing students and its potential integration into health care and nursing education and practice. The results suggest that there is substantial interest in telehealth among health care and nursing students but a lack of knowledge and experience with it. We discovered similarities and differences among various student groups, which will be discussed in detail and with regard to previously proposed telehealth competency frameworks for health care professionals.

Telehealth Interest

There was a generally high level of interest in telehealth across all study programs. The study programs with the highest median interest in telehealth were 3 master's programs (ANC, ANP, and HAE) and 1 bachelor's program (DIE). Interest in telehealth appears to be higher among master's students than among bachelor's students, possibly because of their advanced level of education and experience. Students in master's programs may have gained more professional working experience, which could have raised their awareness of the potential benefits of telehealth, such as increasing access to care [43], improving patient outcomes [44], and reducing health care costs [45]. Moreover, they might have encountered that telehealth has not yet become a ubiquitous component of the health care system. Furthermore, health care master's programs often place greater emphasis on leadership and innovation, which could make students more interested in exploring new methods [46]. Students in master's programs may be more focused on career advancement opportunities and recognize the potential of telehealth to create new roles or expand existing ones in the health care sector. This result also suggests that higher education may play an important role in promoting the adoption and use of telehealth in health care.

However, the students of the bachelor's programs displayed a substantial level of interest in telehealth, which remained consistent across various semesters and generations. The slight differences in interest between the study programs could be because of differences in their professions and the extent to which telehealth is currently integrated into their practices. For example, dietitians may have a stronger focus on counseling and patient education without manual or physical approaches compared with other professions, such as MID, NUR, PT, and OT. Although these professions also have an important educational role, their physical nature may make in-person consultations more essential for their professions, whereas for DIE, telehealth consultations may be a more practical and effective option. In addition, SLT has a strong focus on communication and telehealth, especially in the form of synchronous videoconferencing, and has been successfully used for several years, for example, in rural and remote areas of countries such as Australia and Canada [47].

The high level of interest in telehealth among health care students shows a positive attitude toward the technology, indicating that many perceive it as a beneficial tool in their future professional practice. However, enthusiasm alone is not

XSL•FO
RenderX

sufficient for effective telehealth practice; it must be supplemented by the right competencies as well as the ability to demonstrate successful performance in using telehealth tools and services.

Telehealth Knowledge and Experience

In general, most students showed some level of familiarity with telehealth, with only a small number of students reporting that they had never heard of it. However, there appears to be a profound lack of specific telehealth knowledge, profession-specific applications, and telehealth experience. Moreover, there were significant differences between the professions in terms of the level of knowledge about telehealth. For example, students from the bachelor's programs PT, OT, and SLT and the master's program HAE reported the highest levels of knowledge about telehealth applications in their professions. We can see that in some professions most students had never heard of telehealth or only knew the term but nothing more about it (applies to DIE, MID, NUR, and ANE). While in 3 of the 4 master's programs all students at least knew the term telehealth, this was not the case for 6 of the 7 bachelor's programs.

A relatively low percentage of students already had experience with telehealth applications. In Austria, health care professionals have worked mostly in face-to-face settings, but the COVID-19 pandemic had a significant influence on the attitudes toward telehealth service provision and its implementation [9,48]. Nevertheless, the findings of this study suggest that especially bachelor's students have only rarely come into touch with it. However, increased exposure to telehealth in academic settings and practical experience are important to enhance awareness and adoption of this emerging health care approach [49]. Only 35.4% (86/243) of all students reported direct or indirect experience with phone or video call–based counseling, treatment, or therapy; 29.5% (69/234) had implemented apps for self-management, self-training, or monitoring; and only 8.3% (19/229) had experience with the implementation of virtual reality, exergaming, or sensors over a distance. However, these percentages vary among different health care profession students. The highest percentage of students with experience in phone or video call–based counseling, treatment, or therapy were studying the bachelor's programs SLT, MID, and DIE and the master's programs ANE, ANP, and HAE. While master's students usually already have gained professional experience and thus determine their chosen methods themselves, the experience of bachelor's students mostly is limited to practical training within the curriculum and placements. Therefore, their experience with telehealth methods highly depends on their implementation by their teachers and supervisors. However, it remains unclear why some professions have had greater exposure to telehealth among their students than others. It is possible that the urgency of certain cases, particularly those related to acute therapy, led to greater adoption of telehealth in some professions. Another explanation could be that some professions, such as MID and DIE, might already have used telephone consultations more often before the pandemic than other professions. Students in the HAE program had the highest percentage of experience in implementing apps for self-management, self-training, or monitoring, and the implementation of virtual reality, exergaming, or sensors over a distance. This is not surprising as this program has a focus on health care technology and students might have an affinity for integrating more complex technologies into their practice.

Telehealth Attitudes

Most students in all health care programs expected that telehealth will play an important role in their profession also after the pandemic. This aligns with current research that supports the future relevance of telehealth in health care [5,50-52]. The overall high rates of expectations among health care students regarding the integral role of telehealth in the future of health care emphasizes the need for a stronger integration of telehealth education into health care curricula. Previous surveys in other countries also found a strong belief of students that telehealth services will be strongly integrated in the future [49,52]. Approximately 30% of the participants from MID (11/35) and ORT (6/23) form an exemption and believed that telehealth will not or rather not be relevant in their profession after the pandemic. In ORT, in particular, telehealth practices may not yet be as established as in other health care professions. Similar to other professions, orthoptists heavily rely on hands-on procedures, but they may require even more specialized equipment than others for the assessment and treatment of eye disorders.

Video call consultations seem to be the most widely accepted form of telehealth provision, with a high percentage of students from all study programs agreeing that they have an important role in their profession. Students may recognize the benefits of video integration for observations, capturing interpersonal features, nonverbal cues, and eye contact that may be lost when relying solely on phone calls [51]. Moreover, in recent years, videoconferencing has seen substantial growth across various aspects of daily life, driven by technological advancements, faster internet speeds, and the transition to remote work, particularly during the COVID-19 pandemic [53]. Hence, it is conceivable that students can best envision the use of video calls within the scope of telehealth, as they are familiar with this technology from other contexts.

Self-management apps are also widely accepted, with the exception of MID, which has noticeably lower scores. This is in line with previous research that reported that 58% of the surveyed nonphysicians (including MID) categorically rejected self-monitoring apps in pregnancy [54]. Self-management apps hold potential value for a wide range of contexts and stakeholders, including patients, health care professionals, and caregivers [55] and should therefore be incorporated into education [56]. In addition, the provision of information for self-management via video courses or websites was perceived as relevant by a majority of the students. Moreover, simple phone call consultations were perceived as more relevant compared with treatments or therapy over the phone, which received more skepticism among the surveyed students. This appears to indicate students' uncertainty regarding the feasibility or effectiveness of implementing specific interventions, techniques, or procedures through virtual means. Sensor-based monitoring of vital parameters is relatively well accepted among students in NUR and those in master's programs. Compared

XSL•FO
RenderX

with the other students examined, students in NUR likely already had the most frequent experience with monitoring vital parameters in their current roles and constitute a significant portion of patient care. Consequently, students may perceive a direct potential for alleviating their workload in their professional lives through remote monitoring technologies. Virtual reality or exergaming at home is not widely accepted among students, although it is most accepted among students in the OT, PT, and HAE programs. Exergaming interventions have demonstrated effectiveness in enhancing balance, function, physical activity levels, strength, fatigue, emotions, cognition, and pain relief [57]. Consequently, these interventions hold relevance for professionals and students in related fields.

There was a high level of agreement among students in all study programs that telehealth is important for their education. Only a minority of students of MID and ANP programs thought that this was not important. However, curricula for health care professionals have not yet widely incorporated telehealth [26] and are not consistent in their educational approaches [31]. However, health educators have started to recommend or plan to incorporate telehealth into the curriculum [58]. Furthermore, research is being conducted on which telehealth competencies should be implemented in education and with what didactic means [16,31,59-61].

Students expressed a strong desire for practical training that included hands-on experience with telehealth devices, software, and apps; case examples for telehealth with various target groups; and practical tips and exercises for telehealth provision. Previously published telehealth curricula had similarly presented a strong focus on practical experience [61]. As with any new technology or practice, students often benefit from experiential learning and simulation [62]. Therefore, it is reasonable for inexperienced health care students to prioritize practical training with devices, software, or apps; case examples for telehealth with various patient groups; and practical tips and exercises for telehealth provision. This is in line with a prior study reporting that new graduate physiotherapists perceived exposure to and practical skills training for telehealth as essential for their profession [27]. Another study with new graduate speech and language therapists concluded that they should learn to initiate telepractice service delivery through demonstration and role play to reduce initial anxieties [63].

The students' preference for both interdisciplinary and program-specific courses might be because telehealth is a complex and multidisciplinary field that requires a broad range of knowledge and skills [5,31] but still has profession-specific requirements and applications. Previous research has shown that students benefit from an interprofessional telehealth course [60]. The main reason for preferring electives could be that students want the flexibility to choose courses that align with their specific interests and career goals. Moreover, as mentioned by some students with additional comments, the curricula and timetables are already very intense and dense. Students might fear that the introduction of new content into the curriculum would come at the expense of other relevant study content. On the other hand, the preference for compulsory subjects by 30.6% (77/252) of the students could be because students feel that telehealth is an important topic that should be incorporated into the core curriculum of their program. Most bachelor's students had a preference for learning about telehealth in the third or fourth semester. Master's students also showed a slight preference for telehealth content in the second part of their education. This could be because students have already acquired a foundational knowledge of health care by this time and are better equipped to understand the complex nature of telehealth. In contrast, students in their first or second semester may be overwhelmed with this topic and may not have the necessary foundational knowledge to fully comprehend the nuances of telehealth.

Implications for Telehealth Education

Given the observed high interest and mainly positive attitude, but relatively low levels of perceived knowledge, and experience in telehealth, we conclude that it is important to enhance telehealth education for health care and nursing students. The apparent divide between perceived telehealth competence and importance of telehealth underscores the necessity that telehealth education should be integrated into the core curriculum, despite students having a preference for elective courses when directly asked. On the basis of the limited availability of publicly funded, profession-specific master's programs in Austria [64], we believe that it is important to integrate basic telehealth education at the bachelor's level to reach as many students as possible. However, this might not be applicable to countries with a different educational structure. Curricula should strategically incorporate the principles of Miller pyramid of clinical competence into telehealth education by emphasizing competency across the levels of knowledge, skills, performance, and action and by providing opportunities to form attitudes, as highlighted by other authors [16,60]. We suggest that there is a need for increased knowledge transfer, practical exposure, and training in the use of telehealth applications, especially in professions with lower levels of knowledge about telehealth, to increase their awareness and understanding of the potential benefits of telehealth, their specific skills, and therefore overall competency in their respective fields. It is further crucial to empower educators with the necessary competencies to effectively teach telehealth and to provide organizational framework conditions to integrate telehealth into the curricula [65].

As the students preferred to learn about case examples and hands-on experience with devices, software, and apps used in telehealth, we suggest that they early on can become more familiar and comfortable with using them. Case examples for telehealth with various target groups can help students understand the diverse needs of different patient populations and learn how to adapt their approach accordingly. Practical tips and exercises for telehealth provision can also help students to develop skills and confidence in their ability to provide telehealth services; improve their overall competency; and understand the ethical, clinical, and legal aspects that arise when using them. Furthermore, courses should expand on the essential knowledge details of legal aspects, data protection, technical skills, critical appraisal, and scientific evidence based on or in combination with practical examples. Even if these aspects did not rank highest in the needs analysis, students confirmed their relevance. Students need to build knowledge about the legal

XSL•FO
RenderX

framework in which they will operate, the importance of protecting patient data and how to maintain data privacy, and the potential risks and liabilities involved. As telehealth relies heavily on technology, students need to have the technical skills to use and troubleshoot various telehealth tools and platforms [66]. Furthermore, students need to be able to apply clinical reasoning in a telehealth context and critically appraise the scientific evidence on telehealth, including its benefits and limitations, to make informed decisions about its use [67].

In terms of content, we conclude that future telehealth curricula should focus on teaching the basics and the application of practical training on consultation over the phone with or without video integration, the integration of self-management apps, and the development or integration of video courses or websites for self-management within all study programs. This focus has been previously suggested for nurse practitioner training [5]. Furthermore, specific courses for therapeutical professions (SLT, OT, and PT) could teach the possibilities of direct therapy approaches through video calls and further exergaming and virtual reality. Sensor-based monitoring of vital parameters, movement, and activity might be more appropriate for NUR, PT, and OT students, within specialization courses for students of other bachelor's programs that are interested in this topic, and for master's programs. Guidelines that are specific to each profession and report on implementation, financial, and technical considerations [68] should also be integrated into the development of curricula. For instance, incorporating strategies for executing telehealth practices in fields such as OT [69], musculoskeletal physiotherapy [70], SLT [71], and nursing [72] can be beneficial.

Limitations

This study has several limitations that impact the interpretation of the results. The sample size of 261, representing a small segment of eligible health care students, and the overall low (11%) and variable response rate across programs, may affect the results' generalizability and comparability and raises the possibility of nonresponse bias, whereby the views of those who did not participate may systematically differ from those who did. This issue poses the risk of over- or underestimation of the true distribution of perceived telehealth competencies in the target population. In addition, the cross-sectional design, capturing attitudes at a single point in time, further limits the findings. The generation distribution differed between bachelor's and master's programs, which may confound the perceived importance of telehealth education, knowledge levels, and postpandemic telehealth relevance across generations. Although a large portion (151/206, 73.3%) of students in bachelor's programs belonged to Generation Z, master's programs had a higher representation of Generation Y, Generation X, and baby boomers (49/55, 89%). Therefore, the statistical differences in the perceived importance of telehealth education between generations and differences in telehealth knowledge and perceived relevance of telehealth after the pandemic must be interpreted with caution. It should also be noted that health professionals pursuing a master's degree later in their careers might show more interest in innovation, making these results less generalizable to other health professionals of the same generations. Regarding the statistical analysis, it should be mentioned that multiple comparisons increase the risk of type I errors. Even with the Bonferroni adjustment, which is conservative, there is a tradeoff with statistical power, potentially leading to type II errors [73]. In addition, self-reported measures of telehealth interest and knowledge may be influenced by social desirability bias or inaccurate self-assessment. Furthermore, it was not possible to directly assess student's telehealth skills and actual performance using a web-based survey. In addition, the study's context, focused on students from specific health care programs in 1 Austrian university, restricts the applicability of the findings to other institutions or countries. Finally, a limitation of this study is the potential impact of the COVID-19 pandemic on the participants' attitudes, experiences, and perspectives toward telehealth. The students who participated in this study in May 2022 were probably affected by the pandemic in various ways, including disruptions in their placements and the rapid adoption of telehealth services in health care settings. As a result, their views on telehealth might be influenced by the unique circumstances of the pandemic, which could limit the generalizability of the findings to other periods.

Recommendations for Further Research

Future research should consider several steps to build on this study. First, expanding the study to include a larger, more diverse sample of health care students from different institutions and countries will allow for examining potential variations in knowledge, skills, performance, action, and attitudes in telehealth. In addition, exploring factors that may act as barriers or facilitators to the adoption of telehealth within health care education, such as the interest, skills, and knowledge of educators, technological infrastructure, legal and ethical considerations, or institutional barriers, is crucial. Second, conducting more intervention-based studies that aim at improving telehealth knowledge, competence, and interest among health care students [74,75] will be valuable for investigating the effectiveness of different teaching methods and content that can help identify the most effective strategies for telehealth education. Moreover, conducting longitudinal research would enable tracking changes in students' attitudes, knowledge, and interest in telehealth over time as they progress through their education, providing a comprehensive understanding of the development and potential factors influencing these perspectives, especially in the time after the COVID-19 pandemic. Assessing the impact of telehealth training on clinical practice is important. Investigating the relationship between telehealth training during health care education and its application in clinical practice, as well as evaluating the impact of telehealth knowledge and competence on patient outcomes and health care delivery, can provide valuable insights. Finally, examining the role of interprofessional collaboration in telehealth education and practice and its impact on students' attitudes and knowledge regarding telehealth is essential [76]. Evaluating the effectiveness of interdisciplinary courses in fostering collaboration and improving telehealth competence among health care students can contribute to the development of more efficient telehealth education strategies.

XSL•FO
RenderX

Conclusions

Our study findings underscore the need for structured telehealth education within health care curricula to equip students with the necessary competencies for future practice. Students recognize the importance of telehealth in their future profession and feel that they need to be adequately prepared. However, the study also revealed that the level of telehealth experience and knowledge among participating health care students is currently low. Therefore, there is an urgent need to provide comprehensive telehealth education and training to health care students to prepare them for the future demands in their profession. By incorporating telehealth education into health care curricula, institutions can better prepare students for the evolving landscape of health care and promote the successful integration of telehealth into future practice.

Acknowledgments

The authors would like to thank the students who participated in the web-based survey and Inkscape for its open-source vector graphics editor, Draw Freely.

This work was supported by the City of Vienna, Magistratsabteilung 23, Austria, under grant MA23-338474-2021-2.

Data Availability

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

Authors' Contributions

LR was involved in project administration; LR and LA with the support from all authors conceptualized the study; LR and PP finalized the methodology; LR did data curation; LR and PP were involved in formal analysis and investigation; LR prepared the original draft; all authors reviewed and edited the draft; LR and FW acquired the funding; and SK supervised the study.

Conflicts of Interest

SK is the founder and shareholder of MED.digital. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Original questionnaire (German version).
[PDF File (Adobe PDF File), 103 KB - mededu_v10i1e51112_app1.pdf]

Multimedia Appendix 2
Translated questionnaire (English version).
[PDF File (Adobe PDF File), 232 KB - mededu_v10i1e51112_app2.pdf]

Multimedia Appendix 3
Pairwise comparisons.
[PDF File (Adobe PDF File), 657 KB - mededu_v10i1e51112_app3.pdf]

References

1. Gogia S. Fundamentals of Telemedicine and Telehealth. San Diego, CA: Academic Press; 2019.
2. Orlando JF, Beard M, Kumar S. Systematic review of patient and caregivers' satisfaction with telehealth videoconferencing as a mode of service delivery in managing patients' health. PLoS One 2019;14(8):e0221848 [FREE Full text] [doi: 10.1371/journal.pone.0221848] [Medline: 31469865]
3. Seron P, Oliveros MJ, Gutierrez-Arias R, Fuentes-Aspe R, Torres-Castro RC, Merino-Osorio C, et al. Effectiveness of telerehabilitation in physical therapy: a rapid overview. Phys Ther 2021 Jun 01;101(6):pzab053 [FREE Full text] [doi: 10.1093/ptj/pzab053] [Medline: 33561280]
4. Cinthuja P, Krishnamoorthy N, Shivapatham G. Effective interventions to improve long-term physiotherapy exercise adherence among patients with lower limb osteoarthritis. A systematic review. BMC Musculoskelet Disord 2022 Feb 14;23(1):147 [FREE Full text] [doi: 10.1186/s12891-022-05050-0] [Medline: 35164714]
5. Rutledge C, Kott K, Schweickert P, Poston R, Fowler C, Haney T. Telehealth and eHealth in nurse practitioner training: current perspectives. Adv Med Educ Pract 2017;8:399-409 [FREE Full text] [doi: 10.2147/AMEP.S116071] [Medline: 28721113]
6. Aggarwal K, Patel R, Ravi R. Uptake of telepractice among speech-language therapists following COVID-19 pandemic in India. Speech Lang Hear 2020 Oct 06;24(4):228-234. [doi: 10.1080/2050571x.2020.1812034]

XSL•FO
RenderX

7. Greenwood J, Fragala-Pinkham M, Dakhlian MG, Brennan E, Ploski C, Correia A. A pediatric hospital physical therapy and occupational therapy department's response to COVID-19: an administrative case report. Phys Ther 2021 Sep 01;101(9):pzab164 [FREE Full text] [doi: 10.1093/ptj/pzab164] [Medline: 34174072]
8. Heiskanen T, Rinne H, Miettinen S, Salminen AL. Uptake of tele-rehabilitation in Finland amongst rehabilitation professionals during the COVID-19 pandemic. Int J Environ Res Public Health 2021 Apr 20;18(8):4383 [FREE Full text] [doi: 10.3390/ijerph18084383] [Medline: 33924234]
9. Rettinger L, Klupper C, Werner F, Putz P. Changing attitudes towards teletherapy in Austrian therapists during the COVID-19 pandemic. J Telemed Telecare 2021 Jan 11;29(5):406-414. [doi: 10.1177/1357633x20986038]
10. Wijesooriya NR, Mishra V, Brand PL, Rubin BK. COVID-19 and telehealth, education, and research adaptations. Paediatr Respir Rev 2020 Sep;35:38-42 [FREE Full text] [doi: 10.1016/j.prrv.2020.06.009] [Medline: 32653468]
11. Adams JE, Ecker DJ. Telehealth: from the abstract to necessity to competency. FASEB Bioadv 2021 Jul 03;3(7):475-481 [FREE Full text] [doi: 10.1096/fba.2020-00098] [Medline: 33821234]
12. Chike-Harris KE, Harmon E, van Ravenstein K. Graduate nursing telehealth education: assessment of a one-day immersion approach. Nurs Educ Perspect 2020;41(5):E35-E36. [doi: 10.1097/01.NEP.0000000000000526] [Medline: 31232882]
13. Jones SE, Campbell PK, Kimp AJ, Bennell K, Foster NE, Russell T, et al. Evaluation of a novel e-learning program for physiotherapists to manage knee osteoarthritis via telehealth: qualitative study nested in the PEAK (physiotherapy exercise and physical activity for knee osteoarthritis) randomized controlled trial. J Med Internet Res 2021 Apr 30;23(4):e25872 [FREE Full text] [doi: 10.2196/25872] [Medline: 33929326]
14. Lowe JT, Patel SR, Hao WD, Butt A, Strehlow M, Lindquist B. Teaching from afar: development of a telemedicine curriculum for healthcare workers in global settings. Cureus 2021 Dec;13(12):e20123 [FREE Full text] [doi: 10.7759/cureus.20123] [Medline: 35003963]
15. Mahabamunuge J, Farmer L, Pessolano J, Lakhi N. Implementation and assessment of a novel telehealth education curriculum for undergraduate medical students. J Adv Med Educ Prof 2021 Jul;9(3):127-135 [FREE Full text] [doi: 10.30476/jamp.2021.89447.1375] [Medline: 34277843]
16. Rutledge CM, O'Rourke J, Mason AM, Chike-Harris K, Behnke L, Melhado L, et al. Telehealth competencies for nursing education and practice: the Four P's of telehealth. Nurse Educ 2021;46(5):300-305 [FREE Full text] [doi: 10.1097/NNE.0000000000000988] [Medline: 33481494]
17. Rettinger L, Kuhn S. Barriers to video call-based telehealth in allied health professions and nursing: scoping review and mapping process. J Med Internet Res 2023 Aug 01;25:e46715 [FREE Full text] [doi: 10.2196/46715] [Medline: 37526957]
18. Gunner CK, Eisner E, Watson AJ, Duncan JL. Teaching webside manner: development and initial evaluation of a video consultation skills training module for undergraduate medical students. Med Educ Online 2021 Dec;26(1):1954492 [FREE Full text] [doi: 10.1080/10872981.2021.1954492] [Medline: 34313579]
19. Newcomb AB, Duval M, Bachman SL, Mohess D, Dort J, Kapadia MR. Building rapport and earning the surgical patient's trust in the era of social distancing: teaching patient-centered communication during video conference encounters to medical students. J Surg Educ 2021;78(1):336-341 [FREE Full text] [doi: 10.1016/j.jsurg.2020.06.018] [Medline: 32709566]
20. O'Shea MC, Reeves NE, Bialocerkowski A, Cardell E. Using simulation-based learning to provide interprofessional education in diabetes to nutrition and dietetics and exercise physiology students through telehealth. Adv Simul (Lond) 2019 Dec 20;4(Suppl 1):28 [FREE Full text] [doi: 10.1186/s41077-019-0116-7] [Medline: 31890319]
21. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67 [FREE Full text] [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
22. De Kare-Silver N, Mehay R. Assessment and competence. In: Mehay R, editor. The Essential Handbook for GP Training and Education. New York, NY: Radcliffe Publishing; 2012.
23. Arends R, Gibson N, Marckstadt S, Britson V, Nissen MK, Voss J. Enhancing the nurse practitioner curriculum to improve telehealth competency. J Am Assoc Nurse Pract 2021 May;33(5):391-397. [doi: 10.1097/jxx.0000000000000303]
24. Guenther J, Branham S, Calloway S, Hilliard W, Jimenez R, Merrill E. Five steps to integrating telehealth into APRN curricula. J Nurse Pract 2021 Mar;17(3):322-325 [FREE Full text] [doi: 10.1016/j.nurpra.2020.12.004] [Medline: 33746648]
25. Hilty DM, Maheu MM, Drude KP, Hertlein KM. The need to implement and evaluate telehealth competency frameworks to ensure quality care across behavioral health professions. Acad Psychiatry 2018 Dec 13;42(6):818-824. [doi: 10.1007/s40596-018-0992-5] [Medline: 30426453]
26. Hui KY, Haines C, Bammann S, Hallandal M, Langone N, Williams C, et al. To what extent is telehealth reported to be incorporated into undergraduate and postgraduate allied health curricula: a scoping review. PLoS One 2021 Aug 19;16(8):e0256425 [FREE Full text] [doi: 10.1371/journal.pone.0256425] [Medline: 34411171]
27. Martin R, Mandrusiak A, Russell T, Forbes R. New-graduate physiotherapists' training needs and readiness for telehealth. Physiother Theory Pract 2022 Nov;38(13):2788-2797. [doi: 10.1080/09593985.2021.1955423] [Medline: 34282699]
28. Bouamra B, Chakroun K, Medeiros De Bustos E, Dobson J, Rouge JA, Moulin T. Simulation-based teaching of telemedicine for future users of teleconsultation and tele-expertise: feasibility study. JMIR Med Educ 2021 Dec 22;7(4):e30440 [FREE Full text] [doi: 10.2196/30440] [Medline: 34941553]

XSL•FO
RenderX

29. Camden C, Silva M. Pediatric teleheath: opportunities created by the COVID-19 and suggestions to sustain its use to support families of children with disabilities. Phys Occup Ther Pediatr 2021 Oct 06;41(1):1-17. [doi: 10.1080/01942638.2020.1825032] [Medline: 33023352]
30. Martin-Sanchez F, Lázaro M, López-Otín C, Andreu AL, Cigudosa JC, Garcia-Barbero M. Personalized precision medicine for health care professionals: development of a competency framework. JMIR Med Educ 2023 Feb 07;9:e43656 [FREE Full text] [doi: 10.2196/43656] [Medline: 36749626]
31. Chike-Harris KE, Durham C, Logan A, Smith G, DuBose-Morris R. Integration of telehealth education into the health care provider curriculum: a review. Telemed J E Health 2021 Feb;27(2):137-149. [doi: 10.1089/tmj.2019.0261] [Medline: 32250196]
32. Richmond T, Peterson C, Cason J, Billings M, Terrell EA, Lee AC, et al. American telemedicine association's principles for delivering telerehabilitation services. Int J Telerehabil 2017 Nov 20;9(2):63-68 [FREE Full text] [doi: 10.5195/ijt.2017.6232] [Medline: 29238450]
33. Telehealth implementation playbook. American Medical Association. 2022. URL: https://www.ama-assn.org/system/files/ama-telehealth-playbook.pdf [accessed 2023-07-19]
34. Pollard JS, Karimi KA, Ficcaglia MB. Ethical considerations in the design and implementation of a telehealth service delivery model. Behav Anal Res Pract 2017 Nov;17(4):298-311. [doi: 10.1037/bar0000053]
35. WHO-ITU global standard for accessibility of telehealth services. World Health Organization and International Telecommunication Union. 2022. URL: https://www.who.int/publications/i/item/9789240050464 [accessed 2023-07-19]
36. Gustin TS, Kott K, Rutledge CM. Telehealth etiquette training: a guideline for preparing interprofessional teams for successful encounters. Nurse Educ 2020;45(2):88-92. [doi: 10.1097/NNE.0000000000000680] [Medline: 31022072]
37. Anil K, Freeman JA, Buckingham S, Demain S, Gunn H, Jones RB, et al. Scope, context and quality of telerehabilitation guidelines for physical disabilities: a scoping review. BMJ Open 2021 Aug 12;11(8):e049603 [FREE Full text] [doi: 10.1136/bmjopen-2021-049603] [Medline: 34385253]
38. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
39. Home page. LimeSurvey® Partners. URL: https://www.limesurvey.com/ [accessed 2024-03-05]
40. Ethics committee for research activities. FH Campus Wien's Ethics Committee. URL: https://www.fh-campuswien.ac.at/en/research/ethics-commission-for-research-activities.html [accessed 2023-10-24]
41. Population: census. Australian Bureau of Statistics. URL: https://www.abs.gov.au/statistics/people/population/population-census/latest-release [accessed 2023-05-15]
42. Dimock M. Defining generations: where millennials end and generation Z begins. Pew Research Center. 2019. URL: http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/defgenerations-pew.pdf [accessed 2023-05-15]
43. Shigekawa E, Fix M, Corbett G, Roby DH, Coffman J. The current state of telehealth evidence: a rapid review. Health Aff (Millwood) 2018 Dec;37(12):1975-1982. [doi: 10.1377/hlthaff.2018.05132] [Medline: 30633674]
44. Snoswell CL, Chelberg G, De Guzman KR, Haydon HH, Thomas EE, Caffery LJ, et al. The clinical effectiveness of telehealth: a systematic review of meta-analyses from 2010 to 2019. J Telemed Telecare 2021 Jun 29;29(9):669-684. [doi: 10.1177/1357633x211022907]
45. Jennett PA, Affleck Hall L, Hailey D, Ohinmaa A, Anderson C, Thomas R, et al. The socio-economic impact of telehealth: a systematic review. J Telemed Telecare 2003;9(6):311-320. [doi: 10.1258/135763303771005207] [Medline: 14680514]
46. Seebacher B, Bergmann E, Geimer C, Kahraman T, Reindl M, Diermayr G. Factors influencing the willingness to adopt telerehabilitation among rehabilitation professionals in Austria and Germany: a survey comparing data before and during COVID-19. Disabil Rehabil 2023 Mar 27:1-9. [doi: 10.1080/09638288.2023.2193428] [Medline: 36970941]
47. Van Eerdenbrugh S, Schraeyen K, Leysen H, Mostaert C, D'haenens W, Vandenborre D. Delivery of speech-language therapy and audiology services across the world at the start of the COVID-19 pandemic: a survey. Perspect ASHA SIGs 2022 Apr 14;7(2):635-646. [doi: 10.1044/2021_persp-21-00134]
48. Guggenberger B, Jocham B, Maul L, Jocham AJ. Implementation of telerehabilitation in Austrian outpatient physiotherapy ? A qualitative study / Implementierung von Telerehabilitation in der ambulanten Physiotherapie in Österreich ? Eine qualitative Studie. Int J Health Prof 2022;9(1):78-88 [FREE Full text] [doi: 10.2478/ijhp-2022-0007]
49. Mbada CE, Baderinwa TA, Sanuade CT, Ademoyegun Adekola B, Fatoye C, Maikudi L, et al. Awareness, attitude and expectations of physiotherapy students on telerehabilitation. Med Sci Educ 2021 Apr;31(2):627-636 [FREE Full text] [doi: 10.1007/s40670-021-01234-w] [Medline: 33619445]
50. Abbott-Gaffney CR, Gafni-Lachter L, Cason J, Sheaffer K, Harasink R, Donehower K, et al. Toward successful future use of telehealth in occupational therapy practice: what the COVID-19 rapid shift revealed. Work 2022;71(2):385-394. [doi: 10.3233/WOR-210789] [Medline: 35068409]
51. Edwards-Gaither L, Harris O, Perry V. Viewpoint telepractice 2025: exploring telepractice service delivery during COVID-19 and beyond. Perspect ASHA SIG 2023 Apr 05;8(2):412-417 [FREE Full text] [doi: 10.1044/2022_persp-22-00095]
52. Glinkowski W, Pawłowska K, Kozłowska L. Telehealth and telenursing perception and knowledge among university students of nursing in Poland. Telemed J E Health 2013 Jul;19(7):523-529 [FREE Full text] [doi: 10.1089/tmj.2012.0217] [Medline: 23650941]

XSL•FO
RenderX

53. 100 video conferencing statistics and facts for the 2022 market. Sonary. 2022. URL: https://sonary.com/content/100-video-conferencing-statistics-and-facts-for-the-2022-market/ [accessed 2023-04-04]
54. Grassl N, Nees J, Schramm K, Spratte J, Sohn C, Schott TC, et al. A web-based survey assessing the attitudes of health care professionals in Germany toward the use of telemedicine in pregnancy monitoring: cross-sectional study. JMIR Mhealth Uhealth 2018 Aug 08;6(8):e10063 [FREE Full text] [doi: 10.2196/10063] [Medline: 30089606]
55. Feng S, Mäntymäki M, Dhir A, Salmela H. How self-tracking and the quantified self promote health and well-being: systematic review. J Med Internet Res 2021 Sep 21;23(9):e25171 [FREE Full text] [doi: 10.2196/25171] [Medline: 34546176]
56. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3:14 [FREE Full text] [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
57. Tough D, Robinson J, Gowling S, Raby P, Dixon J, Harrison SL. The feasibility, acceptability and outcomes of exergaming among individuals with cancer: a systematic review. BMC Cancer 2018 Nov 21;18(1):1151 [FREE Full text] [doi: 10.1186/s12885-018-5068-0] [Medline: 30463615]
58. Govender SM, Mars M. The perspectives of South African academics within the disciplines of health sciences regarding telehealth and its potential inclusion in student training. Afr J Health Prof Educ 2018 Apr 09;10(1):38. [doi: 10.7196/ajhpe.2018.v10i1.957]
59. Fenton A, Montejo L, Humphrey KG, Mangano E, Gentry Russell N, Fingerhood M. Development of an integrated telehealth primary care and mental health training program for nurse practitioner students: review of the literature. J Nurse Pract 2023 Nov;19(10):104774. [doi: 10.1016/j.nurpra.2023.104774]
60. Boos K, Murphy K, George T, Brandes J, Hopp J. The impact of a didactic and experiential learning model on health profession students' knowledge, perceptions, and confidence in the use of telehealth. J Educ Health Promot 2022;11(1):232 [FREE Full text] [doi: 10.4103/jehp.jehp_1553_21] [Medline: 36177412]
61. Edirippulige S, Armfield N. Education and training to support the use of clinical telehealth: a review of the literature. J Telemed Telecare 2016 Jul 08;23(2):273-282. [doi: 10.1177/1357633x16632968]
62. Gartz J, O'Rourke J. Telehealth educational interventions in nurse practitioner education: an integrative literature review. J Am Assoc Nurse Pract 2020 Sep 01;33(11):872-878. [doi: 10.1097/JXX.0000000000000488] [Medline: 32890052]
63. Page C, Wahl R, Clements B, Woody R, Napier K. Adapting to telepractice: views of graduate student clinicians. Perspect ASHA SIG 2021 Dec 17;6(6):1876-1888. [doi: 10.1044/2021_persp-20-00275]
64. Mériaux-Kratochvila S. The academization of the health professions in Austria: facts and figures / Akademisierung der Gesundheitsberufe in Österreich: Zahlen und Fakten. Int J Health Prof 2021 Jan;8(1):141-145 [FREE Full text] [doi: 10.2478/ijhp-2021-0018]
65. Kuhn S, Ammann D, Cichon I, Ehlers J, Guttormsen S, Hülsken-Giesler M, et al. Wie revolutioniert die digitale transformation die bildung der berufe im gesundheitswesen? Careum Working Paper. URL: https://goeg.at/sites/goeg.at/files/inline-files/Careum_Working_Paper_8_de_kurz.pdf [accessed 2023-03-04]
66. Galpin K, Sikka N, King SL, Horvath KA, Shipman SA, AAMC Telehealth Advisory Committee. Expert consensus: telehealth skills for health care professionals. Telemed J E Health 2021 Jul;27(7):820-824. [doi: 10.1089/tmj.2020.0420] [Medline: 33236964]
67. Bonney A, Knight-Billington P, Mullan J, Moscova M, Barnett S, Iverson D, et al. The telehealth skills, training, and implementation project: an evaluation protocol. JMIR Res Protoc 2015 Jan 07;4(1):e2 [FREE Full text] [doi: 10.2196/resprot.3613] [Medline: 25567780]
68. Leone E, Eddison N, Healy A, Royse C, Chockalingam N. Exploration of implementation, financial and technical considerations within allied health professional (AHP) telehealth consultation guidance: a scoping review including UK AHP professional bodies' guidance. BMJ Open 2021 Dec 27;11(12):e055823 [FREE Full text] [doi: 10.1136/bmjopen-2021-055823] [Medline: 34969656]
69. Occupational Therapists WFO. World federation of occupational therapists' position statement on telehealth. Int J Telerehab 2014 Sep 03;6(1):37-40. [doi: 10.5195/ijt.2014.6153]
70. Cottrell MA, Russell TG. Telehealth for musculoskeletal physiotherapy. Musculoskelet Sci Pract 2020 Aug;48:102193 [FREE Full text] [doi: 10.1016/j.msksp.2020.102193] [Medline: 32560876]
71. Castillo-Allendes A, Contreras-Ruston F, Cantor-Cutiva LC, Codino J, Guzman M, Malebran C, et al. Voice therapy in the context of the COVID-19 pandemic: guidelines for clinical practice. J Voice 2021 Sep;35(5):717-727 [FREE Full text] [doi: 10.1016/j.jvoice.2020.08.001] [Medline: 32878736]
72. Carius C, Zippel-Schultz B, Schultz C, Schultz M, Helms HM. Developing a holistic competence model for telenursing practice: perspectives from telenurses and managers of telemedical service centres. J Int Soc Telemed eHealth 2016;4(e22):1-17 [FREE Full text]
73. Sullivan GM, Feinn RS. Facts and fictions about handling multiple comparisons. J Grad Med Educ 2021 Aug;13(4):457-460 [FREE Full text] [doi: 10.4300/JGME-D-21-00599.1] [Medline: 34434505]
74. DuBose-Morris R, McSwain SD, McElligott JT, King KL, Ziniel S, Harvey J. Building telehealth teams of the future through interprofessional curriculum development: a five-year mixed methodology study. J Interprof Care 2023 Dec 16;37(1):100-108 [FREE Full text] [doi: 10.1080/13561820.2021.2005556] [Medline: 34915788]

XSL•FO
RenderX

75. Pit SW, Velovski S, Cockrell K, Bailey J. A qualitative exploration of medical students' placement experiences with telehealth during COVID-19 and recommendations to prepare our future medical workforce. BMC Med Educ 2021 Aug 16;21(1):431 [FREE Full text] [doi: 10.1186/s12909-021-02719-3] [Medline: 34399758]
76. Jadotte YT, Noel K. Definitions and core competencies for interprofessional education in telehealth practice. Clinics in Integrated Care 2021 Jun;6:100054. [doi: 10.1016/j.intcar.2021.100054]

Abbreviations

ANC: advanced nursing counseling
ANE: advanced nursing education
ANP: advanced nursing practice
CROSS: Consensus-Based Checklist for Reporting of Survey Studies
DIE: dietetics
HAE: health assisting engineering
MID: midwifery
NUR: health care and nursing
ORT: orthoptics
OT: occupational therapy
PT: physiotherapy
SLT: speech and language therapy

Edited by T de Azevedo Cardoso; submitted 21.07.23; peer-reviewed by H Alshawaf, L Davies, K Drude, P Koppel; comments to author 17.10.23; revised version received 04.12.23; accepted 13.02.24; published 21.03.24.

Please cite as:
Rettinger L, Putz P, Aichinger L, Javorszky SM, Widhalm K, Ertelt-Bach V, Huber A, Sargis S, Maul L, Radinger O, Werner F, Kuhn S
Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Students’ Perceived Knowledge, Skills, Attitudes, and Experience
JMIR Med Educ 2024;10:e51112
URL: https://mededu.jmir.org/2024/1/e51112
doi:10.2196/51112
PMID:38512310

©Lena Rettinger, Peter Putz, Lea Aichinger, Susanne Maria Javorszky, Klaus Widhalm, Veronika Ertelt-Bach, Andreas Huber, Sevan Sargis, Lukas Maul, Oliver Radinger, Franz Werner, Sebastian Kuhn. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.03.2024. 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.

XSL•FO
RenderX

Impact of Health Informatics Analyst Education on Job Role, Career Transition, and Skill Development: Survey Study

Kye Hwa Lee[1,2], MD, PhD; Jae Ho Lee[1,2,3], MD, PhD; Yura Lee[1,2], MD, PhD; Hyunna Lee[4], PhD; Ji Sung Lee[5], PhD; Hye Jeon Jang[4], MS; Kun Hee Lee[4], BS; Jeong Hyun Han[4], BS; SuJung Jang[4,6], BS

1

2

3

4

5

6

Corresponding Author:
Jae Ho Lee, MD, PhD

Abstract

Background: Professionals with expertise in health informatics play a crucial role in the digital health sector. Despite efforts to train experts in this field, the specific impact of such training, especially for individuals from diverse academic backgrounds, remains undetermined.

Objective: This study therefore aims to evaluate the effectiveness of an intensive health informatics training program on graduates with respect to their job roles, transitions, and competencies and to provide insights for curriculum design and future research.

Methods: A survey was conducted among 206 students who completed the Advanced Health Informatics Analyst program between 2018 and 2022. The questionnaire comprised four categories: (1) general information about the respondent, (2) changes before and after program completion, (3) the impact of the program on professional practice, and (4) continuing education requirements.

Results: The study received 161 (78.2%) responses from the 206 students. Graduates of the program had diverse academic backgrounds and consequently undertook various informatics tasks after their training. Most graduates (117/161, 72.7%) are now involved in tasks such as data preprocessing, visualizing results for better understanding, and report writing for data processing and analysis. Program participation significantly improved job performance (*P*=.03), especially for those with a master's degree or higher (odds ratio 2.74, 95% CI 1.08 - 6.95) and those from regions other than Seoul or Gyeonggi-do (odds ratio 10.95, 95% CI 1.08 - 6.95). A substantial number of respondents indicated that the training had a substantial influence on their career transitions, primarily by providing a better understanding of job roles and generating intrinsic interest in the field.

Conclusions: The integrated practical education program was effective in addressing the diverse needs of trainees from various fields, enhancing their capabilities, and preparing them for the evolving industry demands. This study emphasizes the value of providing specialized training in health informatics for graduates regardless of their discipline.

(JMIR Med Educ 2024;10:e54427) doi:10.2196/54427

KEYWORDS

health informatics; health informatics training; informatics training; professional development; training program; digital health technology; informatics workforce; informatics competencies; competencies; job skills; continuing education; data science

Introduction

The field of digital health is rapidly advancing, driven by innovative technologies such as artificial intelligence, big data analytics, digital therapeutics, and embedded medical systems [1]. Amid this dynamic progression, there is a pressing need to cultivate aptitudes and insights to remain abreast of these changes. The growing demand for digital health professionals is evident in the United Kingdom, where almost 90% of health care professionals are projected to require digital proficiencies within the next 2 decades [2]. Additionally, the World Health Organization and World Bank estimate that by 2030, approximately 40 million new health and social care jobs will be created, many of which will be within the field of digital health [3]. Despite the high demand for digital health professionals, there is a substantial gap between the skills health informatics (HI) graduates possess upon graduation and those desired by employers [4,5]. As the health care paradigm shifts

XSL•FO
RenderX

toward digitalization, there is an escalating demand for adept professionals capable of conceptualizing, instituting, and overseeing digital health interventions [6]. Current HI educational frameworks, however, fall short of equipping students with the requisite practical acumen [7], leaving many underprepared for the challenges of the profession.

Globally, substantial progress has been made in developing curricula dedicated to HI, with the United States being at the forefront of these efforts. The American Medical Informatics Association saw the need for expertise beyond traditional academic pathways and established an education committee in 2002 [8], leading to the development of the 10×10 Program, a bold initiative to train 10,000 HI specialists from 2005 to 2010. This 10-week program, covering 10 core topics, was designed to cater to a wide range of professionals, from health information managers to system developers. However, the multifaceted nature of HI challenges the feasibility of a one-size-fits-all curriculum [9]. A substantial proportion of professionals work without formal training and have distinct requirements depending on their role and academic background. In response, the International Medical Informatics Association has created a curriculum that aims to deliver tailored professional education to a diverse cohort [10]. However, despite these notable advances, a consistent strategy for an ideal curriculum or training approach for health care professionals of diverse backgrounds—cultural and educational—wishing to explore HI is lacking.

In recent years, clinical informatics has been incorporated into the primary curriculum of medical and graduate schools in South Korea. However, these trainings also remain somewhat limited. Acknowledging the interdisciplinary essence of HI, the South Korean government initiated multiple intermediate informatics education programs. In 2018, a large-scale educational program was launched in collaboration with 3 major universities, offering the Genomic Specialist Training Program, Advanced Health Informatics Analyst (AHIA) Training Program, and Precision Medicine Workforce Training Program. This 5-year program was taught independently of regular university and graduate courses and was offered free of charge to all participants. The AHIA program, in particular, was important owing to its focus on data analysis using real-world hospital data, a skill essential for medical information analysis experts. To cultivate an HI expert within a postgraduate program, a commitment of at least 60 credits or a year of full-time study is mandatory. This bespoke program entailed an annual workload of 56.5 study hours and mirrored the rigor of a postgraduate curriculum. By March 2023, a total of 206 participants from various academic backgrounds and degree pursuits, some with prior experience in relevant fields, had completed the AHIA program. Given the diverse academic trajectories of the attendees and the in-depth instruction in HI, scrutinizing their posttraining professional applications or shifts in employment status was anticipated to yield intriguing insights. This study was therefore conducted to survey the program alumni, aiming to discern the educational impact and gather valuable perspectives to shape subsequent informatics curricula. The overarching objective of this research is to validate the efficacy of intensive HI education when delivered to individuals across a spectrum of academic and professional backgrounds. Such insights will underpin the development and refinement of future informatics curricula.

Methods

AHIA Course

The survey was conducted among the 206 students who completed the AHIA course over a period of 5 years (2018 - 2022). This advanced course was conducted once a year in 2018 and 2019, and twice a year from 2020 to 2022. On average, each course had 27.4 students enrolled. Each course consisted of 10 sessions. A total of 206 (97.2%) out of 212 students completed the course, and the failure and dropout rate was 2.8% (6/212). The program is designed to provide practice-oriented education, allowing students to experience the practical needs and challenges faced in the medical field. The curriculum encompasses a range of topics, including electronic medical records, medical images, public health, lifelog data, biosignals, and genome data, which are essential in the medical field. A lifelog is a practice where individuals digitally document their daily experiences with different levels of granularity, serving various aims [11]. Theoretical knowledge and practical training were incorporated into the course, which included a 3-week team project to practice problem-solving skills relevant to real-world scenarios. The curriculum and structure of this course are provided in Multimedia Appendix 1.

Survey Design

We implemented a structured survey to systematically evaluate and analyze the insights for future informatics curriculum design and assess the effectiveness of our training approaches. The questionnaire was organized into four categories, as outlined in Table 1: (1) general information on the respondents: this section collected essential demographic data about the survey respondents; (2) changes before and after program completion: participants were asked to reflect on any changes they experienced in their knowledge or skills after completing the professional program; (3) impact of the professional program on professional practice: this section aimed to assess how the program influenced the professional practices of the participants; and (4) continuing education requirements: this section sought to identify any specific needs or preferences for further education among the respondents. The questionnaire incorporated adaptive questioning techniques to minimize the quantity and complexity of the questions presented. The number of questionnaire items per page was from 2 to 4. The questionnaire was distributed over 12 pages. The respondent was able to modify their response by using the back button. The questionnaire is provided in Multimedia Appendix 2.

XSL•FO
RenderX

Table . Questionnaire composition and questions.

Category	Classification	Questions, n
General information	• Sex • Age • Residence • Final education • Major • Employment status • Occupation • Organization • Job before and after completing the program • Duration of work experience • Work related to informatics or health informatics	12
Changes before and after completing the intensive course	• Purpose or reason for program application • Current work impact • Changes in work after completing the program • How helpful (satisfied) was the course? • Positive impact • A change in career or intention to change careers in informatics or health informatics after completing the program	8
Effects of advanced courses on informatics in performing tasks	• Level of change in medical data analysis and processing ability • Performed tasks related to informatics or health informatics • Change in fear of working with informatics or health informatics • Participation in informatics- or health informatics–related activities • Improved skills and abilities related to informatics or health informatics • Change in interest in informatics or health informatics • Change in participation in informatics- or health informatics–related activities • Reasons why informatics- or health informatics–related activities were difficult to do	6
Demand for continuing education	• Continuing education or reinforcement after completing the course • Demand for informatics- or health informatics–related personnel in your organization • Cultivation of specialized personnel related to informatics or health informatics • Improvements and suggestions	5

Evaluation of the Reliability and Validity of the Survey

The survey was designed to align with the educational objectives and learning outcomes of the AHIA course. The purpose was to evaluate the applicability of the theoretical knowledge and practical training in addressing real-world health care challenges. Survey questions were developed to assess shifts in students' knowledge and skills, the impact on their professional practices, and their needs for continuing education. The initial draft of the survey was designed by our research team of HI experts to ensure the relevance and clarity of the questions, thereby enhancing the survey's reliability and validity. To ascertain the reliability and validity of the survey, we engaged a panel of experts in HI to review and critique the initial survey draft. Their invaluable feedback led to the refinement of our survey questions, ensuring they effectively captured the educational outcomes and experiences of our students. This step was crucial in validating the survey instrument and ensuring that the data collected were both reliable and reflective of the course's impact.

Participant Recruitment and Data Collection

From January 30 to February 8, 2023, data were gathered using a web-based questionnaire. Links to this questionnaire were emailed to the 206 students who had completed the AHIA course between 2018 and 2022. The first page of the survey provided

XSL•FO
RenderX

information on the purpose of the research, the length of time needed to complete the survey, and the methods of personal information protection.

Statistical Analysis

The collected survey data underwent a comprehensive statistical analysis. Only completed survey data were analyzed. First, basic descriptive analysis techniques were applied to examine the characteristics of the survey respondents, including computing measures such as mean, median, and SD to summarize the central tendency, dispersion, and distribution of the data. Additionally, frequency tables were used to present categorical variables, providing insights into the demographic composition of the respondents. Furthermore, ANOVA was conducted to determine if there were significant differences in informatics medicine work based on the major of the education graduates. Visualization methods, such as pie graphs, were used to illustrate differences among majors and to depict shifts in proficiency levels before and after the training course. All statistical analyses were performed using R software (version 4.1.1; R Foundation for Statistical Computing).

Ethical Considerations

This research paper was reviewed and approved by the Institutional Review Board of Asan Medical Center (S2022-2671-0001). Before commencing the survey, participants were mandated to provide their informed consent, acknowledging the study's objectives and permitting the collection of their personal data. Participation was voluntary, and upon completing all items in the survey, participants were compensated with a small reward. The collected responses were anonymous and used exclusively for analysis and deriving outcomes, with the confidentiality of individual responses being strictly protected under Article 33 of the Statistics Act.

Results

Participant Demographics

Among the 206 trainees, 161 (78.2%) responded. Due to the anonymous nature of the survey, identifying the nonrespondents was not feasible, and in comparison to prior research [8], achieving a response rate of 78.2% is considered substantial. This sample size was deemed sufficient to draw meaningful conclusions, with the sampling error, expressed as a margin of error, calculated to be ±3.62% at a 95% confidence level. Table 2 presents the demographic characteristics of the study population. The sex distribution shows that of the 161 respondents, 54% (n=87) were female and 46% (n=74) were male, with no statistically significant difference between the 2 groups (*P*=.31). The age distribution was as follows: 52.9% (46/87) of female participants and 35.1% (26/74) of male participants were in the 20‐29 years group; 28.7% (25/87) of female participants and 37.9% (28/74) of male participants were in the 30‐39 years group; and 18.4% (16/87) of female participants and 27% (20/74) of male participants were in the 40+ years group. No significant difference between the age of male and female participants was observed (*P*=.08). There were also no significant educational differences between the sexes (*P*=.25). Most participants had a bachelor's degree (34/87, 39.1% of female participants and 19/74, 25.7% of male participants) or a master's degree (39/87, 44.8% of female participants and 38/74, 51.4% of male participants), while a smaller proportion had a high school diploma or a PhD. Job status differed significantly between female and male participants (*P*=.01). Most female participants (52/87, 59.8%) and male participants (56/74, 64.4%) were employed in full-time jobs, while smaller proportions were engaged in job preparation or postgraduate studies. Lastly, there were no significant differences in the locations between the sexes (*P*=.14). Most participants (62/87, 71.3% of female participants and 42/74, 48.3% of male participants) were from Seoul, followed by Gyeonggi-do and other regions.

Among the 161 respondents, 64% (n=103) had a master's degree, PhD, or higher. These respondents' undergraduate majors and changes in their majors in the master's and doctoral courses after the bachelor's degree were compared. The most common undergraduate majors were computer science (n=30, 18.6%), "others" (n=26, 16.1%), statistics (n=25, 15.5%), medicine (n=20, 12.4%), and nursing (n=15, 9.3%), whereas for the master's degree or doctoral major, "others" was the most common at 42.9% (n=69), followed by biomedical or medical informatics and statistics at 13% (n=21) each, medicine at 9.9% (n=16), computer science at 8.7% (n=14), health service research at 8.1% (n=13), and biomedical engineering at 4.3% (n=7). "Others" includes experimental psychology, molecular science, microbiology, psychology, etc (Figure 1).

The primary motivation for enrolling in the AHIA course and the reason for applying to informatics-related jobs were examined. The highest proportion of respondents (n=120, 74.5%) indicated that they sought to enhance or improve their job competencies, followed by 67.1% (n=108) who applied to enhance their future job prospects. Most respondents (n=120, 74.5%) joined the course to strengthen their job-related skills and performance. Moreover, a substantial proportion (n=70, 43.5%) expressed their interest in learning new technologies with promising opportunities, 36% (n=58) highlighted interest in joint research or collaboration, and 32.3% (n=52) considered leveraging the potential for collaboration, indicating the high expectations for collaboration and growth opportunities in the converging field. Furthermore, some participants (n=47, 29.2%) stated that they pursued the course to obtain a completion certificate, and 6.8% (n=11) indicated that they received recommendations or endorsements from others as motivation for enrollment.

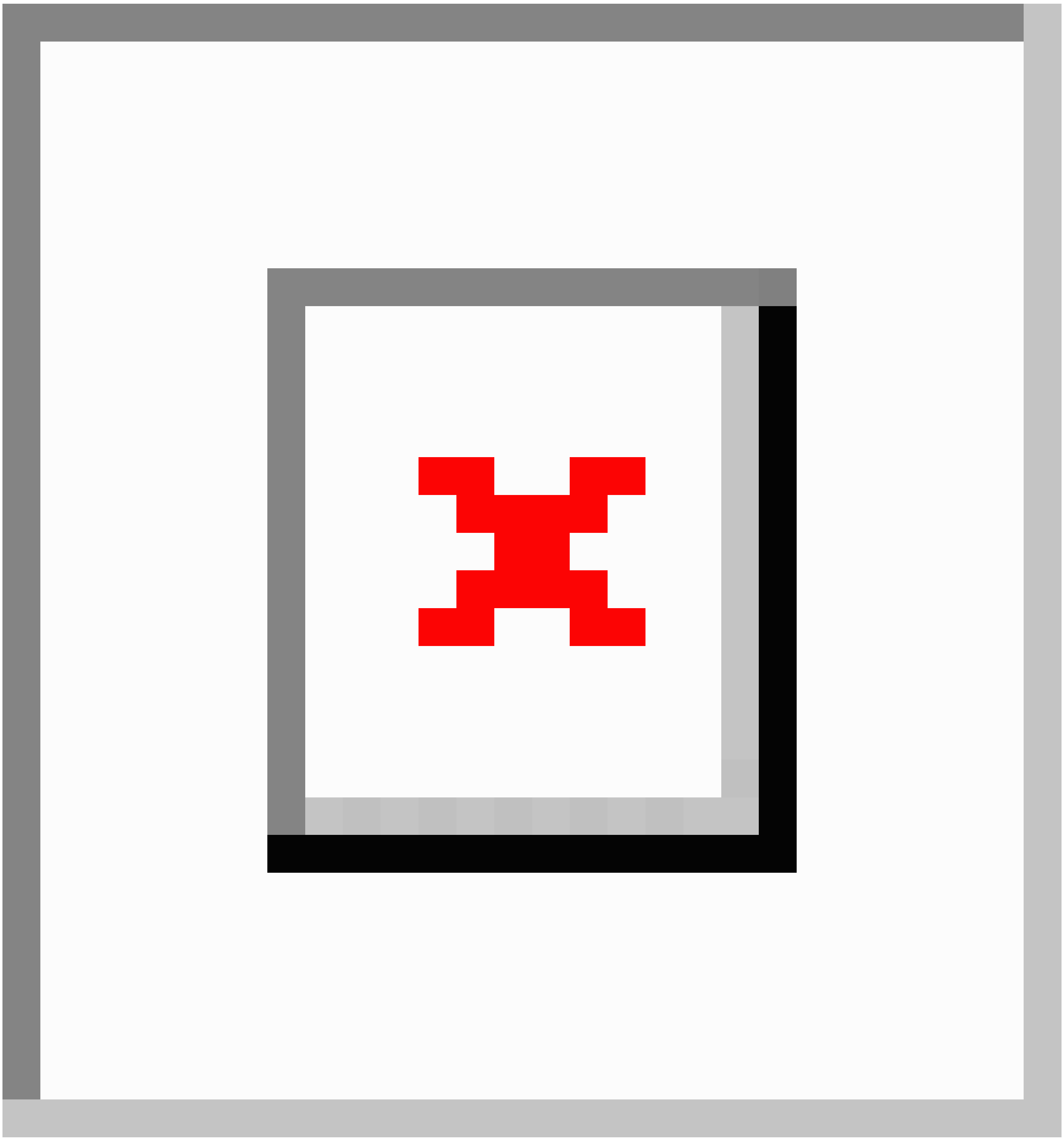
XSL•FO
RenderX

Table . Respondent demographics.

Variables		Female sex (n=87)	Male sex (n=74)	*P* value
Respondents (n=161), n (%)		87 (54)	74 (46)	.31
Age group (years), n (%)				.08
	20 - 29	46 (52.9)	26 (35.1)	
	30 - 39	25 (28.7)	28 (37.9)	
	40+	16 (18.4)	20 (27)	
Education, n (%)				.25
	High school diploma	3 (3.4)	2 (2.7)	
	Bachelor's degree	34 (39.1)	19 (25.7)	
	Master's degree	39 (44.8)	38 (51.4)	
	PhD	11 (12.6)	15 (20.3)	
Job status, n (%)				.01
	Job preparation	13 (14.9)	1 (1.1)	
	Postgraduate student	15 (17.2)	11 (12.6)	
	Full-time job	52 (59.8)	56 (64.4)	
	Part-time job	5 (5.7)	6 (6.9)	
Region, n (%)				.14
	Seoul	62 (71.3)	42 (48.3)	
	Gyeonggi-do	17 (19.5)	20 (23)	
	Other	8 (9.2)	12 (13.8)	

XSL•FO
RenderX

Figure 1. Respondents' undergraduate and graduate majors.



Informatics Ability After the Training Course

Regarding the current medicine-related work of respondents according to their majors, 117 (72.7%) of the 161 respondents answered that they "preprocess the collected data"; 57.1% (n=92) noted that they "visualize the results to help understand the main analysis results"; and 57.1% (n=92) said that they "write an analysis report applying various modeling techniques." To identify a difference in the information medicine work currently performed according to the final major of the education graduates, data were visualized according to majors (ANOVA; $P<.001$). With the exception of graduates with mathematics majors, most graduates (n=117, 72.7%) performed plenty of data preprocessing (Multimedia Appendix 3). Data preprocessing was identified as the main task for graduates of health science research, statistics, and computer science majors. Graduates of medical (doctors and nurses) and double majors often performed tasks related to data utilization plan establishment, and data analysis was performed frequently by graduates with mathematics, statistics, and "others" majors.

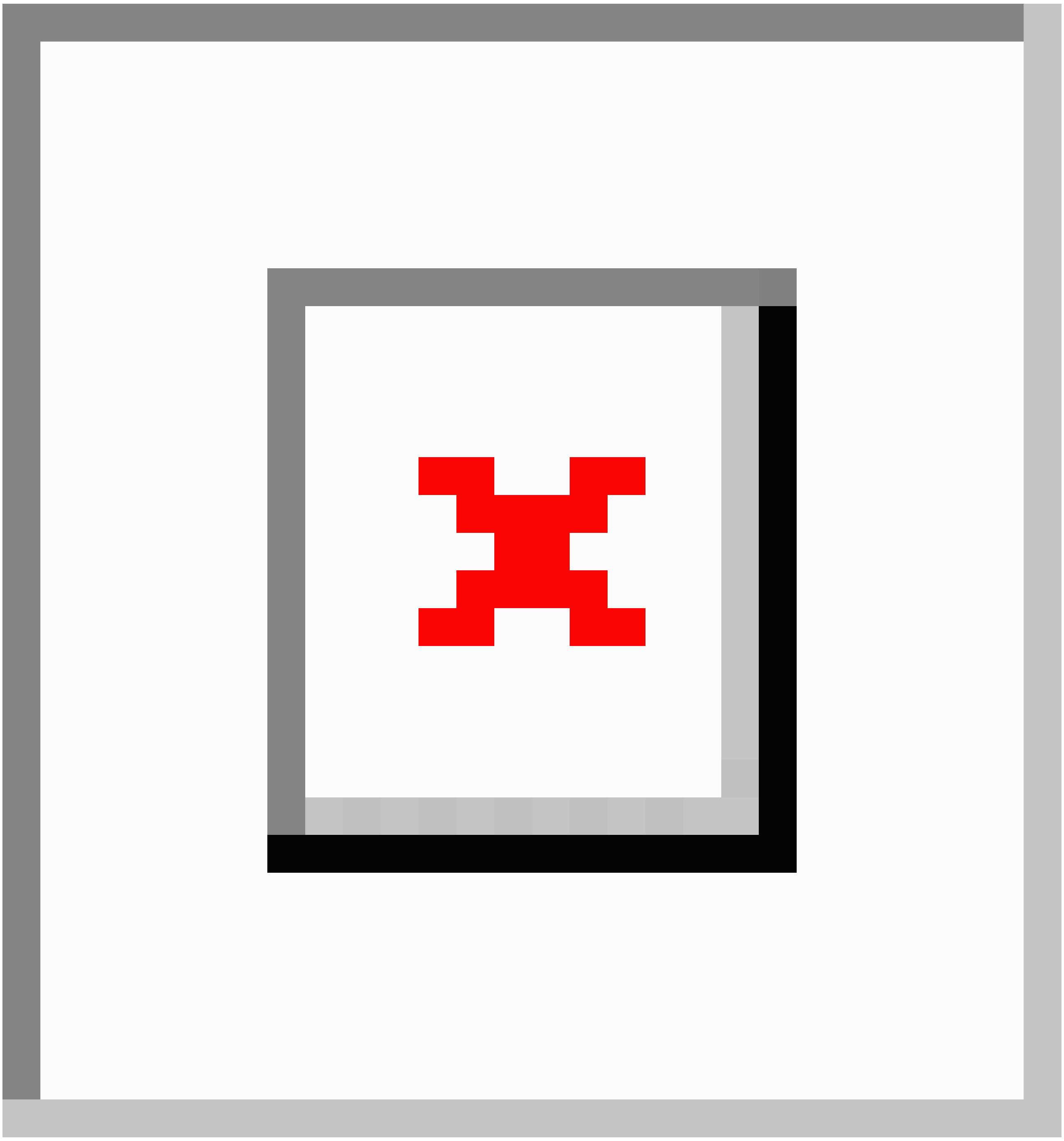
Changes in Technology and Knowledge After the Training Course

Regarding the ability to analyze and process medical data before and after the training course, most (68/161, 42.2%) respondents selected "I can see and follow the analysis method (Step 2)" before the course. This stage primarily involves the ability to visually recognize and replicate given analysis methods based

XSL•FO
RenderX

on instructional guidance, which is crucial for foundational learning and initial engagement with medical data analysis. However, after completing the course, 46% (n=74) said "I know and can design the analysis method (Step 3)," indicating that medical data analysis skills and processing abilities had improved upon completion. This progression signifies a deeper understanding and ability, not just to follow but also to design and conceptualize analysis methods independently. Step 3 encompasses a critical transition from merely executing predefined analysis steps to creating customized analysis frameworks suited to specific medical data challenges. Additionally, of the 6.2% (n=10) of respondents who answered "I did not know at all (Step 0)" before completion, half (5/10, 50%) of them improved to "I understand the concept after hearing the term (Step 1)," while the other half (5/10, 50%) moved to "I can see and follow the analysis method (Step 2)," highlighting the improvement from the beginner to the intermediate step (Figure 2 and Table 3). After completion, 111 (68.9%) students improved their skills by at least 1 step. Excluding the 8 (5%) individuals who began the course at the fifth step, 67.3% (103/153) showed a technical improvement of at least 1 step.

Figure 2. Step of change in medical data analysis and processing capabilities. Steps: 0=I did not know at all, 1=I understand the concept after hearing the term, 2=I can see and follow the analysis method, 3=I know and can design the analysis method, and 4=Expert-level analysis and results can be drawn.



XSL•FO
RenderX

Table . Step of change in medical data analysis and processing capabilities (n=161).

Step	Degree of change	Before training, n (%)	After training, n (%)
0	I did not know at all	10 (6.2)	0 (0)
1	I understand the concept after hearing the term	45 (28)	7 (4.3)
2	I can see and follow the analysis method	68 (42.2)	58 (36)
3	I know and can design the analysis method	30 (18.6)	74 (46)
4	Expert-level analysis and results can be drawn	8 (5)	22 (13.7)

Shift in Job Change Intentions Before and After Training

After completion of the AHIA course, 58.4% (94/161) of respondents in the field of informatics or HI changed jobs or intended to change jobs. Although more than half expressed this intent, actual job change was relatively rare, with only 22.4% (36/161) experiencing a job change. Among those who responded to further questions (n=91), except for 3 who did not respond, 82.4% (n=75) stated that the course influenced their decision. Specifically, “understanding work contents and characteristics in the field of informatics” (n=58, 63.7%) and “interest in the informatics field” (n=56, 61.5%) were cited as significant factors influencing either the experience of a job change or the intention to change jobs (as depicted in Table 4).

Table . Impact of the Advanced Health Informatics Analyst course on job change or the intention to change jobs.

Factors that impacted job change or the intention to change jobs	Respondents (n=91; 91/161, 56.5%)
Understanding work contents and characteristics in the field of informatics	58 (63.7)
Interest in the informatics field	56 (61.5)
Confirmation of the potential for growth in the area of IT	50 (54.9)
Improving individual academic skills and gaining academic qualifications	42 (46.2)
Developing jobs skills (like developing technology, analyzing data, etc)	38 (41.8)
Confirmation that the area of informatics and the person’s skills are a good match	36 (39)
Exchange with people interested in medical information	30 (33)

Impact of the AHIA Course on Informatics Activities and Attitudes

To assess the influence of the AHIA course on activities and attitudes within the field of informatics, we formulated 4 questions. These were rated on a Likert scale across 3 key areas: the type of positive impact the course had, the increase in activities related to informatics, and changes in interest toward or apprehension about the field of informatics. Among the 161 respondents, 72.6% (n=117) answered that the course had a positive effect. Among them, 75.2% (n=121) and 60.2% (n=97) highlighted that they had “enhanced knowledge in informatics for personal improvement” and “gained practical experience in utilizing data for academic or professional purposes,” respectively. Regarding the increase in activities in the informatics field and changes in attitudes, 63.4% (n=102) of the respondents stated that their participation in informatics activities had increased. Additionally, 90.1% (n=145) of respondents reported that their interest in informatics increased and 73.9% (n=119) said their fear had decreased.

Examining the Link Between the Impact of the AHIA Course and Respondent Characteristics

To confirm the relationship between the influence of the AHIA course and the characteristics of the respondents, we performed a multivariable logistic analysis for each outcome by adding the age, sex, education level, and field of work of the respondents. Participants showing a positive correlation in the effect of the AHIA course on job performance were those with a master’s degree, PhD, or higher (odds ratio [OR] 2.74, 95% CI 1.08‐6.95) and workers in regions other than Seoul and Gyeonggi-do (OR 10.95, 95% CI 1.08‐6.95). The respondent factor significantly associated with increased informatics activity was having a master’s degree or higher (OR 2.84, 95% CI 1.27‐6.31; *P*=.01), while other factors did not show any correlation. Regarding the increase in informatics activities, the highest observed proportion (83/161, 51.6%) involved the “use of informatics in actual work,” followed by “completing an additional degree in informatics” at 26.7% (43/161), suggesting that these degree-related matters are closely related factors. Respondent sex was associated with increased interest in informatics, and female participants responded that they were more interested than male participants (OR 3.51, 95% CI

XSL•FO
RenderX

1.10 - 11.21). Age showed no significant difference in all 4 items (Table 5).

Table . Course effects on health informatics work and attitudes.

Variables	Effect		Activity		Interest		Fear decreased	
	OR[a] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Age group (years; reference: 20-29)								
30 - 39	0.43 (0.16-1.14)	.009	0.78 (0.33-1.84)	.57	0.88 (0.19-4.08)	.87	0.77 (0.31-1.94)	.58
40+	0.98 (0.31-3.08)	.10	1.00 (0.38-2.65)	.99	0.63 (0.14-2.93)	.56	0.92 (0.33-2.58)	.88
Sex (reference: male)								
Female	1.59 (0.75-3.36)	.23	1.55 (0.78-3.08)	.21	3.51 (1.10-11.21)	.03	1.40 (0.67-2.91)	.37
Education (reference: bachelor's degree)								
Master's degree or PhD	2.74 (1.08-6.95)	.03	2.84 (1.27-6.31)	.01	3.69 (0.96-14.23)	.06	1.30 (0.55-3.07)	.55
Location (reference: Seoul)								
Gyeong-gi-do	1.22 (0.51-2.89)	.66	0.96 (0.43-2.15)	.92	1.07 (0.30-3.90)	.91	1.11 (0.46-2.69)	.82
Other areas	10.95 (1.34-89.33)	.02	0.73 (0.26-1.99)	.53	1.30 (0.25-6.85)	.76	0.66 (0.24-1.87)	.44

[a]OR: odds ratio.

Discussion

Our study investigated the changes observed in graduates who completed the 56-hour medical information specialist education course, which was offered across a period of 5 years. Graduates from this course had diverse academic backgrounds, and their postcourse informatics work varied accordingly. Graduates of health sciences, statistics, and computer engineering majors primarily focused on data preprocessing. Medical professionals and those with double majors often worked on data utilization plans, while those from mathematics and statistics backgrounds frequently engaged in data analysis. The faculty involved in delivering the course, comprising professors from multiple universities including some authors of this paper, primarily aimed to enhance job competency or equip participants with the necessary skills for future employment opportunities. Regarding the informatics work currently performed by the graduates, most (116/161, 72.2%) cited tasks such as "preprocessing the collected data," "visualizing the results to understand the main analysis outcomes," and "processing the data and preparing an analysis report using various modeling techniques." Additionally, local workers with a master's degree or higher reported experiencing the greatest positive impact from the course. A significant proportion of graduates also indicated that their education impacted their career transitions. Predominantly, comprehension of job roles and characteristics, along with an intrinsic interest in the field, emerged as the primary influencing factors. The results indicate that this convergence practice education program serves as a successful model, effectively addressing the needs of trainees from various fields, enhancing their competencies, and preparing them for the evolving demands of the industry.

This study demonstrates the value of using real-world clinical data for training individuals across various academic fields within HI. The finding that 63.4% (102/161) of participants became more engaged in informatics-related endeavors after the course is compelling evidence for the effectiveness of this educational method in enhancing individual proficiency and involvement. Furthermore, our data suggest that hands-on clinical data education benefits those from diverse academic backgrounds, augmenting their HI activities. This aligns with prior research underscoring the value of hands-on experience in the competencies desired in industry professionals [12]. This could be an important factor to consider in future educational curriculum development or professional training programs. Interestingly, respondents holding a master's degree or higher, or those working in local settings, indicated the most significant benefits from the program. These findings imply that individuals with higher degrees possess a solid grounding in foundational HI concepts. While industry hiring trends often favor individuals with a bachelor's degree [12,13], those with advanced degrees are positioned to derive maximum value from HI training. Their robust foundation in core HI areas, including computer science, statistics, and health sciences, enables them to fully grasp and effectively implement HI concepts. In a field such as HI, wherein cultivating interdisciplinary expertise is crucial, offering specialized courses rooted in real data to master's-level professionals from varied backgrounds can be a potent strategy for developing top-tier talent. Notably, our findings suggest that graduates in rural regions benefited more from the training compared to their counterparts in well-resourced areas such as Seoul or Gyeonggi-do. This highlights the potential of similar targeted programs to elevate the educational standard and bridge the gap in educational opportunities across regions.

In well-resourced nations, medical system computerization is on the rise, with 96% of the general hospitals in South Korea

XSL•FO
RenderX

adopting electronic medical records [14]. These health care digital shifts produce vast data volumes that are essential for decision-making, gauging treatment efficacy, and underpinning evidence-based approaches [15]. To efficiently handle such expansive medical data, experts in medical information are imperative at every phase [16]. As the convergence of health care and technology accelerates, academic institutions are striving to cater to varied aspirants, spanning recent high school graduates to IT and medical professionals [17]. However, there is a scarcity of programs tailored for the HI sector, which is vaguely defined [18,19]. A disparity in the provision of HI degrees exists even in well-equipped nations such as the United States [20]. The 10×10 Program by the American Medical Informatics Association stands out as a robust training model [9]; however, for countries lacking ample educational resources, state-backed initiatives resembling the AHIA course offer a potential solution. The International Medical Informatics Association advocates for dedicated establishments offering continuous education courses [10]. Given the dynamic landscape of health IT, future programs must embed lifelong learning tenets, encouraging graduates to persistently upskill. HI professional development should also extend to current medical staff and those from varied academic realms. Training individuals with profound health care and IT expertise, coupled with a hands-on grasp of the health care structure, remains a formidable challenge. The graduates of the AHIA course, with their varied academic histories, highlight the need for more holistic, adaptive future programs. The assertion by Topol [2] regarding the imminent digital skill requirement for all National Health Services roles emphasizes the paramount importance of digital literacy in health care. Despite the focus of the program being professionals and researchers, 6.2% (10/161) of respondents professed ignorance of informatics jargon. This underscores the pivotal role of informatics education in molding a competent health workforce. Moreover, those possessing an advanced degree reportedly reaped the most benefit from this training, suggesting that upcoming courses should be tailored to the unique educational backgrounds of enrollees.

AHIA graduates hail from a broad array of fields, including computer science, health care, and statistics, with a wide range of focus in their advanced studies as well. While HI skills are crucial, medical schools offer minimal, inconsistent, and rarely updated education, leaving students unprepared for the digital health care landscape [21]. Although the swift evolution of medical technology necessitates continual adaptation of research and educational content within the field, it appears that updates to the undergraduate HI curriculum lag behind these technological advancements. Our findings highlight that, in the fast-evolving domain of medical informatics, implementing a robust and comprehensive curriculum that transcends the confines of specific academic departments or institutions—particularly one that caters to the interdisciplinary education of students from a variety of academic backgrounds—emerges as a crucial strategy. This approach is essential for keeping pace with the rapid developments in the field. We also explored the impact of the program on the career paths of graduates, finding that a majority were willing to consider job changes and that a notable portion had already done so. They attributed a significant influence on their career decisions to the program. Analyzing web-based job postings, a German study revealed a high demand for medical informatics professionals, with half of the jobs concentrated in hospitals [22]. While hands-on experience is crucial, employers and graduates alike find it challenging to find programs that bridge the gap between computer science expertise and practical hospital knowledge. Considering that the AHIA program offers diverse students hands-on training and real-world experience using actual hospital data while working on projects in teams, the AHIA model suggests a positive strategy for strengthening the competencies of students entering the HI job market and meeting job demands.

Despite its contributions to understanding the impact of an AHIA course on graduates' job roles and competencies, this study acknowledges several limitations. First, the survey's broad focus on the overall educational program might have overlooked the nuanced impacts of specific subjects, such as information security and privacy, on graduates' career outcomes. While this approach captures the general effectiveness of the program, it leaves room for further exploration into how individual modules shape professional skills and knowledge. Potential biases in survey responses due to their self-reported nature represent another limitation. While measures were taken to ensure anonymity and encourage candidness, the inherent nature of self-reporting might introduce biases that could affect the interpretation of our findings. This aspect underscores the need for caution in generalizing the results beyond the surveyed population. The generalizability of our findings is also limited by the specific context of the educational program and its participants. This specificity might not fully capture the diverse experiences across the broader field of HI, suggesting that the findings might not be universally applicable without further validation in different settings. Furthermore, the data collection process, focusing on immediate postgraduation outcomes, does not account for the longitudinal development of competencies or career progression. This temporal limitation highlights the potential for more comprehensive, longitudinal studies to understand the lasting impacts of educational programs. Lastly, our study could not evaluate the characteristics of nonrespondents. A total of 21.8% (45/206) of individuals who completed the course did not participate in the subsequent survey. While survey fatigue, time constraints, or privacy concerns might have contributed to this nonresponse rate, the specific reasons for their lack of participation remain unexplored due to the anonymized nature of our data collection process. This limitation prevents us from fully understanding whether the nonrespondents possess distinct demographic or professional characteristics compared to survey participants, which could potentially introduce a bias in our findings. Despite these limitations, our research offers significant insights into the value of HI education and its role in preparing professionals for the field. It underscores the importance of curriculum design in addressing the evolving needs of the HI sector and provides a foundation for future research.

Our survey of AHIA graduates revealed 2 particularly compelling findings that enrich our understanding of the program's effectiveness and potential areas for future research. First, the reported educational impact of the AHIA program

XSL•FO
RenderX

was more pronounced among graduates holding a master's degree or higher. Second, students residing outside the metropolitan area of Seoul perceived a greater benefit from their participation in the program. These results suggest that advanced educational background and geographic diversity play significant roles in the perceived value and impact of HI education. Incorporating these perspectives, we propose additional areas for future research. Further investigation into the influence of prior academic achievements on the outcomes of HI education is warranted. Specifically, understanding the mechanisms through which graduates with higher degrees report greater benefits from their education can inform the tailoring of programs to enhance benefits across various educational levels. This exploration is essential for developing curricula that are responsive to the educational background of students, ensuring that all participants can achieve significant gains from their involvement. Additionally, the greater perceived impact of HI programs among students from nonmetropolitan areas highlights the need for an in-depth analysis of geographic disparities in educational outcomes. Identifying the challenges faced by students in accessing traditional educational resources and the ways in which alternative models such as the AHIA program can effectively bridge these gaps will be crucial. Such research could lead to the development of more accessible and inclusive HI education programs, which are particularly vital for students in resource-limited countries and those residing outside major urban centers.

Timely, project-based, and government-supported education programs similar in rigor to degree programs can provide practical education that helps professionals in HI meet the ever-changing needs of the field and continue to upskill. As a case study of a flexible educational program that can accommodate a variety of educational methods and topics, the experience and results of the AHIA program can act as a reference for the global HI education community in the design of integrated educational modules centered on real data and cases.

In conclusion, this study underscores the necessity of flexible, inclusive, and responsive educational models in HI. By addressing the highlighted areas for future research, the field can move toward developing educational programs that not only cater to the diverse needs of students but also prepare them to meet the challenges and opportunities within the dynamic landscape of HI.

Acknowledgments

The authors would like to thank Bitna Kim and Minseo Kang for their help in the preparation of this paper. This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant HR21C0198).

Authors' Contributions

KHL and JHL contributed to the conceptualization. HJJ and JHH contributed to data curation. YL and JSL contributed to the formal analysis. HL and JSL contributed to the methodology. HJJ, KHL, and JHH contributed to the investigation. KHL contributed to writing—original draft. KHL, YL, JHL, and SJ contributed to writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Advanced Health Informatics Analyst course training modules.
[DOC File, 143 KB - mededu_v10i1e54427_app1.doc]

Multimedia Appendix 2
Survey on changes in informatics competency and demand for continuing education among graduates after completing the Advanced Health Informatics Analyst course.
[DOCX File, 59 KB - mededu_v10i1e54427_app2.docx]

Multimedia Appendix 3
Current informatics work sectors of respondents according to their majors.
[PNG File, 48 KB - mededu_v10i1e54427_app3.png]

References

1. Yeung AWK, Torkamani A, Butte AJ, et al. The promise of digital healthcare technologies. Front Public Health 2023 Sep 26;11:1196596. [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
2. Topol E. The Topol review: preparing the healthcare workforce to deliver the digital future. NHS. 2019 Feb. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2023-07-04]

XSL•FO
RenderX

3. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://apps.who.int/iris/handle/10665/344249 [accessed 2023-07-15]
4. Alzghaibi H. The gap between bachelor's degree graduates in health informatics and employer needs in Saudi Arabia. BMC Med Educ 2023 Jun 26;23(1):475. [doi: 10.1186/s12909-023-04442-7] [Medline: 37365545]
5. Chang H. Recent movement on education and training in health informatics. Healthc Inform Res 2014 Apr;20(2):79-80. [doi: 10.4258/hir.2014.20.2.79] [Medline: 24872905]
6. Kinnunen UM, Heponiemi T, Rajalahti E, Ahonen O, Korhonen T, Hyppönen H. Factors related to health informatics competencies for nurses-results of a national electronic health record survey. Comput Inform Nurs 2019 Aug;37(8):420-429. [doi: 10.1097/CIN.0000000000000511] [Medline: 30741730]
7. Monkman H, Mir S, Borycki EM, Courtney KL, Bond J, Kushniruk AW. Updating professional competencies in health informatics: a scoping review and consultation with subject matter experts. Int J Med Inform 2023 Feb;170:104969. [doi: 10.1016/j.ijmedinf.2022.104969] [Medline: 36572000]
8. Feldman SS, Hersh W. Evaluating the AMIA-OHSU 10x10 program to train healthcare professionals in medical informatics. AMIA Annu Symp Proc 2008 Nov 6;2008:182-186. [Medline: 18999199]
9. Butler-Henderson K, Gray K, Pearce C, et al. Exploring the health informatics occupational group in the 2018 Australian Health Information Workforce Census. Stud Health Technol Inform 2019 Aug 8;266:44-50. [doi: 10.3233/SHTI190771] [Medline: 31397300]
10. Bichel-Findlay J, Koch S, Mantas J, et al. Recommendations of the International Medical Informatics Association (IMIA) on education in biomedical and health informatics: second revision. Int J Med Inform 2023 Feb;170:104908. [doi: 10.1016/j.ijmedinf.2022.104908] [Medline: 36502741]
11. Gurrin C, Smeaton AF, Doherty AR. LifeLogging: personal big data. Found Trends Inf Retr 2014 Jun 16;8(1):1-125. [doi: 10.1561/1500000033]
12. McLane TM, Hoyt R, Hodge C, Weinfurter E, Reardon EE, Monsen KA. What industry wants: an empirical analysis of health informatics job postings. Appl Clin Inform 2021 Mar;12(2):285-292. [doi: 10.1055/s-0041-1726423] [Medline: 33792008]
13. Fenton SH, Marc DT, Kennedy A, et al. Aligning the American Health Information Management Association entry-level curricula competencies and career map with industry job postings: cross-sectional study. JMIR Med Educ 2022 Jul 7;8(3):e38004. [doi: 10.2196/38004] [Medline: 35584188]
14. Lee K, Seo L, Yoon D, et al. Digital health profile of South Korea: a cross sectional study. Int J Environ Res Public Health 2022 May 23;19(10):6329. [doi: 10.3390/ijerph19106329] [Medline: 35627866]
15. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data 2022;9(1):3. [doi: 10.1186/s40537-021-00553-4] [Medline: 35013701]
16. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019 Jun 19;6(1):54. [doi: 10.1186/s40537-019-0217-0]
17. Ashrafi N, Kuilboer JP, Joshi C, Ran I, Pande P. Health informatics in the classroom: an empirical study to investigate higher education's response to healthcare transformation. J Inf Syst Educ 2014 Jan 1;25(4):305-316.
18. Digital education for building health workforce capacity. World Health Organization. 2020. URL: https://apps.who.int/iris/handle/10665/331524 [accessed 2023-07-18]
19. Hersh W. The health information technology workforce: estimations of demands and a framework for requirements. Appl Clin Inform 2010 Jun 30;1(2):197-212. [doi: 10.4338/ACI-2009-11-R-0011] [Medline: 23616836]
20. Patel JS, Vo H, Nguyen A, Dzomba B, Wu H. A data-driven assessment of the U.S health informatics programs and job market. Appl Clin Inform 2022 Mar;13(2):327-338. [doi: 10.1055/s-0042-1743242] [Medline: 35354210]
21. Walpole S, Taylor P, Banerjee A. Health informatics in UK medical education: an online survey of current practice. JRSM Open 2016 Dec 1;8(1):2054270416682674. [doi: 10.1177/2054270416682674] [Medline: 28210492]
22. Schedlbauer J, Raptis G, Ludwig B. Medical informatics labor market analysis using web crawling, web scraping, and text mining. Int J Med Inform 2021 Jun;150:104453. [doi: 10.1016/j.ijmedinf.2021.104453] [Medline: 33862508]

Abbreviations

AHIA: Advanced Health Informatics Analyst
HI: health informatics
OR: odds ratio

XSL•FO
RenderX

Edited by B Lesselroth; submitted 09.11.23; peer-reviewed by B Hoyt, J Jiang; revised version received 20.02.24; accepted 20.06.24; published 25.09.24.

Please cite as:
Lee KH, Lee JH, Lee Y, Lee H, Lee JS, Jang HJ, Lee KH, Han JH, Jang S
Impact of Health Informatics Analyst Education on Job Role, Career Transition, and Skill Development: Survey Study
JMIR Med Educ 2024;10:e54427
URL: https://mededu.jmir.org/2024/1/e54427
doi:10.2196/54427

© Kye Hwa Lee, Jae Ho Lee, Yura Lee, Hyunna Lee, Ji Sung Lee, Hye Jeon Jang, Kun Hee Lee, Jeong Hyun Han, SuJung Jang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 25.9.2024. 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.

XSL•FO
RenderX

Original Paper

Evaluating the Impact of the National Health Service Digital Academy on Participants' Perceptions of Their Identity as Leaders of Digital Health Change: Mixed Methods Study

Amish Acharya[1*], MBBS, BSc; Ruth Claire Black[1*], EdD, JD; Alisdair Smithies[1], PhD; Ara Darzi[1], MD, FRCS

Institute of Global Health Innovation, Imperial College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Amish Acharya, MBBS, BSc
Institute of Global Health Innovation
Imperial College London
10th Floor, St Mary's Hospital
Paddington
London, W2 1NY
United Kingdom
Phone: 44 207 886 2125
Email: aa2107@ic.ac.uk

Abstract

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

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

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

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

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

(JMIR Med Educ 2024;10:e46740) doi:10.2196/46740

XSL•FO
RenderX

KEYWORDS
digital leadership; professional identity; dissertation of practice

Introduction

Background

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

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

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

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

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

Aim

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

Methods

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

Recruitment

Participants in the first 2 cohorts of the NHSDA's flagship Digital Health Leadership program were recruited for the study. All participants had completed both years of the program and were >6 months from completion to avoid recency bias. This could involve overemphasizing the impact of later teaching in course compared to that which occurred earlier. Studies suggest that a later evaluation can provide a more holistic evaluation [18]. It also provided time for candidates to reflect on future career opportunities. No other exclusion criteria were placed

XSL•FO
RenderX

upon participants; therefore, a nonprobabilistic sampling method to reach the necessary sample size was used. Eligible participants were contacted through email by a member of the research team (AA) with no direct link to the NHSDA. Both cohorts were impacted by the COVID-19 pandemic, particularly with respect to their dissertation projects that were undertaken during the pandemic.

Scoping Questionnaire

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

Semistructured Interviews

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

Focus Groups

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

Analysis

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

Audio recordings, obtained with the consent of participants for both interviews and focus groups, were transcribed using the web application Descript (Descript, Inc). The accuracy of the outputs was confirmed by one author (AA), who was present in the interviews and focus groups. Anonymized transcripts were then uploaded to the analysis tool MAXQDA (VERBI GmbH). A deductive thematic analysis was conducted using a technique previously used in similar studies [20]. This involved familiarization with the transcripts by 2 authors (RCB and AA). The transcripts were then coded with the data explored to examine the frequency and relationship of the codes. Similar codes were combined into themes and subthemes, which were aligned with the components of professional identity elicited by Chin et al [19]. This review was selected as a framework on which to base the thematic analysis for 3 reasons: first, because of its comprehensive evaluation of identity with 10 evidence-based facets described; second, the examination of internship or workplace-based learning parallels the educational model of the NHSDA's second year; and finally, the authors' examination of how these components map to other contexts can facilitate cross-discipline comparisons was helpful in understanding the participant's identity across wider teams [21]. Although Chin et al [19] found that only a subset of these components was applicable to higher education internships, this study examined the relevance of all components, as some were more significant in postgraduate studies.

As a means of validation, the anonymized transcripts were reviewed again, and themes were amended until a consensus was attained. All discrepancies within the coding exercise or allocations of themes were discussed until resolution. Themes that were consistently mentioned by different participants, those that aligned with findings from the questionnaire or focus group, and those that were regarded as stronger determinants were considered more impactful influences. A constructivist approach

XSL•FO
RenderX

was used as the basis of this study, which paralleled the active learning undertaken throughout the program. The paradigm focuses on the importance of active learning and its transformation through experience [22,23]. It involves the engagement and reflection of the learner, which can be impacted by context, knowledge, motivation, values, or organizational setting [24]. This is particularly pertinent to identity development, which can be influenced by such intrinsic and extrinsic factors.

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

Ethical Considerations

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

Results

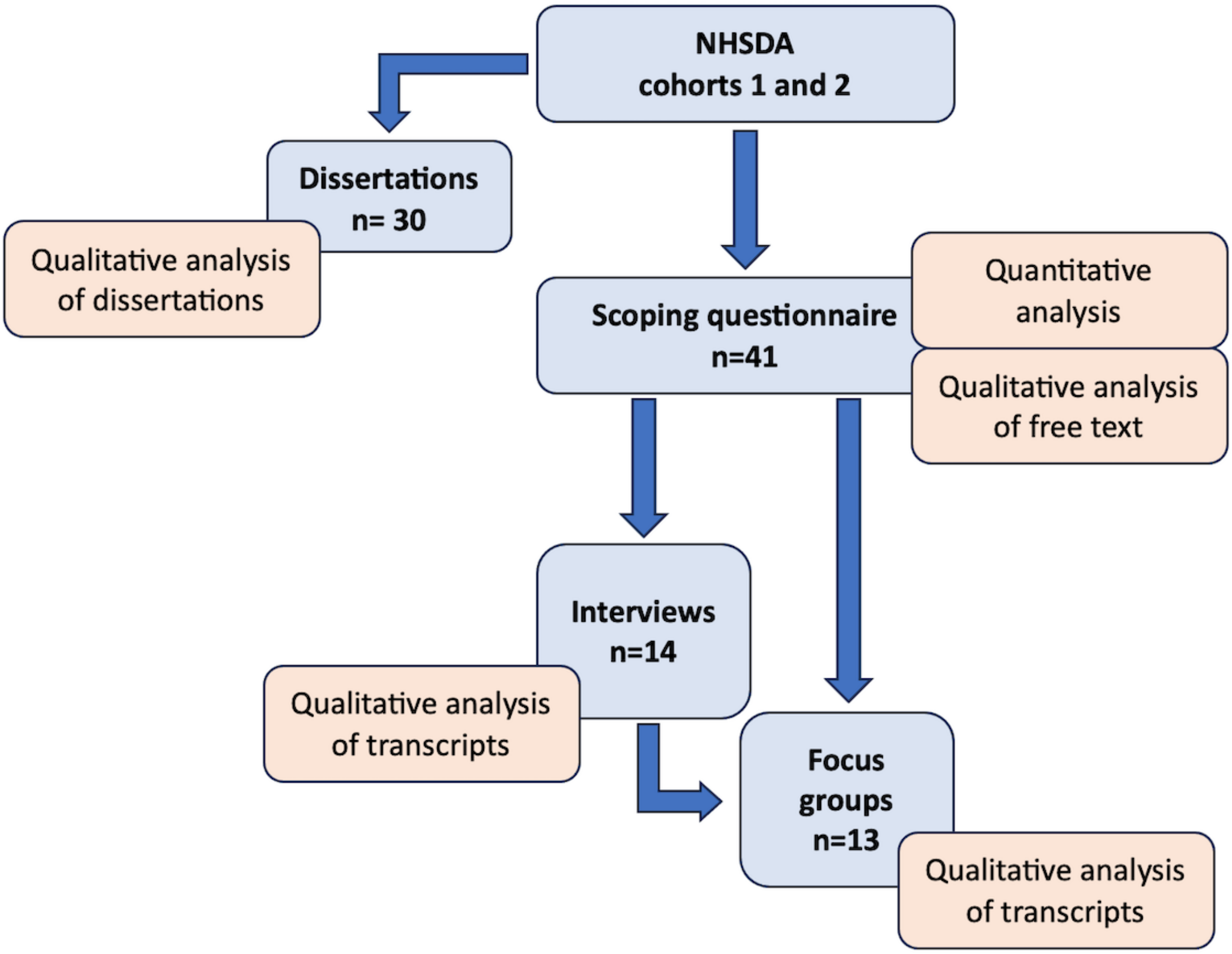
Overview

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

Semistructured interviews were conducted with 34% (14/41) of participants. The demographics of which paralleled those from the wider cohort, with 43% (3/7) of participants identifying as female and 50% (7/14) working in clinical roles. In total, 3 focus groups were held with more than half of the attendees (7/13, 54%) not involved in the preceding interviews. The data sources including the number of participants used in the study is presented in Figure 1.

XSL•FO
RenderX

Figure 1. The flow of participants in the study.



Thematic analysis mapped findings from interviews and focus groups to 8 of the 10 components of professional identity highlighted by Chin et al [19]. Internship experience was not measured, as most of the cohort had been in their roles before the NHSDA and could not be considered entering an internship. However, aspects encompassing skill acquisition during dissertations were covered in other domains. The work environment was also not included, as the participants came from disparate fields, precluding comparisons. However, the findings were mapped to the following domains: commitment to the profession, critical thinking, goal orientation, mentoring, perception of the profession, socialization, reflection, and self-efficacy. When undertaking the thematic analysis and mapping of the highest-scoring dissertation to the components from the framework by Chin et al [19], only 4 were found to be applicable. These included critical thinking, goal orientation, mentoring, and reflection. This is likely because the dissertation was more descriptive of a specific transformation project rather than reflective of the attitudes of participants toward digital health leadership as a whole.

Table 1 demonstrates how the course impacted these areas through quotations from the respondents.

XSL•FO
RenderX

Table 1. Key domains[a,b] of professional identity with quotes from participants on the impact of the course.

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

[a]Work environment not included as relevant components covered in "mentoring," and participants came from disparate environments.

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

Commitment to the Profession

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

XSL•FO
RenderX

also appeared to reaffirm participants' motivation for undertaking digital health leadership roles. One CCIO stated:

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

Critical Thinking

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

Providing a supportive environment for change enabled candidates to put theoretical learning into practice. An interviewee said "I [they] approach things differently, I'm [they are] more strategic, more constructed after the project." These cognitive skills have continued postgraduation with individuals feeling they have "different tools that were picked up during the academy, which I [they] use day-to-day."

XSL•FO
RenderX

Table 2. Topics of the highest-scoring dissertations of the 2 cohorts and components of identity that were incorporated.

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

[a]BYOD: bring your own device.

XSL•FO
RenderX

[b]NHS: National Health Service.

[c]GP: general practitioner.

[d]SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

[e]EHR: electronic health record.

[f]AI: artificial intelligence.

[g]EPR: electronic patient record.

Goal Orientation

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

Mentoring

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

Perception of the Profession

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

Professional Socialization

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

Reflection

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

Self-Efficacy

The development of self-efficacy was found to be a key tenet of the program, with 61% (25/41) of the respondents reporting that the course had positively increased their confidence in their role. One candidate noted that "When I [they] first went on the digital academy there was an element of imposter syndrome,"

XSL•FO
RenderX

being told, "not to think about imposters, you need to think as pioneers." Others had reflected that the digital academy had given them "the confidence to lead in digital" and "empowerment...to recognize that I [they] have the ability to do anything I [they] put my [their] mind to." This has led to several candidates being recognized as leaders within the wider digital health ecosystem, but not necessarily in their own organizations. One clinician noted that they "had taken up a few national unpaid roles"; however, another noted that "they [the director] was not interested...did not recognize the training we had."

Discussion

Principal Findings

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

Professional identity formation has become the focus of a diverse range of fields, including medical education [26]. Among health care professionals, studies have shown that the development of a shared core value set can have substantial benefits, including improving the well-being and resilience of physicians [27]. Professional identity's influence in other areas is less well documented, but some benefits may be appreciable across a range of disciplines. In a study by Meadows and de Braine [28], industry leaders displayed stronger leadership identities during the COVID-19 pandemic to help overcome challenges such as the implementation of new technologies. Digital health leadership teams who were faced with comparable issues are likely to have also relied on stronger leadership identities during this time. Consequently, there is increasing interest in understanding how educational programs can foster professional identity in their cohorts. Some have suggested that to develop identity requires departing from traditional pedagogy and using greater participatory or sociocultural learning opportunities [29]. The NHSDA program uses a mixed approach, blending didactic learning with collaborative work in the first year and a dissertation of the practice model in the second year. Therefore, a breadth of impact, both intended and unintended, upon the identity of participants as digital health leaders were noted.

A principle focus of the course and the dissertation project is on reflective practice. Therefore, it was not surprising that 80% (24/30) of the projects incorporated these skills. Reflective exercises are an important component of professional identity and can help leaders hone their metacognitive and inductive reasoning skills. Studies have shown that through these processes, learners can also identify their own cognitive biases and avoid errors [30]. Moreover, the dissertation was also noted to foster critical thinking, with several participants reporting that they had become more "strategic." Critical thinking is a higher-level cognitive skill in which individuals understand phenomena through their interpretation and inference of contributory factors and variables. Critical thinking enables learners to become more agile [31]. Given that the digital health landscape is continuing to evolve in the United Kingdom following the pandemic and the continued challenge of resource allocation, an ability to acclimatize to these newer contexts would appear integral. In fact, when asked directly "What is a digital health leader?" several respondents referred to this adaptability, noting the need for "fearlessness, curiosity, and being comfortable going into unknown territories."

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

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

XSL•FO
RenderX

of role switching after graduation. Future work should look to examine these findings as well as how accreditation from courses such as the NHSDA can impact organizational buy-in.

Limitations

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

Conclusions

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

Acknowledgments

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

Data Availability

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

Authors' Contributions

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

Conflicts of Interest

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

Multimedia Appendix 1
Survey results and interview topic guide.
[DOCX File , 46 KB - mededu_v10i1e46740_app1.docx]

References

1. What good looks like framework. National Health Service England. 2021 Aug 31. URL: https://transform.england.nhs.uk/digitise-connect-transform/what-good-looks-like/what-good-looks-like-publication/ [accessed 2022-12-09]

XSL•FO
RenderX

2. Building strong integrated care systems everywhere. National Health Service. 2021 Sep 2. URL: https://www.england.nhs.uk/wp-content/uploads/2021/06/B0664-ics-clinical-and-care-professional-leadership.pdf [accessed 2024-02-13]
3. A plan for digital health and social care. United Kingdom Government. 2022 Jun 29. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2023-12-09]
4. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug 07;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
5. Murphy M, Scott LJ, Salisbury C, Turner A, Scott A, Denholm R, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021 Jan 17;71(704):e166-e177. [doi: 10.3399/bjgp.2020.0948]
6. Clarke J, Flott K, Fernandez Crespo R, Ashrafian H, Fontana G, Benger J, et al. Assessing the safety of home oximetry for COVID-19: a multisite retrospective observational study. BMJ Open 2021 Sep 14;11(9):e049235 [FREE Full text] [doi: 10.1136/bmjopen-2021-049235] [Medline: 34521666]
7. Online version of the NHS Long Term Plan. National Health Service. URL: https://www.longtermplan.nhs.uk/online-version/ [accessed 2021-07-04]
8. Digital Health Leadership Programme (DHLP). National Health Service England. URL: https://digital-transformation.hee.nhs.uk/digital-academy/programmes/digital-health-leadership-programme/digital-health-leadership-programme/eligibility [accessed 2023-10-14]
9. NHS digital academy. National Health Service England. URL: https://www.hee.nhs.uk/our-work/nhs-digital-academy [accessed 2021-07-03]
10. Price-Dowd C, Edwards M, Carter A, Hogan C, Martin A, Johnson K. NHS Digital Academy – evaluation scoping report. National Health Service England. 2019 Sep. URL: https://digital-transformation.hee.nhs.uk/binaries/content/assets/digital-transformation/nhs-digital-academy/evaluation-of-the-digital-academy---scoping-report---nov-2019.pdf [accessed 2024-02-13]
11. Acharya A, Black RC, Smithies A, Darzi A. Evaluating the impact of a digital leadership programme on national digital priorities: a mixed methods study. BMJ Open 2022 Apr 29;12(4):e056369 [FREE Full text] [doi: 10.1136/bmjopen-2021-056369] [Medline: 35487747]
12. Matthews J, Bialocerkowski A, Molineux M. Professional identity measures for student health professionals - a systematic review of psychometric properties. BMC Med Educ 2019 Aug 13;19(1):308 [FREE Full text] [doi: 10.1186/s12909-019-1660-5] [Medline: 31409410]
13. Marlowe JM, Appleton C, Chinnery SA, Van Stratum S. The integration of personal and professional selves: developing students' critical awareness in social work practice. Soc Work Educ 2014 Aug 22;34(1):60-73. [doi: 10.1080/02615479.2014.949230]
14. Cornett M, Palermo C, Ash S. Professional identity research in the health professions-a scoping review. Adv Health Sci Educ Theory Pract 2023 May 09;28(2):589-642 [FREE Full text] [doi: 10.1007/s10459-022-10171-1] [Medline: 36350489]
15. Moseley LE, McConnell L, Garza KB, Ford CR. Exploring the evolution of professional identity formation in health professions education. New Dir Teach Learn 2021 Dec 06;2021(168):11-27. [doi: 10.1002/tl.20464]
16. Spehar I, Frich JC, Kjekshus LE. Professional identity and role transitions in clinical managers. J Health Org Manag 2015;29(3):353-366 [FREE Full text] [doi: 10.1108/JHOM-03-2013-0047] [Medline: 25970529]
17. Gordon D, Achuck K, Kempner D, Jaffe R, Papanagnou D. Toward unity and inclusion in the clinical workplace: an evaluation of healthcare workforce belonging during the COVID-19 pandemic. Cureus 2022 Sep;14(9):e29454 [FREE Full text] [doi: 10.7759/cureus.29454] [Medline: 36312604]
18. Dickey D, Pearson C. Recency effect in college student course evaluations. Pract Assess Res Eval 2019 Nov;10(6) [FREE Full text] [doi: 10.7275/8fdy-vr38]
19. Chin D, Phillips Y, Woo MT, Clemans A, Yeong PK. Key components that contribute to professional identity development in internships for Singapore's tertiary institutions: a systematic review. Asian J Scholarsh Teach Learn 2020;10:89-113 [FREE Full text]
20. Blum ER, Stenfors T, Palmgren PJ. Benefits of massive open online course participation: deductive thematic analysis. J Med Internet Res 2020 Jul 08;22(7):e17318 [FREE Full text] [doi: 10.2196/17318] [Medline: 32672680]
21. Mylrea MF, Sen Gupta T, Glass BD. Developing professional identity in undergraduate pharmacy students: a role for self-determination theory. Pharmacy (Basel) 2017 Mar 24;5(2):16 [FREE Full text] [doi: 10.3390/pharmacy5020016] [Medline: 28970428]
22. Kolb DA, Fry R. Towards an applied theory of experiential learning. In: Cooper C, editor. Theories of Group Process. London, UK: John Wiley & Sons; 1975.
23. Lockey A, Conaghan P, Bland A, Astin F. Educational theory and its application to advanced life support courses: a narrative review. Resusc Plus 2021 Mar;5:100053 [FREE Full text] [doi: 10.1016/j.resplu.2020.100053] [Medline: 34223327]
24. Tsai CA, Song MY, Lo YF, Lo CC. Design thinking with constructivist learning increases the learning motivation and wicked problem-solving capability—an empirical research in Taiwan. Think Skills Creat 2023 Dec;50:101385. [doi: 10.1016/j.tsc.2023.101385]

XSL•FO
RenderX

25. Preparing the healthcare workforce to deliver the digital future. National Health Service. 2019 Feb. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2024-02-13]
26. Findyartini A, Greviana N, Felaza E, Faruqi M, Zahratul Afifah T, Auliya Firdausy M. Professional identity formation of medical students: a mixed-methods study in a hierarchical and collectivist culture. BMC Med Educ 2022 Jun 08;22(1):443 [FREE Full text] [doi: 10.1186/s12909-022-03393-9] [Medline: 35676696]
27. Toubassi D, Schenker C, Roberts M, Forte M. Professional identity formation: linking meaning to well-being. Adv Health Sci Educ Theory Pract 2023 Mar 01;28(1):305-318 [FREE Full text] [doi: 10.1007/s10459-022-10146-2] [Medline: 35913664]
28. Meadows S, de Braine R. The work identity of leaders in the midst of the COVID-19 pandemic. Front Psychol 2022 Sep 26;13:958679 [FREE Full text] [doi: 10.3389/fpsyg.2022.958679] [Medline: 36225677]
29. Orsmond P, McMillan H, Zvauya R. It's how we practice that matters: professional identity formation and legitimate peripheral participation in medical students: a qualitative study. BMC Med Educ 2022 Feb 09;22(1):91 [FREE Full text] [doi: 10.1186/s12909-022-03107-1] [Medline: 35139839]
30. Mamede S, Schmidt HG, Rikers R. Diagnostic errors and reflective practice in medicine. J Eval Clin Pract 2007 Feb 24;13(1):138-145 [FREE Full text] [doi: 10.1111/j.1365-2753.2006.00638.x] [Medline: 17286736]
31. Kasalaei A, Amini M, Nabeiei P, Bazrafkan L, Mousavinezhad H. Barriers of critical thinking in medical students' curriculum from the viewpoint of medical education experts: a qualitative study. J Adv Med Educ Prof 2020 Apr;8(2):72-82 [FREE Full text] [doi: 10.30476/jamp.2020.83053.1080] [Medline: 32426391]
32. Horsburgh J, Ippolito K. A skill to be worked at: using social learning theory to explore the process of learning from role models in clinical settings. BMC Med Educ 2018 Jul 03;18(1):156 [FREE Full text] [doi: 10.1186/s12909-018-1251-x] [Medline: 29970052]
33. Bandura A. Social cognitive theory of self-regulation. Organ Behav Hum Decis Process 1991 Dec;50(2):248-287. [doi: 10.1016/0749-5978(91)90022-L]
34. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020 Jun 16;22(6):e19284 [FREE Full text] [doi: 10.2196/19284] [Medline: 32501804]

Abbreviations

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

Edited by T de Azevedo Cardoso, SR Mogali; submitted 23.02.23; peer-reviewed by E Stubbing, M Aanestad, S Ashraf; comments to author 10.07.23; revised version received 14.10.23; accepted 31.01.24; published 21.02.24.

Please cite as:
Acharya A, Black RC, Smithies A, Darzi A
Evaluating the Impact of the National Health Service Digital Academy on Participants' Perceptions of Their Identity as Leaders of Digital Health Change: Mixed Methods Study
JMIR Med Educ 2024;10:e46740
URL: https://mededu.jmir.org/2024/1/e46740
doi:10.2196/46740
PMID:38381477

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

XSL•FO
RenderX

Use of Multiple-Choice Items in Summative Examinations: Questionnaire Survey Among German Undergraduate Dental Training Programs

Lena Rössler[1]; Manfred Herrmann[2], Dr rer nat; Annette Wiegand[1], Prof Dr med dent; Philipp Kanzow[1], MSc, Dr rer medic, PD Dr med dent

1

2

Corresponding Author:
Philipp Kanzow, MSc, Dr rer medic, PD Dr med dent

Abstract

Background: Multiple-choice examinations are frequently used in German dental schools. However, details regarding the used item types and applied scoring methods are lacking.

Objective: This study aims to gain insight into the current use of multiple-choice items (ie, questions) in summative examinations in German undergraduate dental training programs.

Methods: A paper-based 10-item questionnaire regarding the used assessment methods, multiple-choice item types, and applied scoring methods was designed. The pilot-tested questionnaire was mailed to the deans of studies and to the heads of the Department of Operative/Restorative Dentistry at all 30 dental schools in Germany in February 2023. Statistical analysis was performed using the Fisher exact test ($P<.05$).

Results: The response rate amounted to 90% (27/30 dental schools). All respondent dental schools used multiple-choice examinations for summative assessments. Examinations were delivered electronically by 70% (19/27) of the dental schools. Almost all dental schools used single-choice Type A items (24/27, 89%), which accounted for the largest number of items in approximately half of the dental schools (13/27, 48%). Further item types (eg, conventional multiple-select items, Multiple-True-False, and Pick-N) were only used by fewer dental schools (≤67%, up to 18 out of 27 dental schools). For the multiple-select item types, the applied scoring methods varied considerably (ie, awarding [intermediate] partial credit and requirements for partial credit). Dental schools with the possibility of electronic examinations used multiple-select items slightly more often (14/19, 74% vs 4/8, 50%). However, this difference was statistically not significant ($P=.38$). Dental schools used items either individually or as key feature problems consisting of a clinical case scenario followed by a number of items focusing on critical treatment steps (15/27, 56%). Not a single school used alternative testing methods (eg, answer-until-correct). A formal item review process was established at about half of the dental schools (15/27, 56%).

Conclusions: Summative assessment methods among German dental schools vary widely. Especially, a large variability regarding the use and scoring of multiple-select multiple-choice items was found.

(JMIR Med Educ 2024;10:e58126) doi:10.2196/58126

KEYWORDS

alternate-choice; assessment; best-answer; dental; dental schools; dental training; education; educational assessment; educational measurement; examination; German; Germany; k of n; Kprim; K’; medical education; medical student; MTF; Multiple-True-False; multiple choice; multiple-select; Pick-N; scoring; scoring system; single choice; single response; test; testing; true/false; true-false; Type A; Type K; Type K’; Type R; Type X; undergraduate; undergraduate curriculum; undergraduate education

Introduction

Summative examinations of theoretical knowledge are an integral part of university degree programs. As they are intended to assess examinees’ ability regarding predefined learning objectives, they should reflect examinees’ true knowledge as closely as possible. To assess examinees objectively and efficiently, multiple-choice examinations were described as early as 1916 [1,2]. To date, these types of examinations have been expanded by further item types, and multiple-choice examinations are frequently used within higher education including but not limited to dental training programs [3-5]. Multiple-choice items (ie, questions) can be subdivided into single-choice items (eg, Type A, Type K, Type R, and alternate-choice) and multiple-select items (eg, Pick-N and Multiple-True-False [Type K’]) [6]. While dichotomous scoring (ie, 1 full credit point is awarded if examinees mark the correct

XSL•FO
RenderX

answer option or statements, otherwise no credit is awarded) is most commonly proposed for single-choice items [7], scoring methods for multiple-select items are more heterogeneous: Besides dichotomous scoring, further scoring methods resulting in (intermediate) partial credit or even negative points (ie, malus points) have been described [8,9].

Besides paper-based examinations, examinations are nowadays frequently delivered electronically. While electronic examinations are well perceived by examinees [10], comprehensive studies regarding their effectiveness are still lacking [11]. However, the use of different examination software (eg, UCAN's [Umbrella Consortium for Assessment Networks] CAMPUS examination software) might improve the ease of multiple-choice examinations, accelerate the evaluation of examinations and item analysis, and allow for more complex scoring algorithms. Despite the benefits associated with electronic examinations, the availability of hardware and software at the level of individual institutions might limit its use.

In Germany, the revised undergraduate dental curriculum consists of 10 semesters and includes preclinical training (4 semesters), training using simulators or phantom heads (2 semesters), and clinical training (4 semesters). Following the state examinations after each part (ie, after the fourth, sixth, and 10th semester), students receive their license ("Approbation") to practice dentistry. Besides practical skills, theoretical knowledge is taught within the undergraduate dental curriculum, and students' ability is often assessed using written multiple-choice examinations. However, such examinations are not standardized among German dental schools. While general recommendations exist for their design and evaluation [12,13], details such as suitable item types and applied scoring methods are often defined in local examination guidelines at the level of individual dental schools. However, these details might impact examinees' scoring results [5]. To the best of our knowledge, a comprehensive overview regarding the used item types and applied scoring methods at German dental schools does not exist.

Therefore, this study aimed to gain insight into the current use of multiple-choice items in summative examinations in German undergraduate dental training programs. The null hypothesis is that the use of digital examinations does not impact the use of more complex (ie, multiple-select) multiple-choice items.

Methods

Ethical Considerations

The study was designed as a prospective investigation. In preparation for the investigation, the websites of all German dental schools were screened (n=30), and the names of the heads of the Department of Operative/Restorative Dentistry and the deans of studies were noted for later procedures.

The study was performed after approval by the local ethics committee of the University Medical Center Göttingen (approval number 22/1/23). Participation in this study was voluntary, and participants gave their informed consent for the anonymous evaluation of the provided answers by returning the questionnaires. Participants did not receive any incentives or compensation.

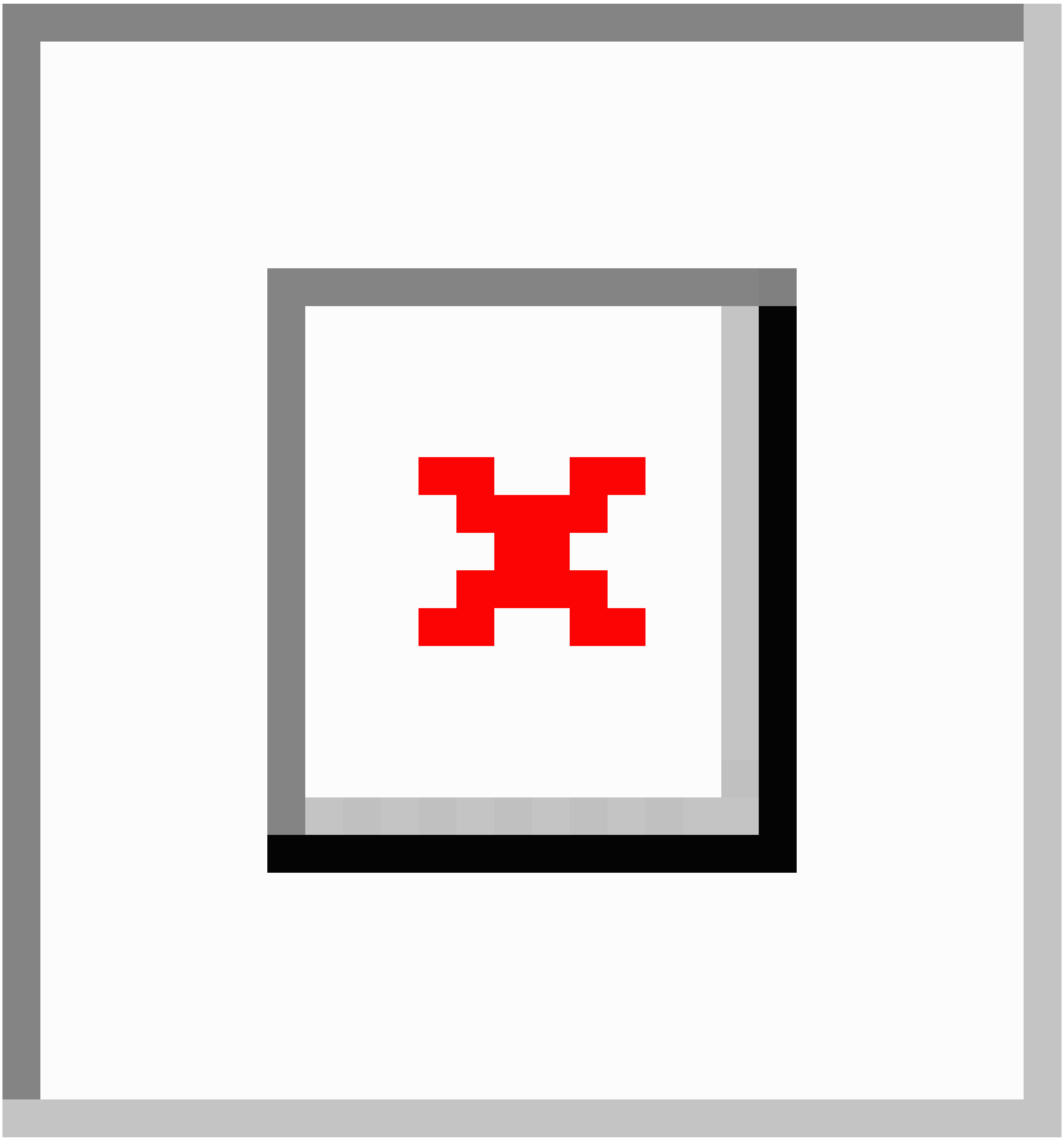
Questionnaire

A paper-based questionnaire, consisting of 10 items about the construction and evaluation of summative examinations, was jointly designed by the authors and pilot-tested in the University Medical Center Göttingen (Multimedia Appendix 1). Both closed and open-ended items were used. The opening questions related to different examination types used for the summative assessment of theoretical knowledge, and whether or not electronic examinations were being used. Additionally, it was asked whether the examination items undergo a formal review process and if so, the participants had the chance to give a brief description of this procedure. The more specific questions related to the types of multiple-choice items used and asked for the relative percentage to which these items were being used. Furthermore, the participants were asked to describe the applied scoring methods for each of the item types used. Finally, participants were provided with a text field open for comments and their contact details (ie, if required for further clarification) and were asked to supply a copy of their local examination guidelines or program regulations.

Following the evaluation of the pilot survey among 5 dentists at the University Medical Center Göttingen, the questionnaire was slightly modified for clarification, printed, and mailed to (1) the heads of the Department of Operative/Restorative Dentistry and to (2) the deans of studies on February 1, 2023. The wording was slightly adjusted for each recipient: (1) "used in your department" versus (2) "permitted at your dental school". Mailings included a personalized cover letter, an overview illustrating different multiple-choice item types (Figure 1), and a stamped return envelope. The survey was closed after 12 weeks. Nonresponders were reminded once 6 weeks after the initial distribution of the questionnaires.

XSL•FO
RenderX

Figure 1. Exemplary presentation of the most commonly used multiple-choice item types referenced in the questionnaire. Round marking boxes represent 1 answer option to be selected (1 out of x), while square marking boxes imply that multiple answer options or statements (x out of X) can be chosen.



Statistical Analysis

First, data were manually transferred into a digital chart using a piloted spreadsheet containing columns for each item of the questionnaire. This step was independently performed by 2 authors (LR and PK). In case of disagreement, data were repeatedly extracted from the returned questionnaires.

In case of disagreement between the heads of the Department of Operative/Restorative Dentistry and the deans of studies, results were based on the responses from the heads of the Department of Operative/Restorative Dentistry. For further clarification, responses were cross-validated with the supplied or publicly available examination guidelines and program regulations. If required, respondents were contacted for further clarification if they had agreed to do so previously.

Second, statistical analysis was performed using the software SPSS Statistics (Macintosh version 29.0.0.0; IBM Corp). The effect of delivering digital examinations on the use of multiple-select items was assessed using the Fisher exact test. The level of significance was set at .05.

XSL•FO
RenderX

Results

Overview

In total, responses from 27 dental schools were received yielding a response rate of 90% (27/30 dental schools). More specifically, 25 Departments of Operative/Restorative Dentistry and 17 deans of studies replied. All dental schools responded that they use written multiple-choice examinations for the assessment of examinees' theoretical knowledge. Therefore, subsequent results are based on the number of respondent dental schools.

Multiple-Choice Items Used

The most commonly used multiple-choice item types at German dental schools were single-choice Type A or Type $A_{negative}$ items with 3 to 6 answer options (24/27, 89%). Pick-N items (ie, the number of answer options to be selected is known to examinees) were reported to contain between 3 and 26 answer options and were used by 67% (18/27) of dental schools. Type K items were reported to contain between 3 and 6 statements and were used by 52% (14/27) of the dental schools. Multiple-True-False (also known under further names such as Kprim, Type K', or Type X) and conventional multiple-select items (ie, the number of answer options to be selected is unknown to examinees) were reported to contain between 4 and 6 statements or answer options and were both used by 44% (12/27) of the dental schools. The use of further item types is shown in Table 1.

Table . Different multiple-choice item types for the assessment of theoretical knowledge at the respondent dental schools (N=27).

Item type	Dental schools, n (%)
Type A	24 (89)
Pick-N	18 (67)
Type K	14 (52)
Conventional multiple-select	12 (44)
Multiple-True-False (Type K')	12 (44)
Type R	6 (22)
Alternate-choice	4 (15)

Examination Setting

Key feature problems consisting of a clinical case scenario followed by a number of items focusing on critical treatment steps were used by approximately half of the dental schools (15/27, 56%). Not a single school used alternative testing methods (eg, answer-until-correct). Also, a formal item review process prior to the delivery of the examination was only established at about half of the dental schools (15/27, 56%).

Delivery of Examinations

The percentage of dental schools that deliver examinations electronically amounted to 70% (19/27). However, the software used by the dental schools differed: a dedicated examination software (ie, UCAN's CAMPUS or tEXAM, Q-Exam [IQUL GmbH]) was used by 8 dental schools, while learning management systems such as Moodle (Moodle Pty Ltd), ILIAS (ILIAS open source e-Learning e.V.), or OpenOLAT (frentix GmbH) were used by 7 dental schools for the purpose of examination delivery. The remaining 4 dental schools did not provide any information regarding the examination software they used.

Dental schools with the possibility of electronic examinations used multiple-select items slightly more often (14/19, 74% vs 4/8, 50%). However, this difference was statistically not significant (P=.38).

Applied Scoring Methods

All dental schools scored single-choice items (ie, Type A, Type $A_{negative}$, Type K, Type R, and alternate-choice) dichotomously (ie, 1 full credit point is awarded if examinees mark the correct answer option or statements, otherwise no credit is awarded).

Scoring of multiple-select items was more heterogeneous and no single scoring method that was commonly used was identified: some dental schools used scoring algorithms resulting in partial (ie, 0.5 credit points) or intermediate partial credit (ie, 1/n partial credit for each correct response) besides dichotomous scoring on multiple-select items. However, scoring methods resulting in negative points (ie, malus points) were not used at any location.

Discussion

Principal Findings

The aim of this study was to gain insight into summative assessment methods that involve the use of multiple-choice items and are used at German dental schools. The purpose of summative assessment is to evaluate examinees' knowledge at the end of a course by comparing their scores to a predefined standard (ie, cutoff score) [14]. Our results demonstrate that all respondent dental schools use multiple-choice examinations for summative assessment of theoretical knowledge. Besides individual items, approximately half of the dental schools also use key feature problems.

Single-choice Type A items are the most popular item types used at German dental schools. These items are used by almost every respondent dental school and often account for the largest number of items at the respective dental schools. This might be explained by the demand for ease of scoring (ie, dichotomous scoring, no partially correct responses).

XSL•FO
RenderX

Multiple-select item types such as Pick-N or Multiple-True-False are used by fewer dental schools. For these item types, the applied scoring methods vary considerably: Some dental schools award partial or even intermediate partial credit for partially correct responses while others do not. However, the exact cutoff levels and scoring methods for partial credit differed. For example, Partial Scoring 50% (PS_{50}) was used by some dental schools for Pick-N items: In these cases, 1 full credit point is awarded if all answer options are marked correctly, and 0.5 credit points are awarded if at least half of the true answer options are marked, otherwise no credit is awarded [9,15]. Furthermore, a similar scoring method named Half-point Scoring was used by some dental schools for Multiple-True-False and conventional multiple-select items: 1 full credit point is awarded if all statements or answer options are marked correctly, 0.5 credit points are awarded if the response to 1 statement or answer option is incorrect, otherwise no credit is awarded [8,16]. In addition, some dental schools awarded intermediate partial credit on multiple-select items: In the case of Partial Scoring 1/n ($PS_{1/n}$), 1/n partial credit was awarded for each correct response [8,9]. Some dental schools also subtracted 1/n partial credit for each incorrect response (Blasberg-Method) [8,9,17].

As a result, the scoring of multiple-select items at different German dental schools can be considered very heterogeneous. This is not surprising, as a vast number of different scoring methods for multiple-select items have been described in the literature [8,9]. As stated previously, it is not possible to suggest a single versatile scoring method. Different requirements as defined in dental schools' local examination guidelines (eg, fixed pass-mark and fixed proportion of true answer options) impact the scoring method to be selected. Regarding jurisdictional requirements, scoring methods resulting in negative points (ie, malus points) must not be used in Germany [13]. Consequently, not a single dental school uses scoring methods resulting in malus points. However, almost half of the dental schools do not use a formal item review process. A formal review process is recommended prior to the delivery of the examinations and might further improve the quality and overall validity of the examinations.

In addition, 70% (19/27) of all dental schools stated to deliver examinations electronically. While the electronic delivery of examinations allows for automatic scoring and more complex scoring methods (ie, within the context of multiple-select items), no statistically significant relation between the type of delivery (paper-based vs electronic) and the use of multiple-select item types was found. Therefore, our results fail to reject the null hypothesis. This might be explained by the software used for the delivery and scoring of electronic examinations: it was found that dental schools use learning management systems such as Moodle, ILIAS, or OpenOLAT besides dedicated examination software such as UCAN's CAMPUS, UCAN's tEXAM, or Q-Exam for the delivery and scoring of summative assessments. This is of relevance, as learning management systems usually support fewer item types and scoring methods than dedicated examination software [8,9]. As a result, electronic delivery of examinations does not necessarily result in an increased use of multiple-select items.

Interestingly, not a single dental school used alternative testing methods that deviate from the standard setting during examinations (ie, examinees mark the answer options or statements they believe to be correct or true but receive no immediate feedback regarding correctly or incorrectly marked answer options or statements). Within multiple-choice examinations, alternative testing methods such as confidence weighting scoring (ie, examinees are requested to indicate the degree of confidence in their marking) [18], elimination scoring (ie, examinees are instructed to mark the incorrect instead of correct answer options) [19], or answer-until-correct [20,21] have been described in the literature. Within the answer-until-correct method, examinees receive immediate feedback and examinees may correct their marking on previously incorrectly marked items, thereby still receiving partial credit. However, the benefit of such testing methods within the field of dental education is questionable. Dental school examinees are becoming future dentists. While treating patients, dentists are required to make informed choices and dentists might not always have a second chance without potentially harming their patients. In addition, such alternative testing methods benefit from the electronic delivery of examinations and set even higher requirements for the used examination software.

Strengths and Limitations

To the best of our knowledge, this is the first study to systematically assess the use and scoring of multiple-choice item types in summative examinations among German dental schools. A number of strengths are present. First, a pretested questionnaire was used. Second, our questionnaire survey study yielded a high response rate of 90% (27/30 dental schools). Third, our results might be considered representative of the current use of multiple-choice items in summative examinations among German dental schools.

Nevertheless, limitations are also present. First, our questionnaire focused on multiple-choice items; therefore, the use of other assessment types (eg, objective structured clinical examinations, oral examinations) remains unknown. Second, the number of dental schools in Germany is limited. Thereby, results from the Fisher exact test might be underpowered despite the high response rate. Furthermore, this study could not control for potential confounders (eg, location, number of students per dental school) due to the overall low number of dental schools. Third, transferability and generalizability to other educational settings might be limited due to different jurisdictional requirements or the overall lower importance of written examinations.

Future Directions

New dental licensing regulations ("Approbationsordnung") have been in effect since 2021, which restructured the undergraduate dental curriculum in Germany. For the first time, a nationwide written board examination with single-choice items takes place at the end of all undergraduate dental programs (ie, after the 10th semester) [22]. Therefore, multiple-choice examinations in general and especially single-choice Type A items will remain a popular format for summative examinations among German undergraduate dental programs. Ideally,

XSL•FO
RenderX

examinees already become familiar with single-choice Type A items during their studies. Therefore, all dental schools should use single-choice Type A items to adequately prepare their students for the final board examination.

Nevertheless, additional examinations (eg, objective structured clinical or practical examinations) are required to test examinees' practical skills [3]. Regardless of the used item type, multiple-choice examinations are not suitable to assess the higher levels Miller's Pyramid of clinical competence (ie, does and shows how) [23].

Conclusion

While students from almost all dental schools can be expected to be familiar with single-choice Type A items, techniques for the summative assessment of theoretical knowledge differ widely among German dental schools. Especially, a large variability regarding the use and scoring of multiple-select multiple-choice items was found. In addition, implementing a formal item review process might further improve the quality and overall validity of the examinations.

Acknowledgments

The authors acknowledge support from the Open Access Publication Funds of Göttingen University. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Generative artificial intelligence was not used in any portion of this manuscript.

Data Availability

All data generated during or analyzed during this study are included in this published article.

Authors' Contributions

LR, MH, AW, and PK contributed to the study's conception and designed the questionnaire. LR and PK transferred the data. PK performed statistical analyses. All authors interpreted the data, critically revised the manuscript, and approved the final version of the manuscript.

Conflicts of Interest

PK is an associate editor of *JMIR Medical Education* at the time of this publication. Other authors have no competing interests to declare.

Multimedia Appendix 1
Authors' translation of the used questionnaire, which was originally distributed in German.
[PDF File, 247 KB - mededu_v10i1e58126_app1.pdf]

References

1. Kelly FJ. The Kansas silent reading tests. J Educ Psychol 1916 Feb;7(2):63-80. [doi: 10.1037/h0073542]
2. Ruch GM, Stoddard GD. Comparative reliabilities of five types of objective examinations. J Educ Psychol 1925 Mar;16(2):89-103. [doi: 10.1037/h0072894]
3. Gerhard-Szep S, Güntsch A, Pospiech P, et al. Assessment formats in dental medicine: an overview. GMS J Med Educ 2016 Aug;33(4):Doc65. [doi: 10.3205/zma001064] [Medline: 27579365]
4. Kanzow P, Schuelper N, Witt D, et al. Effect of different scoring approaches upon credit assignment when using Multiple True-False items in dental undergraduate examinations. Eur J Dent Educ 2018 Nov;22(4):e669-e678. [doi: 10.1111/eje.12372] [Medline: 29934980]
5. Kanzow P, Schmidt D, Herrmann M, Wassmann T, Wiegand A, Raupach T. Use of multiple-select multiple-choice items in a dental undergraduate curriculum: retrospective study involving the application of different scoring methods. JMIR Med Educ 2023 Mar 27;9:e43792. [doi: 10.2196/43792] [Medline: 36841970]
6. Krebs R. Prüfen mit Multiple Choice: Kompetent planen, entwickeln, durchführen und auswerten [Testing With Multiple Choice: Plan, Develop, Implement, and Evaluate Competently]: Hogrefe; 2019.
7. Kanzow AF, Schmidt D, Kanzow P. Scoring single-response multiple-choice items: scoping review and comparison of different scoring methods. JMIR Med Educ 2023 May 19;9:e44084. [doi: 10.2196/44084] [Medline: 37001510]
8. Schmidt D, Raupach T, Wiegand A, Herrmann M, Kanzow P. Relation between examinees' true knowledge and examination scores: systematic review and exemplary calculations on Multiple-True-False items. Educ Res Rev 2021 Nov;34:100409. [doi: 10.1016/j.edurev.2021.100409]
9. Schmidt D, Raupach T, Wiegand A, Herrmann M, Kanzow P. Relation between examinees' true knowledge and examination scores: systematic review and exemplary calculations on Pick-N items. Educ Res Rev 2022 Nov;37:100483. [doi: 10.1016/j.edurev.2022.100483]

XSL•FO
RenderX

10. Nardi A, Ranieri M. Comparing paper‐based and electronic multiple‐choice examinations with personal devices: impact on students' performance, self‐efficacy and satisfaction. Brit J Educ Tech 2019 May;50(3):1495-1506. [doi: 10.1111/bjet.12644]
11. Way WD, Davis LL, Keng L, Strain-Seymour E. Increasing the accessibility of assessments through technology. In: Drasgow F, editor. Technology and Testing: Improving Educational and Psychological Measurement: Routledge; 2015:217-234.
12. Jünger J, Just I. Recommendations of the German Society for Medical Education and the German Association of Medical Faculties regarding university-specific assessments during the study of human, dental and veterinary medicine. GMS Z Med Ausbild 2014 Aug;31(3):Doc34. [doi: 10.3205/zma000926] [Medline: 25228936]
13. Kubinger KD. Gutachten zur Erstellung „gerichtsfester" Multiple-Choice-Prüfungsaufgaben. Psychol Rundschau 2014 Jul;65(3):169-178. [doi: 10.1026/0033-3042/a000218]
14. Scriven M. The methodology of evaluation. In: Tyler RW, Gagné RM, Scriven M, editors. Perspectives of Curriculum Evaluation: Rand McNally; 1967.
15. Bauer D, Holzer M, Kopp V, Fischer MR. Pick-N multiple choice-exams: a comparison of scoring algorithms. Adv Health Sci Educ Theory Pract 2011 May;16(2):211-221. [doi: 10.1007/s10459-010-9256-1] [Medline: 21038082]
16. Vorkauf H. Teilpunktbewertung bei K'-Items [Partial credit scoring of Multiple-True-False items]. In: Jahresbericht 1986 der Gruppe Medizinalprüfungen und der Gruppe Statistik und EDV: Institut für Ausbildungs- und Examensforschung, Medizinische Fakultät der Universität Bern; 1987:44-48.
17. Blasberg R, Güngerich U, Müller-Esterl W, Neumann D, Schappel S. Erfahrungen mit dem Fragentyp „k aus n" in Multiple-Choice-Klausuren [Experiences with item type "k from n" in multiple-choice-tests]. Med Ausbild 2001;18(S1):73-76.
18. Hevner K. A method of correcting for guessing in true-false tests and empirical evidence in support of it. J Soc Psychol 1932 Aug;3(3):359-362. [doi: 10.1080/00224545.1932.9919159]
19. Collet LS. Elimination scoring: an empirical evaluation. J Educ Meas 1971 Sep;8(3):209-214. [doi: 10.1111/j.1745-3984.1971.tb00927.x]
20. Gilman DA, Ferry P. Increasing test reliability through self-scoring procedures. J Educ Meas 1972 Sep;9(3):205-207. [doi: 10.1111/j.1745-3984.1972.tb00953.x]
21. Hanna GS. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. J Educ Meas 1975 Sep;12(3):175-178. [doi: 10.1111/j.1745-3984.1975.tb01019.x]
22. Schmitz UJ, Daubländer M. Übertragung papierbasierter Multiple Choice Aufgaben in die digitale Welt – ein Weg zur Verbesserung der Prüfungsqualität? [Transferring paper-based multiple-choice items to the digital world – a way to improve examination quality?] [Abstract]. In: Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA): German Medical Science GMS Publishing House; 2022.
23. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]

Abbreviations

$PS_{1/n}$: Partial Scoring 1/n
PS_{50}: Partial Scoring 50%
UCAN: Umbrella Consortium for Assessment Networks

Edited by B Lesselroth; submitted 07.03.24; peer-reviewed by E Feofanova, F Melzow, K Scholz; revised version received 15.04.24; accepted 07.05.24; published 27.06.24.

Please cite as:
Rössler L, Herrmann M, Wiegand A, Kanzow P
Use of Multiple-Choice Items in Summative Examinations: Questionnaire Survey Among German Undergraduate Dental Training Programs
JMIR Med Educ 2024;10:e58126
URL: https://mededu.jmir.org/2024/1/e58126
doi:10.2196/58126

© Lena Rössler, Manfred Herrmann, Annette Wiegand, Philipp Kanzow. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.6.2024. 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.

XSL•FO
RenderX

Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study

Zelin Wu[1,*], PhD; Wenyi Gan[2,*], PhD; Zhaowen Xue[1,*], PhD; Zhengxin Ni[3,*], MD; Xiaofei Zheng[1], PhD; Yiyi Zhang[1], PhD

1

2

3

[*]these authors contributed equally

Corresponding Author:
Yiyi Zhang, PhD

Abstract

Background: The creation of large language models (LLMs) such as ChatGPT is an important step in the development of artificial intelligence, which shows great potential in medical education due to its powerful language understanding and generative capabilities. The purpose of this study was to quantitatively evaluate and comprehensively analyze ChatGPT's performance in handling questions for the National Nursing Licensure Examination (NNLE) in China and the United States, including the National Council Licensure Examination for Registered Nurses (NCLEX-RN) and the NNLE.

Objective: This study aims to examine how well LLMs respond to the NCLEX-RN and the NNLE multiple-choice questions (MCQs) in various language inputs. To evaluate whether LLMs can be used as multilingual learning assistance for nursing, and to assess whether they possess a repository of professional knowledge applicable to clinical nursing practice.

Methods: First, we compiled 150 NCLEX-RN Practical MCQs, 240 NNLE Theoretical MCQs, and 240 NNLE Practical MCQs. Then, the translation function of ChatGPT 3.5 was used to translate NCLEX-RN questions from English to Chinese and NNLE questions from Chinese to English. Finally, the original version and the translated version of the MCQs were inputted into ChatGPT 4.0, ChatGPT 3.5, and Google Bard. Different LLMs were compared according to the accuracy rate, and the differences between different language inputs were compared.

Results: The accuracy rates of ChatGPT 4.0 for NCLEX-RN practical questions and Chinese-translated NCLEX-RN practical questions were 88.7% (133/150) and 79.3% (119/150), respectively. Despite the statistical significance of the difference (*P*=.03), the correct rate was generally satisfactory. Around 71.9% (169/235) of NNLE Theoretical MCQs and 69.1% (161/233) of NNLE Practical MCQs were correctly answered by ChatGPT 4.0. The accuracy of ChatGPT 4.0 in processing NNLE Theoretical MCQs and NNLE Practical MCQs translated into English was 71.5% (168/235; *P*=.92) and 67.8% (158/233; *P*=.77), respectively, and there was no statistically significant difference between the results of text input in different languages. ChatGPT 3.5 (NCLEX-RN *P*=.003, NNLE Theoretical *P*<.001, NNLE Practical *P*=.12) and Google Bard (NCLEX-RN *P*<.001, NNLE Theoretical *P*<.001, NNLE Practical *P*<.001) had lower accuracy rates for nursing-related MCQs than ChatGPT 4.0 in English input. English accuracy was higher when compared with ChatGPT 3.5's Chinese input, and the difference was statistically significant (NCLEX-RN *P*=.02, NNLE Practical *P*=.02). Whether submitted in Chinese or English, the MCQs from the NCLEX-RN and NNLE demonstrated that ChatGPT 4.0 had the highest number of unique correct responses and the lowest number of unique incorrect responses among the 3 LLMs.

Conclusions: This study, focusing on 618 nursing MCQs including NCLEX-RN and NNLE exams, found that ChatGPT 4.0 outperformed ChatGPT 3.5 and Google Bard in accuracy. It excelled in processing English and Chinese inputs, underscoring its potential as a valuable tool in nursing education and clinical decision-making.

(JMIR Med Educ 2024;10:e52746) doi:10.2196/52746

KEYWORDS

artificial intelligence; ChatGPT; nursing licensure examination; nursing; LLMs; large language models; nursing education; AI; nursing student; large language model; licensing; observation; observational study; China; USA; United States of America; auxiliary tool; accuracy rate; theoretical

XSL•FO
RenderX

Introduction

The large language model (LLM) technology is a stepping stone in the evolution of artificial intelligence (AI) [1,2]. Through the analysis of a large database, the primary module generates a logical and plain text response to the user's query promptly following the user's textual input [3]. Currently, popular AI software includes ChatGPT 4.0, ChatGPT 3.5, and Google Bard, and research indicates that these 3 AI algorithms perform well when answering queries about lung cancer [4]. AI tools are the result of the advancement of science and technology, and the advent of revolutionary tools will alter the way people learn and work, which is an irreversible trend.

ChatGPT has been controversial since its public release in November 2022 due to its powerful text generation capabilities, and attention has been focused on students using ChatGPT for essay writing and assignment plagiarism [5-7]. With the birth of regulatory software such as GPTZero, AI-Text-Classifier, and ChatGPT Detector, people gradually focused on the application of ChatGPT, trying to explore and expand the application field of ChatGPT. The study found that ChatGPT showed both professionalism and empathy in answering general public health questions [8]. ChatGPT not only showed strong expertise in answering basic research directions but also followed evidence-based clinical decision-making [9,10]. Nevertheless, there may be some ethical problems in clinical application, and it is necessary to consider whether the use of ChatGPT will violate the rights and interests of patients [11-13]. Therefore, more and more researchers have placed the application field of ChatGPT in education [14]. The studies found that ChatGPT performed well on multiple-choice questions (MCQs) about otolaryngology and gynecology [15,16]. In addition, ChatGPT software can pass the Plastic Surgery Inservice Training Examination [17], the American Heart Association Basic Life Support Examinations [18], and the Taiwanese Pharmacist Licensing Examination [19]. ChatGPT is also able to solve higher-order problems related to medical biochemistry while also achieving satisfactory performance in surgical education and training [20,21]. However, ChatGPT is not a training tool for all exams, with the exception of the American Heart Association's Advanced Cardiovascular Life Support (ACLS) exams and Taiwan's Family Medicine Board Exam [18,22]. This might suggest that ChatGPT's application areas may be limited by language and region in addition to speciality.

Both the United States and China have instituted licensing exams to regulate the qualifications of registered nurses [23]. China uses the National Nursing Licensure Examination (NNLE) [23], whereas the United States uses the National Council Licensure Examination for Registered Nurses (NCLEX-RN) [24], both of which seek to standardize the theoretical and practical foundations of nurses through standardized assessment procedures to ensure the professionalism of nurses who are entering the medical field. The content of nursing studies is not medically specialized but rather interdisciplinary and multidisciplinary [25]. On the basis of their nursing work, nurses are frequently required to comprehend clinical decisions made by physicians. As a result, it is easy for society to disregard the difficulty of nursing education and training, that is, the necessity of a medical foundation for the development of nursing expertise [26]. Presently, there are no professional nursing learning aids to assist nurses in gaining a better understanding of the professional medical issues encountered during the clinical learning process. Huge and intricate, the medical knowledge system necessitates repeated learning, even for specialists, in order to master specialized knowledge [27]. Despite the fact that many researchers attempt to implement various review strategies to increase the passage rate of nursing professional examinations, it is frequently difficult to popularize a single review strategy due to varying local practical policies [28]. No single revision method is appropriate for all individuals. How to assist nurses in gaining a deeper understanding of medical knowledge, enhancing their stockpile of professional theoretical knowledge, and increasing their exam pass rate is a pressing issue for nurses today.

The design of this research is cross-sectional. By incorporating NCLEX-RN and NNLE questions, we evaluated the precision of responses from ChatGPT 4.0, ChatGPT 3.5, and Google Bard. Concurrently, the translation feature of ChatGPT 3.5 was used to convert between Chinese and English, while an examination was conducted into the disparity in the rate of accurate responses provided by ChatGPT across various languages. The aim of this study is to offer a conceptual framework that supports the implementation of ChatGPT and advances nursing education and clinical application.

Methods

Design

With reference to Zong et al [29], we designed a cross-sectional study. The experimental data from our study had been recorded in an Excel file and uploaded as Multimedia Appendix 1. The STROBE Initiative [30] was used in this study and the STROBE Initiative checklist is available in Multimedia Appendix 2.

Ethical Considerations

As this study does not involve interventional experiments on humans or animals, the research does not require approval per the Ethics Committee of the First Affiliated Hospital of Jinan University guidelines.

Data Source

NCLEX-RN practice questions were compiled at the website "nurseslabs" [31]. There were no set questions on the official NCLEX-RN test; instead, a computer produced new questions with a minimum of 75 and a maximum of 265 depending on how accurate the preceding questions were. Thus, we got the most recent 2 sets of practice questions for the NCLEX-RN exam from the internet. In 2 practice sets, we compiled a total of 150 MCQs.

The NNLE question categories were divided into 2 sections: nursing theory and nursing practice, each containing 120 MCQs. On the website "baidu" [32], we used the most current 480 NNLE-MCQs from the 2022 and 2021 exams that were accessible. According to the classification of nursing theory examination and nursing practice, the questions for 2022 and

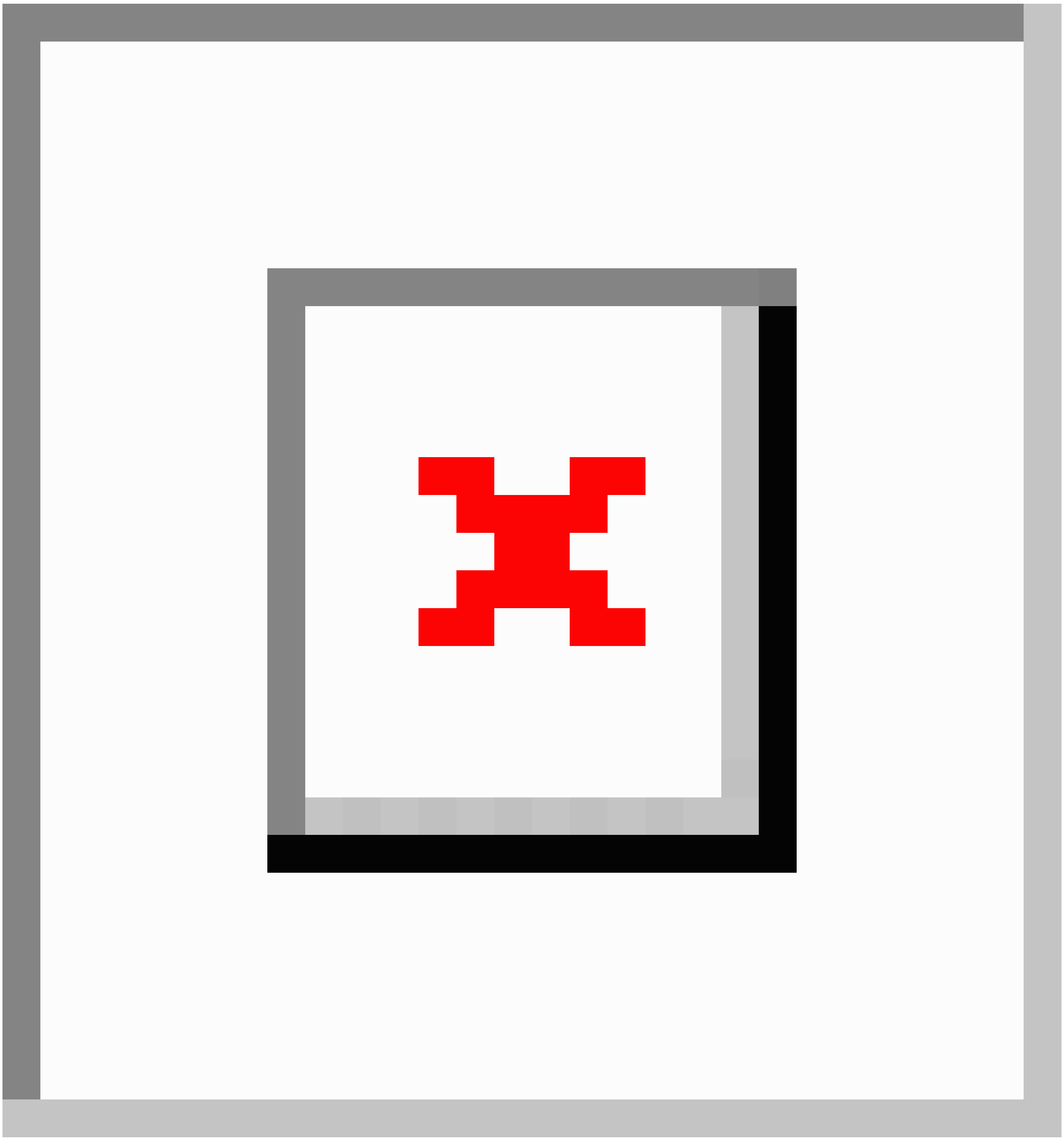
XSL•FO
RenderX

2021 were merged and then separated into NNLE Theoretical MCQs (n=240) and NNLE Practical MCQs (n=240).

Procedures

According to the research stages (Figure 1), we translated the original English NCLEX-RN-MCQs into the Chinese version of the NCLEX-RN-MCQs. The original NNLE queries were written in Chinese, and we also translated them into English. To avoid systematic errors induced by differences in translation quality during the translation process, ChatGPT 3.5 was used to translate both from Chinese to English and from English to Chinese. We checked the language both before and after translating using ChatGPT 3.5 to translate between Chinese and English, as well as English and Chinese. About some clear translation mistakes, we entered the incorrect translation points in ChatGPT 3.5's dialog box and requested that ChatGPT 3.5 retranslate the text.

Figure 1. Diagrammatic representation of the progression of exploratory application experiments. MCQ: multiple-choice question; NCLEX-RN: National Council Licensure Examination for Registered Nurses; NNLE: National Nursing Licensure Examination.

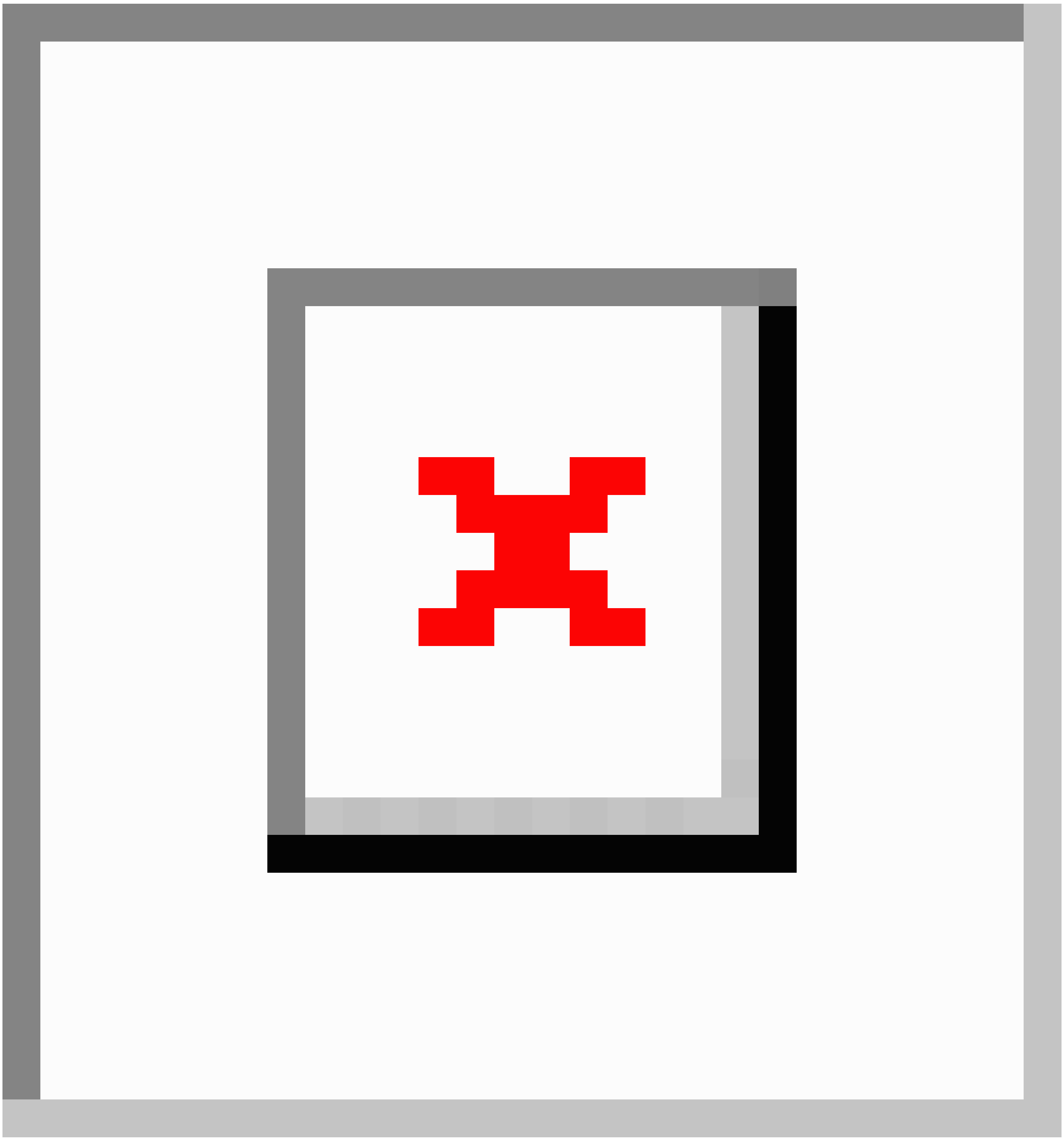


Entered all questions on ChatGPT 4.0 (Figure 2Aand C) [33] and ChatGPT 3.5 (Figure 2B and D) [34] as well as Google Bard (Figure 2E) [35], then recorded the responses. Both ChatGPT 4.0 and ChatGPT 3.5 support text input in non-English languages, whereas Google Bard only supports text input in English at this time. The use of "New chat" for each inquiry ensured the independence of each response because it prevented the AI from using context from previous interactions, thereby

XSL•FO
RenderX

eliminating any learning or bias that may have been carried over from earlier questions. Additionally, no plugins were used with ChatGPT, and the "Chat history & training" option was deactivated to preserve the objectivity of each response.

Figure 2. (A) English multiple-choice questions (MCQs) input in ChatGPT 4.0. (B) English MCQs input in ChatGPT 3.5. (C) Chinese MCQs input in ChatGPT 4.0. (D) Chinese MCQs input in ChatGPT 3.5. (E) English MCQs input in Google Bard.



Data Analysis

SPSS program (version 26.0; IBM Corp) was used for statistical analysis. With reference to Zong et al [29]. Collected the responses from ChatGPT 4.0, ChatGPT 3.5, and Google Bard and converted them to the binary variables "true" or "false." Pearson The χ^2 test was used to compare the differences between various LLM software or the same software input in various languages. A difference was considered statistically significant when the *P* value was less than .05. Used the web-based VENN diagram drawing website "bioinfogp" [36] to draw VENN diagrams to display different AI software's results for the same type of subject with various linguistic inputs. Last, bar charts were constructed from a portion of the data using GraphPad Prism 8.

Results

Overview

We collected 150 NCLEX-RN-MCQs in total. We excluded the image questions from the compiled NNLE-MCQs because the picture analysis of ChatGPT and Google Bard required the

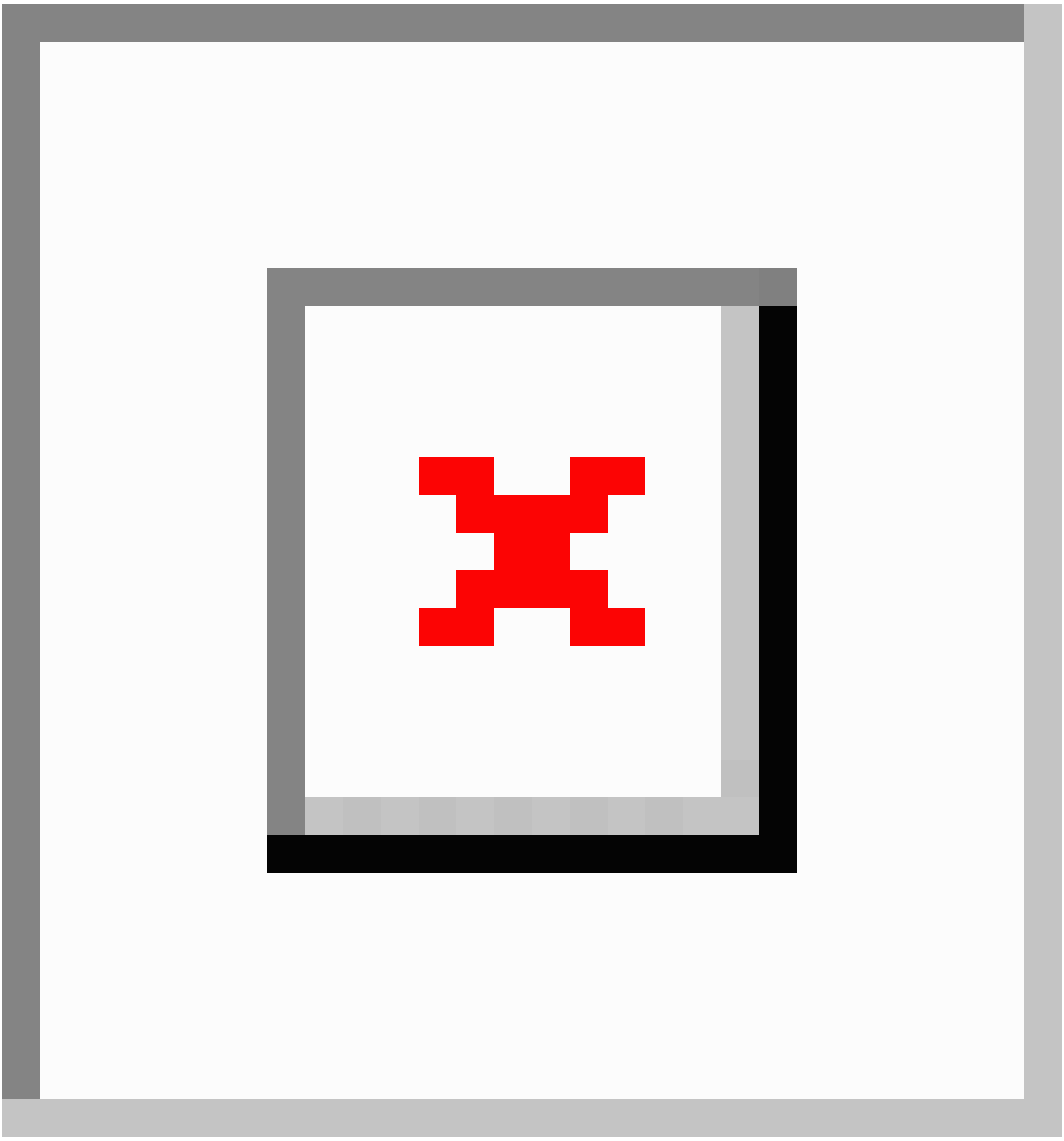
XSL•FO
RenderX

use of external plug-ins. After eliminating the image questions, there were a total of 235 NNLE Theoretical MCQs and 233 NNLE Practical MCQs left. Then, ChatGPT 3.5 converted NCLEX-RN-MCQs for English questions into the Chinese version and NNLE-MCQs into the English version.

Performance of LLMs in Responding to English NCLEX-RN MCQs

ChatGPT 4.0 had an accuracy rate of 88.67% (133/150) when answering NCLEX-RN MCQs in English, which was higher than ChatGPT 3.5 (113/150, 75.3%) and Google Bard (96/150, 64%) (Figure 3C). Statistically, ChatGPT 4.0 performed significantly better than the other 2 categories (ChatGPT 4.0 vs ChatGPT 3.5, *P*=.003; ChatGPT 4.0 vs Google Bard, *P*<.001) (Figure 3C). ChatGPT 3.5 was more accurate than Google Bard and the difference was statistically significant (*P*=.03) (Figure 3C).

Figure 3. (A,B) VENN diagram shows the correct and incorrect intersection of NCLEX-RN practical questions in different large language models. (C) The correct rate of NCLEX-RN practical questions in various large language models. MCQ: multiple-choice question; NCLEX-RN: National Council Licensure Examination for Registered Nurses.

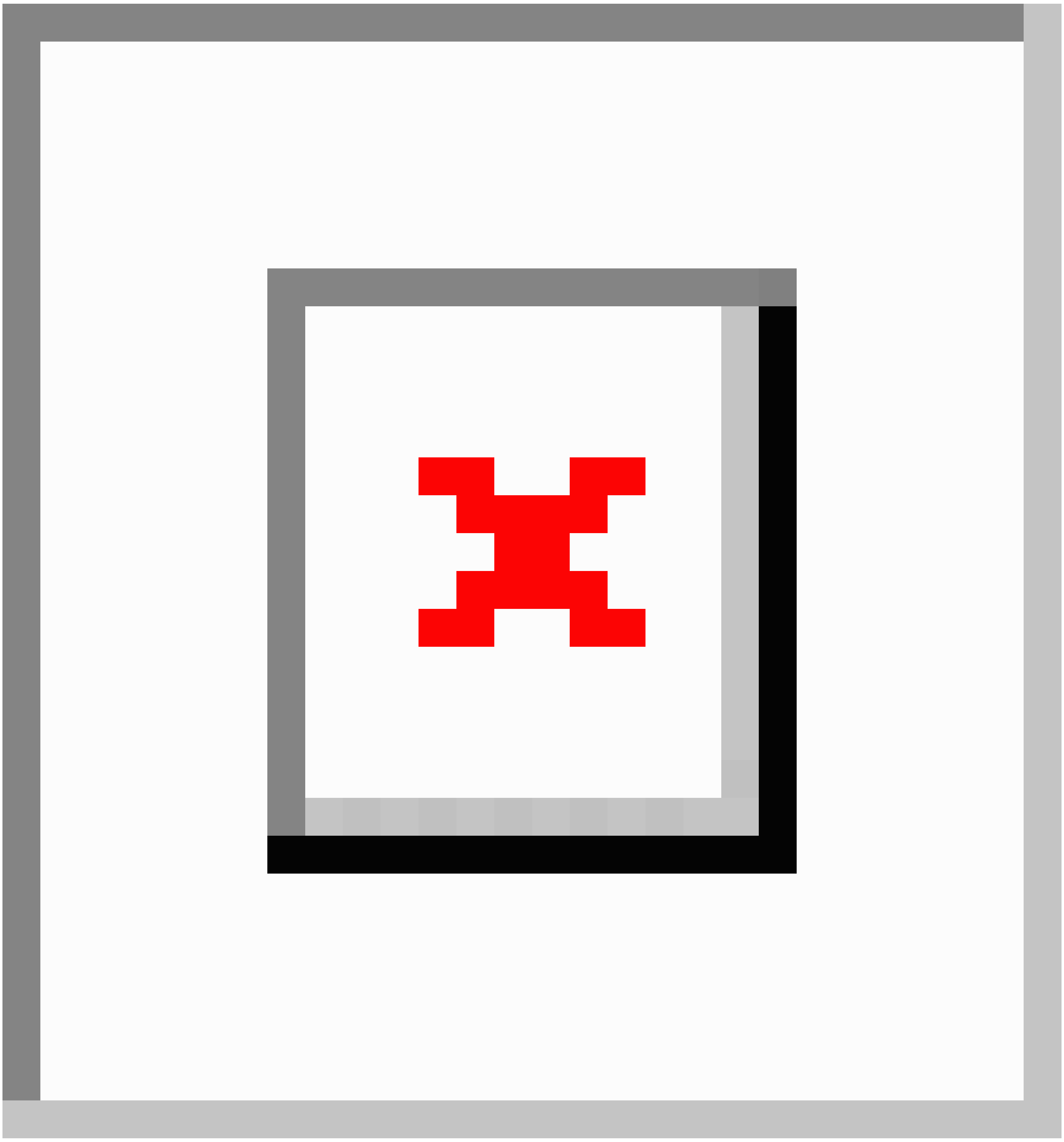


XSL•FO
RenderX

Performance of LLMs in Responding to Chinese NNLE-MCQs

The difference between the correct rates of ChatGPT 4.0 and ChatGPT 3.5 in answering the Chinese version of NNLE theoretical MCQs (*P*<.001) and NNLE practical MCQs (*P*<.001) was statistically significant (Figure 4E and F). The correct rates of ChatGPT 4.0 answering NNLE theoretical MCQs and NNLE practical MCQs were 71.9% (169/235) and 69.1% (161/233), respectively, compared with 53.2% (125/235) and 50.2% (117/233) for ChatGPT 3.5 (Figure 4E and F).

Figure 4. (A,B) VENN diagram shows the correct and incorrect intersection of NNLE theoretical MCQs in different large language models (LLMs). (C,D) VENN diagram shows the correct and incorrect intersection of NNLE practical MCQs in different LLMs. (E) The correct rate of NNLE theoretical MCQs in various LLMs. (F) The correct rate of NNLE practical MCQs in various LLMs. MCQ: multiple-choice question; NNLE: National Nursing Licensure Examination.



Performance and Variations of MCQs Input Into LLMs in Various Languages

After entering the Chinese-translated version of NCLEX-RN-MCQs into ChatGPT 4.0 and ChatGPT 3.5, we discovered that the accuracy rates were 79.3% (119/150) and 63.3% (95/150), respectively, with a statistically significant difference between the two (*P*=.002) (Figure 3C).

Then, we fed the English-translated version of NNLE Theoretical MCQs into ChatGPT 4.0, ChatGPT 3.5, and Google Bard and determined that their respective accuracy rates were

XSL•FO
RenderX

71.5 % (168/235), 55.7% (131/235), and 49.8% (117/235) (Figure 4E). ChatGPT 4.0 had a higher accuracy rate than ChatGPT 3.5 (P<.001) and Google Bard (P<.001) for the English-translated version of NNLE Theoretical MCQs while the difference was statistically significant (Figure 4E). ChatGPT 3.5 had a higher accuracy rate than Google Bard, but the difference was not statistically significant (P=.20) (Figure 4E).

The accuracy rates of ChatGPT 4.0, ChatGPT 3.5, and Google Bard were 67.8% (158/233), 60.9% (142/233), and 46.8% (109/233), respectively, when the English-translated version of NNLE Practical MCQs was inputted (Figure 4F). In terms of the English-translated version of NNLE Practical MCQs, the accuracy rates of both ChatGPT 4.0 (P<.001) and ChatGPT 3.5 (P=.002) were higher than those of Google Bard, and the difference was statistically significant; however, unlike before, the difference in accuracy rates between ChatGPT 4.0 and ChatGPT 3.5 was not statistically significant (P=.12) (Figure 4F).

When processing NCLEX-RN-MCQs, the accuracy of inputs in the original English version was statistically significantly higher than that of inputs translated into Chinese for both ChatGPT 4.0 (P=.03) and ChatGPT 3.5 (P=.02) (Figure 3C). The difference was not statistically significant between the accuracy of inputs in the original Chinese version and the inputs of the translated English version for both ChatGPT 4.0 (P=.92) and ChatGPT 3.5 (P=.58) when processing NNLE Theoretical MCQs (Figure 4E). The accuracy of ChatGPT 4.0's inputs in the original Chinese version was higher than that of inputs translated into English when processing NNLE Practical MCQs, but this difference was not statistically significant (P=.77) (Figure 4F). Surprisingly, the accuracy of ChatGPT 3.5's inputs in the original Chinese version was lower than that of inputs translated into English while dealing with NNLE Practical MCQs, and this difference was statistically significant (P=.02) (Figure 4F).

Figure 3A and B depicts, respectively, the intersection of correct and incorrect questions when NCLEX-RN practical questions were inputted into various LLMs in various languages. Similarly, Figure 4A and B depicts NNLE Theoretical MCQs, while Figure 4C and D depicts NNLE Practical MCQs. When the same questions were input into ChatGPT 4.0, ChatGPT 3.5, and Google Bard in English, ChatGPT 4.0 had the highest number (n for NCLEX-RN MCQs=14; n for NNLE Theoretical MCQs=33; n for NNLE Practical MCQs=26) of uniquely correct answers and the lowest number (n for NCLEX-RN MCQs=2; n for NNLE Theoretical MCQs=6; n for NNLE Practical MCQs=7) of uniquely incorrect answers among the 3 engines. Instead, Google Bard had a lower number (n for NCLEX-RN MCQs=2; n for NNLE Theoretical MCQs=10; n for NNLE Practical MCQs=6) of uniquely correct answers than ChatGPT 4.0 and the highest number (n for NCLEX-RN MCQs=26; n for NNLE Theoretical MCQs=34; n for NNLE Practical MCQs=36) of uniquely incorrect answers among the 3 engines when the MCQs were input into 3 engines in English. Likewise, after the questions were submitted in Chinese, we found that ChatGPT 4.0 (n for NCLEX-RN MCQs=35; n for NNLE Theoretical MCQs=61; n for NNLE Practical MCQs=63) gives more uniquely accurate responses than ChatGPT 3.5(n for NCLEX-RN MCQs=11; n for NNLE Theoretical MCQs=17; n for NNLE Practical MCQs=19) does.

Discussion

Principal Findings

This study is a cross-sectional study that collected a total of 618 nursing-related MCQs, including 150 NCLEX-RN practice questions and 468 NNLE actual exam questions. To observe differences between inputs in different languages, ChatGPT 3.5 was used exclusively for Chinese-to-English and English-to-Chinese translations. The results revealed that ChatGPT 4.0 had a significantly higher accuracy rate when handling English input for NCLEX-RN practical MCQs compared with ChatGPT 3.5 and Google Bard. Similarly, ChatGPT 4.0 also outperformed ChatGPT 3.5 in accuracy when processing the Chinese input of NNLE exam MCQs. Therefore, ChatGPT 4.0 has the potential to be an effective learning assistance software for ChatGPT users, and due to its powerful real-time text generation capabilities, it can also provide additional sources of information and reference for nursing decisions in clinical nursing work.

Despite being a tool that accepts input in different languages, ChatGPT has linguistic bias while processing text input, as this research has shown. ChatGPT 3.5 translates NCLEX-RN practical MCQs from English to Chinese. Following input, it was discovered that while interacting with English, ChatGPT 4.0 and ChatGPT 3.5 had accuracy rates that were noticeably greater than Chinese. When NNLE MCQs were input into ChatGPT in English, ChatGPT 4.0's accuracy of the response was only somewhat less accurate than the Chinese input, while ChatGPT 3.5's English input was even more accurate than the Chinese input. Although there may be some linguistic distortion when translating between languages using software, the findings of our cross-sectional investigation indicated that ChatGPT processes English input more accurately than Chinese input. I asked ChatGPT, an AI program that facilitates real-time communication, questions in an attempt to comprehend the logic behind handling input in various languages. In response, ChatGPT said that it can assess and respond to queries in several languages depending on the language of input. This capability stems from its training of various input kinds in various languages. As a result, the current discrepancy in accuracy caused by input in Chinese and English may be the result of ChatGPT receiving different amounts of training in different languages. This discrepancy may disappear with an increase in language training once ChatGPT becomes more well-known worldwide.

The low passage rate of nursing examinations is partly attributed to the lack of fundamental theoretical and clinical knowledge among nursing staff [24,37]. Researchers have tried to reform and innovate nursing education models within certain limits to improve knowledge levels and exam pass rates [28]. However, due to differences in language and local policies, it is challenging to widely implement a single educational model. MCQs are an effective method to assess student knowledge [38], but existing learning resources often require students to conduct independent searches to expand knowledge, adding to

XSL•FO
RenderX

learning pressure and affecting the coherence of the learning process. ChatGPT's big data analysis and rapid text feedback can help students consolidate and expand knowledge points while completing MCQ exercises [39]. Besides, ChatGPT 4.0 not only enhances the efficiency of nursing education [40] but also provides clinicians and nurses with objective information support based on evidence-based medicine and big data analysis in complex clinical scenarios [41]. For instance, the research discovered that ChatGPT 4.0 not only analyzed imaging data with acceptable accuracy and sensitivity but also assisted physicians in thinking outside the box and offering several helpful recommendations when making individualized clinical treatment choices for tumor patients [41]. Furthermore, ChatGPT may provide nurses with a customized and immersive learning experience, bolster their competence and self-assurance in overseeing remote patient care, and furnish them with the necessary abilities for remote patient monitoring, all of which can contribute to the enhancement of patient outcomes and care quality [42]. Additionally, ChatGPT may assist doctors in streamlining patient data organization and easing the burden of interpreting medical records in order to improve patient communication while doing therapeutic procedures [43].

According to this study and previous research findings, ChatGPT 4.0 is currently the most accurate and repeatable AI software among many LLMs. In answering questions related to electrocardiogram images [44], the Multi-Specialty Recruitment Assessment exam [45], dental professional issues [46], and analyzing radiology data [47], ChatGPT 4.0 provides more accurate and comprehensive responses compared with ChatGPT 3.5 and Google Bard. Since ChatGPT 4.0 is currently the only paid AI software compared with free-to-use LLMs like ChatGPT 3.5, Google Bard, and Bing, it is essential to compare its functionality with these free LLMs when exploring its real-world application value. The economic cost of use is also a factor that must be considered in the popularization and promotion of its application [48].

Assessing ChatGPT's clinical application value in a manner that aligns with the training of experienced clinical workers is the same approach; upon passing the theory test, candidates will be deemed to possess fundamental medical theoretical knowledge and be capable of managing simple clinical scenarios [49]. The intricacy of clinical issues will then continuously increase as a result of ongoing training that corrects incorrect theoretical knowledge and clinical reasoning. Last, they get training to become highly repeatable and capable self-correcting clinical practitioners. ChatGPT has shown that it has a theoretical foundation for supporting clinical practice with its outstanding success in the qualifying exams of many clinical professions [15-22,45,46,49]. However, whether it is used as an auxiliary tool for self-learning and education, to support patient communication, or to aid in the analysis of complicated clinical circumstances, a commensurate regulatory system must be developed. In order to limit the circumstances in which ChatGPT is used, schools, hospitals, and publishing companies must first create pertinent policies [50]. Some examples of these policies include forbidding the use of ChatGPT during exams [51] and obtaining patient consent before using ChatGPT as an auxiliary tool in real clinical settings [52]. Authors must state that ChatGPT was not directly engaged in the creation of the text for the paper and are forbidden from claiming ChatGPT as an independent author [53]. Furthermore, the most immediate regulators of ChatGPT are its users. ChatGPT can assist with data collection and content integration, but the user has to take part in the quality review process of the content that ChatGPT generates, identify any problems in the responses that ChatGPT generates, and finish training ChatGPT via error correction and continuous input and output. Although many companies developing LLMs claim to avoid the collection and leakage of private information, as users of these software, it is also essential to ensure the content and quality of the input information. Users should intentionally avoid and delete personal and private information, thereby enhancing their personal oversight function during the use of the software. It is also crucial to seek the informed permission of other participants and make suitable declarations while using ChatGPT in public to prevent unwanted confrontations between doctors and patients, moral and ethical disagreements, and concerns with writing integrity.

Implication

Our study has demonstrated that ChatGPT 4.0 exhibits a satisfactory accuracy rate in handling MCQs for the NCLEX-RN and NNLE exams, outperforming 2 other AI engines, ChatGPT 3.5 and Google Bard. Although there were differences in accuracy rates when the same questions were inputted in different languages, the overall accuracy of ChatGPT 4.0 remains commendable. Combined with conclusions from previous research, it can be inferred that ChatGPT 4.0 possesses the knowledge reserve necessary for application in medical education, learning, and clinical scenarios, with the potential to assist in managing complex clinical situations. To promote the rational application of ChatGPT 4.0 in the medical field, it is imperative for relevant authorities to develop effective and reasonable regulatory mechanisms and supervisory bodies in the future. This will ensure that ChatGPT 4.0, a powerful auxiliary AI software, is used appropriately within the health care sector.

Limitation

This study is a cross-sectional analysis, and the findings suggest that ChatGPT 4.0 possesses a certain level of nursing professional knowledge. However, high-quality prospective randomized controlled trials are still required to validate the actual effectiveness of ChatGPT 4.0 in nursing education, learning, and clinical application. Besides, since the logic behind how AI processes questions is part of the company's "black box," we can only understand its logic in processing inputs in different languages by interacting with the AI software. Therefore, we infer that the differences in handling Chinese and English inputs are due to variations in the amount of training between languages.

Conclusion

This cross-sectional study collected and analyzed 618 nursing-related MCQs, including NCLEX-RN practice questions and NNLE actual exam questions, to evaluate the performance of ChatGPT 4.0 in processing different language inputs. The study exclusively used ChatGPT 3.5 for Chinese-to-English

XSL•FO
RenderX

and English-to-Chinese translations and found that ChatGPT 4.0 demonstrated a significantly higher accuracy rate than ChatGPT 3.5 and Google Bard, particularly in handling English input for NCLEX-RN Practice MCQs and Chinese input for NNLE exam MCQs. These findings suggest that ChatGPT 4.0 has substantial potential as an effective learning assistance tool for nursing education and can provide valuable information and reference in clinical nursing settings due to its advanced real-time text generation capabilities.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Authors' Contributions

ZW, WG, ZX, and ZN contributed equally. ZW, WG, ZX, and ZN conceived the study, performed the statistical analysis, interpreted the results, and drafted the manuscript. YZ and XZ supervised the entire study. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
File for the original dataset.
[XLSX File, 55 KB - mededu_v10i1e52746_app1.xlsx]

Multimedia Appendix 2
STROBE checklist cross-sectional.
[DOCX File, 32 KB - mededu_v10i1e52746_app2.docx]

References

1. Mesko B. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. J Med Internet Res 2023 Jun 22;25:e48392. [doi: 10.2196/48392] [Medline: 37347508]
2. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023 May 30;9(1):44. [doi: 10.1038/s41523-023-00557-8] [Medline: 37253791]
3. Perera Molligoda Arachchige AS. Large language models (LLM) and ChatGPT: a medical student perspective. Eur J Nucl Med Mol Imaging 2023 Jul;50(8):2248-2249. [doi: 10.1007/s00259-023-06227-y] [Medline: 37046082]
4. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023 Jun;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]
5. Graham A. ChatGPT and other AI tools put students at risk of plagiarism allegations, MDU warns. BMJ 2023 May 17;381:1133. [doi: 10.1136/bmj.p1133] [Medline: 37197782]
6. Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? Nature 2022 Dec 9. [doi: 10.1038/d41586-022-04397-7] [Medline: 36494443]
7. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023 Mar;5(3):e102. [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
8. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
9. Zhou Z, Wang X, Li X, Liao L. Is ChatGPT an evidence-based doctor? Eur Urol 2023 Sep;84(3):355-356. [doi: 10.1016/j.eururo.2023.03.037] [Medline: 37061445]
10. Miao H, Ahn H. Impact of ChatGPT on interdisciplinary nursing education and research. Asian Pac Isl Nurs J 2023 Apr 24;7:e48136. [doi: 10.2196/48136] [Medline: 37093625]
11. Kao HJ, Chien TW, Wang WC, Chou W, Chow JC. Assessing ChatGPT's capacity for clinical decision support in pediatrics: a comparative study with pediatricians using KIDMAP of rasch analysis. Medicine (Baltimore) 2023 Jun 23;102(25):e34068. [doi: 10.1097/MD.0000000000034068] [Medline: 37352054]
12. Liao Z, Wang J, Shi Z, Lu L, Tabata H. Revolutionary potential of ChatGPT in constructing intelligent clinical decision support systems. Ann Biomed Eng 2024 Feb;52(2):125-129. [doi: 10.1007/s10439-023-03288-w] [Medline: 37332008]
13. Secor AM, Célestin K, Jasmin M, et al. Electronic medical record data missingness and interruption in antiretroviral therapy among adults and children living with HIV in Haiti: retrospective longitudinal study. JMIR Pediatr Parent 2024 Mar 6;7:e51574. [doi: 10.2196/51574] [Medline: 38488632]

XSL•FO
RenderX

14. Torales J, O'Higgins M. ChatGPT and social psychiatry: a commentary on the article 'Old dog, new tricks? exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry'.. Int J Soc Psychiatry 2023 Jun 30:207640231178488. [doi: 10.1177/00207640231178488] [Medline: 37392002]
15. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023 Sep;280(9):4271-4278. [doi: 10.1007/s00405-023-08051-4] [Medline: 37285018]
16. Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023 Aug;229(2):172. [doi: 10.1016/j.ajog.2023.04.020] [Medline: 37088277]
17. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J 2023 Nov 16;43(12):NP 1078-NNP1082. [doi: 10.1093/asj/sjad128] [Medline: 37128784]
18. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
19. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc 2023 Jul 1;86(7):653-658. [doi: 10.1097/JCMA.0000000000000942] [Medline: 37227901]
20. Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus 2023 Apr;15(4):e37023. [doi: 10.7759/cureus.37023] [Medline: 37143631]
21. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
22. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's family medicine board exam. J Chin Med Assoc 2023 Aug 1;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
23. Hou J, Chen S, Sabharwal S, Fan V, Yan M, Wang W. Comparison of RN licensure examination: China and the United States. Int J Nurs Sci 2019 Jan 10;6(1):111-116. [doi: 10.1016/j.ijnss.2018.11.002] [Medline: 31406876]
24. Muirhead L, Cimiotti JP, Hayes R, et al. Diversity in nursing and challenges with the NCLEX-RN. Nurs Outlook 2022;70(5):762-771. [doi: 10.1016/j.outlook.2022.06.003] [Medline: 35933180]
25. O'Reilly P, Lee SH, O'Sullivan M, Cullen W, Kennedy C, MacFarlane A. Assessing the facilitators and barriers of interdisciplinary team working in primary care using normalisation process theory: an integrative review. PLoS One 2017 May 18;12(5):e0177026. [doi: 10.1371/journal.pone.0177026] [Medline: 28545038]
26. Horsley TL, Reed T, Muccino K, Quinones D, Siddall VJ, McCarthy J. Developing a foundation for interprofessional education within nursing and medical curricula. Nurse Educ 2016;41(5):234-238. [doi: 10.1097/NNE.0000000000000255] [Medline: 26963036]
27. Gan W, Mok TN, Chen J, et al. Researching the application of virtual reality in medical education: one-year follow-up of a randomized trial. BMC Med Educ 2023 Jan 3;23(1):3. [doi: 10.1186/s12909-022-03992-6] [Medline: 36597093]
28. Cobourne K. Strategies to increase NCLEX pass rates: from 68% to 92% in 1 year. Nurse Educ 2023;48(4):220-222. [doi: 10.1097/NNE.0000000000001382] [Medline: 36857572]
29. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143. [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
30. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
31. NCLEX practice questions test bank for free. Nurseslabs. 2024. URL: https://nurseslabs.com/nclex-practice-questions [accessed 2024-09-23]
32. National nursing licensure examination. Baidu. 2024. URL: https://wenku.baidu.com/ [illegible] [accessed 2024-09-23]
33. ChatGPT 4.0. OpenAI. 2024. URL: https://chat.openai.com/?model=gpt-4 [accessed 2024-09-23]
34. ChatGPT 3.5. OpenAI. 2024. URL: https://chat.openai.com/?model=text-davinci-002-render-sha [accessed 2024-09-23]
35. Google Bard. Google. 2024. URL: https://bard.google.com [accessed 2024-09-23]
36. Venny2.1. Bioinfogp. 2024. URL: https://bioinfogp.cnb.csic.es/tools/venny/index.html [accessed 2024-09-23]
37. Flowers M, Olenick M, Maltseva T, Simon S, Diez-Sampedro A, Allen LR. Academic factors predicting NCLEX-RN success. Nurs Educ Perspect 2022;43(2):112-114. [doi: 10.1097/01.NEP.0000000000000788] [Medline: 35192289]
38. Levant B, Zückert W, Paolo A. Post-exam feedback with question rationales improves re-test performance of medical students on a multiple-choice exam. Adv Health Sci Educ Theory Pract 2018 Dec;23(5):995-1003. [doi: 10.1007/s10459-018-9844-z] [Medline: 30043313]
39. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus 2023 Aug;15(8):e43271. [doi: 10.7759/cureus.43271] [Medline: 37692629]

XSL•FO
RenderX

40. Ahmed SK. The impact of ChatGPT on the nursing profession: revolutionizing patient care and education. Ann Biomed Eng 2023 Nov;51(11):2351-2352. [doi: 10.1007/s10439-023-03262-6] [Medline: 37266721]
41. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023 Nov 1;6(11):e2343689. [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
42. Sharma M, Sharma S. A holistic approach to remote patient monitoring, fueled by ChatGPT and metaverse technology: the future of nursing education. Nurse Educ Today 2023 Dec;131:105972. [doi: 10.1016/j.nedt.2023.105972] [Medline: 37757713]
43. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. J Am Acad Orthop Surg 2024 Feb 1;32(3):123-129. [doi: 10.5435/JAAOS-D-23-00474] [Medline: 37976385]
44. Fijačko N, Prosen G, Abella BS, Metličar Š, Štiglic G. Can novel multimodal chatbots such as bing chat enterprise, ChatGPT-4 Pro, and Google Bard correctly interpret electrocardiogram images? Resuscitation 2023 Dec;193:110009. [doi: 10.1016/j.resuscitation.2023.110009] [Medline: 37884222]
45. Tsoutsanis P, Tsoutsanis A. Evaluation of large language model performance on the multi-specialty recruitment assessment (MSRA) exam. Comput Biol Med 2024 Jan;168:107794. [doi: 10.1016/j.compbiomed.2023.107794] [Medline: 38043471]
46. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580. [doi: 10.2196/51580] [Medline: 38009003]
47. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023 Nov;309(2):e232561. [doi: 10.1148/radiol.232561] [Medline: 37987662]
48. Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology 2023 Jul;308(1):e230970. [doi: 10.1148/radiol.230970] [Medline: 37489981]
49. Sahin MC, Sozer A, Kuzucu P, et al. Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. Comput Biol Med 2024 Feb;169:107807. [doi: 10.1016/j.compbiomed.2023.107807] [Medline: 38091727]
50. Zhu Z, Ying Y, Zhu J, Wu H. ChatGPT's potential role in non-english-speaking outpatient clinic settings. D Health 2023 Jun 26;9:20552076231184091. [doi: 10.1177/20552076231184091] [Medline: 37434733]
51. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
52. Adhikari K, Naik N, Hameed BZ, Raghunath SK, Somani BK. Exploring the ethical, legal, and social implications of ChatGPT in urology. Curr Urol Rep 2024 Jan;25(1):1-8. [doi: 10.1007/s11934-023-01185-2] [Medline: 37735339]
53. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023 Jan;613(7945):612. [doi: 10.1038/d41586-023-00191-1] [Medline: 36694020]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCQ: multiple-choice question
NCLEX-RN: National Council Licensure Examination for Registered Nurses
NNLE: National Nursing Licensure Examination

Edited by TDA Cardoso; submitted 14.09.23; peer-reviewed by H Kabir, I Bojic, JJ Beunza; revised version received 12.06.24; accepted 15.06.24; published 03.10.24.

Please cite as:
Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y
Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study
JMIR Med Educ 2024;10:e52746
URL: https://mededu.jmir.org/2024/1/e52746
doi:10.2196/52746

© Zelin Wu, Wenyi Gan, Zhaowen Xue, Zhengxin Ni, Xiaofei Zheng, Yiyi Zhang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 3.10.2024. 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

XSL•FO
RenderX

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.

XSL•FO
RenderX

Original Paper

Development of a Clinical Simulation Video to Evaluate Multiple Domains of Clinical Competence: Cross-Sectional Study

Kiyoshi Shikino[1,2*], MD, MHPE, PhD; Yuji Nishizaki[3*], MD, MPH, PhD; Sho Fukui[4], MD, MPH; Daiki Yokokawa[2], MD, PhD; Yu Yamamoto[5], MD; Hiroyuki Kobayashi[6], MD, PhD; Taro Shimizu[7], MD, MPH, PhD; Yasuharu Tokuda[8,9], MD, MPH, PhD

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

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

Abstract

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

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

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

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

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

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e54401) doi:10.2196/54401

KEYWORDS

discrimination index; General Medicine In-Training Examination; clinical simulation video; postgraduate medical education; video; videos; training; examination; examinations; medical education; resident; residents; postgraduate; postgraduates; simulation; simulations; diagnosis; diagnoses; diagnose; general medicine; general practice; general practitioner; skill; skills

Introduction

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

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

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

Designed by a committee of experienced attending physicians organized by the JAMEP, the 2-hour GM-ITE comprises 80 MCQs covering multiple domains [9]. The scores range from 0 to 80, with higher scores indicating better performance and knowledge of internal medicine. The content and validity of each question undergo review by JAMEP's question-development committee comprising experienced physicians from various fields, an independent peer-review committee, and examination-analysis experts [10]. The GM-ITE is not used as a pass or fail test for training advancement but only as a source of education feedback. The test is strictly voluntary, and approximately one-third of residents take the examination each year (7669 in the 2020 academic year, 6869 in the 2019 academic year, 6133 in the 2018 academic year, 5593 in the 2017 academic year, and 4568 in the 2016 academic year) [11,12].

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

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

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

XSL•FO
RenderX

Methods

Study Design

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

Study Participants

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

Procedures

Innovative Examination Using High-Quality Patient-Simulated Video

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

Extended Matching Questions

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

XSL•FO
RenderX

Textbox 1. Clinical simulation video (CSV) innovative examination questions.

- Q1. Which 3 physical findings would you expect to be positive in this patient? Please choose 3 of the following:
  - Pallor of the eyelid conjunctiva
  - Pupil irregularity
  - Angry external jugular vein
  - Cervical vascular murmur
  - Thyroid gland enlargement
  - "Fixed" splitting of the second heart tone
  - Loud P2
  - Systolic murmur
  - Diastolic murmur
  - Torsion sound at the base of the lung
  - Tender points in the abdomen
  - Fresh blood in stool on rectal examination
  - Barre sign positive
  - Muscle stiffness
  - Loss of tendon reflexes
- Q2. Please state the most likely diagnosis for this patient (free text).
- Q3. Following the SBAR (situation, background, assessment, and recommendation) format, please prepare a patient handoff record for the internal medicine physician in charge of admission.
  - Q3-1. Situation (free text, 100 words maximum)
  - Q3-2. Background (free text, 100 words maximum)
  - Q3-3. Assessment (free text, 100 words maximum)
  - Q3-4. Recommendation (free text, 100 words maximum)
- Q4-1. Do you think the simulated patient-examination video was better suited to assessing your clinical competence than the traditional all-text format?
- Q4-2. Was the video simulation realistic enough for you to assess the patient?
- Q4-3. Did this experience increase your motivation to learn?

Modified Essay Questions

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

Anonymous Posttest Questionnaire

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

Measurements

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

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

Statistical Analyses

We conducted these analyses using SPSS Statistics for Windows (version 26.0; IBM Corp), following the Strengthening the Reporting of Observational Studies in Epidemiology guidelines. Two authors (KS and SF) independently assessed the answers and then discussed, identified, and agreed on them. We measured the interrater reliability with the κ coefficient (0.8-1.0=almost perfect, 0.6-0.8=substantial, 0.4-0.6=moderate,

XSL•FO
RenderX

and 0.2-0.4=fair) [18]. The Angoff method was used to define the cutoff for the DI calculation [19].

Ethical Considerations

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

Results

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

A total of 6 (11%) out of 56 participants answered Q2 correctly, and all the correct answers came from PGY 2 residents. The DI for the entire CSV innovative examination portion of the GM-ITE indicated high discriminatory power in all domains.

Figure 1 shows the DI for the MP (8 questions) domain, with 6 innovative questions scoring a DI of ≥0.20, indicating its robustness in differentiating examinee proficiency.

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

In Figure 3, the PP (18 questions) domain is analyzed, with 5 innovative questions achieving a DI of ≥0.20, demonstrating its effectiveness in assessing the examinees' clinical skillset.

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

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

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

Table 1. Discrimination index[a].

Domain (questions, n)	Question 1	Question 2	Question 3-1	Question 3-2	Question 3-3	Question 3-4
Medical interview or professionalism (8)	0.48	0.38	0.94	0.74	0.30	0.61
Clinical diagnosis (18)	0.50	0.40	0.77	0.56	0.27	0.18
Physical examination or procedure (18)	0.52	0.35	0.39	0.19	0.22	0.39
Disease knowledge (36)	–0.09	0.58	0.13	0.04	0.27	–0.10
Total (80)	0.06	0.47	0.10	–0.06	0.01	–0.12
Question type	MC[b]	FD[c]	FD	FD	FD	FD

[a]A discrimination index of ≥0.20 indicates that the question had high discriminatory power; a discrimination index of >0.40 indicates that the question was a very good measure of the participant's qualifications.

[b]MC: multiple choice.

[c]FD: free description (<100 words).

XSL•FO
RenderX

Figure 1. DIs of the examination scores of the General Medicine In-Training Examination: medical interview or professionalism (8 questions). DI: discrimination index; Q: question.

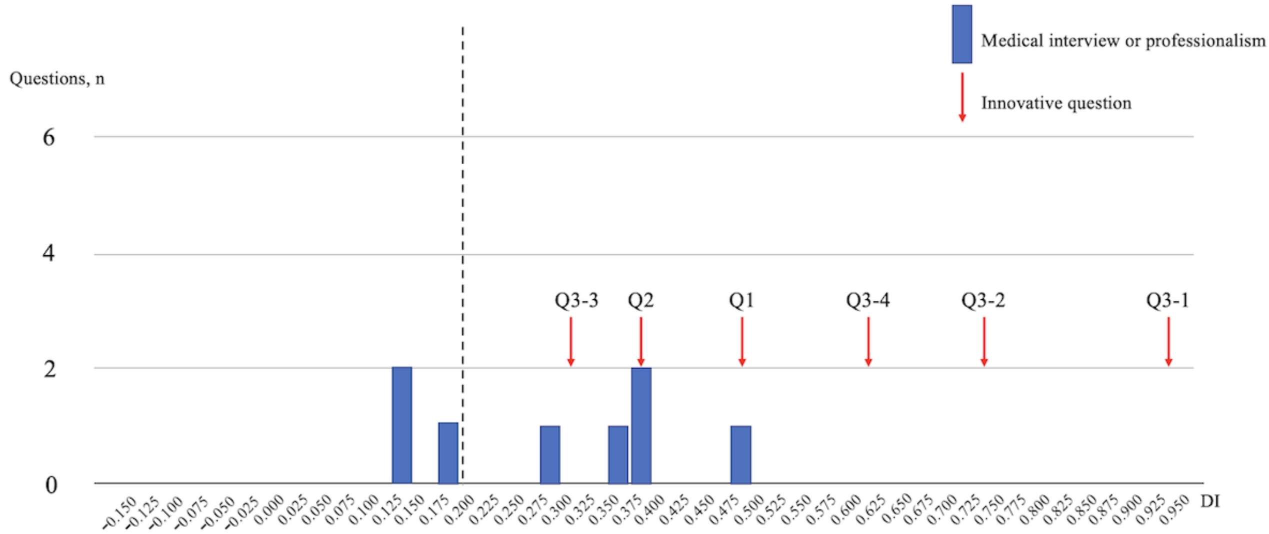
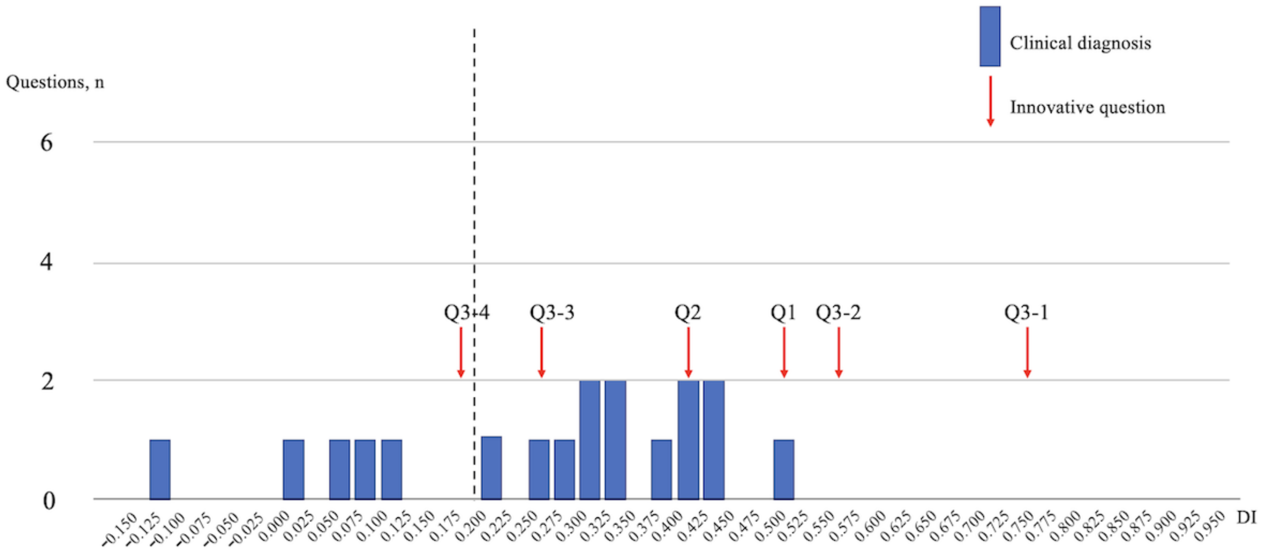


Figure 2. DIs of the examination scores of the General Medicine In-Training Examination: clinical diagnosis (18 questions). DI: discrimination index; Q: question.



XSL•FO
RenderX

Figure 3. DIs of the examination scores of the General Medicine In-Training Examination: physical examination or procedure (18 questions). DI: discrimination index; Q: question.

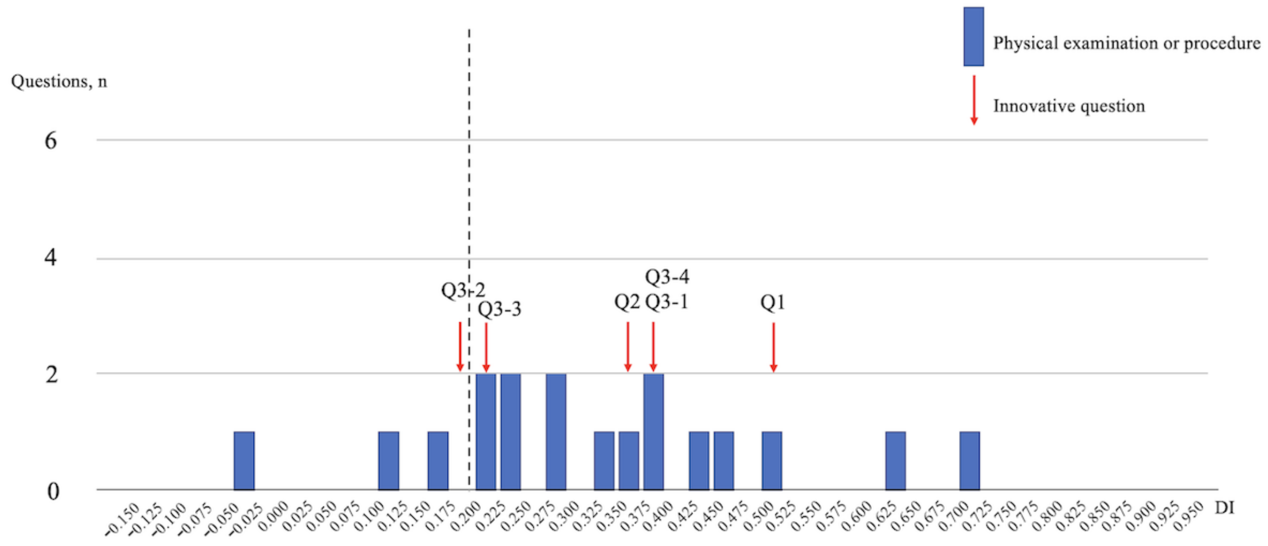
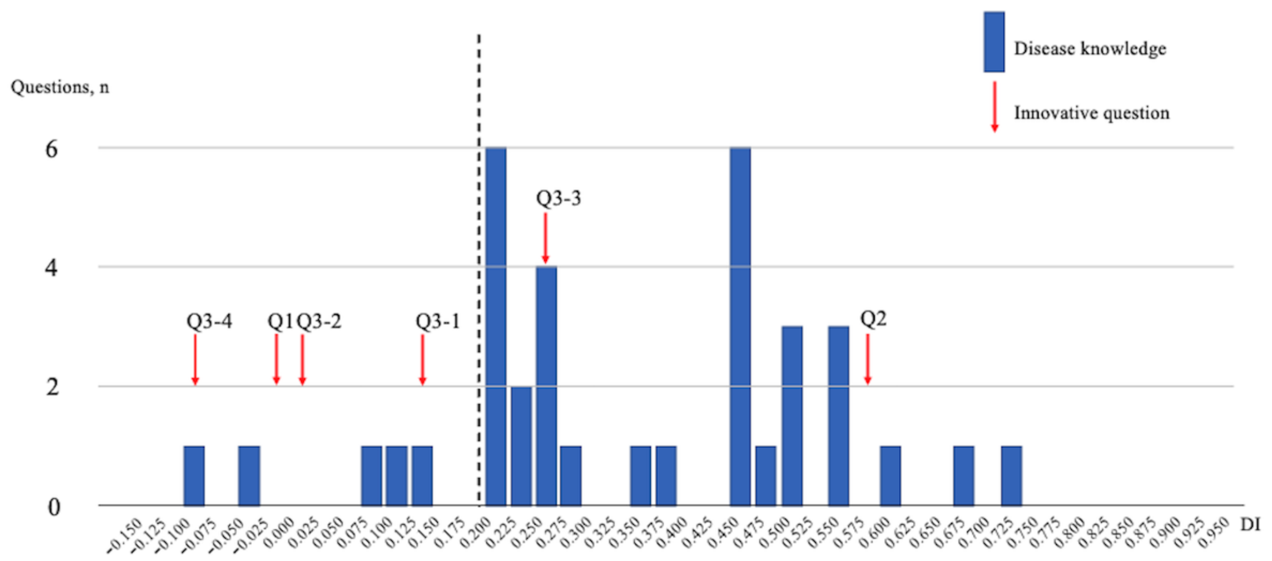


Figure 4. DIs of the examination scores of the General Medicine In-Training Examination: disease knowledge (36 questions). DI: discrimination index; Q: question.



Discussion

Principal Findings

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

Education, including medical education, has increasingly embraced computer-based testing. Today, students are accustomed to answering questions and writing essays via

XSL•FO
RenderX

computer-based testing. This study designed a single video simulation to assess the knowledge and skills of residents from the nonverbal information portion of the national medical licensing examination domains, particularly general theory. We included information from 3 domains in a single question, and the participants obtained high scores. This finding suggests that a single CSV module could test multiple skills and knowledge areas of residents. In other words, using innovative CSV-based questions could provide more realistic assessments while making the examinations more efficient.

The 3 domains covered in the CSV innovative examination Q1 (MP, CD, and PP) indicated DI of 0.4 or higher; the GM-ITE means were 0.32 (SD 0.13), 0.32 (SD 0.16), and 0.31 (SD 0.18), respectively. Therefore, the successful participants (based on GM-ITE scores) had higher scores on these domains in the CSV innovative examination question than in the GM-ITE. Q1 required participants to select 3 options from the MCQ (2 cutoffs per question). We found that the CSV could cover 3 separate domains in a single MCQ.

CSV innovative examination Q2 required a descriptive response; specifically, the participants needed to name the most likely diagnosis. Two physicians (KS and SF) independently assessed the diagnoses and achieved an agreement rate of 1.00. The DI of Q2 was 0.4 or higher for symptomatology or clinical reasoning and diseases and 0.3 or higher for general theory, physical examination, and clinical techniques. The overall GM-ITE scores had a high identification index of 0.47. Specifically, the CSV innovative examination Q2's requirement for participants to provide a definitive diagnosis allowed for a comprehensive assessment across all domains included in the GM-ITE. Furthermore, Q2 was distinguished as the sole question that demonstrated high DIs across individual disease categories. In addition, Q2 was the only question that also presented a high DI in each disease category.

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

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

Comparison to Prior Work

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

Limitations

Although this study reveals important findings, it has several limitations. First, the number of participants included in the study was low. For the data to be more valid, the number of examinees needs to be increased. However, adding more participants would also increase the test-scoring burden, which calls the viability of CSV-based testing into question. In this study, 2 physicians (KS and SF) scored the written questions. Increasing the number of examinees would also increase the time and effort required to score the results. If all of the approximately 8000 examinees who took the GM-ITE completed the CSV innovative examination module, the scoring time required would be untenable, and adding more CSV-based modules would compound the problem. One way to overcome this limitation could be the use of a morphological analysis or to only score a statistically significant sampling.

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

XSL•FO
RenderX

Conclusions

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

Acknowledgments

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

Data Availability

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

Authors' Contributions

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

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

Conflicts of Interest

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

Multimedia Appendix 1
Innovative examination.
[MP4 File (MP4 Video), 137419 KB - mededu_v10i1e54401_app1.mp4]

References

1. Kozu T. Medical education in Japan. Acad Med 2006;81(12):1069-1075 [FREE Full text] [doi: 10.1097/01.ACM.0000246682.45610.dd] [Medline: 17122471]
2. Teo A. The current state of medical education in Japan: a system under reform. Med Educ 2007;41(3):302-308. [doi: 10.1111/j.1365-2929.2007.02691.x] [Medline: 17316216]
3. Tago M, Shikino K, Hirata R, Watari T, Yamashita S, Tokushima Y, et al. General medicine departments of Japanese universities contribute to medical education in clinical settings: a descriptive questionnaire study. Int J Gen Med 2022;15:5785-5793 [FREE Full text] [doi: 10.2147/IJGM.S366411] [Medline: 35774114]
4. Garibaldi RA, Subhiyah R, Moore ME, Waxman H. The In-Training Examination in Internal Medicine: an analysis of resident performance over time. Ann Intern Med 2002;137(6):505-510. [doi: 10.7326/0003-4819-137-6-200209170-00011] [Medline: 12230352]
5. Kinoshita K, Tsugawa Y, Shimizu T, Tanoue Y, Konishi R, Nishizaki Y, et al. Impact of inpatient caseload, emergency department duties, and online learning resource on General Medicine In-Training Examination scores in Japan. Int J Gen Med 2015;8:355-360 [FREE Full text] [doi: 10.2147/IJGM.S81920] [Medline: 26586961]
6. Nundy S, Kakar A, Bhutta ZA. Developing learning objectives and evaluation: multiple choice questions/objective structured practical examinations. In: How to Practice Academic Medicine and Publish from Developing Countries?: A Practical Guide. Singapore: Springer; 2022.
7. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 2019;94(6):902-912 [FREE Full text] [doi: 10.1097/ACM.0000000000002618] [Medline: 30720527]

XSL•FO
RenderX

8. Cleary TJ, Battista A, Konopasky A, Ramani D, Durning SJ, Artino AR. Effects of live and video simulation on clinical reasoning performance and reflection. Adv Simul (Lond) 2020;5:17 [FREE Full text] [doi: 10.1186/s41077-020-00133-1] [Medline: 32760598]
9. Shimizu T, Tsugawa Y, Tanoue Y, Konishi R, Nishizaki Y, Kishimoto M, et al. The hospital educational environment and performance of residents in the General Medicine In-Training Examination: a multicenter study in Japan. Int J Gen Med 2013;6:637-640 [FREE Full text] [doi: 10.2147/IJGM.S45336] [Medline: 23930077]
10. Nagasaki K, Nishizaki Y, Nojima M, Shimizu T, Konishi R, Okubo T, et al. Validation of the General Medicine In-Training Examination using the Professional and Linguistic Assessments Board examination among postgraduate residents in Japan. Int J Gen Med 2021;14:6487-6495 [FREE Full text] [doi: 10.2147/IJGM.S331173] [Medline: 34675616]
11. Nagasaki K, Nishizaki Y, Shinozaki T, Shimizu T, Yamamoto Y, Shikino K, et al. Association between mental health and duty hours of postgraduate residents in Japan: a nationwide cross-sectional study. Sci Rep 2022;12(1):10626 [FREE Full text] [doi: 10.1038/s41598-022-14952-x] [Medline: 35739229]
12. Nishizaki Y, Shinozaki T, Kinoshita K, Shimizu T, Tokuda Y. Awareness of diagnostic error among Japanese residents: a nationwide study. J Gen Intern Med 2018;33(4):445-448 [FREE Full text] [doi: 10.1007/s11606-017-4248-y] [Medline: 29256086]
13. Tiffin PA, Illing J, Kasim AS, McLachlan JC. Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. BMJ 2014;348:g2622 [FREE Full text] [doi: 10.1136/bmj.g2622] [Medline: 24742539]
14. Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psychol 1939;30(1):17-24. [doi: 10.1037/h0057123]
15. Kanna B, Gu Y, Akhuetie J, Dimitrov V. Predicting performance using background characteristics of international medical graduates in an inner-city university-affiliated internal medicine residency training program. BMC Med Educ 2009;9:42 [FREE Full text] [doi: 10.1186/1472-6920-9-42] [Medline: 19594918]
16. Perez JA, Greer S. Correlation of United States Medical Licensing Examination and Internal Medicine In-Training Examination performance. Adv Health Sci Educ Theory Pract 2009;14(5):753-758. [doi: 10.1007/s10459-009-9158-2] [Medline: 19283500]
17. Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. Int J Appl Basic Med Res 2016;6(3):170-173 [FREE Full text] [doi: 10.4103/2229-516X.186965] [Medline: 27563581]
18. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33(2):363-374 [FREE Full text] [doi: 10.2307/2529786]
19. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, editor. Educational Measurement, 2nd Edition. Washington, DC: American Council on Education; 1971:508-600.
20. Epstein RM. Assessment in medical education. N Engl J Med 2007;356(4):387-396 [FREE Full text] [doi: 10.1056/NEJMra054784] [Medline: 17251535]
21. Hift RJ. Should essays and other 'open-ended'-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ 2014;14:249 [FREE Full text] [doi: 10.1186/s12909-014-0249-2] [Medline: 25431359]
22. Bird JB, Olvet DM, Willey JM, Brenner J. Patients don't come with multiple choice options: essay-based assessment in UME. Med Educ Online 2019;24(1):1649959 [FREE Full text] [doi: 10.1080/10872981.2019.1649959] [Medline: 31438809]
23. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach 2011;33(6):478-485. [doi: 10.3109/0142159X.2011.565828] [Medline: 21609177]
24. Kasai H, Shikino K, Saito G, Tsukamoto T, Takahashi Y, Kuriyama A, et al. Alternative approaches for clinical clerkship during the COVID-19 pandemic: online simulated clinical practice for inpatients and outpatients—a mixed method. BMC Med Educ 2021;21(1):149 [FREE Full text] [doi: 10.1186/s12909-021-02586-y] [Medline: 33685442]
25. Grissa MH, Dhaoui R, Ali KBH, Sekma A, Toumia M, Sassi S, et al. Comparison of simulation and video-based training for acute asthma. BMC Med Educ 2023;23(1):873 [FREE Full text] [doi: 10.1186/s12909-023-04836-7] [Medline: 37974223]
26. Saeed S, Khan MH, Siddiqui MMU, Dhanwani A, Hussain A, Ali MM. Hybridizing video-based learning with simulation for flipping the clinical skills learning at a university hospital in Pakistan. BMC Med Educ 2023;23(1):595 [FREE Full text] [doi: 10.1186/s12909-023-04580-y] [Medline: 37605200]

Abbreviations

CD: clinical diagnosis
CSV: clinical simulation video
DI: discrimination index
DK: disease knowledge
GM-ITE: General Medicine In-Training Examination
JAMEP: Japan Institute for Advancement of Medical Education Program
MCQ: multiple-choice question
MP: medical interview or professionalism

XSL•FO
RenderX

PGY: postgraduate year
PP: physical examination or procedure
SBAR: situation, background, assessment, and recommendation

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 08.11.23; peer-reviewed by R Khanmohammadi; comments to author 09.12.23; revised version received 14.12.23; accepted 28.01.24; published 29.02.24.

Please cite as:
Shikino K, Nishizaki Y, Fukui S, Yokokawa D, Yamamoto Y, Kobayashi H, Shimizu T, Tokuda Y
Development of a Clinical Simulation Video to Evaluate Multiple Domains of Clinical Competence: Cross-Sectional Study
JMIR Med Educ 2024;10:e54401
URL: https://mededu.jmir.org/2024/1/e54401
doi:10.2196/54401
PMID:38421691

©Kiyoshi Shikino, Yuji Nishizaki, Sho Fukui, Daiki Yokokawa, Yu Yamamoto, Hiroyuki Kobayashi, Taro Shimizu, Yasuharu Tokuda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Teaching Digital Medicine to Undergraduate Medical Students With an Interprofessional and Interdisciplinary Approach: Development and Usability Study

Annabelle Mielitz[1], MSc; Ulf Kulau[2], Dr.-Ing.; Lucas Bublitz[2], MSc; Anja Bittner[3], MD; Hendrik Friederichs[4], MD, PhD, MA; Urs-Vito Albrecht[1], MD, DrPH, PhD

[1]Department of Digital Medicine, Medical School OWL, Bielfeld University, Bielefeld, Germany

[2]Department of Smart Sensors, Hamburg Technical University, Hamburg, Germany

[3]Dean's Office for Academic Affairs, Medical School OWL, Bielfeld University, Bielefeld, Germany

[4]Department for Medical Education, Medical School OWL, Bielfeld University, Bielefeld, Germany

Corresponding Author:
Urs-Vito Albrecht, MD, DrPH, PhD
Department of Digital Medicine
Medical School OWL
Bielfeld University
Universitätsstraße 25
Bielefeld, 33615
Germany
Phone: 49 521 106 ext 86714
Email: urs-vito.albrecht@uni-bielefeld.de

Abstract

Background: An integration of digital medicine into medical education can help future physicians shape the digital transformation of medicine.

Objective: We aim to describe and evaluate a newly developed course for teaching digital medicine (the Bielefeld model) for the first time.

Methods: The course was held with undergraduate medical students at Medical School Ostwestfalen-Lippe at Bielefeld University, Germany, in 2023 and evaluated via pretest-posttest surveys. The subjective and objective achievement of superordinate learning objectives and the objective achievement of subordinate learning objectives of the course, course design, and course importance were evaluated using 5-point Likert scales (1=*strongly disagree*; 5=*strongly agree*); reasons for absences were assessed using a multiple-choice format, and comments were collected. The superordinate objectives comprised (1) the understanding of factors driving the implementation of digital medical products and processes, (2) the application of this knowledge to a project, and (3) the empowerment to design such solutions in the future. The subordinate objectives comprised competencies related to the first superordinate objective.

Results: In total, 10 undergraduate medical students (male: n=4, 40%; female: n=6, 60%; mean age 21.7, SD 2.1 years) evaluated the course. The superordinate objectives were achieved well to very well—the medians for the objective achievement were 4 (IQR 4-5), 4 (IQR 3-5), and 4 (IQR 4-4) scale units for the first, second, and third objectives, respectively, and the medians for the subjective achievement of the first, second, and third objectives were 4 (IQR 3-4), 4.5 (IQR 3-5), and 4 (IQR 3-5) scale units, respectively. Participants mastered the subordinate objectives, on average, better after the course than before (presurvey median 2.5, IQR 2-3 scale units; postsurvey median 4, IQR 3-4 scale units). The course concept was rated as highly suitable for achieving the superordinate objectives (median 5, IQR 4-5 scale units for the first, second, and third objectives). On average, the students strongly liked the course (median 5, IQR 4-5 scale units) and gained a benefit from it (median 4.5, IQR 4-5 scale units). All students fully agreed that the teaching staff was a strength of the course. The category *positive feedback on the course or positive personal experience with the course* received the most comments.

Conclusions: The course framework shows promise in attaining learning objectives within the realm of digital medicine, notwithstanding the constraint of limited interpretability arising from a small sample size and further limitations. The course concept aligns with insights derived from teaching and learning research and the domain of digital medicine, albeit with identifiable areas for enhancement. A literature review indicates a dearth of publications pertaining to analogous courses in Germany. Future

XSL•FO
RenderX

investigations should entail a more exhaustive evaluation of the course. In summary, this course constitutes a valuable contribution to incorporating digital medicine into medical education.

(JMIR Med Educ 2024;10:e56787) doi:10.2196/56787

KEYWORDS

medical education; digital medicine; digital health

Introduction

Background

Digital health, the field in which health care is linked to technology [1], encompasses digital medicine, which refers to the specific use of digital technology products in health care [1]. Following the suggestion by Bahagon and Jacobson [2] that the successful implementation of such products is "the expertise of tailoring knowledge and leadership capabilities in multidisciplinary areas: clinical, ethical, psychological, legal, comprehension of patient and medical team engagement etc...," *digital medicine* is defined in this paper as follows: digital medicine is concerned with the holistic development and application of digital health applications, where *holistic* means that nonmedical issues and aspects relevant to such development and application are taken into account.

The World Health Organization recommends implementing digital health—and, consequently, digital medicine—technologies in the health care sector due to the potential positive impact of such technologies on health care [3]. This includes products such as health apps and smart devices for monitoring purposes and the use of technologies such as artificial intelligence, big data analysis, advanced computing, and robotics. Promising fields of application are, for example, for cardiovascular diseases [4,5], diabetes [6], and mental health [7]. Successful implementation of digital technologies is underway in many areas of health care. Still, it depends on a culture of change that results in the reorganization of services based on the public's health needs. Among other authors, Iyamu et al [8] refer to this process as the digital transformation of medicine. As relevant stakeholders, physicians have a responsibility to shape this transformation. To prepare and enable future physicians to do so, they need to acquire relevant competencies during their studies. For the first time, skills in the field of digital medicine are classified as core competencies for the medical profession in the draft bill of the medical licensing regulations of June 2023 [9,10]. Training should enable future physicians to take an "active and self-designing role to develop a goal-oriented approach to current technological solutions on their responsibility" [11]. In Germany, the National Competency-Based Catalog of Learning Objectives (NKLM), which lists the competencies that medical students should acquire during their studies, specifies that medical students should acquire the competence of knowing and being able to reflect on the areas of application of digital medicine and the significance of digitalization. For example, they should be able to explain the application scenarios for telemedicine applications and their framework conditions [12]. Courses in digital medicine are necessary to acquire these competencies. Against this background, an integration of topics of digital health and medicine into medical education is recommended [3,11,13-15]. It has already been shown that medical students feel like the early integration of such topics into medical education can help prepare them for their future work environment [16].

Courses on digital health are already offered as part of medical studies in various countries, including Germany [17,18], but the range of classes dealing with digital medicine must be further expanded to enable a true digital transformation of the health care system. Studies such as those by Jacob et al [11], Schreiber et al [15], and Machleid et al [16] show that digital medicine has so far not been given sufficient consideration in medical education. Thus, it can be assumed that future physicians do not yet have the required digital medicine competencies. Many students feel unprepared for the future regarding the digitalization of health care by their teachers and the courses they attended and wish for better integration of digital medicine and digital health topics into their studies [11,16,19]. Against the background of the integration of competencies in digital medicine in the draft bill of the medical licensing regulations of June 2023 [10], an expansion of courses related to digital medicine in Germany may be expected as soon as the new medical licensing regulations come into force.

Following the call for corresponding adjustments, the newly founded Medical School Ostwestfalen-Lippe (OWL; the region of Ostwestfalen-Lippe, Germany) at Bielefeld University has set up a Digital Medicine working group within the faculty and included Digital Medicine in the regular curriculum of medical studies. In addition, a Technological Transformation in Medicine profile is offered for continuing education. In this profile, the Digital Medicine working group contributes with a course called Digital Medicine, referred to in this paper as the "Bielefeld model." The overarching educational objective is to empower students to feel able to actively engage as prospective architects of digital medicine, transcending their role as mere users or consumers of digital products or participants in digital processes. This means that students will need to be equipped to develop and use digital applications comprehensively.

This overarching educational objective translates to 3 specific superordinate learning objectives of the course (Textbox 1).

XSL•FO
RenderX

Textbox 1. The 3 superordinate learning objectives of the course.

Superordinate learning objective 1

Students will know the factors influencing the sustainable implementation of digital medical products and processes.

Superordinate learning objective 2

Students will be able to apply their knowledge of the factors influencing the sustainable implementation of digital medical products and processes in developing a specific project.

Superordinate learning objective 3

Students will feel empowered to design sustainable digital products and processes in future projects.

The implementation of digital products and processes is considered to be sustainable if it is independent of individual motivated implementers. To design sustainable solutions, it is important to consider a wide range of influencing factors. Only if these are taken into account will products and processes be able to achieve a high level of acceptance in the health care system in the long term, develop their full positive impact, and remain in place.

The first of these 3 superordinate learning objectives, in turn, translates to 17 subordinate learning objectives (Textbox 2). These subordinate learning objectives include abstract competencies related to the first superordinate learning objective (subordinate learning objectives 1 to 6) as well as specific competencies in relation to the individual factors that are subsumed under the first superordinate learning objective (subordinate learning objectives 7 to 17). Thus, these subordinate learning objectives should capture aspects of the first superordinate learning objective.

Textbox 2. The 17 subordinate learning objectives of the course.

Subordinate learning objective ID and description

- Sub01: the students can identify players and stakeholders in the field of digital medicine.
- Sub02: the students can name and reflect on overarching themes of digital medicine.
- Sub03: the students can identify, address, and discuss problem areas (medical, technical, legal, ethical, and social) of digital medicine.
- Sub04: the students can critically assess digital medicine and evaluate it based on its opportunities and challenges.
- Sub05: the students can explain the product life cycle of digital medicine and plan an applied project based on it.
- Sub06: the students can plan a concrete project with basic knowledge of project management.
- Sub07: the students can identify markets for digital medicine and discuss challenges of market entry for digital applications.
- Sub08: the students can identify and communicate the necessary information required for commissioning a technical development of digital applications.
- Sub09: the students can explain the legal challenges of using digital tools and digital communication media in relation to digital medicine.
- Sub10: the students can explain the regulatory difference between a medical device and other products and explain legal consequences.
- Sub11: the students can identify and discuss the ethical implications of digital medicine for patients, physicians, contributors, society, and the environment.
- Sub12: the students can explain interoperability and know what characterizes interoperability.
- Sub13: the students can assess the quality of digital medicine applications.
- Sub14: the students can recognize the quality of the usability of a digital application and know which factors can influence the usability.
- Sub15: the students can discuss elements of data science and its tools and requirements in data preparation and data analysis.
- Sub16: the students can discuss and evaluate digital medicine with regard to specific aspects in the context of gender and sex.
- Sub17: the students can explain data protection and data security challenges in the development and use of digital tools and digital communication media.

The decision to design an own course concept instead of adopting an existing one was made to be able to design a course that is very specifically geared toward achieving these learning objectives.

Objectives

The aim of this paper is to introduce the Bielefeld model for the first time and undertake an initial exploratory evaluation.

In particular, the aim of the course evaluation is to determine the following:

1. How well the participants objectively and subjectively achieved the 3 superordinate learning objectives (as presented in Textbox 1) and how well they objectively achieved the 17 subordinate learning objectives of the course (as presented in Textbox 2).

XSL•FO
RenderX

2. Whether the course was well designed. This encompasses how suitable the participants considered the course to be for achieving the superordinate learning objectives, whether they enjoyed the course and benefitted from it, and which aspects they considered as strengths of the course.
3. How important the course was to the participants. This encompasses how important the achievement of the superordinate learning objectives was to them personally and the reasons for the participants' potential absences during the course. Those reasons might indicate whether taking part in the course was important for the participants.

Methods

The Digital Medicine Course

Setting

Undergraduate medical students at the Medical School OWL go through a model curriculum with an interdisciplinary specialization in their first 3 years of studies. The students can choose, or, in rare cases, are assigned to 1 of 5 profile options with Technological Transformation in Medicine (TeTraMed) being one of them. Digital Medicine is the third of the 6 mandatory modules the TeTraMed profile comprises. The module is completed with an examination. Students can earn a bachelor's degree in Interdisciplinary Medical Sciences alongside their medical studies, with course credits counting toward both degrees.

Course Concept

The instructors of the course designed the course and developed the learning objectives based on their experience in research and teaching in digital medicine and in the development of digital medicine applications and based on considerations for the design of an effective course concept. The main instructor is a physician, computer scientist, and public health scientist and a professor of digital medicine at Bielefeld University. Coinstructors (a psychologist, physician, and medical engineer) were members of this working group. The overarching goals and learning objectives of the course were to be achieved by the students by developing a digital medicine application in interdisciplinary collaboration with electrical engineering students from Hamburg University of Technology (TUHH; when students are mentioned hereafter in connection with the course, this refers to medical students. Electrical engineering students are only referred to when they are specifically mentioned. When mentioning the course, unless otherwise stated, this refers to the course for medical students). The course was co-designed with a professor of smart sensors who is skilled in computer science and electrical engineering and teaches the electrical engineering students at TUHH. Consistent with the 3 superordinate learning objectives (Textbox 1), the didactic emphasis of the course was directed toward investigating the manifold factors associated with and influencing the domain of digital medicine. Adhering to project-based learning methodologies, medical students collaboratively engaged in small groups in their own projects related to telemedicine (using technologies for providing health care over a distance [20]), an area of digital medicine. The students were divided into 3 groups, and each group was assigned 1 of 3 cases. The cases described circumstances in which individuals required monitoring due to medical conditions or circumstances (for more detail, see Multimedia Appendix 1). The aim of the students was to systematically solve the challenge described in the case by using mobile sensor technology, exemplifying its application in a corresponding product and process. They were also free to design an accompanying app. A course that was separate but linked to the course for medical students was held at TUHH with the electrical engineering students. Their course followed a research-based learning design. The focus of the electrical engineering students was primarily on developing the sensor systems that the medical students designed in their projects. The sensor should have been designed by the end of the course, but the aim was not to actually implement it on the market. The collaboration between the courses at TUHH and Bielefeld University was designed to emulate the fact that, when working on projects such as that in the Bielefeld model in real life outside the university, a collaboration between medical professionals and technicians is often required to technically implement what medical professionals design. The central concept involved both medical and electrical engineering students collaborating without silos, working together as equals. This approach fosters interdisciplinary cooperation, promotes precise and accurate communication, and equips students for project work in professional settings across both fields. Apart from this collaboration, the medical students and the electrical engineering students were taught separately. The electrical engineering students did not take part in the regular sessions of the Bielefeld model.

The students experienced the entire product life cycle within the Digital Medicine course, from brainstorming to planning, developing, evaluating, implementing, operating, and decommissioning digital products and processes. They were confronted with challenges posed by the diverse interests of the stakeholders, frameworks, resources, and settings involved and had to take into account the factors influencing the sustainable implementation of digital medical products and processes. This encompassed an exploration of their interrelationships and their consequential impact on the development and enduring implementation of digital medical products and processes. By doing so, they practiced directly acknowledging, accepting, and addressing these factors.

On the basis of their experience, the instructors identified and addressed the following multidimensional factors in the course: project management, market entry, technical development, quality, data science, interoperability, usability, law, regulation and ethics, sex and gender sensitivity, and data protection and security. These factors result from the interfaces between medicine and health with the fields of technology, informatics, society, law, regulation, ethics, economics, and psychology.

In each session, one or several of these factors were addressed by the course instructors and often also by external experts who were invited as guest lecturers (Table 1; see Multimedia Appendix 2 for more detailed information about the topics of the sessions and the guest lecturers). Table 1 shows which of the 17 subordinate learning objectives relate to the specific sessions. It was assumed that subordinate learning objectives 1

XSL•FO
RenderX

to 6 (Textbox 2) would be achieved throughout the course. The subordinate learning objectives that relate specifically to the factors subsumed under the first superordinate learning objective (subordinate learning objectives 7 to 17; Textbox 2) should be achieved in the corresponding sessions. After the lecture in which the factor or factors were addressed, the students connected the topic to their own projects in a practical unit, reflecting on its impact and identifying what considerations were necessary for their specific work. For example, in the session on law, regulation, and ethics, students were able to consider whether their sensor would be a medical device from a legal point of view, or in the session on usability, for example, they were able to reflect on how usable their product was and how usability could be increased. The students could exchange ideas, discuss, develop further content, and converse about general and project-related questions with the lecturers alone or in a group. In this process, the students identified and named requirements that their sensors should fulfill. They communicated the technical requirements to the electrical engineering students in Hamburg, who implemented these specifications for the sensors in consultation with the medical students. Progress and outcomes of the small groups were systematically documented in a project outline (as illustrated in Multimedia Appendix 3) subject to continuous updates and revisions throughout the course.

Table 1. Digital Medicine curriculum within the TeTraMed profile.

Session	Topic	Life cycle phase	Subordinate learning objective to be achieved in the session	Guest lecturer
1	Introduction	Idea generation	—[a]	No
2	Project management	Planning	6	Yes
3	Market entry	Planning	7	Yes
4	Technical development	Development	8	Yes
5	Quality[b]	Development	13	No
6	Data science[c]	Development	15	Yes
Additional session A	Meeting with the electrical engineering students	Development	8	—
7	Interoperability	Development	12	Yes
8	Usability	Premarket evaluation	14	Yes
9	Law, regulation, and ethics	Implementation	9, 10, and 11	Yes
10	Sex and gender sensitivity	Postmarket evaluation	16	Yes
11	Data protection and data security	Use	17	No
12	End	Decommissioning	—	No
Additional session B[d]	Ceremony	Memorial	—	—

[a]Not applicable.

[b]Deviation from the original curriculum (Law, Regulation, Ethics) was changed following the students' wishes as it conflicted with other study responsibilities.

[c]Deviation from the original curriculum (Interoperability) was changed following the students' wishes as it collided with other study liabilities.

[d]In this session, it was planned to hand the participants certificates that confirmed participation in the course. This session had to be canceled due to situational circumstances, but the certificates were handed out anyway.

By confronting new issues and challenges, students were encouraged to reflect on their own project on an ongoing basis and then revise it. The course instructors and external experts from the different interface areas provided knowledge transfer and support for the projects. The course concept provided self-management skills (managing oneself and the available resources) as well as project and team management skills, such as preparing, implementing, and recording project meetings. Work, communication, interaction, negotiation, and conflict resolution in interdisciplinary, international, and intercultural teams and exchange and work in English were also part of this. The course was primarily offered in an analog-oriented format; essentially, only the meetings with the electrical engineering students took place in web-based live video formats. The course was supported by web-based materials provided on the university e-learning platform LernraumPlus. Due to external organizational circumstances, some sessions were only available on the web for self-study.

Course Structure

The Digital Medicine course was structured in 12 sessions, usually held weekly, distributed over 4 months in the summer semester of 2023 (April to July). Each session consisted of a 45-minute lecture, a 90-minute workshop during which the students worked on their projects, and a 90-minute guided self-study for which no contact time with the teaching staff was scheduled but was offered at times. Thus, there were a total of 9 hours of lectures, 18 hours of workshops, and 18 hours of guided independent study. In addition, times without contact with the teachers were set aside for preparation and follow-up

XSL•FO
RenderX

work. In total, 3 further scheduled dates were canceled due to holidays. The structure of each course day is shown in Multimedia Appendix 4.

Evaluation and Analysis

Overview

The course concept was evaluated by recruiting undergraduate medical students at the Medical School OWL who took part in the course between April and July 2023. The students were asked to complete a web-based presurvey at the beginning of the course and a postsurvey after the course had ended (after the 12th session) in German using a customized web-based survey tool. For this, they first had to create a pseudonym. The mapping between the pseudonyms and the participants was unknown to the study staff, so anonymity was maintained. However, using the pseudonyms made it possible to merge the individual data sets of the pre- and postsurveys. A customized instrument (using the formr survey framework [21]) was provided for the survey by the Medical Education working group at the Medical School OWL at Bielefeld University. The presurvey was administered on April 3, 2023, and the postsurvey was administered between July 10, 2023, and August 1, 2023. Students were invited anonymously via email and by the instructors during the course. A total of 2 follow-up actions were conducted via email at an interval of 2 weeks after the beginning of the postevaluation.

Instruments

Overview

The presurvey consisted of 27 items: the demographic variables of age and gender, the items related to the objective achievements of the subordinate learning achievements (pre- and postsurvey), as well as the open comments (again, pre- and post-survey, see corresponding tables in the Results section), and an additional 10 items that will not be described or analyzed in this paper but will be focused on in a separate publication. These additional items were administered to assess the achievement of learning goals following a catalog of learning outcomes as described by Foadi et al [22] that partially corresponds to the NKLM. The postsurvey consisted of 47 items and additional 10 items based on the learning outcomes as described by Foadi et al [22] that are neither described nor evaluated in this paper. A comprehensive overview of all items that were used can be found in Multimedia Appendix 5.

All responses to the surveys (except regarding items AB01 and COM01, as described later in the manuscript when these items are presented) were recorded on an ordinal 5-point Likert scale, which was recoded after the study to be more consistent with the usual coding system (the coding in the questionnaire was as follows: 1=*strongly agree*, 2=*rather agree*, 3=*rather neutral*, 4=*rather disagree*, and 5=*strongly disagree*; the scale was recoded as follows for the analyses: 1=*strongly disagree*, 2=*rather disagree*, 3=*rather neutral*, 4=*rather agree*, and 5=*strongly agree*). All items (except those by Foadi et al [22]) were developed in-house, as described in detail in the following sections.

Achievement of the Super- and Subordinate Learning Objectives

The objective and subjective achievement of the 3 superordinate learning objectives was assessed using 1 item each in the postsurvey (for objective achievement: items SUPER1_OA, SUPER2_OA, and SUPER3_OA; for subjective achievement: items SUPER1_SA, SUPER2_SA, and SUPER3_SA). The objective achievement of the subordinate learning objectives was assessed using 17 items (items sub01 to sub17). By recording the objective achievement of the subordinate learning objectives in both the pre- and the postsurvey, it was possible to determine whether the participants would have achieved these objectives better after the course than before.

As the subordinate learning objectives should capture aspects of the first superordinate learning objective, items measuring the objective achievement of these subordinate objectives should capture aspects of the objective achievement of this first superordinate learning objective. In that case, changes in the median of items measuring the objective achievement of the subordinate learning objectives from the pre- to the postsurvey would also indirectly provide information about how the objective achievement of the first superordinate learning objective developed from the pre- to the postsurvey.

Design of the Course

One item per superordinate learning objective was used to record the extent to which the participants considered the course concept suitable for achieving this objective (items SUPER1_SUIT, SUPER2_SUIT, and SUPER3_SUIT). In total, 2 items were used to measure whether the participants enjoyed the course (item FUN01) and whether they felt that they benefited from it (item BEN01). In 4 questions, participants were able to indicate the extent to which they perceived 4 potential benefits as actual strengths of the course (items STR01 to STR04).

Importance of the Course to the Participants

One item per superordinate learning objective was used to record how important it was to the participants to achieve this objective (items SUPER1_IMP, SUPER2_IMP, and SUPER3_IMP).

One item was used to measure the reasons for participants' absences (item AB01) during the course. This item did not use a Likert scale but instead offered various answer options for participants to choose from, allowing them to give multiple answers.

Open Comments

Participants could write comments in an open-response format at the end of the pre- and postsurvey (item COM01).

Development, Validity, and Reliability of the Instruments

All items were developed through an iterative process informed by the insights of 3 authors. The items assessing the objective achievement of the superordinate (items SUPER1_OA, SUPER2_OA, and SUPER3_OA) and subordinate (all items) learning objectives were developed by formulating the learning objectives in first-person singular format and pairing them with a 5-point Likert scale. The development of the learning

XSL•FO
RenderX

objectives and the items based on them was carried out by 2 (in the case of the learning objectives) to 3 (in the case of the items) authors in an iterative process. The consistency of the formulations with the course's conceptual objectives was regularly reviewed to ensure strong content validity. The reason for the development of these new items was that, until now, no instrument has reflected the course concept closely enough to capture what was to be learned in the course. The content validity of the other items was also ensured by matching the formulations in an iterative process with the underlying idea of what should be captured using these items. The items assessing whether the participants enjoyed the course (item FUN01) and whether they benefitted from it (item BEN01) have been used with similar or related wordings in other studies [23,24]. Therefore, content validity was assumed.

The items did not capture common things but were instead considered individually, so they were not combined as a scale. Thus, no internal consistencies were calculated. Interrater reliability was assumed for all items that were recorded and statistically evaluated on the Likert scale as the evaluation was objectively independent of the rater.

Analysis

All items were analyzed descriptively. The mean and SD were calculated to describe the age-related data. The information on gender was analyzed through the calculation of occurrence frequencies. The Spearman ρ was calculated to investigate a correlation between the item measuring the objective achievement of the first superordinate learning objective and the 17 items measuring the objective achievement of the subordinate learning objectives in the postsurvey. A positive correlation would indicate that the items measuring the objective achievement of the subordinate learning objectives are well suited to capture aspects of the objective achievement of the first superordinate learning objective. The items that were recorded using the ordinal 5-point Likert scales were evaluated by calculating the median and IQR. Using a descriptive account of the aggregated median values of the students for the items assessing the objective achievement of the subordinate learning objectives, we compared a possible change in this achievement from the pre-to the postsurvey. Textbox 3 shows how the changes in the median of these items from the pre- to the postsurvey will be evaluated qualitatively. The qualitative evaluation was based on considerations of how changes are to be evaluated. Due to the small sample, an inferential statistical analysis was not performed. The comments were analyzed by dividing them into main categories and subcategories based on the structuring qualitative content analysis according to Kuckarzt and Rädiker [25]. The frequency of occurrence of each subcategory was determined.

Textbox 3. Qualitative evaluation of the changes in the median of the objective achievement of the subordinate learning objectives.

Size of change and qualitative evaluation

- 0: no change
- >0 to 0.4: minimal change
- 0.5 to 0.9: small change
- 1 to 1.4: rather big change
- 1.5 to 1.9: big change
- >1.9: extensive change

Ethical Considerations

The study was approved by the Ethics Committee Ethik-Kommission Westfalen-Lippe, located in Münster, Germany, under the chairmanship of Professor Dr Wolfgang E Berdel, on May 11, 2023 (2023-233-f-S).

Results

All 15 course participants were recruited for the study. A total of 10 participants (n=4, 40% male and n=6, 60% female; mean age 21.7, SD 2.1 years) completed both the pre- and the postsurvey and were included in the analysis.

Achievement of the Super- and Subordinate Learning Objectives

Table 2 shows the results regarding the objective and subjective achievement of the superordinate learning objectives (for information on how frequently the individual response categories were selected by the participants, see Multimedia Appendix 6). Regarding the objective achievement of these objectives, medians of 4 (IQR 4-5), 4 (IQR 3-5), and 4 (IQR 4-4) scale units were found for the first, second, and third superordinate learning objectives, respectively. Concerning the question of whether the participants also subjectively achieved these learning objectives, medians of 4 (IQR 3-4), 4.5 (IQR 3-5), and 4 (IQR 3-5) scale units were found for the first, second, and third learning objectives, respectively. Therefore, the median values varied between the scale points *rather agree* (scale value of 4) and *strongly agree* (scale value of 5). Thus, on average, the results indicate a rather to very good achievement of the 3 superordinate learning objectives both objectively and subjectively.

XSL•FO
RenderX

Table 2. Objective and subjective achievement of the 3 superordinate learning objectives, the suitability of the course for achieving them, and the importance of achieving them, administered in the postsurvey (N=10)[a].

Superordinate learning objective and item ID	Item	Scale units, median (IQR)
Superordinate learning objective 1		
SUPER1_OA[b]	I know the factors that influence the sustainable implementation of digital medical products and processes.	4 (4-5)
SUPER1_SUIT[c]	The course concept was suitable for teaching these factors.	5 (4-5)
SUPER1_IMP[d]	It was important to me to achieve this learning goal.	4 (3-4)
SUPER1_SA[e]	I have achieved this learning goal from a personal perspective.	4 (3-4)
Superordinate learning objective 2		
SUPER2_OA[f]	I can apply my knowledge regarding the factors that influence the sustainable implementation of digital medical products and processes in developing a concrete project.	4 (3-5)
SUPER2_SUIT	The course concept was suitable to apply my knowledge regarding these factors in developing a concrete project.	5 (4-5)
SUPER2_IMP	It was important to me to achieve this learning goal.	4 (3-5)
SUPER2_SA	I have achieved this learning goal from a personal perspective.	4.5 (3-5)
Superordinate learning objective 3		
SUPER3_OA[f]	I feel empowered to design sustainable digital products and processes in future projects.	4 (4-4)
SUPER3_SUIT	The course concept was suitable to enable me to design sustainable digital products and processes in future projects.	5 (4-5)
SUPER3_IMP	It was important to me to achieve this learning goal.	3.5 (3-5)
SUPER3_SA	I have achieved this learning goal from a personal perspective.	4 (3-5)

[a]The “superordinate learning objective” column indicates to which superordinate learning objective the respective items belong.

[b]OA: objective achievement (of the respective superordinate learning objective).

[c]SUIT: suitability (of the course concept for achieving the respective superordinate learning objective).

[d]IMP: importance (of achieving the respective superordinate learning objective).

[e]SA: subjective achievement (of the respective superordinate learning objective).

[f]This item was answered by only 90% (9/10) of the participants.

Regarding the objective achievement of the subordinate learning objectives, on average, the participants performed better on all items after the course than before (Table 3; for information on how frequently the individual response categories were selected by the participants, see Multimedia Appendix 7). While an average median of 2.5 (IQR 2-3) scale units was achieved across the items in the presurvey, it increased by 1.5 scale units to 4 (IQR 3-4) in the postsurvey. This can be rated as a big change (Textbox 3). While, in the presurvey, the median values varied between the scale points *strongly disagree* (scale value of 1) and *rather agree* (scale value of 4), in the postsurvey, they varied between the scale points *rather neutral* (scale value of 3) and *strongly agree* (scale value of 5). While the presurvey results, therefore, indicate a strong nonachievement to rather good achievement on average of the subordinate learning objectives, the postsurvey results indicate a partial to very good achievement of those objectives on average. When analyzing the change in the median for each session, there was an improvement from the pre- to the postsurvey in 15 of the 17 items (Table 3; range of change in the median between 0.5 and 2 scale units). For items sub03 and sub04, no change in the median values could be found. Still, when looking at how often the individual response categories were selected (Multimedia Appendix 7), it can be seen that, in these items, there was also an improvement from the pre- to the postsurvey. While only 60% (6/10) of the participants strongly or rather agreed with items sub03 and sub04 in the presurvey, 100% (10/10) of the participants strongly or rather agreed with these items in the postsurvey. It can be summarized that, on average, all subordinate learning objectives were achieved better after the course than before by the participants.

XSL•FO
RenderX

Table 3. Objective achievement of the subordinate learning objectives, administered in the pre- and postsurvey (N=10).

Item ID	Item	ρ[a]	Presurvey score, median (IQR)[b]	Postsurvey score, median (IQR)[c]	ΔMedian[d]
Sub01	I can identify players and stakeholders in the field of digital medicine.	0.37	2 (1-3)	4 (3-5)	2
Sub02	I can name and reflect on overarching themes of digital medicine.	0.56	3 (2-4)	4.5 (4-5)	1.5
Sub03	I can identify, address, and discuss problem areas (medical, technical, legal, ethical, and social) of digital medicine.	0.11	4 (3-5)	4 (4-5)	0
Sub04	I can critically assess digital medicine and evaluate it based on its opportunities and challenges.	0.30	4 (3-4)	4 (4-5)	0
Sub05	I can explain the product life cycle of digital medicine and plan an applied project based on it.	0.73	2 (1-2)	4 (3-4)	2
Sub06	I can plan a concrete project with basic knowledge of project management.	0.85	2 (1-2)	4 (3-5)	2
Sub07	I can identify markets for digital medicine and discuss challenges of market entry for digital applications.	0.14	3 (1-3)	4 (3-5)	1
Sub08	I can identify and communicate necessary information required for commissioning a technical development of digital applications.	0.42	2 (2-3)	4 (4-4)	2
Sub09	I can explain the legal challenges of using digital tools and digital communication media in relation to digital medicine.	0.56	3 (2-3)	4 (4-4)	1
Sub10	I can explain the regulatory difference between a medical device and other products and explain legal consequences.	0.52	3 (2-3)	4 (4-5)	1
Sub11	I can identify and discuss the ethical implications of digital medicine for patients, physicians, contributors, society, and the environment.	0.65	3 (3-4)	3.5 (3-4)	0.5
Sub12	I can explain interoperability and know what characterizes interoperability.	0.79	1.5 (1-3)	3.5 (3-4)	2
Sub13	I can assess the quality of digital medicine applications.	0.60	3 (2-3)	4 (4-5)	1
Sub14	I can recognize the quality of the usability of a digital application and know which factors can influence the usability.	0.38	2 (1-3)	4 (4-5)	2
Sub15	I can discuss elements of data science and its tools and requirements in data preparation and data analysis.	0.62	2 (1-3)	3 (2-4)	1
Sub16	I can discuss and evaluate digital medicine with regard to specific aspects in the context of gender and sex.	0.50	2 (1-3)	4 (3-4)	2
Sub17	I can explain data protection and data security challenges in the development and use of digital tools and digital communication media.	0.44	2.5 (2-3)	4 (3-4)	1.5

[a]Spearman ρ (correlation between the item measuring the objective achievement of the first superordinate learning objective and the items measuring the objective achievement of each subordinate learning objective as measured in the postsurvey).

[b]Average median 2.5 (IQR 2-3).

[c]Average median 4 (IQR 3-4).

[d]Average change in the median from the pre- to the postsurvey: 1.5 (IQR 1-2).

We found a mostly medium to high correlation between the item measuring the objective achievement of the first superordinate learning objective and the 17 items measuring the objective achievement of the subordinate learning objectives (Table 3). This suggests that these items are well suited to capture aspects of the objective achievement of the first

XSL•FO
RenderX

superordinate learning objective. Therefore, these results indirectly suggest that achieving the first superordinate learning objective might have also improved from the pre- to the postsurvey.

Regarding the size of the changes in the median, rather big, big, or extensive changes were found for 14 of the 17 items (Tables 3 and 4). The smallest (small) change was observed in item sub11 (Tables 3 and 4). The greatest (extensive) changes were observed in items sub01, sub05, sub06, sub08, sub12, sub14, and sub16 (Tables 3 and 4).

Table 4. Evaluation of the size of the change in the median from the pre- to the postsurvey regarding the 17 items measuring the objective achievement of the subordinate learning objectives (N=17).

Qualitative evaluation	Items, n (%)	Item ID
No change	2 (12)	Sub03 and sub04
Minimal change	0 (0)	—[a]
Small change	1 (6)	Sub11
Rather big change	5 (29)	Sub07, sub09, sub10, sub13, and sub15
Big change	2 (12)	Sub02 and sub17
Extensive change	7 (41)	Sub01, sub05, sub06, sub08, sub12, sub14, and sub16

[a]Not applicable.

In addition, as seen in Table 5, there was an average improvement of the median across all 17 items for each participant from the pre- to the postsurvey (range of change in the median between 1 and 3 scale units; in addition, Multimedia Appendix 8 shows the raw data for each participant for each item in the pre- and postsurvey and how those values changed from the pre- to the postsurvey). To summarize, this means that all participants improved from the pre- to the postsurvey across all these learning objectives.

Table 5. Intraindividual differences in the median from the pre- to the postsurvey in the objective achievement of the 17 subordinate learning objectives (items sub01 to sub17; N=10).

Participant number	Presurvey score, median (IQR)	Postsurvey score, median (IQR)	ΔMedian[a]
1	3 (2-3)	4 (4-4)	1
2	3 (2-4)	5 (4-5)	2
3	1 (1-1)	3 (2-4)	2
4	2 (1-4)	4 (4-5)	2
5	3 (2-4)	4 (4-5)	1
6	3 (3-4)	5 (4-5)	2
7	1 (1-1)	4 (3-4)	3
8	3 (2-4)	4 (3-4)	1
9	3 (2-3)	4 (3-4)	1
10	2 (2-3)	4 (3-4)	2

[a]Change in the median from the pre- to the postsurvey.

Design of the Course

Suitability of the Course for Achieving the Superordinate Learning Objectives

Concerning the question of whether the course concept was suitable for achieving the 3 superordinate learning objectives, we found a median of 5 (IQR 4-5) scale units for the first, second, and third superordinate learning objectives (Table 2; for information on how frequently the individual response categories were selected by the participants, see Multimedia Appendix 6). The median values correspond to the scale point *strongly agree* (scale value of 5). Thus, the results indicate that, based on the average ratings, the course concept was strongly suited for achieving the 3 superordinate learning objectives.

Enjoyment of the Course

Concerning the question of whether the participants enjoyed the course, the median was 5 (IQR 4-5) scale units; for information on how frequently the individual response categories were selected by the participants, see Multimedia Appendix 6. This corresponds to the scale point *strongly agree* (scale value of 5). Therefore, this result indicates that, on average, participants enjoyed the course very much.

Benefits Obtained From the Course

Concerning the question of whether the participants felt that they obtained a benefit from having taken part in the course, the median was 4.5 (IQR 4-5) scale units; for information on how frequently the individual response categories were selected

XSL•FO
RenderX

by the participants, see Multimedia Appendix 6. This median varies between the scale points *rather agree* (scale value of 4) and *strongly agree* (scale value of 5). Thus, this result indicates that, on average, participants obtained a benefit rated as rather to very well from taking part in the course.

Strengths of the Course

With regard to the question of which of several proposed aspects were strengths of the course, the median for all proposed aspects fluctuated between the scale points *rather agree* (scale value of 4) and *strongly agree* (scale value of 5; Table 6); medians of 5 (IQR 5-5), 5 (IQR 4-5), 5 (IQR 3-5), and 4 (IQR 4-5) scale units for items STR01, STR02, STR03, and STR04, respectively; for information on how frequently the individual response categories were selected by the participants, see Multimedia Appendix 6. Regarding the teaching staff, all the students strongly agreed that this was a strength, as shown in Multimedia Appendix 6. Thus, in total, the results indicate that, on average, all the characteristics mentioned were considered to be strengths of the course either somewhat or strongly.

Table 6. Strengths of the Digital Medicine course, administered in the postsurvey (N=10).

Item stem and ID	Item	Scale units, median (IQR)
Strengths of the Digital Medicine course		
STR01[a]	The teaching staff (friendliness, openness, appreciation, professionalism, and interdisciplinarity)	5 (5-5)
STR02	The design of the course (content preparation, interaction, material, and equipment)	5 (4-5)
STR03	Timing of the classes (punctuality and time frame for lectures and seminars).	5 (3-5)
STR04	Offline content and preparation in the LernraumPlus platform	4 (4-5)

[a]STR: strengths (of the course).

Importance of the Course to the Participants

Importance of Achieving the Superordinate Learning Objectives

Regarding the question of whether it was important to the participants to achieve the 3 superordinate learning objectives, medians of 4 (IQR 3-4), 4 (IQR 3-5), and 3.5 (IQR 3-5) scale units were found for the first, second, and third learning objectives, respectively (Table 2; for information on how frequently the individual response categories were selected by the participants, see Multimedia Appendix 6). These median values vary between the scale points *rather neutral* (scale value of 3) and *rather agree* (scale value of 4). Therefore, the results indicate that the participants varied, on average, between being neutral about the importance of the course to them and rather agreeing that it was important to them to achieve the 3 superordinate learning objectives.

Reasons for Potential Absences During the Course

The reasons for not attending individual course sessions (Table 7) comprised insufficient time capacity due to other study-related requirements (10/10, 100% of the participants), illness (5/10, 50% of the participants), inadequate time capacity due to personal requirements (2/10, 20% of the participants), and perceived irrelevance for the individual participant (1/10, 10% of the participants). No participants named another reason.

Table 7. Reasons for nonparticipation, administered in the postsurvey (N=10).

Item ID, question, and item	Participants, n (%)
AB01[a]: If I was unable to attend course days, it was largely due to the following reasons (multiple choices possible):	
Too little time capacity due to other study-related requirements	10 (100)
Insufficient time capacity due to personal demands	2 (20)
Illness	5 (50)
The course was irrelevant to me	1 (10)
Other	0 (0)

[a]AB: absence (during course sessions).

Open Comments

There was 1 open comment in the presurvey and 5 open comments in the postsurvey. Table 8 shows the categories that appeared in the comments and the frequency of the occurrence of each subcategory. Most of the comments belonged to the main category (*positive feedback on the course or positive personal experience with the course*).

XSL•FO
RenderX

Table 8. Open comments (COM), administered in the pre- and postsurvey (N=5).

Item ID, item, survey, main category, and subcategory	Participants, n (%)
COM01: Here you can enter further comments, suggestions, or proposals	
Presurvey	
Request for similar workload to that of other courses of the profile	1 (20)
Postsurvey	
Positive feedback on the course or positive personal experience with the course	
The course was fun	1 (20)
The teaching was good	1 (20)
The teachers were friendly	1 (20)
The course contributed positively to personal and professional development or getting something out of the course	2 (40)
Good well-being	1 (20)
Gratitude for the course or for the opportunity to participate in the course	1 (20)
Negative feedback on the course or negative personal experience of the course	
The course was only offered in person but not in a hybrid format	1 (20)
Stress, frustration, and a guilty conscience toward the TUHH[a] due to the fact that one could not invest more time in the course because of limited capacity	1 (20)
Demotivation due to the absence of other members of their own small group	1 (20)
Other	
Wish that one could have participated more, but this was not possible	2 (40)
Frequent nonparticipation due to illness	1 (20)

[a]TUHH: for Hamburg University of Technology.

Discussion

Achievement of the Learning Objectives and Design of the Course

This paper presents a newly developed course concept for teaching digital medicine in medical education (the Bielefeld model) and an initial evaluation of this concept. This evaluation was based on feedback provided by undergraduate medical students who took part in this course. By developing and implementing this course, we responded to the calls [3,11,13-15] for integrating the topic of digital health or digital medicine into medical education. According to the students' self-report in the evaluation, they were overall able to achieve the super- and subordinate learning objectives of the course and felt that the course was largely suitable for achieving them.

The evaluation results indicate that, on average, the participants achieved the super- and subordinate learning objectives of the course rather to very well, as recorded after the course. The average change in the objective achievement of the subordinate learning objectives from before to after the course could be rated as big.

The fact that there was only a small change in item sub11 could be related to the circumstance that the participants were supposed to achieve the learning objective captured in this item by autonomously working on it using web-based materials, which they may not have done.

The findings on the achievement of the learning objectives indicate that the course concept was well suited for achieving them and could be interpreted as the course being well designed. This aligns with the finding that, on average, participants perceived the course concept as well suited for attaining the 3 superordinate learning objectives. On the basis of these outcomes, it can be assumed that the course effectively fulfills its overarching objective: empowering participants to feel able to actively engage as future architects of digital medicine, transcending the role of mere participants in digital processes or users of digital products to being able to develop and use such products comprehensively.

The assumption that the course was well designed is also supported by the fact that most to all participants enjoyed the course, benefited from it, and agreed that the course's design and teaching staff were strengths of the course. As outlined, the teaching staff consisted of interdisciplinary experts on the topics covered in the course in addition to the instructors from the Digital Medicine working group. Whether it was this interdisciplinarity, the appearance of the teaching staff (eg, friendliness, openness, and appreciation), or both that the students experienced as a strength is not clear from the results. These results are supported by positive statements in the open comments that relate precisely to the aforementioned points. The statements that the participants felt comfortable in the course and were appreciative of their participation support the overall positive evaluation of the course. However, there was also criticism that the course was not offered in a hybrid form.

XSL•FO
RenderX

In the future, this may be resolved by offering hybrid courses if necessary. Reasons for the statements that participation in the course resulted in stress, frustration, or demotivation were a lack of time and the unplanned absence of course members, over which the teachers had no influence. However, should this negative experience also occur in future courses, the lecturers and students should jointly consider measures to remedy the situation. A further quality assurance evaluation by the Medical School OWL supports the overall positive assessment of the course concept. A total of 50% (5/10) of the students took part in this evaluation, and they agreed that the instructors fostered a positive learning environment, acknowledged the participants' previous knowledge, and explained concepts clearly. These students offered comments that, in terms of content, closely aligned with those from the evaluation that is the subject of this paper, addressing both the positive and negative aspects. Unsystematic, spontaneous observations of the instructors during the course support the finding that the course was suitable for achieving the learning objectives and showed that it was also suitable for the participants to improve their self-, project, and team management skills. Overall, the results indicate that the course concept was well designed, implemented, and perceived and accepted positively.

Perceived Importance of the Course to the Participants

The results showed that many students perceived the achievement of the superordinate learning objectives as important to them and that only 10% (1/10) of the participants felt that the course was irrelevant. This perception aligns with the aforementioned finding that many medical students want the topic of digital medicine to be (more extensively) integrated into their studies [11,16], supporting the relevance of offering this course. However, as the perceived importance of the course was not further recorded or classified, the findings should be interpreted with caution.

Course Concept

Alignment of the Course With Findings From Teaching and Learning Research

From a pedagogical research perspective, the Bielefeld model integrates elements from multiple learning theories, including project-based learning, which has demonstrated efficacy [26], and authentic learning [27]. These approaches offer significant benefits as they may enhance the ability to recall knowledge in real-world problem-solving contexts [28] while also positively impacting student motivation and engagement [29]. In addition, in the Bielefeld model, theoretical and practical components were interwoven. Tempelman and Pilot [30] argue against the background of the theory of constructivism that this might help build knowledge and skills upon one another. In addition, structuring a course in which each unit builds on the previous one, as seen in the Bielefeld model, can promote a more cohesive learning experience. This can be beneficial for learning [31]. Medical education research has shown that the mix of knowledge and skill teaching in simulation-based medical education can be successful, with high effect sizes for learning outcomes of both aspects [32,33]. In light of these findings, the direct link between the theoretical and practical units in the Bielefeld model may enable participants to practically apply newly acquired knowledge directly and, thus, achieve the goal of holistically developing digital applications. An advantage of developing content in practical sessions, as implemented in the Bielefeld model, is that it makes learning outcomes more tangible, as Kuhn et al [34] point out. This approach may help in demonstrating learning success in a more verifiable and substantial way. The positive effect of including guest lecturers in the course has also been scientifically proven (eg, [35]). The idea of having medical students work on solutions for practical use cases at the interface between medicine and technology and in collaboration with students from other disciplines has already been implemented in a course described by Breil et al [36]. However, this was not geared toward digital medicine. In that course, medical students collaborated with computer science students. For this to work, the students had to develop soft skills such as communication skills and subject-specific knowledge. Almost all the participants in the aforementioned course rated the integration of theory and practice as good or excellent. Although it is unclear what exactly supported the good integration of theory and practice in that course, it can be assumed that this is generally related to the course concept. Therefore, the results suggest that the Bielefeld model, which followed a similar concept, facilitated a good integration of theory and practice as well. However, some aspects could be improved to align with findings from teaching and learning research. For example, although it was pointed out to which life cycle phases the topics of the individual sessions could be assigned in the Bielefeld model, this could have been explained more clearly. This might have helped make the course even more coherent (eg, [31]).

Alignment of the Course With Recommendations for Digital Medicine Courses

The design of the course is supported by recommendations and perspectives in the literature for implementing courses in digital medicine. The interdisciplinary design of the course aligns with the suggestion by Foadi and Varghese [37] that courses in digital medicine should be interdisciplinary due to the significant role that interdisciplinarity plays in the field of digital medicine. The Bielefeld model covered various topics, such as ethical aspects, legal frameworks, and entrepreneurial opportunities concerning digital medicine applications. Bahagon and Jacobson [2] state that successful implementation of eHealth solutions requires expertise in such areas. Students would like for courses on digital health to include precisely these aspects as well as practical training, for example, on developing apps, as Machleid et al [16] found out. Even though it cannot be said with certainty that these were also the wishes of the students enrolled in our course, it can be assumed based on this research finding. Therefore, in that case, these desires were well met by the course concept. Such a match between the students' wishes and the course concept could explain why many of the students perceived the course as important to them and enjoyed and benefitted from it. According to Goldsack and Zanetti [38], clinical and technical expertise must be considered together for a successful digital transformation. In the Bielefeld model, medical students do not receive in-depth technical training. Nonetheless, they learn how to collaborate with people from the technical field and gain some insights into the technical

XSL•FO
RenderX

implementation of digital medicine concepts. The University Digitalization Forum (Hochschulforum Digitalisierung [9]) points out that it is insufficient to teach only technical skills to improve digital transformation and that qualifications in areas such as communication and leadership or constitutional corporate design should also be provided. Although the Bielefeld model does not focus on this, it can be assumed that communication and leadership skills are also being learned during the teamwork units of the course.

As the course concept was geared precisely toward designing a telemedicine application and taking the framework conditions—various interdisciplinary factors—into consideration, it can be assumed that the learning objective of the NKLM stating that students should acquire the competencies to explain the application scenarios for telemedicine applications and their framework conditions [12] was achieved. In addition to this, other authors have described different frameworks regarding what students should learn about digital medicine. Brunner et al [39], for example, describe a framework that stipulates that students in the health care sector should achieve learning objectives related to the areas of (1) digital technologies, systems, and policies; (2) clinical practice and applications; (3) data analysis and knowledge creation; and (4) system and technology implementation. The Bielefeld model covers these areas at least to some extent, albeit more on a theoretical than a practical level. Kuhn [40] points out that, against the backdrop of digital transformation, physicians must be able to understand and categorize the change processes and new digital treatment concepts associated with change, and in addition to learning practical skills, they must also reflect on an attitude toward digital medicine. Due to the wide range of topics covered by the Bielefeld model, this is assumed to have been achieved at least to some extent. Recently, experts identified 40 specific digital health topics from the areas of knowledge, skills, and attitudes that they believe should be taught during medical school [41]. The Bielefeld model covers some of these knowledge and attitude topics.

In total, the Bielefeld model has many similarities with recommendations for the design of teaching in digital medicine. However, there are some aspects that could be given greater consideration in the Bielefeld model in the future to be even more in line with the recommendations. Including other experts or professional groups could be useful. In line with the suggestion that clinical and technical expertise is important for digital transformation [38], including cybersecurity experts or hardware and software engineers, for example, has been proposed [38]. It could also be useful to involve other groups, such as data scientists, ethicists, and patients [38]. To acquire and develop further knowledge and skills and a deeper attitude regarding digital medicine, it might make sense to include more content in the course and allow more time for a critical examination of digital medicine. However, as this is covered by other compulsory courses within the Bielefeld medical degree program, there is no need for the Bielefeld model to cover these aspects as well.

Furthermore, the question arises on whether the didactic formats used in the course correspond to what students want in digital medicine courses. Vossen et al [19] found that many students would like to be taught about technological developments through real-life scenarios and case descriptions. The Bielefeld model aims in this direction. Teaching in the form of lectures, which was also integrated into the Bielefeld model, was, on average, neither strongly supported nor strongly rejected in the study by Vossen et al [19]. Although, based on these results, this is not the preferred teaching format, it was necessary for the Bielefeld model to impart basic knowledge by offering lectures. The teaching format that received the most support in the study by Vossen et al [19] was remotely following a real-life patient under the supervision of a physician. However, such a teaching format is inappropriate for what needs to be taught and learned in the Bielefeld model. Therefore, the teaching format in the Bielefeld model cannot be meaningfully compared with it.

On the basis of the positive and improvable aspects mentioned in this section, the Bielefeld model will be continued, expanded, and further developed. Following the suggestion that the rapid pace of transformation processes must be taken into account and that it must be possible to adapt the specific curriculum, the Bielefeld model will be continuously refined and updated to adequately reflect current developments in digital medicine and general teaching and learning research.

Student Education in Digital Medicine in Germany

Comparison of the Course With Courses in Digital Medicine or Digital Health in Germany

In this section, the similarities and differences between the Bielefeld model and courses in digital medicine or digital health at other locations in Germany will be analyzed. The Bielefeld model was only compared with courses with a similar focus in terms of content and for which a publication could be found in a literature search. It should be noted that works such as those by Aulenkamp et al [17] or Behrends et al [42] and internet research indicate that there are also other courses on digital health and medicine. As no publications were available for these courses or they were unsuitable for the comparison due to different framework conditions, they were omitted. Although some course concepts described by Aulenkamp et al [17] appear to be related to the courses included in the comparison, it cannot be said with certainty that this is the case. Courses that focus on artificial intelligence or machine learning or place a clear emphasis on topics from the field of medical informatics (eg, hospital information systems) are regarded as connected to digital medicine but were not included in the comparison because they have a completely different focus in terms of content from that of the Bielefeld model. The Bielefeld model was compared to the courses described by the following authors: (1) Ehlers et al [43]; (2) Werner et al [44]; (3) Chaltikyan et al [45]; (4) Behrends et al [42]; (5) Offergeld et al [46]; (6) Nitsche et al [47,48], whereby both works share the same course concept but differ in the names and number of modules and, therefore, probably in the specific content; (7) Poncette et al [49] and Seemann et al [50]; and (8) Kuhn et al [34], Kuhn [40], Kuhn and Jungmann [51], and Kuhn et al [52-54], whereby they all share the same course concept but differ slightly in the names and number of modules and, therefore, possibly in the specific content.

XSL•FO
RenderX

If several publications were found for a course, only one is discussed here if the information used for the comparison was sufficiently clear.

The comparison of the Bielefeld model with other courses did not reveal any immediately comparable courses. However, there were certain overlaps in the design. As in the Bielefeld model, a common feature of many course concepts is the use of guest lecturers from different areas [34,43,44,47,49] (eg, app developers and representatives of the state data protection authority [34], an industry panel with representatives from the biomedical field [44], and lecturers from hospital departments and academic and nonacademic institutions who lectured on interdisciplinary topics [44]). Similarly to the Bielefeld model, topics covering economic, legal, or ethical aspects were incorporated into some courses, such as those described by Poncette et al [49] or Ehlers et al [43]. Another feature that the Bielefeld model has in common with some other courses is combining theoretical and practical units [34,44,49]. A difference between the Bielefeld model and many courses with which it was compared is that many courses followed a structure in which different application areas and examples of digital medicine were often considered in various sessions [34,43,44,47,49]. In contrast, the Bielefeld model is a course in which participants design a digital application independently and continuously throughout the course. The objective was to enable students to develop holistic digital applications independently and become active as future designers of digital medicine. Concepts with which the Bielefeld course can be most closely compared in terms of this focus are the modules described by Poncette et al [49] or Seemann et al [50] and an orientation module described by Werner et al [44]. Although these modules also follow a basic structure that includes numerous different areas of application and examples, one similarity to our course concept is that the focus is on students designing a product for a specific problem in the health care system [49] or developing a business model for digital transformation [44]. This allows them to take on the role of various stakeholders in the health care system [49] or company representatives [44]. A difference between the Bielefeld model and the course described by Poncette et al [49] is that, in the latter, the practical units were individual sessions instead of being integrated into each session as in the Bielefeld model. They seem to be rather separate from the theoretical units. There is no information on this with regard to the module presented in the work by Werner et al [44]. In light of the previously discussed possible advantages that could arise from interweaving theoretical and practical units and building units on each other [30,31], in our opinion, directly linking theoretical and practical units might be more advantageous. Examining numerous application areas and examples of digital medicine, as provided for in the aforementioned courses, is certainly enriching for students to gain a broad overview of digital medicine. However, at Bielefeld University, the students already deal with these aspects as part of an earlier course in digital medicine during their studies.

In summary, it can be said that the Bielefeld model is a newly developed part of a series of courses that contributes to fulfilling the call to train medical students in digital medicine. The Bielefeld model follows a detailed concept that has not yet been found in other evaluated courses in Germany.

Challenges and Initiatives Regarding the Implementation of Digital Medicine Courses in Germany

Various challenges that can make the integration of digital health and medicine into medical education more difficult [55] must be addressed to offer even more courses in digital medicine nationwide. For example, a sufficient teaching staff with the relevant expertise is needed. However, unlike Bielefeld University, few universities currently have a professorship or chair for this area [55]. There are various approaches to solving these challenges [56]. Initiatives such as the Digitalization of Departments-Medicine working group (Digitalisierung in den Fächern – Medizin), which is part of the University Digitalization Forum, are, for example, developing or presenting solutions [9]. Several initiatives also drive the implementation of educational projects. For example, a reform project in digital medicine [40] was funded by the Stifterverband initiative as part of its joint Curriculum 4.0 program with the Carl Zeiss Foundation. In the field of medical informatics associated with digital medicine, educational programs are also being offered as part of the HiGHmed initiative, which began as part of the Medical Informatics Initiative [57]. Such offers can also enable the teaching of digital medicine. The creation and implementation of initiatives is seen as an important measure [56].

Limitations

The interpretability of the results of evaluating the Bielefeld model is limited due to several aspects. The evaluation questionnaire was not subjected to a pilot test. Although it was designed based on largely careful considerations and reflected upon between the authors, a pilot study would have been favorable to determine its quality and suitability for evaluating the course concept. Although an attempt was made to ensure the validity of the items by carefully formulating them based on detailed considerations, the criterion and construct validity in particular were not statistically tested. Therefore, no reliable statement can be made regarding the existence of criterion and construct validity and, thus, external validity overall. The existence of retest reliability could have been examined to better determine the stability of the instruments, a facet of reliability. As this did not happen, it cannot be said with certainty whether the instruments are reliable in this regard. One limitation with regard to the design of the survey is that the achievement of the superordinate learning objectives was only recorded in the postsurvey. Therefore, no direct statement can be made as to whether they were achieved better after the course than before. Because the data were not analyzed using inferential statistics, the results cannot be reliably transferred to the general population. The interpretability and generalizability of the results are also limited by the small sample size and by potential biases that may have occurred. With regard to the sample, it should be taken into account that the Medical School OWL was only just commissioned in 2021 and that the Digital Medicine course presented in this paper was the first of its kind ever offered at this medical school. Therefore, it was not possible to form a larger sample by summarizing and analyzing data from several

XSL•FO
RenderX

courses. In addition, the teachers had no influence on the number of course participants. Therefore, an evaluation based on the existing sample of 10 participants was the only option.

It was found that only two-thirds of the students who took part in the presurvey (10/15, 67%) also took part in the postsurvey. It is possible that the third who dropped out had little interest in digital medicine and, therefore, were not motivated enough to participate in the postsurvey. The sample would then only consist of students who were interested in the topic. In this case, the results would only be transferable to students with an interest in digital medicine instead of all medical students. It is also possible that the students with a particular interest in the topic also achieved the learning objectives better and rated the course better or that, in general, only students who felt that they had achieved the learning objectives participated in the postsurvey. Both would result in a systematic bias of the results. This is aggravated by the fact that the data were collected via the self-report of the students. If an objective measure of data collection had been used that did not involve any extra effort for the students, it might have been easier to motivate all students to provide data for the postevaluation. Such an objective measure could have been, for example, an analysis of the project outlines.

In addition to the aforementioned aspects, it was observed during numerous sessions of the course that some students were absent, and some sessions were only held on the web and the instructors do not know whether the participants worked on the topics of those sessions by themselves. It is questionable whether the students who were not physically present also improved in the learning objectives of these sessions. If this were the case, the question of what, if not the course, could have led to this improvement would arise. The data showed that not every learning objective was achieved better by all students after the course than before. This is particularly understandable if the students who were not present in the session covering this learning objective did not deal with the topic. However, these questions were not investigated.

The results of this evaluation should only be seen as initial indications of the effectiveness of the course concept in achieving the learning objectives due to the named limitations.

In addition, as part of the study, participants were asked to what extent they perceived various potential strengths of the course as such. This format made it possible to ask about specific aspects but has the disadvantage that other possible strengths of the course could remain hidden. Although we included an optional free-text field at the end of the surveys meant to allow participants to name strengths and additional aspects related to the course, it cannot be assumed that all participants used this option. A final statement on whether a course similar to the Bielefeld model already exists at other locations in Germany cannot be made as the course was only compared with course concepts for which a publication was found.

Implications

Our results indicate that using the Bielefeld model might improve self-sufficiency in digital medicine instead of letting future physicians passively consume digital medicine's offers. However, due to the small sample size and other limitations, this warrants additional evaluations. The observation made during the course that the medical and electrical engineering students learned what is essential when working with students from another discipline suggests that the course can be offered across disciplines and universities.

Outlook

As inferential statistics were not performed, no direct conclusions can be drawn from this study on whether the same results can also be expected in the general population. In future studies, the concept should be implemented with and analyzed based on more participants and then subjected to an evaluation including inferential statistical procedures. Bielefeld University plans to continue conducting and evaluating the concept over the next few years to obtain a longitudinal sample with a larger number of cases and carry out inferential statistical analyses. To better assess whether the course contributes to the achievement of the superordinate learning objectives, their achievement should be recorded in future studies in the pre- and postsurvey. A more objective assessment of the achievement of the learning objectives could be made by analyzing the students' project outlines. The project outlines make it clear whether the students know the factors that are relevant in the development of a digital application and whether they were able to apply them to their specific project.

Conclusions

A novel approach to teaching digital medicine was conceived, executed, and assessed for the first time. The evaluation outcomes suggest that the course framework has the potential to effectively facilitate the transformation of participating students into future architects of digital medicine. These results signify that the course is poised to play a pivotal role in fostering digital transformation and seamlessly incorporating digital medicine into the medical curriculum, aligning with the aspirations of diverse stakeholders.

Acknowledgments

The authors thank Ute von Jan for the fruitful discussion about the educational aspects of medical informatics related to undergraduate medical education. The authors acknowledge support for the publication costs from the Open Access Publication Fund of Bielefeld University and the German Research Foundation (DFG).

Authors' Contributions

UVA and AM created the concept, developed the methodology, conducted the formal analysis and investigation, contributed to data curation, prepared and wrote the original draft, contributed to review and editing of the draft, and created the visualizations.

XSL•FO
RenderX

UVA and AM validated the data. UVA provided supervision and undertook project administration and funding acquisition. HF programmed the surveys using software and contributed to reviewing and editing the draft. UK and LB contributed to data curation and review and editing of the draft. AB contributed to review and editing of the draft. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of the three instructional cases for medical students, created by the course instructor.
[DOCX File , 15 KB - mededu_v10i1e56787_app1.docx]

Multimedia Appendix 2
Overview of web-based and in-person course sessions, including guest lecturers and main content, with additional materials and exercises provided.
[DOCX File , 20 KB - mededu_v10i1e56787_app2.docx]

Multimedia Appendix 3
Template of questions and content that students can consider when writing the project outline.
[DOCX File , 17 KB - mededu_v10i1e56787_app3.docx]

Multimedia Appendix 4
Schedule of a standard course session. The exact sequence of the individual sessions could be flexibly adapted.
[DOCX File , 16 KB - mededu_v10i1e56787_app4.docx]

Multimedia Appendix 5
Evaluation items overview. Items sub01 to sub17 and COM01 were included in both pre- and postsurveys, while others were postsurvey only.
[DOCX File , 18 KB - mededu_v10i1e56787_app5.docx]

Multimedia Appendix 6
Postsurvey response frequencies for evaluating learning objectives, course suitability, importance, enjoyment, benefit, and strengths.
[DOCX File , 18 KB - mededu_v10i1e56787_app6.docx]

Multimedia Appendix 7
Response frequencies for items evaluating objective achievement of subordinate learning objectives in the pre- and postsurvey.
[DOCX File , 23 KB - mededu_v10i1e56787_app7.docx]

Multimedia Appendix 8
Raw data for each participant on objective achievement of subordinate learning objectives in the pre- and postsurvey, including intraindividual changes.
[DOCX File , 35 KB - mededu_v10i1e56787_app8.docx]

References

1. Dang A, Arora D, Rane P. Role of digital therapeutics and the changing future of healthcare. J Family Med Prim Care 2020 May;9(5):2207-2213 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_105_20] [Medline: 32754475]
2. Bahagon Y, Jacobson O. e-Health, m-Health and healthier social media reform: the big scale view. Int J Integr Care 2012 Jun 08;12(4):13. [doi: 10.5334/ijic.927]
3. Global strategy on digital health 2020-2025. World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2024-04-29]
4. Hernandez MF, Rodriguez F. Health techequity: opportunities for digital health innovations to improve equity and diversity in cardiovascular care. Curr Cardiovasc Risk Rep 2023 Nov 28;17(1):1-20 [FREE Full text] [doi: 10.1007/s12170-022-00711-0] [Medline: 36465151]
5. Santo K, Redfern J. Digital health innovations to improve cardiovascular disease care. Curr Atheroscler Rep 2020 Oct 03;22(12):71 [FREE Full text] [doi: 10.1007/s11883-020-00889-x] [Medline: 33009975]

XSL•FO
RenderX

6. Kaufman N, Ferrin C, Sugrue D. Using digital health technology to prevent and treat diabetes. Diabetes Technol Ther 2019 Feb;21(S1):S79-S94. [doi: 10.1089/dia.2019.2506] [Medline: 30785320]
7. Sasseville M, LeBlanc A, Tchuente J, Boucher M, Dugas M, Gisèle M, et al. The impact of technology systems and level of support in digital mental health interventions: a secondary meta-analysis. Syst Rev 2023 May 04;12(1):78 [FREE Full text] [doi: 10.1186/s13643-023-02241-1] [Medline: 37143171]
8. Iyamu I, Xu AX, Gómez-Ramírez O, Ablona A, Chang H, Mckee G, et al. Defining digital public health and the role of digitization, digitalization, and digital transformation: scoping review. JMIR Public Health Surveill 2021 Nov 26;7(11):e30399 [FREE Full text] [doi: 10.2196/30399] [Medline: 34842555]
9. Arbeitsgruppe DIF-Medizin. Digitale Transformation in der medizinischen Ausbildung: Eine Handreichung der Arbeitsgruppe Digitalisierung in den Fachbereichen: Medizin. Hochschulforum Digitalisierung. 2023 Nov 01. URL: https://hochschulforumdigitalisierung.de/wp-content/uploads/2023/11/HFD_AP_74_Medizin.pdf [accessed 2024-04-29]
10. des Bundesministeriums für Gesundheit: Verordnung zur Neuregelung der ärztlichen Ausbildung. Bundesgesundheitsministerium. 2023. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/A/AEApprO_RefE_ueberarbeitet.pdf [accessed 2024-01-29]
11. Jacob R, Kopp J, Fellinger P. Berufsmonitoring Medizinstudierende. Kassenärztliche Bundesvereinigung. 2018. URL: https://www.kbv.de/media/sp/Berufsmonitoring_Medizinstudierende_2018.pdf [accessed 2024-04-29]
12. Nationaler Kompetenzbasierter Lernzielkatalog Medizin. Version 2.0. Charité – Universitätsmedizin Berlin. URL: https://nklm.de/zend/menu [accessed 2024-01-29]
13. Neustrukturierung des Medizinstudiums und Änderung der Approbationsordnung für Ärzte. Wissenschaftsrat. URL: https://www.wissenschaftsrat.de/download/archiv/7271-18.pdf?__blob=publicationFile&v=4 [accessed 2024-01-29]
14. Mosch L, Machleid F, von Maltzahn F, Kaczmarczyk R, Nokhbatolfoghahai F, Balciunas J, et al. Digital health in the medical curriculum: addressing the needs of the future health workforce. European Medical Students' Association. URL: https://emsa-europe.eu/wp-content/uploads/2021/06/Policy-2019-04-Digital-Health-in-the-Medical-Curriculum_-Addressing-the-Needs-of-the-Future-Health-Workforce.pdf [accessed 2024-04-29]
15. Schreiber S, Jacob R, Kopp J. Berufsmonitoring Europäische Medzinstudierende. Kassenärztliche Bundesvereinigung. URL: https://www.kbv.de/media/sp/KBV_Berufsmonitoring-Bericht2022.pdf [accessed 2024-04-29]
16. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
17. Aulenkamp J, Mikuteit M, Löffler T, Schmidt J. Overview of digital health teaching courses in medical education in Germany in 2020. GMS J Med Educ 2021;38(4):Doc80 [FREE Full text] [doi: 10.3205/zma001476] [Medline: 34056069]
18. Tudor Car L, Kyaw BM, Nannan Panday RS, van der Kleij R, Chavannes N, Majeed A, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275 [FREE Full text] [doi: 10.2196/28275] [Medline: 34287206]
19. Vossen K, Rethans J, van Kuijk SM, van der Vleuten CP, Kubben PL. Understanding medical students' attitudes toward learning eHealth: questionnaire study. JMIR Med Educ 2020 Oct 01;6(2):e17030 [FREE Full text] [doi: 10.2196/17030] [Medline: 33001034]
20. Field MJ. Telemedicine: a guide to assessing telecommunications in healthcare. J Digit Imaging 1997 Aug;10(3 Suppl 1):28 [FREE Full text] [doi: 10.1007/BF03168648] [Medline: 9268830]
21. Arslan RC, Walther MP, Tata CS. formr: a study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. Behav Res Methods 2020 Feb 1;52(1):376-387 [FREE Full text] [doi: 10.3758/s13428-019-01236-y] [Medline: 30937847]
22. Foadi N, Koop C, Mikuteit M, Paulmann V, Steffens S, Behrends M. Defining learning outcomes as a prerequisite of implementing a longitudinal and transdisciplinary curriculum with regard to digital competencies at Hannover Medical School. J Med Educ Curric Dev 2021 Jul 21;8:23821205211028347 [FREE Full text] [doi: 10.1177/23821205211028347] [Medline: 34368455]
23. Friederichs H, Marschall B, Weissenstein A. Practicing evidence based medicine at the bedside: a randomized controlled pilot study in undergraduate medical students assessing the practicality of tablets, smartphones, and computers in clinical life. BMC Med Inform Decis Mak 2014 Dec 05;14:113 [FREE Full text] [doi: 10.1186/s12911-014-0113-7] [Medline: 25477073]
24. Holling M, Stummer W, Friederichs H. Teaching the concept of brain death in undergraduate medical education. J Surg Educ 2015;72(3):504-508 [FREE Full text] [doi: 10.1016/j.jsurg.2014.10.012] [Medline: 25467732]
25. Kuckarzt U, Rädiker S. Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. 5th edition. Berlin, Germany: Beltz Juventa; 2022.
26. Krajcik JS, Shin N. Project-based learning. In: Sawyer RK, editor. The Cambridge Handbook of the Learning Sciences. Cambridge, MA: Cambridge University Press; 2022:727-792.

XSL•FO
RenderX

27. Herrington J. Introduction to authentic learning. In: Bozalek V, Ng'ambi D, Wood D, Herrington J, Hardman J, Amory A, editors. Activity Theory, Authentic Learning and Emerging Technologies: Towards a Transformative Higher Education Pedagogy. Boca Raton, FL: Taylor & Francis Group; 2014.
28. Herrington J, Reeves TC, Oliver R. A Guide to Authentic e-Learning. New York, NY: Routledge; 2010.
29. Nachtigall V, Shaffer DW, Rummel N. Stirring a secret sauce: a literature review on the conditions and effects of authentic learning. Educ Psychol Rev 2022 Apr 25;34(3):1479-1516. [doi: 10.1007/s10648-022-09676-3]
30. Tempelman E, Pilot A. Strengthening the link between theory and practice in teaching design engineering: an empirical study on a new approach. Int J Technol Des Educ 2010 Mar 31;21(3):261-275. [doi: 10.1007/s10798-010-9118-4]
31. Hammerness K. From coherence in theory to coherence in practice. Teach Coll Rec 2022 Feb 01;108(7):1241-1265. [doi: 10.1177/016146810610800704]
32. Cook DA. How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. Med Educ 2014 Aug 09;48(8):750-760. [doi: 10.1111/medu.12473] [Medline: 25039731]
33. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 07;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
34. Kuhn S, Müller N, Kirchgässner E, Ulzheimer L, Deutsch KL. Digital skills for medical students - qualitative evaluation of the curriculum 4.0 "medicine in the digital age". GMS J Med Educ 2020;37(6):Doc60 [FREE Full text] [doi: 10.3205/zma001353] [Medline: 33225052]
35. Zou P, Sun W, Hallowell SG, Luo Y, Lee C, Ge L. Use of guest speakers in nursing education: an integrative review of multidisciplinary literature. Adv Med Educ Pract 2019;10:175-189 [FREE Full text] [doi: 10.2147/AMEP.S196456] [Medline: 31118860]
36. Breil B, Fritz F, Thiemann V, Dugas M. Multidisciplinary education in medical informatics--a course for medical and informatics students. Stud Health Technol Inform 2010;160(Pt 1):581-584. [Medline: 20841753]
37. Foadi N, Varghese J. Digital competence - a key competence for todays and future physicians. J Eur CME 2022 Jan 02;11(1):2015200 [FREE Full text] [doi: 10.1080/21614083.2021.2015200] [Medline: 34992949]
38. Goldsack JC, Zanetti CA. Defining and developing the workforce needed for success in the digital era of medicine. Digit Biomark 2020 Nov 26;4(Suppl 1):136-142 [FREE Full text] [doi: 10.1159/000512382] [Medline: 33442586]
39. Brunner M, McGregor D, Keep M, Janssen A, Spallek H, Quinn D, et al. An eHealth capabilities framework for graduates and health professionals: mixed-methods study. J Med Internet Res 2018 May 15;20(5):e10229 [FREE Full text] [doi: 10.2196/10229] [Medline: 29764794]
40. Kuhn S. Medizin im digitalen Zeitalter: transformation durch Bildung. Deutsches Ärzteblatt. 2018. URL: https://www.aerzteblatt.de/archiv/197293/Medizin-im-digitalen-Zeitalter-Transformation-durch-Bildung [accessed 2024-04-29]
41. Khurana MP, Raaschou-Pedersen DE, Kurtzhals J, Bardram JE, Ostrowski SR, Bundgaard JS. Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Med Educ 2022 Feb 26;22(1):129 [FREE Full text] [doi: 10.1186/s12909-022-03163-7] [Medline: 35216611]
42. Behrends M, Paulmann V, Koop C, Foadi N, Mikuteit M, Steffens S. Interdisciplinary teaching of digital competencies for undergraduate medical students - experiences of a teaching project by medical informatics and medicine. Stud Health Technol Inform 2021 May 27;281:891-895. [doi: 10.3233/SHTI210307] [Medline: 34042802]
43. Ehlers JP, Herrmann M, Mondritzki T, Truebel H, Boehme P. Digitale Transformation der Medizin – Erfahrungen mit einem Kurs, um die Handlungsfähigkeit Studierender in der digitalen Welt zu fördern. GMS Med Inform Biom Epidemiol 2019;15(1):Doc06 [FREE Full text] [doi: 10.3205/mibe000200]
44. Werner R, Henningsen M, Schmitz R, Guse AH, Augustin M, Gauer T. Digital Health meets Hamburg integrated medical degree program iMED: concept and introduction of the new interdisciplinary 2 track Digital Health. GMS J Med Educ 2020;37(6):Doc61 [FREE Full text] [doi: 10.3205/zma001354] [Medline: 33225053]
45. Chaltikyan G, Fernandes FA, Pfeiffer J. Digital health education: determining competences and piloting innovative study course. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe JD, Pelayo S, et al, editors. Challenges of Trustable AI and Added-Value on Health. New York, NY: IOS Press; 2019:825-826.
46. Offergeld C, Neudert M, Emerich M, Schmidt T, Kuhn S, Giesler M. Vermittlung digitaler Kompetenzen in der curricularen HNO-Lehre: abwartende Haltung oder vorauseilender Gehorsam? HNO 2020 Apr;68(4):257-262. [doi: 10.1007/s00106-019-00745-8] [Medline: 31538215]
47. Nitsche J, Busse TS, Kernebeck S, Ehlers JP. Virtual classrooms and their challenge of interaction-an evaluation of chat activities and logs in an online course about digital medicine with heterogeneous participants. Int J Environ Res Public Health 2022 Aug 17;19(16):10184 [FREE Full text] [doi: 10.3390/ijerph191610184] [Medline: 36011818]
48. Nitsche J, Busse TS, Ehlers JP. Teaching digital medicine in a virtual classroom: impacts on student mindset and competencies. Int J Environ Res Public Health 2023 Jan 22;20(3):2029 [FREE Full text] [doi: 10.3390/ijerph20032029] [Medline: 36767393]
49. Poncette AS, Glauert DL, Mosch L, Braune K, Balzer F, Back DA. Undergraduate medical competencies in digital health and curricular module development: mixed methods study. J Med Internet Res 2020 Oct 29;22(10):e22161 [FREE Full text] [doi: 10.2196/22161] [Medline: 33118935]

XSL•FO
RenderX

50. Seemann RJ, Mielke AM, Glauert DL, Gehlen T, Poncette AS, Mosch LK, et al. Implementation of a digital health module for undergraduate medical students: a comparative study on knowledge and attitudes. Technol Health Care 2023;31(1):157-164 [FREE Full text] [doi: 10.3233/THC-220138] [Medline: 35754241]
51. Kuhn S, Jungmann F. Telemedizin in der studentischen Lehre. Radiologe 2018 Mar 9;58(3):236-240. [doi: 10.1007/s00117-017-0351-7] [Medline: 29318348]
52. Kuhn S, Kadioglu D, Deutsch K, Michl S. Data Literacy in der Medizin. Onkologe 2018 Feb 13;24(5):368-377. [doi: 10.1007/S00761-018-0344-9]
53. Kuhn S, Kirchgässner E, Deutsch K. Medizin im digitalen Zeitalter – „do it by the book… but be the author!“. Synergie. 2017. URL: https://www.synergie.uni-hamburg.de/de/media/ausgabe04/synergie04.pdf [accessed 2024-04-29]
54. Kuhn S, Jungmann F, Deutsch K, Drees P, Rommens PM. Digitale Transformation der Medizin: Brauchen wir ein Curriculum 4.0 für die Aus-, Fort- und Weiterbildung? Oxf Univ J 2018;7:458. [doi: 10.3238/oup.2018.0453?0458]
55. Aulenkamp J, Mosch L, Schmidt J, Kaufmann M, Wirbelauer L. Digitale bildung. In: Matusiewicz D, Henningsen M, Ehlers P, editors. Digitale Medizin. Kompendium für Studium und Praxis. Berlin, Germany: MWV Verlag; 2020:75-94.
56. Haag M, Igel C, Fischer MR, German Medical Education Society (GMA), Committee “Digitization – Technology-Assisted Learning and Teaching”, Joint working group “Technology-enhanced Teaching and Learning in Medicine (TeLL)” of the German Association for Medical Informatics, Biometry and Epidemiology (gmds) and the German Informatics Society (GI). Digital teaching and digital medicine: a national initiative is needed. GMS J Med Educ 2018;35(3):Doc43 [FREE Full text] [doi: 10.3205/zma001189] [Medline: 30186953]
57. HiGHmed Medical Informatics. URL: https://www.highmed.org/de/home [accessed 2024-01-30]

Abbreviations

NKLM: National Competency-Based Catalog of Learning Objectives
OWL: Ostwestfalen-Lippe
TUHH: Hamburg University of Technology

Edited by B Lesselroth; submitted 05.02.24; peer-reviewed by R Patel, M Roos, D Lungeanu; comments to author 26.04.24; revised version received 28.05.24; accepted 14.08.24; published 30.09.24.

Please cite as:
Mielitz A, Kulau U, Bublitz L, Bittner A, Friederichs H, Albrecht UV
Teaching Digital Medicine to Undergraduate Medical Students With an Interprofessional and Interdisciplinary Approach: Development and Usability Study
JMIR Med Educ 2024;10:e56787
URL: https://mededu.jmir.org/2024/1/e56787
doi:10.2196/56787
PMID:39189929

©Annabelle Mielitz, Ulf Kulau, Lucas Bublitz, Anja Bittner, Hendrik Friederichs, Urs-Vito Albrecht. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.09.2024. 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.

XSL•FO
RenderX

Original Paper

Bridging the Telehealth Digital Divide With Collegiate Navigators: Mixed Methods Evaluation Study of a Service-Learning Health Disparities Course

Zakaria Nadeem Doueiri[1], BS; Rika Bajra[2], MD; Malathi Srinivasan[2], MD; Erika Schillinger[2], MD; Nancy Cuan[2], MS, MD

[1]Department of Epidemiology and Population Health, Stanford University School of Medicine, Palo Alto, CA, United States
[2]Division of Primary Care and Population Health, Stanford University School of Medicine, Palo Alto, CA, United States

Corresponding Author:
Nancy Cuan, MS, MD
Division of Primary Care and Population Health
Stanford University School of Medicine
211 Quarry Road, Suite 402
Palo Alto, CA, 94304
United States
Phone: 1 650 724 1800
Email: cuannan@stanford.edu

Abstract

Background: Limited digital literacy is a barrier for vulnerable patients accessing health care.

Objective: The Stanford Technology Access Resource Team (START), a service-learning course created to bridge the telehealth digital divide, trained undergraduate and graduate students to provide hands-on patient support to improve access to electronic medical records (EMRs) and video visits while learning about social determinants of health.

Methods: START students reached out to 1185 patients (n=711, 60% from primary care clinics of a large academic medical center and n=474, 40% from a federally qualified health center). Registries consisted of patients without an EMR account (at primary care clinics) or patients with a scheduled telehealth visit (at a federally qualified health center). Patient outcomes were evaluated by successful EMR enrollments and video visit setups. Student outcomes were assessed by reflections coded for thematic content.

Results: Over 6 academic quarters, 57 students reached out to 1185 registry patients. Of the 229 patients contacted, 141 desired technical support. START students successfully established EMR accounts and set up video visits for 78.7% (111/141) of patients. After program completion, we reached out to 13.5% (19/141) of patients to collect perspectives on program utility. The majority (18/19, 94.7%) reported that START students were helpful, and 73.7% (14/19) reported that they had successfully connected with their health care provider in a digital visit. Inability to establish access included a lack of Wi-Fi or device access, the absence of an interpreter, and a disability that precluded the use of video visits. Qualitative analysis of student reflections showed an impact on future career goals and improved awareness of health disparities of technology access.

Conclusions: Of the patients who desired telehealth access, START improved access for 78.7% (111/141) of patients. Students found that START broadened their understanding of health disparities and social determinants of health and influenced their future career goals.

(JMIR Med Educ 2024;10:e57077) doi:10.2196/57077

KEYWORDS

service learning; medical education; access to care; telehealth; telemedicine; health disparities; social determinants of health; digital literacy; vulnerable populations; community engagement; value-added medical education; digital health; digital divide; health equity; collegiate navigator; experimental; education; student; qualitative analysis; technology; mobile phone

XSL•FO
RenderX

Introduction

Telehealth is an emergent tool to connect patients with their health care providers [1]. During the COVID-19 pandemic, the number of video visits increased by nearly 50% in the first quarter of 2020 compared to that of 2019 [2]. During this time frame, there was a 63-fold increase in telehealth visits among Medicare patients [3]. Telehealth use in the United States continued to rise, with nearly 38 times as many patients using it in 2021 compared to prior to the onset of the COVID-19 pandemic in 2019 [2,4]. Ensuring that patients have access to their electronic medical records (EMRs) is important for communication and coordination with their health care team and to promote patient self-advocacy [5]. While access to video visits and EMR accounts positively impacts access to health care, fewer than half of adults aged 65 years and older own a smartphone or have reliable internet access in the United States [6]. Additionally, racial minority individuals and patients with low socioeconomic status are less likely to have access to essential devices and high-speed internet and face increased obstacles to digital health literacy [7-10]. Consequently, while telehealth can be a conduit to more accessible health care, many at-risk patients experience heightened challenges in connecting with their providers remotely, which may magnify health disparities [6,7].

In response to shelter-in-place orders during the COVID-19 pandemic, medical practices strove to expand telemedicine options and increase resources to aid patients with telehealth [11]. In an effort to streamline this process, several clinics created video tutorials and step-by-step telehealth training guides for patients [12]. However, these static resources quickly proved to be insufficient, and the need for additional assistance from support personnel became evident [11,12]. One promising approach to making telehealth more accessible is the integration of "value-added medical education," whereby students' service learning improves health care [13]. Recognizing patients' need for individualized technology support and students' desire to serve the community, we created the Stanford Technology Access Resource Team (START) to help bridge the digital divide.

START is an undergraduate and graduate student course and volunteer program that offers immersive, meaningful patient-engaged experiences to promote patient health equity. START is a quarter-long course that teaches students about social determinants of health and health technologies and then connects them with patients with low digital technology literacy. Since digital literacy has been identified as a determinant of health, we hypothesized that this service-learning course would improve patient access to telemedicine to decrease their digital divide while also improving professional development, learner communication, and understanding of health equity and patient challenges. We conducted a comprehensive program evaluation to assess patient outcomes (connection to digital health and telemedicine and satisfaction) and learner outcomes (relationship-building skills, awareness of patient experiences, and future career life goals or purposes).

Methods

Ethical Considerations

The Stanford University Institutional Review Board determined this was nonhuman participant research (protocol #73006). All data have been anonymized and deidentified.

Study Design

In the fall of 2020, amid the COVID-19 pandemic, we developed START for undergraduate and graduate students. The course addressed social determinants of health for vulnerable populations and included a service-learning component focused on digital literacy for patients in 2 health systems. The study evaluated the impact on student learning through thematic analysis of student reflections and the impact on patients through data on outreach, enrollment in EMR accounts, access to telehealth, and patient reflections.

Study Context

Stanford University School of Medicine and Stanford Healthcare form an integrated academic medical center in Northern California, situated within an undergraduate and graduate campus. Within the surrounding counties (37 cities with large unincorporated rural areas), patients' social determinants of health vary widely. Surrounding county median income is about US $100,000, with racial or ethnic distribution of approximately 3% African American, 30% Asian, 25% Latinx, 50% White, and 14% other, with overlap [14]. We collaborated with a federally qualified health center (FQHC) to expand our service-learning project to additional underserved community patients. The FQHC serves a population of whom 82% are living at or below 200% of the federal poverty level; 27% are uninsured; and 81% of patients identify as Latinx, African American, or Pacific Islander [15]. Stanford University encompasses 7841 undergraduate and 9688 graduate students [16]. The second most popular major for students is human biology, and approximately 10% of the campus' student body apply to medical school each year [17,18].

Course Development

We conducted needs assessments of students (Multimedia Appendix 1), patients (Multimedia Appendix 2), and health care teams (Multimedia Appendix 3) to inform the design of a course that would enable students to support patients who needed help accessing digital health. We distributed the student needs assessment through premed advisors, dormitory email lists, and flyers posted throughout the campus. All except 1 of the 107 student respondents agreed that the proposed course would meet a community need, and 93.5% (n=100) wished to have more information about the course. We distributed patient needs assessments to residents at 2 assisted living facilities in the community, and 125 patients completed the assessment. In total, 28% (n=35) of patients stated they would like to do telehealth visits, while 25.6% (n=32) said they had barriers to using telehealth. The top 3 reasons cited were not knowing how to connect to the telehealth platform, lack of familiarity with the internet or technology, and not having a stable internet connection. We distributed the needs assessments for health care teams to patient care coordinators at 3 of the institutional

XSL•FO
RenderX

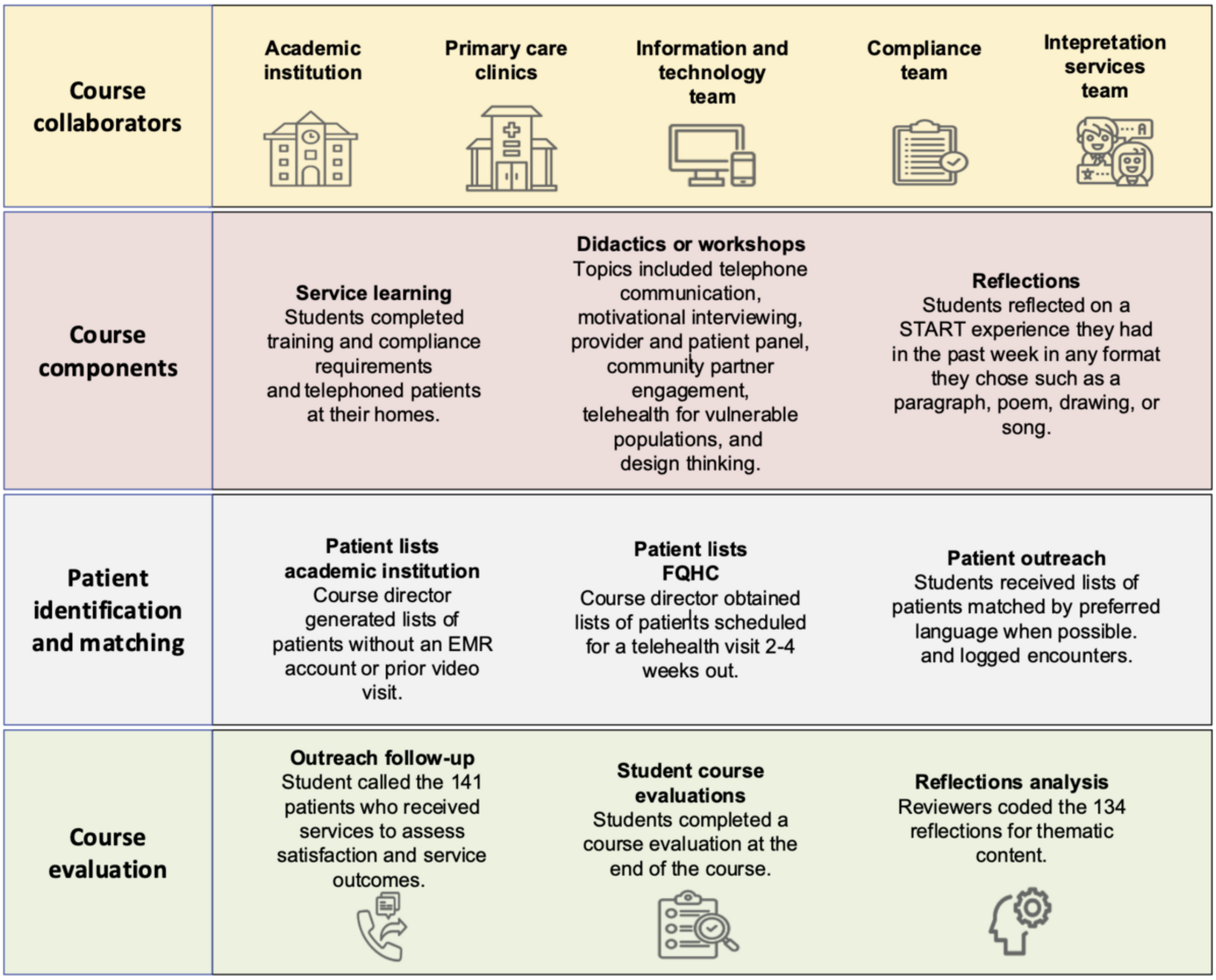
primary care clinics. In total, 6 (55%) of the 11 care coordinators recalled that more than 10 patients requested help with telehealth connection. The care coordinators cited the following barriers to providing the help: the lack of a resource line to help troubleshoot (n=8, 73%), time constraints (n=6, 55%), and the lack of training (n=5, 46%). Based on these findings, we developed a 1-unit, repeatable, 10-week-long course consisting of weekly 1-hour didactics combined with service learning to directly support patients in need of digital health support.

This course structure was chosen since the instructor wanted to balance the amount of time spent on didactics, workshops, and reflection with time reaching out to patients. The 10-week format followed the quarter system of the institution. Students are allowed to repeat the course or continue to provide service as volunteer without repeating the course. Due to the tight schedule of 10 weeks, the beginning 2 weeks of the course needed to focus on completing HIPAA (Health Insurance Portability and Accountability Act) training and training on the technical aspects of helping patients set up an EMR account and video visits. Since the course involved only telephone communication with patients, training on telephone communication skills was deemed essential. Education on the social determinants of health was key to understanding how to decrease the digital divide. Design thinking gave students a practical framework for thinking about how to solve a problem. The availability of speakers to do repeated digital sessions with the students during the COVID-19 pandemic also influenced the topics chosen.

Didactics or Workshops

Weekly didactics and workshops (Figure 1) focused on teaching students inclusivity when interacting with diverse patients and caregivers. Guest speakers provided insight into topics such as the telehealth experience from patient and provider perspectives, social drivers of health, motivational interviewing, and design thinking. These collaborative sessions challenged students to think critically about creating more effective solutions to address community determinants of health that impact health care access.

Figure 1. Structure of START course. EMR: electronic medical record; FQHC: federally qualified health center; START: Stanford Technology Access Resource Team.



XSL•FO
RenderX

Figure 2. Examples of current telehealth or service-learning educational programs. DBMI: Department of Biomedical Informatics; DNP: Doctor of Nursing Practice; MHA: Master of Health Administration; MS-1: medical school year 1; MS-3: medical school year 3; USC: University of Southern California.

Institution or organization	Course structure	Curriculum content
Florida Atlantic University	**Christine E. Lynn College of Nursing (DNP):** - 4-week course (web-based module) open for general public ☑ Didactic ☐ Experiential ☐ Community-based ☐ Longitudinal	**Telehealth certificate** "Learning about health equity; improving access to health care, health disparities, and inter-professional practice with specific populations in relationship to the application and implementation of telehealth platforms."
Stanford University	**Stanford Technology Access Resource Team:** - 10-week course for Stanford undergraduate, graduate, and medical students ☑ Didactic ☑ Experiential ☑ Community-based ☑ Longitudinal	**A primary care effort to bridge the telehealth divide** "Exploring concepts in communication, community-building, design thinking, and team-based patient care while providing a service that will connect vulnerable patients and their caregivers to health care providers through video visits."
Thomas Jefferson University	**College of Health Professions:** - 12-month course (web-based module) open for Thomas Jefferson graduate students ☑ Didactic ☑ Experiential ☐ Community-based ☐ Longitudinal	**Connected care: telehealth and digital health innovation** "Improving healthcare delivery & outcomes; engaging patients in prevention medicine, early diagnosis, and managing chronic conditions through the successful implementation and practice of digital health."
Uniformed Services University of the Health Sciences	**Defence Health Agency:** - Asynchronous or classroom course (9 hours) open for MS-3 students ☑ Didactic ☑ Experiential ☐ Community-based ☐ Longitudinal	**Introduction to telehealth and service learning** "Covering telehealth history, applications, ethics, safety, military uses, etiquette, and patient considerations; engaging in faculty-supervised mock patient telehealth encounters; practicing advanced surgical procedures using telehealth equipment."
University of New Hampshire	**The Telehealth Practice Center: Student Learning and Professional Training:** - 17-week course open for third and fourth year undergraduate students ☑ Didactic ☐ Experiential ☐ Community-based ☐ Longitudinal	**HHS 798 or HHS 898: Introduction to telehealth** "Learning about rural health, digital divides, assistive technology, innovations, and current practices in telehealth; understanding the roles of healthcare providers as they offer digital visits; working with classmates on a telehealth-related project."
University of Southern California	**Keck Service Learning Program:** - 1-year course (4-6 hours) open for USC MS-1 students ☑ Didactic ☑ Experiential ☑ Community-based ☐ Longitudinal	**Course in health justice and systems of care curriculum** "Learning about structural, social, and psychological influences on health, engaging in a service-learning experience; discussing ways clinicians can partner with community health entities, reflecting on the role of medical students as health professionals and citizens."
University of Utah	**University of Utah Health:** - 12-week web-based synchronous course open for DBMI and MHA students ☑ Didactic ☐ Experiential ☐ Community-based ☐ Longitudinal	**BMI 6050 or MHA 6050** "Thinking through ways technology can solve healthcare access challenges facing patients and providers, learning about the legal, ethical, service delivery, and financial aspects of telehealth services."
Veterans Health Administration Palo Alto General Medicine Clinic	**Asynchronous Prevention Clinic:** - Volunteer program for high school, college, and medical students ☑ Didactic ☑ Experiential ☑ Community-based ☑ Longitudinal	**Decreasing racial disparities in hypertension** "Connecting students with medical teams to make telephone calls; instructing veterans how to take accurate home blood pressures; designing an intervention that reaches Black patients and other poorly resourced veterans to improve clinical outcomes for all."

Service Learning

We partnered with the institution's informatics and technology, compliance, and interpretation services teams to develop a protocol for students to assist patients in establishing accounts in the EMR portal and video visits. The START course is open to any Stanford undergraduate or graduate student interested in service learning around medically vulnerable populations and is listed in the Stanford course directory as MED 258: Stanford Technology Access Resource Team: A Primary Care Effort to Bridge the Telehealth Divide. Once enrolled, students completed HIPAA training and other institutional compliance requirements. The institution's informatics and technology team provided students with technical training. Students used Doximity Dialer,

XSL•FO
RenderX

Cisco Jabber, and Google Voice to keep their personal phone numbers private.

We included patients from Stanford and one of our affiliated FQHCs. We identified and recruited Stanford patients for this program through EMR queries for patients seen in our primary care system in the last 2 years who lacked an EMR account or who never had a video appointment. We discussed a similar recruitment process with our FQHC partners. Many of their patients were using phone visits, and the FQHC leadership was interested in ensuring that these phone-visit patients could participate in video visits, which increased the patient-provider connection and diagnostic potential of the encounter. As such, they preferred to reach out to patients who had been seen by their primary care partner within the past 1 month and who were on the schedule for any remote clinical visit—which included both telephone and video visits within 2 to 4 weeks. We also informed the primary care clinicians of START services and encouraged referrals. After completing the necessary training, START students were each given a list of 10 patients without digital health access and called all the "remote clinical visit" patients, offering technology help to patients who did not have current video visit services. The course director prioritized language concordance between the students and patients when possible.

START students reached out to patients by telephone using scripts and resources tailored to the 2 institutions (Multimedia Appendix 4). The FQHC provided a student guidebook that included scripts and screenshots for helping patients set up EMR accounts and video visits. The institution's technology team also developed a PowerPoint presentation to show students how to set up EMR accounts and video visits. Among the various protocols and resources, students had access to sample introductory scripts and university IT phone numbers. We have included the protocols for assistance in Multimedia Appendix 4. Following each call, START students logged each of their patient encounters using a web-based survey tool (Multimedia Appendix 5). If a patient did not pick up their phone, the student left a detailed voicemail and attempted 2 more calls. Students requested additional lists of 10 patients when ready to outreach to additional patients. Students had access to interpretation services when needed and could reach the course director outside of class time for questions or patient-specific concerns. For the nonresponders who we could not reach after multiple contact attempts, we could not draw conclusions about why they did not respond to our phone calls.

Learner Impact

We asked students to submit weekly reflections in response to the following prompt: "Please upload a reflection of any kind - poem, art, paragraph, recording, etc. on an aspect of START you encountered this past week." Students were given opportunities to share their reflections during class. At the end of 6 quarters, 2 independent reviewers (NC and ZND) coded the reflections for thematic content. Thematic saturation was achieved after reviewing 20 reflections. We randomly selected another 5 reflections to review to ensure no additional themes arose. During coding of the remaining 114 reflections, no new themes were identified. Two reviewers (NC and ZND) then independently assigned themes to all reflections (multiple themes could be assigned to a single reflection if applicable). A third reviewer (RB) resolved any discrepancies through consensus discussion [19,20].

Patient Impact

Students completed patient encounter logs via a survey tool (Multimedia Appendix 5) documenting details of the call, including whether the outcome of the call was successful. In July 2023, we conducted follow-up calls with all 141 patients who expressed interest in accessing their EMR accounts a year and a half after the start of the course and video visits, and the respondents were given 4 prompts (Multimedia Appendix 6) to elicit their satisfaction with the student encounter and their success in using telehealth. We took detailed interview notes including patient quotes and conducted a thematic analysis of the responses. We did not follow-up with the 88 patients who declined support when the START student first connected with them.

Results

Patient Outcomes

Over 6 academic quarters (2 years), 57 students reached out to 1185 patients (n=711, 60% from an academic medical center and n=474, 40% from an FQHC). Of the 229 patients reached, 83.8% (n=192) spoke English; 10.9% (n=25) spoke Spanish; 3.9% (n=9) spoke Mandarin; and 1.3% (n=3) spoke Farsi, Korean, or Russian. Two students accessed our institution's interpretation services. The amount of time spent on the calls ranged from less than 5 minutes to more than 60 minutes, and the majority of calls took less than 15 minutes.

Of the 229 patients reached, 141 patients desired telehealth visits but lacked access (Table 1). START students successfully established an EMR portal account and video visit setup in 78.7% (111/141) of patients who desired telehealth. The remaining 21.3% (30/141) experienced barriers to establishing access, which included lack of access to Wi-Fi or a device (n=11), absence of an interpreter (n=4), disabilities precluding the patient's use of video visits (n=2), and unknown reasons (n=13).

XSL•FO
RenderX

Table 1. Outcomes from Stanford Technology Access Resource Team patient encounters (N=229).

Outcomes	Patient encounters (N=229), n (%)
Successful outreaches	111 (48.5)
EMR[a] account established	50 (21.8)
Video visit via Doximity	36 (15.7)
Video visit via EPIC platform	18 (7.9)
Video visit via Zoom	7 (3.1)
Connection but patient declined support	88 (38.4)
Declined assistance or did not specify	44 (19.2)
Account already set up	23 (10)
Preferred in-person or phone visit	18 (7.9)
No longer in health care system	3 (1.3)
Connection but inability to establish telemedicine access	30 (13.1)
Unknown	13 (5.7)
No device or no Wi-Fi	11 (4.8)
Interpreter not available	4 (1.7)
Disability precluding use	2 (0.9)

[a]EMR: electronic medical record.

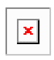
Over a third (88/229, 38.4%) of contacted patients did not engage with START students for the following reasons: 19.2% (44/229) declined assistance for unknown reasons, 10% (23/229) already had accounts or video visits setup, 7.9% (18/229) preferred in-person or phone appointments, and 1.3% (3/229) were no longer part of the health care system.

START Student Outcomes

Over the course of 6 academic quarters, multiple cohorts of students submitted a total of 139 reflections. From the reflections, we identified 7 themes: real-world communication or relationship-building (n=67, 48.2%), logistics of communication (n=67, 48.2%), value of service, (n=60, 43.2%), improved awareness of patient experiences (n=38, 27.3%), patient demographics affecting telehealth access (n=35, 25.2%), future career life goal or purpose (n=19, 13.7%), and graphical or pictorial reflections (n=18, 12.9%; Table 2). On the course evaluations, 76% (39/51) of students stated they learned “a great deal” or “a lot” from the course. Future pre- and postcourse student evaluations could provide quantitative measures of student learning including any changes in communication skills and knowledge of health care systems, community-building, team-based care, and design thinking.

XSL•FO
RenderX

Table 2. Samples of Stanford Technology Access Resource Team (START) student reflections and illustrated themes.

Theme	Mentions (n=139), n (%)	Description	Representative quote
Real-world communication or relationship-building	67 (48.2)	Highlights the development of skills in communication or in building rapport	• "Wanting to go into medicine, communicating with others is a very important skill to have. Many times, it feels like this has been lost in the age of social media, but it has been nice to have this class to sharpen my skills." • "Although the structure was scripted, I was inspired by the glimpses into the lives of the participants that they provided through their answers—from trips to the beach with a pet, to spending time in a garden."
Logistics of communication	67 (48.2)	Documents the successes and areas for improvement regarding the START course	• "This past week has been eventful as I called all my patients again that did not answer before. I finalized all my survey responses and submitted them. There were several patients who did not answer and/or had disconnected phone numbers." • "Being able to effectively communicate with people without providing them with visuals or physical assistance is often difficult."
Value of service	60 (43.2)	Examines the student's efforts to serve the patient or the community	• "Every ring of the phone / Every click of the mouse, / Every warm "Hello," / Every "I can't access this page," / Every "Oh! I see it now," / Every appreciative "Thank you for understanding" / Every concluding "Your help is always cherished." / Only fuels my drive, my purpose / I feel significant in the never-ending telehealth battle, / Slowly building a bridge of compassion / Within a world filled with divides."
Improved awareness of patient experiences	38 (27.3)	Explores the student's exposure to novel patient challenges and the effects of digital literacy on health care	• "In working with patients with varied technological experience, access, and support, I have developed a strong appreciation for the privilege that we, as members of a tech-fortified community are immersed in daily, and most importantly, gained significant empathy and understanding of the challenges that many patients face in accessing essential and quality healthcare, especially technology-reliant healthcare like telemedicine."
Patient demographics affecting telehealth access	35 (25.2)	Centers on the student's recognition of the impact of social determinants of health on patients	• "Seeing how certain industries and people responded to the pandemic shows the need for telemedicine for people who are more at risk due to age, disability, and other conditions that have made it extremely difficult for them to navigate the world." • "A recurring theme for me is how the rise of advancements in telehealth has enabled expansion of care to populations who have historically experienced high barriers to technology entry."
Future career life goal or purpose	19 (13.7)	Focuses on the course's impact on the student's future career or life goal	• "Med258 was a really important experience in confirming my passion to become a physician!" • "Right now, I'm thinking about going into primary care—in large part because of this course." • "The most common thing I have noticed with the doctors I have shadowed is they are in and out when seeing patients. If this class has taught me anything, it is the importance of taking the time to listen to patients' stories."
Graphical or pictorial reflections	18 (12.9)	Illustrates START patient encounters through images and diagrams	

Program Evaluation by Patients

After working with a START student for 1.5 years or more, we reviewed the records from the student logs of the encounters and the patient charts of 50 patients and found that 31 (62%) patients have had successful video visits, 8 (16%) had died, and 5 (10%) established an EMR account but did not have subsequent video visits. The status of 6 (12%) patients could not be determined due to data entry error. We successfully contacted 19 (13.5%) of the 141 patients from our primary care clinics who were helped by START students to conduct a follow-up survey. Almost all (n=18, 95%) patients reported that START students were either very (n=14, 74%) or somewhat (n=4, 21%) helpful in assisting them with health technology. The majority of patients (n=14, 74%) reported that they successfully gained access to their EMR accounts as a result of student help, while 58% (n=11) said that they connected successfully to video visits. Of the 5 patients who did not gain EMR account access, 3 patients forgot the steps and 2 patients had disabilities hindering their abilities to use the EMR or lacked access to a device.

XSL•FO
RenderX

A detailed analysis of notes regarding patient experiences, including quotes, from the follow-up calls produced 5 themes: patient satisfaction and improvement suggestions (n=19, 100%), navigation of technology logistics (n=14, 74%), activation of electronic health portal access (n=9, 47%), preparation for video visit (n=8, 42%), and support of patient scheduling (n=5, 26%; Table 3). All patients reached (n=19, 100%) expressed appreciation for student help in increasing their familiarity with the EMR. Several patients reported that students demystified the digital health experience, which included conversations about data privacy. Some patients expressed gratitude for the companionship from the student calls, highlighting the students' "patience," "kindness," and "effort."

Table 3. Sample patient quotes from follow-up calls and themes illustrated.

Theme (number of mentions)	Mentions (n=19), n (%)	Description	Representative quote
Patient satisfaction and improvement suggestions	19 (100)	Analyzes patient feedback on the support provided by students	• "The student was very, very helpful in answering all of my questions. However, when we hung up, I forgot the steps to log into my account. I would have liked it if the student would have followed up with me again." • "The young lady was very very helpful in setting up my EMR portal account." • [When asked "How useful was your meeting with the START student in helping you connect with your doctor by EMR portal? (Scale 1-5) 5 – Very helpful, 1 – Not helpful] Patient replied: "I would rate it more than a 5, it's a 10!"
Navigation of technology logistics	14 (74)	Explores patient struggles with telehealth technology and the assistance they receive, including step-by-step guidance from students	• "When the pandemic started, I regularly used Zoom to go to church. But for some reason I couldn't figure out how to connect with my doctor [via telehealth]. The nurses tried helping me, but I would forget the steps. So, I decided to just do voice calls with my doctor each visit ... But when the [START] student called, they walked me through the step-by-step process of how to use Zoom and was really patient with me. After a week of working with the student, I was finally able to connect with my doctor using video. The student also helped me set up my fax machine, which I was very grateful for."
Activation of electronic health portal access	9 (47)	Examines patients gaining access to electronic health portals	• "For the longest time, I wasn't able to get an activation code set up for my [EMR portal] account. When the student called me, she was also having trouble getting my account activated ... She [the student] hung up to follow up with my primary care team. The next day, she called me back and we were able to set up my account by using my credit card number."
Preparation for video visit	8 (42)	Focuses on the reminders patients receive from students	• "I cannot fully recall the interaction with the student, but I do remember a kind young lady who called me and went through a checklist of items I should have next to me in preparation for my video visit appointment."
Support of patient scheduling	5 (26)	Delves into the assistance provided to patients in scheduling appointments	• "My mother is 94 years old ... When the pandemic was really bad, the student helped me set up a vaccine appointment for both my mom and me. Since my mom is 94, she prefers to go in person for her appointments, so we didn't use the video visit services. However, I appreciate the student's effort."

Discussion

Principal Findings

We found that undergraduate and graduate students can function as technology navigators for patients with low digital literacy levels while gaining valuable experience in health disparities. Our students successfully guided over 100 patients through the steps necessary for digital health access, thereby connecting patients with their health care teams, supporting patients in setting up appointments, reminding patients of items to bring to their visits, and providing an empathetic perspective. Overall, patients expressed deep satisfaction with their interactions with START students, highlighting students' positive attitudes and demystification of technology (Table 3).

Digital health literacy is an emergent social determinant of health [21]. While telehealth and technology-based tools can make health care more equitable and accessible, they can also create barriers to quality care for patients with low digital literacy [22-24]. The eHealth literacy theoretical framework by Norman and Skinner [25] explains that users of electronic health tools must be able to "seek, find, understand, and appraise" information in order to be technology literate [21]. Our study reinforces previous findings, showing patients with limited digital health literacy may successfully participate in video visits

XSL•FO
RenderX

with proper education in the evolving technology landscape [21,26,27]. While asynchronous web-based tutorials may benefit many patients with technology navigation issues, others require a more personalized, hands-on approach to technology education [11,12,28]. For instance, older adults and patients with cognitive impairments require more individualized resources to increase health literacy [21]. For patients in our cohort, lack of telehealth access was more a consequence of lack of technological support or education, as opposed to lack of internet access or technology ownership.

Treating digital literacy analogously to other social determinants of health can inform both policy and health care practice. Health care providers, educators, and policy makers are well-positioned to integrate digital literacy into patient care and medical education. To address telehealth and other health technology adoption as a social determinant of health, technological support should be individualized and patient-centered [28-30]. Patient-centered digital literacy can be enhanced through interdisciplinary collaboration between technologists, public health experts, and health care providers (social work, etc) to develop new programs such as START.

In addition to positively impacting patient outcomes, integrating service learning and civic engagement in our course enhanced student learning as future health care providers [13,31-35]. Currently existing service-learning programs incorporate learner practicums on direct patient access and offer valuable opportunities for student immersion in community-based experiential work [36-39]. Building on the foundations of existing community engagement courses and programs (Table 4) [40-47], the START curriculum embeds hands-on experiences in a classroom setting to address technology literacy and access.

XSL•FO
RenderX

Table 4. Examples of current telehealth or service-learning educational programs.

Institution or organization	Course structure	Format	Curriculum content
Florida Atlantic University	*Christine E. Lynn College of Nursing (DNP):* 4-week course (web-based module) open for the general public	• Didactic	*Telehealth certificate:* "Learning about health equity; improving access to health care, health disparities, and inter-professional practice with specific populations in relationship to the application and implementation of telehealth platforms."
Stanford University	*Stanford Technology Access Resource Team:* 10-week course for Stanford undergraduate, graduate, and medical students	• Didactic • Experiential • Community based • Longitudinal	*A primary care effort to bridge the telehealth divide:* "Exploring concepts in communication, community-building, design thinking, and team-based patient care while providing a service that will connect vulnerable patients and their caregivers to health care providers through video visits."
Thomas Jefferson University	*College of Health Professions:* 12-month course (web-based module) open for Thomas Jefferson graduate students	• Didactic • Experiential	*Connected care: telehealth and digital health innovation:* "Improving healthcare delivery & outcomes; engaging patients in prevention medicine, early diagnosis, and managing chronic conditions through the successful implementation and practice of digital health."
Uniformed Services University of the Health Sciences	*Defence Health Agency:* Asynchronous or classroom course (9 hours) open for MS-3 students	• Didactic • Experiential	*Introduction to telehealth and service learning:* "Covering telehealth history, applications, ethics, safety, military uses, etiquette, and patient considerations; engaging in faculty-supervised mock patient telehealth encounters; practicing advanced surgical procedures using telehealth equipment."
University of New Hampshire	*The Telehealth Practice Center:Student Learning and Professional Training:* 17-week course open for third- and fourth-year undergraduate students	• Didactic	*HHS 798 or HHS 898: Introduction to telehealth:* "Learning about rural health, digital divides, assistive technology, innovations, and current practices in telehealth; understanding the roles of healthcare providers as they offer digital visits; working with classmates on a telehealth-related project."
University of Southern California	*Keck Service Learning Program:* 1-year course (4-6 hours) open for USC[a] MS-1 students	• Didactic • Experiential • Community based	*Course in health justice and systems of care curriculum:* "Learning about structural, social, and psychological influences on health, engaging in a service-learning experience; discussing ways clinicians can partner with community health entities, reflecting on the role of medical students as health professionals and citizens."
University of Utah	*University of Utah Health:* 12-week web-based synchronous course open for DBMI[b] and MHA[c] students	• Didactic	*BMI 6050 or MHA 6050:* "Thinking through ways technology can solve healthcare access challenges facing patients and providers, learning about the legal, ethical, service delivery, and financial aspects of telehealth services."
Veterans Health Administration of Palo Alto General Medicine Clinic	*Asynchronous Prevention Clinic:* Volunteer program for high school, college, and medical students	• Didactic • Experiential • Community based • Longitudinal	*Decreasing racial disparities in hypertension:* "Connecting students with medical teams to make telephone calls; instructing veterans how to take accurate home blood pressures; designing an intervention that reaches Black patients and other poorly resourced veterans to improve clinical outcomes for all."

[a]USC: University of Southern California.

[b]DBMI: Department of Biomedical Informatics.

[c]MHA: Master of Healthcare Administration.

Comparison to the Literature

Service-learning programs vary in terms of duration, degree of service requirements, direct versus indirect interactions with patients and caregivers, and topics covered in their curricula (Table 4) [40-47]. We found service-learning programs varying in length between a few weeks to a year, usually with a specific scientific or clinical focus, such as telehealth, equity or disparities, community health, and social determinants of health. Programs had relevant experiential components, based on their learning objectives, such as video simulations of patient encounters for video visit experiences. One service-learning program for medical students involved community-based clinical work [45].

Our START program teaches students about the upstream factors that may hinder patients' abilities to access health care and offers students the training resources to support patients in connecting with their health care teams via telehealth. As a result, students can apply the didactic lessons they learn in classroom settings immediately to first-hand experiences working directly with patients. Given the dynamic and experiential nature of the curriculum, to our knowledge, START is currently one of the only courses or programs offered to undergraduate and graduate students that combines the best

XSL•FO
RenderX

practices of service learning with direct support for patients seeking assistance in accessing their EMRs and video visits.

Strengths and Limitations

This study has several strengths and limitations. During program development and implementation, we had access to health equity experts, EMR support, and administrative support. However, even with access to patient lists generated from EMR searches, 80.7% (n=956) of patients did not answer calls or had disconnected numbers, illustrating the need to find better ways of reaching our target patient population. When eventually connecting with patients, respondents shared that they were initially unavailable and did not recognize the phone number on the caller ID and therefore did not initially answer the call. Subsequently, calls were made through a service (Cisco Jabber) that provided the institution's caller identification. Future efforts would benefit from appropriate caller identification, when possible, to signal to patients that the caller is an extension of the care team. Our students had the opportunity to work with patients in 2 diverse populations (an academic medical center and FQHC). While successful, student supervision and trust-building within the FQHC community took additional time and resources. This may pose a challenge for institutions without such capacities. Some communities may have additional barriers such as larger portions of patients who are not technology native. It is also worth noting that the START program is an ongoing partnership with the institution's IT team such that the training session for the students would be updated according to changes in the institution-wide digital health platforms. Similarly, the guidebook for the FQHC site would be updated according to updates in digital health technology.

Our results are reliant on student encounter logs, which may introduce consistency bias and the Hawthorne effect in self-reported data. Given the course structure, we did not follow patients over time. Future initiatives may consider having students follow up with patients soon after their scheduled doctor appointments using the same phone numbers. Patients contacted after course may have recall bias, have forgotten specifics of the interaction, or have been influenced by social desirability bias. Subsequent assessment of interaction impact will include patient follow-up within a week regarding their interaction with START students. A total of 4.8% (n=11) of our patients were unable to get digitally connected with their health care providers due to lack of access to essential devices or Wi-Fi. Future programs may consider developing partnerships with technology providers or community organizations to send devices or Wi-Fi routers to such patients (see Multimedia Appendix 4 for examples of "Links to Local Free Wi-Fi Resources" used in this study). Students had suggested a variety of program improvements, such as collaborative debriefing during sessions and providing in-person services when possible, to mitigate the challenge of navigation solely by telephone. When technical challenges faced by patients and students arose, students connected with patients' family members or caretakers, who were able to troubleshoot. Future initiatives should consider providing students with additional phone numbers to alternate or emergency contacts. Additionally, establishing a schedule for students to enter the clinic at designated times and obtaining direct referrals from clinicians' primary care providers could also streamline the referral process. Finally, we recommend gathering alternative contact information from students (eg, personal emails) to track their career trajectories after completing the course or program. Future studies may also consider using pre- and postcourse student surveys assessing changes in empathy and cultural humility, which could provide quantitative and qualitative measurements of program utility on these qualities.

Conclusions

Helping patients navigate complex health systems has called attention to an arising social determinant of health—digital literacy. Approaching the needs of patients with low digital literacy is crucial to helping them reap the benefits of high-quality and efficient health care. START is an applied, experiential, longitudinal, and community-based program that couples students' desire for service learning with curricula on social determinants of health, health disparities, and patient-centered communication skills. Programs such as START may empower students to serve as digital health navigators and may foster more culturally humble and compassionate health care professionals. START presents a valuable opportunity for our next generation of clinicians to work with patients who are vulnerable and to develop empathy and communication skills earlier on in their academic journeys while addressing digital literacy as an emergent social determinant of health. Future service-learning programs could adapt the START model of integrating a student course with partnerships with the institution's information and technology, compliance, and interpretation and translation services to provide services beyond connecting to video visits such as helping with applications for social services, completing advance directives for health care, navigating transportation barriers, and bridging to community partners.

Acknowledgments

The authors would like to express their gratitude to the students who provided support to the patients and contributed their reflections during the Stanford Technology Access Resource Team course, especially Grace Toluwanimi Adebogun, Sohayla Eldeeb, Catherine Gao, Brianna Gamble, Jennifer John, Japnoor Kaur, Ashley Danielle Nies, Misha Raffiee, and Lea Wenting Rysavy for their featured reflections. The authors recognize Ashley Danielle Nies and Misha Raffiee for their help in reviewing reflections. The authors are indebted to Alice Mau, MD, for her help with the patient needs assessment and Audrey Xu, for her help with the student needs assessment.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Stanford Technology Access Resource Team student needs assessment.
[DOCX File , 20 KB - mededu_v10i1e57077_app1.docx]

Multimedia Appendix 2
Patient needs assessment.
[DOCX File , 16 KB - mededu_v10i1e57077_app2.docx]

Multimedia Appendix 3
Health care teams needs assessment.
[DOCX File , 14 KB - mededu_v10i1e57077_app3.docx]

Multimedia Appendix 4
Tips or resource sheet for Stanford Technology Access Resource Team students.
[DOCX File , 68 KB - mededu_v10i1e57077_app4.docx]

Multimedia Appendix 5
Stanford Technology Access Resource Team patient encounter log template (built into a survey app).
[DOCX File , 14 KB - mededu_v10i1e57077_app5.docx]

Multimedia Appendix 6
Follow-up survey questionnaire.
[DOCX File , 16 KB - mededu_v10i1e57077_app6.docx]

References

1. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
2. Koonin LM, Hoots B, Tsang CA, Leroy Z, Farris K, Jolly B, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic—United States, January-March 2020. MMWR Morb Mortal Wkly Rep 2020 Oct 30;69(43):1595-1599 [FREE Full text] [doi: 10.15585/mmwr.mm6943a3] [Medline: 33119561]
3. Samson LW, Tarazi W, Turrini G, Sheingold S. Medicare beneficiaries' use of telehealth in 2020: trends by beneficiary characteristics and location. US Department of Health and Human Services. 2021. URL: https://aspe.hhs.gov/reports/medicare-beneficiaries-use-telehealth-2020 [accessed 2021-12-03]
4. Charleson K. Telehealth statistics and trends: a 2021 report. The Checkup. 2021. URL: https://www.singlecare.com/blog/news/telehealth-statistics/ [accessed 2023-01-04]
5. Manca DP. Do electronic medical records improve quality of care? Yes. Can Fam Physician 2015 Oct;61(10):846-847 [FREE Full text] [Medline: 26472786]
6. Price JC, Simpson DC. Telemedicine and health disparities. Clin Liver Dis (Hoboken) 2022 Apr;19(4):144-147 [FREE Full text] [doi: 10.1002/cld.1171] [Medline: 35505914]
7. Qian L, Sy LS, Hong V, Glenn SC, Ryan DS, Morrissette K, et al. Disparities in outpatient and telehealth visits during the COVID-19 pandemic in a large integrated health care organization: retrospective cohort study. J Med Internet Res 2021 Sep 01;23(9):e29959 [FREE Full text] [doi: 10.2196/29959] [Medline: 34351865]
8. Lau KHV, Anand P, Ramirez A, Phicil S. Disparities in telehealth use during the COVID-19 pandemic. J Immigr Minor Health 2022 Dec;24(6):1590-1593 [FREE Full text] [doi: 10.1007/s10903-022-01381-1] [Medline: 35976473]
9. McGinley MP, Ontaneda D, Wang Z, Weber M, Shook S, Stanton M, et al. Teleneurology as a solution for outpatient care during the COVID-19 pandemic. Telemed J E Health 2020 Dec;26(12):1537-1539 [FREE Full text] [doi: 10.1089/tmj.2020.0137] [Medline: 32552509]
10. Truong M, Yeganeh L, Cook O, Crawford K, Wong P, Allen J. Using telehealth consultations for healthcare provision to patients from non-Indigenous racial/ethnic minorities: a systematic review. J Am Med Inform Assoc 2022 Apr 13;29(5):970-982 [FREE Full text] [doi: 10.1093/jamia/ocac015] [Medline: 35150266]
11. Shaver J. The state of telehealth before and after the COVID-19 pandemic. Prim Care 2022 Dec;49(4):517-530 [FREE Full text] [doi: 10.1016/j.pop.2022.04.002] [Medline: 36357058]
12. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117 [FREE Full text] [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]

XSL•FO
RenderX

13. Lin SY, Schillinger E, Irby DM. Value-added medical education: engaging future doctors to transform health care delivery today. J Gen Intern Med 2015 Feb;30(2):150-151 [FREE Full text] [doi: 10.1007/s11606-014-3018-3] [Medline: 25217209]
14. Stanford health care community partnership program 2021 community benefit report. Stanford Medicine. 2021. URL: https://stanfordhealthcare.org/content/dam/SHC/about-us/public-services-and-community-partnerships/docs/fy21-cb-oshpd-report-1-7-22-final.pdf [accessed 2024-08-24]
15. Health care that cares: 21 years of service in the community. Ravenswood Family Health Network. 2021. URL: https://ravenswoodfhn.org/wp-content/uploads/2018/10/21_22_AR-Final_singlepages.pdf [accessed 2024-08-24]
16. Stanford Facts. Stanford University. URL: https://facts.stanford.edu/ [accessed 2023-11-15]
17. Other undergraduate education facts—facts. Stanford University. URL: https://facts.stanford.edu/academics/undergraduate-facts/ [accessed 2023-11-15]
18. How to succeed as a Stanford premed student. Shemmassian Academic Consulting. 2023. URL: https://www.shemmassianconsulting.com/blog/stanford-premed [accessed 2023-11-15]
19. Demirören M, Atılgan B. Impacts of service learning-based social responsibility training on medical students. Adv Physiol Educ 2023 Jun 01;47(2):166-174 [FREE Full text] [doi: 10.1152/advan.00049.2022] [Medline: 36701494]
20. Fries KS, Bowers DM, Gross M, Frost L. Service learning in Guatemala: using qualitative content analysis to explore an interdisciplinary learning experience among students in health care professional programs. J Multidiscip Healthc 2013;6:45-52 [FREE Full text] [doi: 10.2147/JMDH.S35867] [Medline: 23430865]
21. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023 Oct;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
22. DeHart D, King LB, Iachini AL, Browne T, Reitmeier M. Benefits and challenges of implementing telehealth in rural settings: a mixed-methods study of behavioral medicine providers. Health Soc Work 2022 Jan 31;47(1):7-18. [doi: 10.1093/hsw/hlab036] [Medline: 34910158]
23. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
24. Simon DA, Shachar C. Telehealth to address health disparities: potential, pitfalls, and paths ahead. J Law Med Ethics 2021;49(3):415-417. [doi: 10.1017/jme.2021.62] [Medline: 34665098]
25. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
26. Busse TS, Nitsche J, Kernebeck S, Jux C, Weitz J, Ehlers JP, et al. Approaches to improvement of digital health literacy (eHL) in the context of person-centered care. Int J Environ Res Public Health 2022 Jul 07;19(14):8309 [FREE Full text] [doi: 10.3390/ijerph19148309] [Medline: 35886158]
27. Truong M, Fenton SH. Understanding the current landscape of health literacy interventions within health systems. Perspect Health Inf Manag 2022;19(Spring):1h [FREE Full text] [Medline: 35692852]
28. Sattar S, Kuperman R. Telehealth in pediatric epilepsy care: a rapid transition during the COVID-19 pandemic. Epilepsy Behav 2020 Oct;111:107282 [FREE Full text] [doi: 10.1016/j.yebeh.2020.107282] [Medline: 32759065]
29. Zhou X, Snoswell CL, Harding LE, Bambling M, Edirippulige S, Bai X, et al. The role of telehealth in reducing the mental health burden from COVID-19. Telemed J E Health 2020 Apr;26(4):377-379. [doi: 10.1089/tmj.2020.0068] [Medline: 32202977]
30. Parsons D, Cordier R, Lee H, Falkmer T, Vaz S. A randomised controlled trial of an information communication technology delivered intervention for children with autism spectrum disorder living in regional Australia. J Autism Dev Disord 2019 Feb;49(2):569-581. [doi: 10.1007/s10803-018-3734-3] [Medline: 30209645]
31. Gonzalo JD, Dekhtyar M, Hawkins RE, Wolpaw DR. How can medical students add value? Identifying roles, barriers, and strategies to advance the value of undergraduate medical education to patient care and the health system. Acad Med 2017 Sep;92(9):1294-1301. [doi: 10.1097/ACM.0000000000001662] [Medline: 28353500]
32. Leeferink H, Koopman M, Beijaard D, Schellings GLM. Overarching professional identity themes in student teacher workplace learning. Teach Teach 2018 Oct 04;25(1):69-89. [doi: 10.1080/13540602.2018.1527762]
33. Leep Hunderfund AN, Starr SR, Dyrbye LN, Gonzalo JD, George P, Miller BM, et al. Value-added activities in medical education: a multisite survey of first- and second-year medical students' perceptions and factors influencing their potential engagement. Acad Med 2018 Oct;93(10):1560-1568. [doi: 10.1097/ACM.0000000000002299] [Medline: 29794526]
34. Menear M, Blanchette MA, Demers-Payette O, Roy D. A framework for value-creating learning health systems. Health Res Policy Syst 2019 Aug 09;17(1):79 [FREE Full text] [doi: 10.1186/s12961-019-0477-3] [Medline: 31399114]
35. Pacho T. Service-Learning: An Innovative Approach to Education in Africa. Africa: Paulines Publications Africa; 2019:232-259.
36. Mason MR, Dunens E. Service-learning as a practical introduction to undergraduate public health: benefits for student outcomes and accreditation. Front Public Health 2019;7:63 [FREE Full text] [doi: 10.3389/fpubh.2019.00063] [Medline: 31001507]
37. Walk with Me: early clinical experiences for medical students. Stanford Medicine, Department of Medicine. URL: https://medicine.stanford.edu/2019-report/walk-with-me.html [accessed 2023-11-15]

XSL•FO
RenderX

38. Bamdas JAM, Averkiou P, Jacomino M. Service-learning programs and projects for medical students engaged with the community. Cureus 2022 Jun;14(6):e26279 [FREE Full text] [doi: 10.7759/cureus.26279] [Medline: 35898383]
39. Rath C, Tillman F, Stickel J, Jones M, Armistead L. Implementation of a student-developed, service-based internship for pharmacy students. Innov Pharm 2019;10(2):10.24926/iip.v10i2.1550 [FREE Full text] [doi: 10.24926/iip.v10i2.1550] [Medline: 34007550]
40. Telehealth certificate program. Florida Atlantic University—Christine E. Lynn College of Nursing. URL: https://nursing.fau.edu/admissions/certificates/telehealth/ [accessed 2023-12-08]
41. Stanford University Bulletin. URL: https://explorecourses.stanford.edu/search?view=catalog&filter-coursestatus-Active=on&page=0&catalog=&academicYear=20222023&q=MED+258%3A+Stanford+Technology+Access+Resource+Team+-+a+primary+care+effort+to+bridge+the+telehealth+divide&collapse= [accessed 2023-12-08]
42. Education & training. Thomas Jefferson University. URL: https://research.jefferson.edu/connected-care-center/education-and-training.html [accessed 2023-12-08]
43. Jonas CE, Durning SJ, Zebrowski C, Cimino F. An interdisciplinary, multi-institution telehealth course for third-year medical students. Acad Med 2019 Jun;94(6):833-837. [doi: 10.1097/ACM.0000000000002701] [Medline: 30870152]
44. Student learning and professional training. University of New Hampshire, College of Health and Human Services. 2019. URL: https://chhs.unh.edu/telehealth-practice-center/focus-areas/student-learning-professional-training [accessed 2023-12-08]
45. Keck service-learning program. KSOM Department of Medical Education. URL: https://medstudent.usc.edu/keck-service-learning-program/ [accessed 2023-12-08]
46. Telehealth education. The University of Utah. 2021 Aug 11. URL: https://attheu.utah.edu/facultystaff/telehealth-education/ [accessed 2024-09-05]
47. Korshak L, Hausmann LRM. Reducing racial disparities in blood pressure control in veterans with severe hypertension. Office of Health Equity Veterans Health Administration Department of Veterans Affairs. URL: https://www.va.gov/HEALTHEQUITY/docs/REDUCING_RACIAL_DISPARITIES_BP_CONTROL_IN_VETERANS_WITH_SEVERE_HYPERTENSION.pdf [accessed 2024-12-07]

Abbreviations

EMR: electronic medical record
FQHC: federally qualified health center
HIPAA: Health Insurance Portability and Accountability Act
START: Stanford Technology Access Resource Team

Edited by B Lesselroth; submitted 07.02.24; peer-reviewed by S Mitra, K Knowles, A Hassan, C Janse van Vuuren; comments to author 06.05.24; revised version received 01.07.24; accepted 15.08.24; published 01.10.24.

Please cite as:
Doueiri ZN, Bajra R, Srinivasan M, Schillinger E, Cuan N
Bridging the Telehealth Digital Divide With Collegiate Navigators: Mixed Methods Evaluation Study of a Service-Learning Health Disparities Course
JMIR Med Educ 2024;10:e57077
URL: https://mededu.jmir.org/2024/1/e57077
doi:10.2196/57077
PMID:

©Zakaria Nadeem Doueiri, Rika Bajra, Malathi Srinivasan, Erika Schillinger, Nancy Cuan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 01.10.2024. 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.

XSL•FO
RenderX

Original Paper

Psychological Safety Competency Training During the Clinical Internship From the Perspective of Health Care Trainee Mentors in 11 Pan-European Countries: Mixed Methods Observational Study

Irene Carrillo[1], MSc, PhD; Ivana Skoumalová[2], PhD; Ireen Bruus[3], MSc; Victoria Klemm[4]; Sofia Guerra-Paiva[5,6], MsD; Bojana Knežević[7], PhD; Augustina Jankauskiene[8], PhD; Dragana Jocic[9], PhD; Susanna Tella[10], PhD; Sandra C Buttigieg[11], PhD; Einav Srulovici[12], PhD; Andrea Madarasová Gecková[2,13], PhD; Kaja Põlluste[14], PhD; Reinhard Strametz[4], MD; Paulo Sousa[5,6], PhD; Marina Odalovic[15], PhD; José Joaquín Mira[1,16], MPH, PhD

[1]Department of Health Psychology, Miguel Hernández University of Elche, Elche, Spain
[2]Department of Health Psychology and Research Methodology, Faculty of Medicine, Pavol Jozef Šafárik University, Kosice, Slovakia
[3]Tartu Health Care College, Tartu, Estonia
[4]Wiesbaden Institute for Healthcare Economics and Patient Safety (WiHelP), Wiesbaden Business School, RheinMain University of Applied Sciences, Wiesbaden, Germany
[5]Public Health Research Centre, National School of Public Health, NOVA University Lisbon, Lisbon, Portugal
[6]Comprehensive Health Research Center, National School of Public Health, NOVA University Lisbon, Lisbon, Portugal
[7]University Hospital Centre Zagreb, University of Zagreb, Zagreb, Croatia
[8]Pediatric Center, Institute of Clinical Medicine, Faculty of Medicine, Vilnius University, Vilnius, Lithuania
[9]BENU Pharmacy, PHOENIX Group Serbia, Belgrade, Serbia
[10]Faculty of Social and Health Care, LAB University of Applied Sciences, Lappeenranta, Finland
[11]Department of Health Systems Management and Leadership, Faculty of Health Sciences, University of Malta, Malta, Malta
[12]Cheryl Spencer Department of Nursing, University of Haifa, Haifa, Israel
[13]Institute of Applied Psychology, Faculty of Social and Economic Sciences, Comenius University Bratislava, Bratislava, Slovakia
[14]Institute of Clinical Medicine, University of Tartu, Tartu, Estonia
[15]Faculty of Pharmacy, University of Belgrade, Belgrade, Serbia
[16]Foundation for the Promotion of Health and Biomedical Research of the Valencia Region (FISABIO), Sant Joan d'Alacant, Spain

Corresponding Author:
Irene Carrillo, MSc, PhD
Department of Health Psychology
Miguel Hernández University of Elche
Avenida de la Universidad s/n
Elche, 03202
Spain
Phone: 34 966658350
Email: icarrillo@umh.es

Related Article:

This is a corrected version. See correction statement: https://mededu.jmir.org/2024/1/e68503

Abstract

Background: In the field of research, psychological safety has been widely recognized as a contributing factor to improving the quality of care and patient safety. However, its consideration in the curricula and traineeship pathways of residents and health care students is scarce.

Objective: This study aims to determine the extent to which health care trainees acquire psychological safety competencies during their internships in clinical settings and identify what measures can be taken to promote their learning.

XSL•FO
RenderX

Methods: A mixed methods observational study based on a consensus conference and an open-ended survey among a sample of health care trainee mentors from health care institutions in a pan-European context was conducted. First, we administered an ad hoc questionnaire to assess the perceived degree of acquisition or implementation and significance of competencies (knowledge, attitudes, and skills) and institutional interventions in psychological safety. Second, we asked mentors to propose measures to foster among trainees those competencies that, in the first phase of the study, obtained an average acquisition score of <3.4 (scale of 1-5). A content analysis of the information collected was carried out, and the spontaneity of each category and theme was determined.

Results: In total, 173 mentors from 11 pan-European countries completed the first questionnaire (response rate: 173/256, 67.6%), of which 63 (36.4%) participated in the second consultation. The competencies with the lowest acquisition level were related to warning a professional that their behavior posed a risk to the patient, managing their possible bad reaction, and offering support to a colleague who becomes a second victim. The mentors' proposals for improvement of this competency gap referred to training in communication skills and patient safety, safety culture, work climate, individual attitudes, a reference person for trainees, formal incorporation into the curricula of health care degrees and specialization pathways, specific systems and mechanisms to give trainees a voice, institutional risk management, regulations, guidelines and standards, supervision, and resources to support trainees. In terms of teaching methodology, the mentors recommended innovative strategies, many of them based on technological tools or solutions, including videos, seminars, lectures, workshops, simulation learning or role-playing with or without professional actors, case studies, videos with practical demonstrations or model situations, panel discussions, clinical sessions for joint analysis of patient safety incidents, and debriefings to set and discuss lessons learned.

Conclusions: This study sought to promote psychological safety competencies as a formal part of the training of future health care professionals, facilitating the translation of international guidelines into practice and clinical settings in the pan-European context.

(JMIR Med Educ 2024;10:e64125) doi:10.2196/64125

KEYWORDS

psychological safety; speaking up; professional competence; patient safety; education; adverse event

Introduction

Theoretical Background

Overview

Clinical errors are a common occurrence in health care, and many of them are preventable [1,2]. Adverse events involving health care professionals, including medical students during their internships, are unfortunately widespread [3]. These events have negative consequences for patients, health care professionals, and the health care system, commonly referred to as the first, second, and third victims, respectively [4-6]. Hence, prioritizing error prevention is crucial at both the local and national health care system levels. This focus serves to mitigate harm, reduce costs, and restore trust in health care [7].

While patient safety management in health care institutions was originally based on a reactive approach focused on acting after an incident and detecting what failed behind it, over time, this way of pursuing safety has evolved to a more positive one in which the objective is no longer just avoiding something going wrong (Safety I) but, above all, ensuring that everything goes right (Safety II). Far from being antagonistic, these 2 perspectives complement each other and allow for a more balanced and flexible approach to the reality of health care [8]. Thus, safety is linked not only to concepts such as incidents, errors, failures, or liability but also to other concepts, such as occupational well-being, resilience, or psychological safety.

Second Victims

Improved patient safety, in turn, has a positive impact on patient outcomes [9,10] and helps prevent health care professional burnout [11]. Health care professionals and even medical students involved in adverse events often experience symptoms akin to those of second victims, with profound implications for their emotional well-being as well as their professional and personal lives [9,12,13]. Recently, Vanhaecht et al [14], based on the literature and the expert consensus, defined a second victim as follows: "Any health care worker, directly or indirectly involved in an unanticipated adverse patient event, unintentional healthcare error, or patient injury and who becomes victimized in the sense that they are also negatively impacted."

Patient Safety Culture

Despite the recognition of the need for safer patient care, there remain various barriers [15-17]. Moreover, the effective processes after the adverse event, such as reporting of an adverse event, the analysis of the event, and open disclosure, can be hindered. These effective practices stem from and contribute to the psychological safety climate within the organization, encompassing elements such as a blame-free culture and a just culture [18,19].

Psychological Safety

Psychological safety is an important part of the safety culture in health care organizations. It is defined as a belief that individuals can feel safe to disagree, ask questions, and report mistakes without negative consequences and that they can cooperate as a team with mutual respect and trust [20]. Psychological safety contributes to improved outcomes in clinical training (eg, willingness to report adverse events [21] and speak up [22]) and, therefore, may be associated with better patient outcomes and improved safety culture [23]. Thus, comprehensive patient safety management combining Safety I

XSL•FO
RenderX

and Safety II approaches requires psychologically safe clinical environments [8].

Speaking Up for Ensuring Patient Safety

Psychological safety is pivotal in transitioning from a proactive to a generative patient safety culture [24]. It enables health care professionals to speak up without fear of consequences, which is crucial for overcoming barriers to safe care practices, including effective adverse event reporting and analysis [25]. Institutions must also prioritize the support and consideration of second victims as failing to do so can compromise patient safety [24].

Creating safe learning environments is essential for fostering psychological safety among future health care professionals. Training programs emphasizing open communication, mutual respect, and psychological safety equip students and professionals with the skills necessary for a safer health care system. By prioritizing these elements, health care organizations can establish a culture that supports both patient and provider well-being, ultimately enhancing safety and quality of care [24,25].

In summary, integrating psychological safety into health care practices creates a comprehensive safety framework, emphasizing a supportive culture, proactive error management, and the development of resilient health care professionals. This approach not only addresses immediate safety concerns but also contributes to the long-term sustainability of health care systems by nurturing a safety culture and psychological well-being [24].

Current Gap

Poor attention has been paid to promoting psychological safety in health care and medical education [26,27]. Future health care professionals' perceptions, attitudes, skills, and knowledge regarding safety culture cocreate a safe health care environment and contribute to better patient safety as well as a psychologically safer climate [28,29]. Strategies to establish psychological safety in clinical supervision were introduced to improve future health care professionals' training [30]. Little is known about how health care professionals are trained in psychological safety in European countries, which competencies and skills promoting psychological safety they obtain during their clinical internship, and which of them are considered important from the perspective of health care professionals in practice.

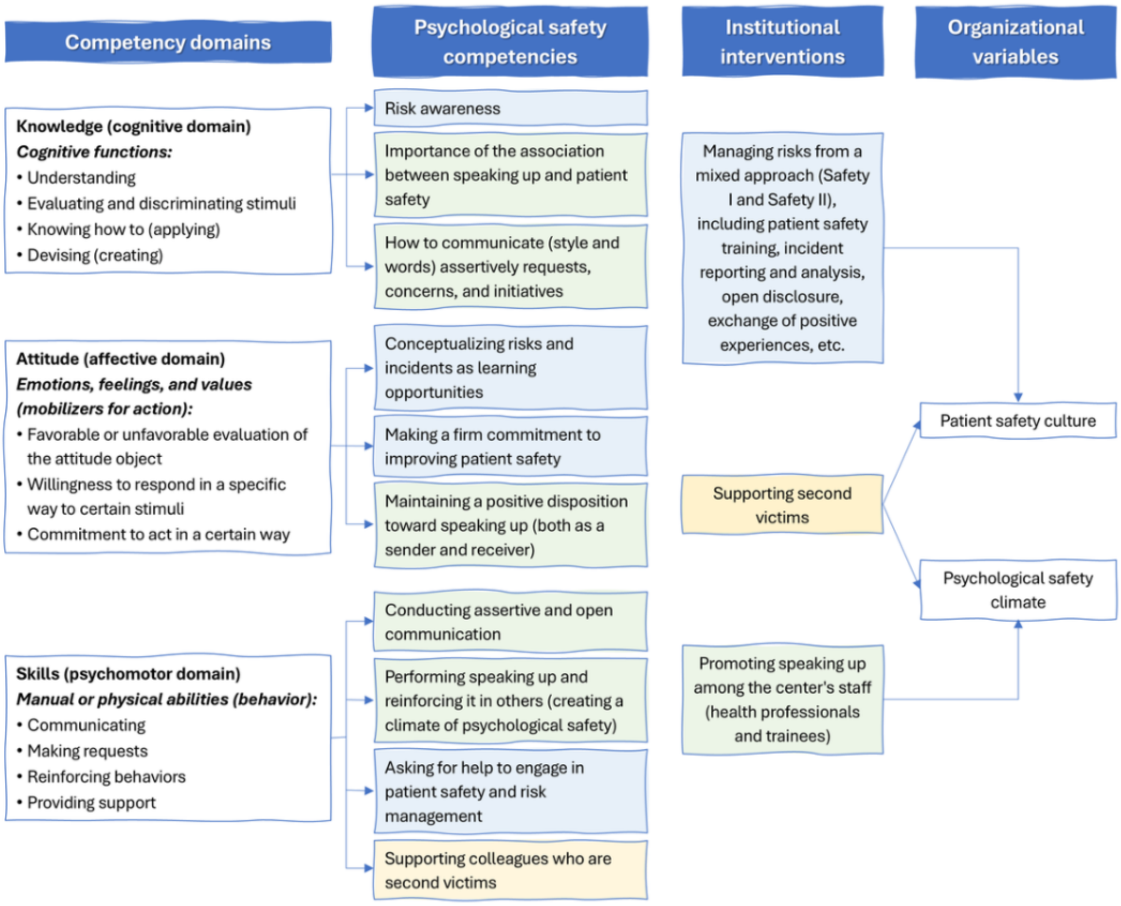
Methodological Framework

In the health care field, competence is defined as "the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice to benefit both individuals and the community" [31]. This study's methodology aligns with the taxonomy by Bloom et al [32], which identifies 3 domains of competencies: cognitive (knowledge), affective (attitudes), and psychomotor (skills), collectively referred to as knowledge, skills, and attitudes or knowledge, attitudes, and skills (KAS) [33].

Figure 1 [32] shows the integration of this study's theoretical background and methodological framework. To assess the professional competence of trainees in psychological safety, they must demonstrate that they know (knowledge), know how to be (attitudes), and know how to do (skills) regarding the topics related to psychological safety described in the theoretical framework. To our knowledge, none of the instruments or frameworks explicitly define competencies in psychological safety. Consequently, this study proposed their measurement based on integrating the KAS model with the variables considered in the instruments available to evaluate speaking up, psychological safety, and support for second victims. Moreover, this approach finds that the culture and climate of the organization play a crucial role in implementing psychological safety competencies. For this reason, we proposed evaluating institutional interventions to promote psychological safety in clinical settings in parallel to assessing these competencies.

XSL•FO
RenderX

Figure 1. Integrated proposal of the study's theoretical background on psychological safety and the methodological framework based on the taxonomy of competencies by Bloom et al [32] and the knowledge, skills, and attitudes model.



Objectives and Hypotheses

The aim of this study was to determine the extent to which health care trainees acquire psychological safety competencies during their internships in clinical settings and identify what measures can be taken to promote their learning.

Our previous hypothesis was that mentors perceive that the level of training and acquisition of psychological safety competencies (as defined in our framework) among European trainees during their internships in clinical settings is low. This hypothesis is supported by the absence of a previous framework that explicitly conceptualizes psychological safety competencies, the relatively low frequency with which trainees speak up on patient safety [34,35], and the generalized lack of formal content on patient safety in the curricula of health care studies in Europe [36]. Regarding measures to promote the learning of these competencies, we expect mentors to identify some of the barriers that currently limit their acquisition by trainees, among which are organizational variables. Therefore, their suggestions will incorporate not only specific initiatives of an educational nature but also others oriented toward cultural forms and patterns of institutional behavior that promote a climate of psychological safety in clinical settings.

Methods

Study Design

We combined quantitative and qualitative research methods and approaches to understand the wideness and depth of the problem of psychological safety training in health care research with mentors of residents and students in health care disciplines from 11 pan-European countries. This study was conducted in 2 phases from February 2022 to July 2023. First, we conducted a web-based consensus conference using a structured questionnaire to prioritize psychological safety competencies and institutional interventions to foster their acquisition among trainees. Subsequently, we performed a second consultation based on an open-ended survey to explore recommended actions to enhance the development of those psychological safety competencies that, according to the results of the consensus conference, were considered significant but the training on them was still lacking or incomplete during internships.

This manuscript was developed in accordance with the Checklist for Reporting Results of Internet E-Surveys [37] (phase 1, survey study) and the COREQ (Consolidated Criteria for Reporting Qualitative Research) [38] (phase 2, qualitative study).

Ethical Considerations

We followed national regulations to obtain ethics approvals in Estonia (Research Ethics Committee of the University of Tartu,

XSL•FO
RenderX

approval 364/T-11; May 16, 2022), Israel (Ethics Committee of the Faculty of Social Welfare and Health Sciences, University of Haifa, approval 036/22; December 21, 2021), Slovakia (Ethics Committee of Pavol Jozef Šafárik University in Košice, approval 11N/2022; March 28, 2022), and Spain (Research Ethics Committee of the San Juan de Alicante University Hospital, approval 22/012; February 23, 2022). In other countries, previous ethics approvals were also considered valid.

Before the participants were registered on the web platform, informed consent was obtained. Likewise, the questionnaire only allowed access to the questions after the participants had explicitly confirmed consent to participate in the study. Participants were given a contact email to exercise their right to withdraw from participation at any time during the study. To allow traceability of the data in the different phases of the study, but at the same time, to guarantee the confidentiality of the participants and their responses, these were deidentified. No form of financial compensation was provided for participation or recruitment.

Recruitment

We invited 256 health care trainee mentors from 11 pan-European countries (20-25 mentors per country), including Croatia, Estonia, Finland, Germany, Israel, Lithuania, Malta, Portugal, Serbia, Slovakia, and Spain, to participate in this study. The invitation to these countries allowed for the collection of information representative of educational models from Northern, Eastern, Southern, and Central Europe. We enrolled professionals assigned to health care institutions (inpatient or outpatient, community, and social care settings) associated with universities or other formal training institutions who were responsible for mentoring and supervising residents and students during their clinical internships in the following areas: family medicine, obstetrics, midwifery, pharmacy, surgery, and other medical fields such as pediatrics. These specialties were chosen based on the main rotation areas of the training and residency programs. Academic-only mentors were excluded. Participants were recruited through convenience sampling by 1 or 2 members of the European Researchers' Network Working on Second Victims (ERNST) Consortium (European Cooperation in Science and Technology Action 19113) [39] from each participating country, who acted as local coordinators. The ERNST includes 28 European countries, integrating experienced research teams focused on patient safety issues, with most working in clinical and academic settings. This network was the vehicle for coordinating the study but not an inclusion criterion for recruiting participants. Thus, all national coordinators were network members, but not all participants involved in the study belonged to ERNST. Only the countries in this European consortium (11/28, 39%) that voluntarily decided to join the study were involved.

In each participating country, trainee mentors were contacted by the national study coordinator through an invitation letter providing information about the study and its objective, content, and procedure. Those participants who agreed to be involved in the study voluntarily completed the questionnaire or survey used for data collection depending on the study phase.

Phase 1: Degree of Acquisition or Implementation and Significance of Psychological Safety Competencies and Institutional Interventions (Consensus Conference)

Procedure

For the first phase of the study, consisting of a web-based consensus conference, we developed an ad hoc questionnaire whose content was structured into 4 blocks of items, the first 3 describing KAS related to psychological safety competencies and the last one describing institutional actions to promote the acquisition of these competencies among trainees. For the development of the questionnaire items, we relied on the validated instrument by Richard et al [40] on psychological safety and speaking up behavior and on those by Lee et al [41] and Schnall et al [42], which propose the assessment of patient safety competencies based on the measurement of KAS that make up each core competency.

The national coordinators from Croatia, Estonia, Finland, Germany, Israel, Portugal, Serbia, Slovakia, and Spain were responsible for the translation and back translation process of the English questionnaire into the language of their country (Multimedia Appendix 1). Participants from Malta and Lithuania completed the English questionnaire.

Before the consensus conference, the questionnaire's readability and face validity were tested between February 2022 and March 2022. In total, 3 mentors per country completed the questionnaire. Given the readability test results, several changes to the questionnaire were made, including rewording some items, reducing the response scale levels, deleting 1 original item, and adding 2 new items for clarity. The overall evaluation of the questionnaire (scale of 0-5) was positive in terms of satisfaction (mean 4.0, SD 0.9), usefulness of the information (mean 4.6, SD 0.7), and usefulness for curricular improvement (mean 4.4, SD 0.6) and somewhat less favorable in terms of length (mean 3.6, SD 0.7).

The final version of the questionnaire consisted of 29 items grouped into the 4 original blocks (Textbox 1). Mentors were asked to assess the degree of acquisition among the trainees of their institution of knowledge (7 items), attitudes (6 items), and skills (7 items) in psychological safety and their degree of significance from their point of view using a 5-point Likert response scale (1=no acquisition at all or not important at all; 5=fully acquired or very important). For institutional interventions (9 items), participants were asked to determine the degree of implementation and significance using a 5-point Likert response scale (1=not yet implemented or not important at all; 5=fully implemented or very important). In the questionnaire instructions, we provided respondents with the definition of the psychological safety concept. We also added a brief conceptual clarification in those items that referred to other secondary terms (such as *second victim*). The questionnaire was distributed via email and using a web-based platform owned by the research group for conducting opinion studies (password-protected survey) between October 2022 and December 2022. A total of 3 reminders were sent during this period to ensure an adequate response rate.

XSL•FO
RenderX

Textbox 1. Final version of the questionnaire to evaluate the degree of acquisition or implementation and significance of psychological safety competencies and institutional actions based on the knowledge, skills, and attitudes framework (phase 1).

Knowledge—in my opinion, internships in my work environment provide trainees with the competency to:

- Understand that an open and direct expression of concerns about patient safety can prevent the occurrence of incidents that could cause harm to the patient.
- Know how to communicate assertively a concern about patient safety to another health care professional (of the same level or higher; eg, what words to choose, how to start and finish the conversation, and what tone of voice or gestures to use).
- Distinguish between situations that could cause avoidable harm to the patient and those that do not represent a high risk for patient safety.
- Choose the best moment to communicate specific concerns about patient safety to another health care professional (of the same level or higher).
- Know how to assertively warn another health care professional (of the same level or higher) of the risk of ignoring an important patient safety rule (eg, words to choose, how to start and finish the conversation, and what tone of voice or gestures to use).
- Know how to deal constructively with the possible negative reaction of a health care professional (of the same level or higher) after having warned them that they were overlooking an important rule for patient safety.
- Know how to express specific proposals that could improve the patient safety in the unit.

Attitude—in my opinion, internships in my work environment provide trainees with the competency to:

- Commit to the identification and prevention of risks for patient safety.
- Perceive risk situations in daily work as an opportunity to highlight the risk and take appropriate measures to prevent harm to patients.
- Respond positively to the expression of warnings or concerns by other health care professionals (of the same level or higher) in relation to patient safety.
- Maintain a positive attitude toward warning other health care professionals if, with their actions, they are ignoring an important patient safety rule.
- Be willing to openly and directly share specific proposals to improve patient safety.
- Be willing to learn from mistakes and patient safety incidents in which other professionals have been involved instead of judging them.

Skills—in my opinion, internships in my work environment provide trainees with the competency to:

- Communicate openly and directly to other professionals (of the same level or higher) specific concerns about patient safety by presenting information, asking questions, or expressing opinions.
- Request the responsible professionals' advice to report, in the appropriate system, the occurrence of a patient safety incident that has been witnessed and make the report (if necessary).
- Warn assertively another health care professional (of the same level or higher) that, with their actions, they are ignoring an important patient safety rule.
- Respond assertively to the negative reaction of a health care professional (of the same level or higher) whom they have warned about ignoring an important patient safety rule.
- Verbally support and reinforce the initiative of other health care professionals (of the same level or higher) to share their specific concerns about patient safety with the rest of the team.
- Set and communicate concrete proposals to improve patient safety in their own unit or service.
- Offer peer support to a colleague involved in an adverse event to reduce the second victim syndrome (characterized by feelings of guilt, inadequacy, anxiety, shame, hypervigilance, or grief).

Interventions—my health care institution:

- Implements a training program for new staff (especially trainees) to foster a positive patient safety culture and a psychological safety climate.
- Appoints an influential group of people to design an intervention plan to foster a trusting climate among health care professionals to ensure patient safety.
- Holds regular clinical sessions with trainees to share patient safety concerns and lessons learned. This measure translates into the set of shared spaces to exchange experiences on patient safety incidents, devise barriers to minimize risks, and provide emotional and instrumental support among peers.
- Raises awareness among the center's professionals, with the collaboration of heads of service, of the need to encourage trainees and colleagues to express their concerns regarding patient safety openly and directly and warn other professionals of the risks they identify in their daily work.
- Raises awareness among the center's professionals, with the collaboration of heads of service, of the importance of responding positively to warnings from other professionals regarding compliance with relevant patient safety rules and reinforcing the open expression of specific patient safety concerns by trainees.
-

XSL•FO
RenderX

- Provides trainees with the opportunity to participate as observers during the planning of adverse event disclosure conversations with the affected patient and family.
- Allows trainees to have the opportunity to be present during the discussion and analysis following a patient safety incident.
- Provides trainees with specific training on reporting patient safety incidents through appropriate means.
- Offers institutional support to health care professionals involved in an adverse event to contribute to better safety at the workplace.

Statistical Analysis

For each item, the mean, SD, and coefficient of variability (CV) were obtained. We considered that those competencies with a score of <3.4 in acquisition required more effort to teach during trainees' internships. These areas were selected to explore possible recommended actions for fostering the acquisition of psychological safety competencies among health care students and residents in a second consultation.

Phase 2: Exploration of Measures and Actions to Promote the Acquisition of Psychological Safety Competencies Among Trainees (Open-Ended Survey Questions)

Procedure

On the basis of the results of the web-based consensus conference, a survey of 3 open-ended questions was developed to explore recommended actions to foster the learning of the psychological safety competencies necessary for trainees to be able to communicate their patient safety concerns or initiatives, observe another health care professional ignoring an important patient safety rule and assertively warn them of the risks of their behavior, and support a colleague who is emotionally affected after being involved in an adverse event (Multimedia Appendix 2). The survey was administered on the web between January 2023 and July 2023 to mentors who had participated in the consensus conference.

Information Categorization and Analysis

For the processing of the information obtained in this second phase, we used the qualitative methodology of content analysis, which consists of systematically transforming a large amount of text into a highly organized and concise summary of the key findings through a process of abstracting the data in consecutive steps to move from manifest and literal content to latent meanings [43]. In total, 2 researchers were involved in the coding process. The first step consisted of reading and rereading the survey data to obtain a sense of the whole and gain a general understanding of what the participants were referring to in their responses. Then, the raw meaning units or ideas were coded into categories agreed upon by the 2 researchers. In a second round, the researchers grouped these categories into themes. Thus, the coding hierarchy used, from least to most abstract, was meaning units, categories, and themes. The coding of the raw data was carried out independently for each of the questions as, although they were related, they focused on different competencies and problem situations. The formation of themes was carried out through an iterative process and was determined by the productivity of ideas and the weight of categories in the set of each separate question. Thus, the same topic could be represented in one question by a single theme and in another by 2. However, once all the ideas per question had been coded, a consistency analysis was carried out to prioritize the themes for the study as a whole and establish interquestion consistency. For each category and theme, we specified the spontaneity (understood as the number of times that the same idea was repeated independently by more than one participant) and the number of countries whose participants agreed on the same idea. We also calculated the overall productivity of ideas and the relative importance of the different themes identified within each question.

Results

Phase 1: Degree of Acquisition or Implementation and Significance of Psychological Safety Competencies and Institutional Interventions (Consensus Conference)

A total of 173 mentors (response rate: 173/256, 67.6%) from 11 countries in the pan-European environment participated in the study. Table 1 shows the participants' sociodemographic characteristics and other variables of interest for the study.

XSL•FO
RenderX

Table 1. Characteristics of participating mentors and their work centers (N=173).

	Values
Country, n (%)	
Croatia	15 (8.7)
Estonia	24 (13.9)
Finland	12 (6.9)
Germany	18 (10.4)
Israel	4 (2.3)
Lithuania	17 (9.8)
Malta	10 (5.8)
Portugal	15 (8.7)
Serbia	15 (8.7)
Slovakia	20 (11.6)
Spain	23 (13.3)
Sex, n (%)	
Female	129 (74.6)
Male	44 (25.4)
Age (y), mean (SD)	43.6 (9.9)
Years being responsible for trainees, mean (SD)	10.7 (7.5)
Number of personally supervised or mentored trainees in the last 3 years (2019-2021), mean (SD)	23.1 (45.4)
Professional profile, n (%)	
Medicine	72 (41.6)
Nursing	60 (34.7)
Pharmacy	24 (13.9)
Midwifery	4 (2.3)
Physiotherapy	3 (1.7)
Psychology	4 (2.3)
Dentistry	1 (0.6)
Microbiology	1 (0.6)
Occupational health care	1 (0.6)
Public health and organization	1 (0.6)
Radiology	2 (1.2)
Setting of clinical and mentoring performance, n (%)	
Primary care	37 (21.4)
Specialized care (hospital)	132 (76.3)
Social care	4 (2.3)
Specific patient safety training program in place at the center, n (%)	
Yes	48 (27.7)
No	125 (72.3)

Regarding the degree of acquisition of psychological safety competencies among the trainees during their internships, the mean scores assigned by the mentors ranged from 2.9 to 3.8 points, suggesting a medium to low level of assimilation (global mean 3.4, SD 0.8; knowledge mean 3.4, SD 0.8; attitude mean 3.5, SD 0.8; skills mean 3.3, SD 1.0). Concerning knowledge, mentors believed that trainees had difficulties in knowing how to assertively warn another health care professional (of the same or a higher level) of the risk of ignoring an important patient safety rule (mean 3.1, SD 1.0; CV=33.1) and how to deal constructively with their possible bad reaction (mean 2.9, SD 1.1; CV=38.6). Mentors perceived among their trainees a

XSL•FO
RenderX

medium level of development of attitudes favorable to creating a psychological safety climate. The aforementioned knowledge gap was expected to be reflected in the trainees' lack of skills to warn others about unsafe practices (mean 3.2, SD 1.1; CV=35.6) and manage the possible interpersonal conflict arising from such verbalization (mean 3.1, SD 1.2; CV=37.8). In the mentors' opinion, another of the least trained skills was offering peer support to a colleague involved in an adverse event to prevent or minimize the second victim response (mean 3.1, SD 1.3; CV=41.5).

Without being high, the most widespread knowledge among trainees in the eyes of their mentors was the understanding that open and direct communication of patient safety concerns could prevent adverse events (mean 3.8, SD 1.0; CV=25.0) and the distinction between situations that could cause avoidable harm to the patient and those that pose no risk (mean 3.7, SD 0.8; CV=22.3). Consequently, mentors noted that the most prevalent attitude among residents and students was the commitment to identify and prevent patient safety risks (mean 3.8, SD 1.0; CV=25.3). Of all the skills explored, consulting and sharing patient safety concerns obtained the highest score from mentors, although not reflecting a high acquisition level (mean 3.5, SD 1.1; CV=30.5).

In general, the significance assigned by mentors to the different psychological safety competencies was high, with mean scores ranging from 4.4 to 4.8 points (global mean 4.6, SD 0.6; knowledge mean 4.5, SD 0.6; attitude mean 4.6, SD 0.6; skills mean 4.6, SD 0.7). According to the mentors' opinions, the most relevant components of psychological safety competencies included understanding the importance of openly communicating patient safety concerns to prevent adverse events (mean 4.8, SD 0.6; CV=25.0), adopting a firm commitment to identify and proactively manage patient safety risks (mean 4.7, SD 0.7; CV=14.9), showing a willingness to learn from safety incidents involving other professionals in a nonjudgmental manner (mean 4.8, SD 0.6; CV=13.2), and openly communicating one's patient safety concerns regardless of the hierarchical level of the recipient (mean 4.7, SD 0.7; CV=15.0).

According to the mentors, the level of implementation of actions aimed at promoting the acquisition of psychological safety competencies among trainees in their institutions was low (mean 2.6, SD 1.1; range 2.4-2.8). Despite its perceived importance, mentors indicated that the least implemented action was holding regular clinical sessions with trainees to share patient safety concerns and lessons learned (mean 2.4, SD 1.3; CV=56.9). Overall, mentors rated the 9 institutional interventions explored in the questionnaire as highly significant (mean 4.4, SD 0.8; range 4.2-4.6). Among the interventions assessed as most significant were implementing a patient safety and psychological safety training program for new trainees (mean 4.6, SD 0.9; CV=19.0) and offering institutional support to trainees involved in an adverse event (mean 4.6, SD 0.9; CV=20.0).

The results of the descriptive analysis (mean, SD, and CV) for each questionnaire item are shown in the tables in Multimedia Appendix 3.

Phase 2: Exploration of Measures and Actions to Promote the Acquisition of Psychological Safety Competencies Among Trainees (Open-Ended Survey Questions)

Overview

In this second consultation, 36.4% (63/173) of the mentors who completed the phase 1 questionnaire participated. The participating countries were Croatia, Estonia, Germany, Portugal, Serbia, Slovakia, and Spain. The overall productivity was 353 ideas or meaning units grouped into 9 joint themes. More than half (210/353, 59.5%) were related to training activities, environmental and structural conditions (safety culture and work climate), and individual attitudes. The rest of the proposals referred to the trainees' reference person, institutional resources to support the second victim, curricula content and training pathways, systems and mechanisms to give trainees a voice, institutional management of clinical risks (incident reporting and analysis), regulations and standards, supervision, and resources to support trainees (Table 2).

XSL•FO
RenderX

Table 2. Productivity and distribution of meaning units by theme overall and per question (N=353).

Theme	Spontaneity (relative priority), n (%)			
	Question 1: communicate patient safety concerns (n=154)	Question 2: warn about an unsafe behavior (n=101)	Question 3: support a second victim (n=98)	Total
Training	46 (29.9)	52 (51.5)	29 (29.6)	127 (36)
Individual attitudes and environmental conditioning factors[a]	35 (22.7)	23 (22.8)	25 (25.5)	83 (23.5)
Person of reference	19 (12.3)	6 (5.9)	15 (15.3)	40 (11.3)
Institutional resources to support second victims	—[b]	—	27 (27.6)	27 (7.6)
Curricula	16 (10.4)	7 (6.9)	—	23 (6.5)
Systems and mechanisms to give a voice to trainees	20 (13)	—	—	20 (5.7)
Institutional risk management, patient safety reporting systems, and incident analysis	8 (5.2)	5 (5)	—	13 (3.7)
Regulations, guidelines, protocols, standards, and policies	10 (6.5)	—	2 (2)	12 (3.4)
Supervision and resources to support trainees	—	8 (7.9)	—	8 (2.3)

[a]In question 1, this theme comprises themes 1.2—safety culture (spontaneity: n=26) and 1.7—organizational structure and culture (spontaneity: n=9).

[b]This theme did not emerge among the ideas proposed by the participants in response to this question.

Independent analysis of each of the questions yielded an individual productivity of 154 ideas for the first question (communication of patient safety concerns or initiatives) classified into 8 themes and 26 categories, 101 ideas for the second question (observing another health care professional ignoring an important patient safety rule and assertively warning them about the risks of their behavior) classified into 6 themes and 22 categories, and 98 ideas for the third question (offer support to a colleague suffering emotionally after being involved in an adverse event) classified into 5 themes and 16 categories. Multimedia Appendix 4 presents a summary figure with the main results of the qualitative analysis and a set of tables with the coding of the ideas per question with the specification of spontaneity, the number of participating countries per theme and category, and some examples of meaning units.

The following is a brief presentation of the themes that emerged from the analysis of the information provided by the mentors in this second phase of the study. With a practical vision and purpose, the relevant information was transformed into recommendations for fostering the learning of psychological safety competencies among trainees in clinical settings (Textbox 2).

XSL•FO
RenderX

Textbox 2. Recommendations for fostering the learning of psychological safety competencies among trainees in clinical settings extracted from the information provided by mentors in the study's second phase.

Training

- What:
  - Focus on patient safety, interpersonal communication, assertiveness, active listening, and social support skills.
  - Include emotional intelligence, critical thinking, reflective skills, argument formation, and achievement recognition.
  - Train in clinical interviewing skills, using open and closed questions and standardized questionnaires.
- How:
  - Implement innovative strategies such as seminars, lectures, workshops, simulation learning, role-playing, case studies, videos, panel discussions, clinical sessions, and debriefings.
  - Provide ongoing training tailored to the health care context rather than one-off sessions.
  - Practice these skills in real-world scenarios, including daily professional exchanges, challenging conversations, and conflict resolution.
  - Address communication in the context of patient safety events to prevent risks and ensure safety.
- For whom:
  - Sensitize and train the entire health care organization, including teams, mentors, and management.
  - Implement top-down training strategies, starting with senior management. Ensure that managers and team leaders receive regular compulsory training in patient safety, leadership, and staff management.
- Who:
  - Use psychologists for communication and interpersonal support skill training.
  - Engage patient safety experts for specialized training.
  - Supervision by mentors should cover communication skills, interactions with professionals, patient safety information collection, and reflection on clinical risks.
- When:
  - Begin introductory training in patient safety, communication, and teamwork skills upon joining the center.
  - Ensure comprehensive and continuous training throughout the professional career regardless of professional profile to maintain a shared understanding among all team members.

Individual attitudes and environmental conditioning factors

- *Supportive, respectful, and trustful climate:*
  - Leaders and middle managers should foster positive relationships, teamwork, openness, trust, honest communication, and mutual support.
  - Inclusive and ethical leadership helps integrate trainees into work teams, allowing them to ask questions without fear and express views assertively.
  - Organize social and informal activities to improve team relationships.
- *Positive attitudes toward patient safety and peer support:*
  - Cultivate respect and equality among professionals to facilitate the understanding that everyone's contributions are relevant for quality and safe care.
  - Foster a positive attitude toward seeking and providing peer support in challenging situations, addressing the second victim phenomenon.
- *Just and nonpunitive safety culture:*
  - Move away from a punitive and blame culture; adopt a systemic approach to adverse events and honest mistakes to prevent recurrences.
  - Recognize human fallibility and the multifactorial origins of patient safety incidents.
  - Encourage open communication without fear of punishment, judgment, stigma, or rejection.
  - Middle managers, mentors, supervisors, and senior professionals must act as role models and change agents, promoting a just safety culture.
  - Launch internal campaigns to create a just culture and make patient safety a regular topic of conversation.
- *Open and constructive communication channels:*

XSL•FO
RenderX

- Promote assertive expression and acceptance of constructive criticism as learning opportunities.
- Establish direct communication channels between all team members.

- *Leadership and management involvement:*
  - Leaders and managers should actively seek feedback and suggestions before initiating procedures.
  - Encourage managers and superiors to share personal experiences with adverse events to demonstrate that these issues are a reality in clinical practice.
  - Appoint an institutional patient safety officer and ensure that trainees are aware of their presence.
- *Structural measures to enhance patient safety:*
  - Promote a culture in which all team members regardless of hierarchy are encouraged to contribute to patient safety.
  - Use cocreated checklists to standardize practice and establish a common language across different roles.

A reference person for trainees

- *Role and responsibilities of mentors:*
  - Mentors are ideal figures to establish trust and promote patient safety supported by institutional resources.
  - Act as role models, explaining safety rules, managing risks, and using tools (eg, reporting systems).
  - Define standard behaviors in critical scenarios: safety concerns, adverse events, mistakes, noncompliance with safety rules, unsafe acts, and emotional distress.
- *Support and accessibility:*
  - Mentors should be approachable, providing support, information, and guidance.
  - Link trainees with their teams and institutional resources, guiding interactions with other professionals.
  - Support trainees as peer supporters during emotional distress from incidents.
- *Training and resources:*
  - Mentors need specialized training, resources, and mechanisms for supervising trainees.
  - Legislation may be needed to create specific supervisor positions in medical institutions.

Institutional resources to support the second victim

- *Support program and format:*
  - Raise awareness and train professionals to act as peer supporters providing emotional first aid through individual meetings or confidential group debriefings.
  - Focus on active listening, emotional validation, positive language, empathy, and companionship without investigating the event.
  - Establish multidisciplinary peer teams with personal experience of such events for a sense of identification.
  - Organize Balint groups for critical and self-reflective discussions on emotional reactions to challenging situations facilitated by a psychotherapist [44].
  - Ensure availability of psychological support through mental health, occupational health, or risk prevention structures.
  - Provide clear referral channels to connect grieving trainees with resources, with mentors acting as bridges.
- *Institutional programs for staff well-being:*
  - Implement programs promoting employee well-being and personal development.
  - Host social and recreational activities such as service dinners and hiking.

Curricula

- *Incorporate patient safety into education:*
  - Integrate patient safety and communication skill training into health care degree programs, postgraduate studies, specialization pathways, and doctoral studies.
  - Foster cooperation between academia and health care training institutions for comprehensive competency development.

XSL•FO
RenderX

- *Guidelines and regulations:*
  - Establish European and national guidelines to systematically regulate and implement patient safety training in all health faculties in alignment with World Health Organization recommendations [29].
- *Assessment and certification:*
  - Include patient safety competency assessment, including communication skills, in state and certification tests using methodologies such as the objective structured clinical examination.

Systems and mechanisms to give trainees a voice (highlights)

- *Create common spaces or forums:*
  - Hold regular meetings to discuss patient safety cases, significant issues, and preventive initiatives.
  - Involve all relevant parties: supervisors, mentors, professors, residents, and students.
  - Use forums for supervisors to share personal experiences and learning.
- *Encourage patient safety ownership:*
  - Foster ownership and meaning in trainees' work to encourage communication on patient safety.
  - Involve trainees in supportive supervision roles and patient safety projects.
- *Collect trainee feedback:*
  - Conduct regular one-to-one mentor-trainee interviews.
  - Create an anonymous mailbox for concerns and suggestions.
  - Maintain a trainee diary to document experiences and feedback.

Institutional risk management, patient safety reporting systems, and incident analysis

- *Anonymous reporting systems:*
  - Create systems for anonymously reporting adverse events.
  - Focus on proposing improvement plans with corrective, not punitive, measures.
  - Ensure legal certainty for reporters to mitigate fear of legal repercussions.
- *Trainee involvement in incident analysis:*
  - Allow trainees to be participant observers in the incident analysis process.
  - Ensure that trainees receive feedback on their reports and the resulting preventive and corrective measures.
- *Debriefing and seminars:*
  - Conduct joint and individual debriefings or seminars for trainees involved in critical incidents to review what happened.

Regulations, guidelines, protocols, standards, and policies

- *Institutional level:*
  - Ensure the availability of standardized protocols for safe clinical procedures, emphasizing risk detection and prevention measures.
  - Establish formal channels and communication procedures to enhance patient safety.
  - Structure patient safety rules hierarchically: elemental and mandatory for all professionals and complementary for specific procedures or situations.
  - Reinforce the center's quality and safety policy supporting the health care quality unit.
- *National level:*
  - Advocate for national legislative changes to ensure legal protection for professionals and trainees involved in patient safety processes (eg, incident reporting and open disclosure).
  - Develop and implement accreditation standards for certifying health care institutions as patient safety promoters in teaching and training.

Supervision and resources to support trainees' learning process

XSL•FO
RenderX

- Conduct regular interviews, knowledge checks, and feedback sessions during joint practice.
- Use clinical scenarios to prepare and execute procedures under safe conditions.
- Encourage trainees to detect errors and practice providing feedback.

Training

Most of the proposals made by the mentors to encourage trainees to acquire psychological safety competencies and develop skills in communicating patient safety concerns, warning a professional that their behavior compromises patient safety, and supporting a colleague who is a second victim were related to training actions.

The proposed ideas focused on the content, target audience, teaching methodologies, career stage, and parties involved in training.

Individual Attitudes and Environmental Conditioning Factors (Organizational Structure, Safety Culture, and Work Climate) That Determine Behavioral Patterns

The acquisition of psychological safety competencies by trainees and their commitment to patient safety seems to have a circular and bidirectional relationship with a positive safety culture and a work climate based on trust and respect in the institution. Most mentors identified these 2 contextual factors as prerequisites for trainees to openly discuss adverse events and honest mistakes and support each other.

A Reference Person for Trainees

In environments as changing, novel, and challenging as clinical settings, mentors mentioned the need for trainees to have a permanent reference person with whom they can establish a trusting bond from the moment they join the health care center.

Institutional Resources to Support the Second Victim

Concerning the approach to the second victim phenomenon, the mentors' contributions referred to implementing institutional resources and programs to minimize the impact of adverse events on health care professionals, including trainees. Most of the proposals were along the lines of the reference programs based on the Scott Three-Tiered Interventional Model of Second Victim Support [45].

Curricula

The mentors agreed on the need to incorporate a subject on patient safety in the curricula of health care degree programs.

Systems and Mechanisms to Give Trainees a Voice

The mentors made several proposals to facilitate bottom-up communication and the expression of patient safety concerns and initiatives among trainees.

Institutional Risk Management, Patient Safety Reporting Systems, and Incident Analysis

The mentors agreed on the appropriateness of involving trainees in adverse event reporting and analysis to promote awareness of health care risks and the adoption of a clinical practice style committed to patient safety.

Regulations, Guidelines, Protocols, Standards, and Policies

As additional measures, the mentors highlighted the need to reinforce the institutions' patient safety through regulations, protocols, and policies at the institutional (meso) and national (macro) levels.

Supervision and Resources to Support Trainees' Learning Process

Mentors attached particular importance to supervision as a mechanism for training students and residents in psychological safety competencies.

Discussion

Principal Findings

Our study's first objective was to find out whether health care trainees in a pan-European environment acquire, from the point of view of their mentors, competencies in psychological safety during their internships in clinical settings. This study's results show that, according to the mentors, the competency acquisition level is moderate to low, so in their opinion, the training currently offered in the pan-European context does not guarantee the systematic acquisition of the competencies needed to foster a psychological safety climate in health care institutions. A total of 40% (8/20) of the competency elements analyzed presented a low acquisition value, being more pronounced in the psychomotor domain (skills). Such lack of KAS reported by mentors can prevent residents and health care students from engaging in challenging conversations in clinical settings (eg, warning a senior professional that their behavior poses a risk to the patient, communicating concerns and initiatives, or asking questions related to patient safety) or supporting a colleague who is suffering emotionally after their involvement in an event that caused or could have caused harm to a patient.

A second question not initially raised explicitly in our study but that emerged from the first-phase findings and connects them to the second objective concerns the reasons for the low acquisition of psychological safety competencies among pan-European health care trainees. In the opinion of the mentors, the optimal development of these competencies is hampered by deficiencies in formal patient safety content in health care curricula and training pathways, modeling by trainers, trust, cohesion, team spirit (safe people and safe environments), decentralization of the health care institutions' structure, specific patient safety policies, and institutional resources to support the creation of a psychological safety climate and a proactive patient safety culture.

In response to what needs to be done to promote the training and learning of psychological safety competencies, the mentors offered an extensive list of wide-ranging measures and recommendations to address the deficiencies identified. These

XSL•FO
RenderX

proposals ranged from specific practices, methodologies, models, and content for training in psychological safety competencies to higher-level measures related to the organization's structural, strategic, cultural, and environmental aspects.

These findings support a conceptualization of psychological safety competencies based on the integration of purely academic and formative actions with others of an institutional nature aimed at fostering cultural patterns and work climates that encourage the manifestation of these competencies. In their proposals, the mentors emphasized that both pillars are essential for trainees to show optimal performance in psychological safety. In this sense, an exclusively academic approach is doomed to failure as, if the health care environment does not allow for the implementation of what has been learned, competence acquisition will only be possible in the cognitive and affective domains but not the psychomotor one. The framework that we present in this study for measuring competencies in psychological safety is, to our knowledge, the first to specify what a competent trainee in psychological safety would look like, exemplifying not only what they should know and feel (aspects that are not directly observable) but also what they should do and what others can observe to evaluate their competence.

Comparison With Prior Work

The perception of the mentors in our study is congruent with the levels of psychological safety reported by pediatric nurses and residents in the American context (mean of 3.4 points on a scale of 1-5) [46]. They identified the following as the main barriers to a psychological safety climate in the team: difficulties in interpersonal relationships between professionals of different disciplines and statuses, unsatisfactory communication style and frequency, inadequate resolution of disagreements, work overload accompanied by lack of collaboration, and interpersonal disrespect. The mentors in our study addressed some of these issues in their proposal for measures to promote psychological safety competency training.

As our study suggests, psychological safety and patient safety are closely linked elements of the clinical environment and practice. The mentors who participated in this study's second phase related the trainees' low competence in psychological safety to the lack of formal incorporation of patient safety into the curricula of health care degrees. This result is in line with the findings of Sánchez-García et al [36] in the pan-European context, which show that there is still a long way to go in adapting the curricula as half of the nursing schools and 60% of the medical schools analyzed did not cover any topics related to patient safety. In those cases in which the curricula did cover patient safety aspects, interpersonal communication, quality of care, and other elementary aspects were the most widespread topics. The second victim phenomenon was formally present in only 1 of the 206 curricula reviewed.

The relationship between psychological safety and patient safety has been extensively studied. Research with nurses has shown that psychological safety predicts the intention to report safety incidents and a greater willingness to engage in open communication, which, in turn, may lead to higher job satisfaction, lower turnover intention, and improved patient safety [47,48]. In this line, Dietl et al [25] observed that the positive effect of psychological safety on patient safety is mediated by interpersonal communication in the team. Therefore, not surprisingly, training in communication skills and teamwork was one of the targets for action identified by the mentors who participated in our study.

Other studies have also suggested that this relationship may run in the opposite direction (ie, a positive safety culture in the health care institution may contribute to developing a psychologically safe work climate). Along these lines, O'Donovan and McAuliffe [49] identified the enablers of psychological safety across the individual, team, and organizational levels of health care institutions, including safety culture and continuous improvement. Our mentors provided solutions related to some of these facilitators. On an individual level, participants highlighted the role of mentors and supervisors in fostering professional responsibility among trainees, recognizing them as full team members. This aspect relates at the team level to the factors of leader behavioral integrity and status, hierarchy, and inclusiveness that our mentors recommend fostering by creating shared spaces free of hierarchical differences. For this purpose, Ulmer et al [50] present the "Mistake Of The Week" initiative, which consists of a weekly semistructured conference in which health care staff are encouraged to voluntarily disclose their mistakes and near misses based on 4 pillars of success, namely, exemplification, fixed time slots and clearly defined dynamics, absence of fear of punishment, and trusting atmosphere. This initiative recalls the morbidity and mortality conferences (M&Ms) widely implemented in North America and included in US residency programs. These conferences aim to critically analyze and discuss safety incidents in a safe environment. When these conferences are cross-cutting, they allow for the participation of different institutional agents in the discussion of incidents, including from students and junior staff to senior leaders or administrators. This model increases the likelihood that M&Ms will become a tool for system-wide improvement [51]. Although presented as an opportunity for learning and improving the quality of care, their inappropriate use can also lead to undesirable results [52]. In Europe, although less widespread, M&Ms are present in surgical services, intensive care units, and emergency departments in some countries, such as Germany, France, the Netherlands, and the United Kingdom [53-56]. Although stakeholders value the M&Ms' implementation positively, there remains a lack of objective outcome measures to determine their impact on patient safety and system-wide improvement and some challenges that jeopardize the effectiveness of these conferences [57]. Although expectations are clear (focus on education and quality improvement, lack of blame, being mandatory for residents and attendings, and orientation toward changes in clinical practices), excessive heterogeneity and lack of structure often limit their impact [53,55]. The European experience suggests better results when M&Ms are interprofessional; incorporate a moderator; are supported by a quality committee; are incorporated as part of the Plan, Do, Check, Act cycle; and include the use of validated instruments for collecting data on complications [53,55-57]. As occurs in the case of psychological safety

XSL•FO
RenderX

competency training, clinician engagement, patient safety culture, and organizational governance and leadership are identified as contributors to effective M&Ms [52]. Facilitators regarding peer support and familiarity of the leader and team members are closely related to the proposals of the mentors, who suggested the formal designation of a reference contact for the trainees with whom they can establish a trusting relationship, the creation of positive team dynamics, and peer support in case of an adverse event. Recently, Seys et al [58] proposed an international multidimensional action plan for second victim support structured in five levels: (1) prevention at the individual health care professional and organizational level, (2) self-care of the health care individual and team, (3) support through peers and triage, (4) structured professional support (eg, mental health or specialized support), and (5) clinical support (pharmacological treatment and long-term psychotherapy). The results of the study by Lyman et al [59] with newly graduated nurses reflect well on how individual factors facilitating psychological safety in clinical settings are built on team experience. Thus, the nurses said that their self-confidence was preserved when team members approached mistakes as opportunities for improvement. Finally, O'Donovan and McAuliffe [49] point out that, when a just safety culture exists in institutions, it is possible to create spaces where trainees and new staff find the courage to speak up. The mentors in our study identified the still widespread existence of "*name-blame-shame*" cultures as a critical barrier to acquiring and implementing psychological safety competencies by any health care team member, especially trainees. In this sense, safety culture is understood as a prerequisite for a psychological safety climate, with culture conceptualized as a more stable element and climate as a more situational and local outcome largely dependent on leadership styles and team dynamics that, in turn, are influenced by the organization's values, norms, and behavioral patterns.

In addition to the attenuation of hierarchies and the establishment of high-quality relationships, McClintock and Fainstad [60] consider other aspects as core features of psychologically safe environments, such as a trainee-driven and flexible learning agenda, the absence of formal assessment, and time for debriefing. In the same direction, the mentors in our study recommended merging supervision and competency assessment with daily clinical practice and creating Balint groups [44] as team debriefings.

Apart from the importance of cultural aspects and the work environment already discussed, the acquisition of psychological safety competencies by trainees requires investing efforts and resources in wide-ranging training actions. According to the categories proposed by O'Donovan and McAuliffe [61] in their systematic review of interventions to improve psychological safety, speaking up, and voice behavior, the proposals of the mentors in our study were based on simulations, video presentations, case studies, workshops, forums, and meetings. The mentors particularly stressed the importance of adopting an organization-wide approach to training on psychological safety and patient safety. To have competent trainees, first, it is necessary to ensure that mentors are trained in patient safety, clinical risk management, communication, leadership, and supervisory skills. Minehart et al [62] implemented an educational intervention to improve the quality of feedback provided by anesthesia teachers. Those who received the training performed better in maintaining a psychologically safe environment and identifying and exploring trainees' performance gaps.

Given this training gap, it is essential to draw up a general plan with specific guidelines to adapt the incorporation of this content into the training plans and specialization pathways of future health care professionals. As the mentors of our study pointed out, more than a decade ago, the World Health Organization [29] published guidelines for incorporating patient safety into curricula. However, national policies have not yet ensured the widespread implementation of these recommendations in pan-European countries, and as of now, only a few isolated universities have taken the initiative on a discretionary basis [36].

Although psychological safety has been widely recognized as part of successive patient safety and quality improvement processes, it remains a relatively unknown construct among many educators and trainees. Consequently, it is often relegated to the hidden curriculum that becomes tangible through mentors' exhibited norms, values, and behaviors. This hidden curriculum can have both positive and negative effects on professional development. The positive effects manifest through empathy, resilience, perseverance, and psychological safety [63]. The way to prevent the negative consequences of the hidden curriculum and enhance psychological safety is to formalize the teaching of those values and norms that support safe practice and a clinical learning environment based on openness, trust, and respect.

This study provides an overview of how psychological safety competencies are being taught in pan-European clinical learning environments from the mentors' perspective. It also highlights the gaps across the board in the competency training of health care trainees. The recommendations proposed by this group seek to reinforce the formal teaching of patient safety and psychological safety in a multifactorial and multilevel manner, including contextual, attitudinal, and educational elements. Psychological safety is fundamental to achieving learning health care organizations and functions as an enabler of the ability of the system and its teams to remain in a continuous improvement cycle that contributes to safer clinical environments for patients [49]. Therefore, efforts should be synergistically directed toward the simultaneous and bidirectional improvement of patient safety and psychological safety. This synergy is already envisaged in the World Health Organization curriculum guide [29] and Global Patient Safety Action Plan 2021 to 2030 [7], in which preserving the psychological safety of health care professionals by preventing harm to their well-being (eg, burnout) is linked to goal 3 of the United Nations Sustainable Development Goals. Furthermore, fostering psychological safety (eg, speaking up and stopping the line) is considered an enabling competency in the Canadian Patient Safety Institute's proposed framework [64] for making patient safety a reality in health care institutions. Along these lines, the Institute for Healthcare Improvement has been working for decades on improving patient safety with a holistic approach, making a wide range of tools, training, and documentation available to health care institutions and

XSL•FO
RenderX

professionals. One of the most noteworthy initiatives is its Certified Professional in Patient Safety credential, which establishes core standards for certifying a proficiency level of professionals in patient safety [65].

The mentors' proposals in this study may be helpful to materialize in concrete actions the global solution that this problem needs. Some of these interventions have shown improvements in psychological safety, speaking up, and voice behavior; however, longitudinal and multifaceted studies to determine their effectiveness are still required [61]. On the other hand, the items of the ad hoc questionnaire used in the first phase of this study can serve as a prescriptive proposal of the KAS to be trained in residents and students for the development of psychological safety competencies and also as an instrument to assess their degree of acquisition.

Our study adds to previous research that supports that acquiring psychological safety and patient safety competencies requires a global and integrated effort that stems from the coordinated involvement of academia and the health care system and extends beyond merely training actions [48,50,61]. Educating those still in the training process on psychological safety means tackling the problem from the ground up with a commitment to the future. It is important to remember that trainees are not solely responsible for creating a psychological safety climate, at least not initially. However, they can still have a significant impact as a driving force for change. The synergistic combination of training and structural measures at the individual, team, and organizational (culture) levels is the key to a psychologically safe environment in clinical settings [48].

Limitations

Despite the merits of this study, it is important to acknowledge a few limitations. Study participants were recruited through convenience sampling. They were selected by national coordinators who are members of the European ERNST Consortium, so their sensitivity to patient safety and psychological safety topics may be higher than that of the average health care trainee mentor in the pan-European context. Therefore, the sample may not have been representative of the study population. Most of the study participants (129/173, 74.6%) were women. Although this large gender discrepancy in the respondent distribution might suggest a possible representation bias, according to data from the European Commission, 78% of health care workers in the third quarter of 2020 were female [66]. Thus, our sample reflects the current picture as far as gender distribution is concerned. Similarly, there may have been discrepancies in how mentors understood psychological safety or their level of familiarity with the topic. On the other hand, we did not control for possible differences between countries in curricula or the structure and functioning of the academic and health care systems. When generalizing these results, it is necessary to consider the impact of international accreditations on medical programs that include curriculum elements focused on patient safety. These aspects may have affected the mentors' experience and familiarity with patient safety, psychological safety, and the second victim phenomenon. For some participants in this study's first phase, the questionnaire items were too specific, which indicated difficulty in discriminating between the core components of competencies (KAS) and may have affected the quality of the responses. In the second phase, the response rate dropped drastically, and only 55% (6/11) of the original countries participated. To prevent the representativeness of the results from being affected by experimental mortality, only those countries that ensured the participation of at least 5 mentors were encouraged to be involved in the second phase of this study. Furthermore, the higher productivity of ideas in the first question of the final survey may be because participants were more familiar with the communication aspects, whereas the issue of second victims and initiatives to address it are less known. Alternatively, there may have been a demotivation effect that caused lower productivity in the later questions. As for the study population, the choice of mentors, although justified, only offers a partial view of the level of acquisition of psychological safety competencies by health care trainees. The vision of the students and residents, who are the protagonists of the learning process, may differ, and future studies should directly survey this group. Finally, although the study offers an extensive and exhaustive list of actions to improve psychological safety and patient safety in clinical settings, the proposal was not prioritized or ordered sequentially, so centers wishing to improve these aspects may find it challenging to decide where to start. The answer to this question can be affected by multiple secondary issues such as resource availability, leadership involvement, or facility and staff resistance. While cultural change at the institutional level is essential to ensure psychologically safe clinical environments, changing values and beliefs takes time and may be more effectively achieved through concrete actions and behavioral adjustments even if these are initially short range. In any case, future research should address prioritizing actions and establishing indicators and compliance standards involving the different stakeholders.

Conclusions

To our knowledge, this study is the first to address how psychological safety competencies are being taught to future health care professionals from the point of view of mentors and with a pan-European scope. Our study results showed a medium to low level of acquisition of psychological safety competencies among health care trainees in a pan-European setting as perceived by their mentors. According to this group, the solution to this competency gap should be comprehensive and consider the following aspects: training in communication skills and patient safety, environmental conditioning factors (safety culture and work climate) and individual attitudes, a reference person for trainees, formal incorporation into the curricula of health care degree programs and specialization pathways, specific systems and mechanisms to give trainees a voice, institutional risk management, regulations, guidelines and standards, supervision, and resources to support trainees.

The results of our study emphasize the importance of taking multiple actions to establish psychological safety in clinical environments. Academia should seek to formally teach psychological safety competencies during formal training by incorporating them into the curriculum and using innovative teaching methodologies based on technological tools and solutions. It should also strengthen communication and

XSL•FO
RenderX

coordination mechanisms with the clinical institutions where trainees perform their internships and maintain a contact person in academia to assist them during the process and supervise their experience. Health care institutions, for their part, should actively promote a just safety culture free of blame and punishment by providing training in patient safety and psychological safety to all members of their staff with the commitment and direct involvement of the management. Mentors should receive specialized education to train trainees in patient safety and promote psychological safety behaviors such as speaking up. They should also supervise the development of these competencies in the trainees under their charge and be their reference person. As for trainees, they should develop from the beginning of their careers a solid commitment to patient safety and a willingness to speak openly about their patient safety concerns and initiatives and stay in a cycle of learning and improvement. They must learn that this commitment is not to themselves but to patients and the delivery of quality care.

The proposal for measures described in this study aims to facilitate the translation of international guidelines into practice and clinical settings in the pan-European context. Further research on the combined effectiveness of these measures is needed to achieve competent trainees and health care professionals in psychological safety and patient safety. Psychological safety is critical in creating learning health care organizations and safer clinical environments for patients.

Acknowledgments

The authors thank all health care trainee mentors who contributed to the study, without whom this research would not have been feasible. This article is based upon work from COST Action, TheERNSTGroup, CA19113, supported by COST (European Cooperation in Science and Technology). During the conduct of this study, JJM enjoyed a research activity intensification contract funded by the Carlos III Health Institute (reference INT22/00012).

Authors' Contributions

JJM conceived and designed the study, and IC coordinated it. IC and JJM developed the questionnaire and survey for data collection, which were revised and improved by the rest of the authors. IS, IB, VK, SGP, BK, AJ, DJ, ST SCB, ES, AMG, KP, RS, PS, and MO acted as local coordinators in their respective countries and were responsible for the processes of translation and back translation of the instruments into their native languages, study dissemination, and recruitment of participants. IC and JJM conducted the data analysis and prepared a manuscript draft. IS, IB, VK, SGP, BK, AJ, DJ, ST SCB, ES, AMG, KP, RS, PS, and MO revised and improved the manuscript. All authors read and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire used in the first phase of the study (web-based consensus conference).
[DOCX File , 377 KB - mededu_v10i1e64125_app1.docx]

Multimedia Appendix 2
Open-ended survey used in the second phase of the study.
[DOCX File , 897 KB - mededu_v10i1e64125_app2.docx]

Multimedia Appendix 3
Descriptive analysis (mean, SD, and coefficient of variability) of the degree of acquisition or implementation and significance of psychological safety competencies and institutional interventions (consensus conference).
[DOCX File , 25 KB - mededu_v10i1e64125_app3.docx]

Multimedia Appendix 4
Content analysis: themes, categories, and examples of meaning units by core questions.
[DOCX File , 511 KB - mededu_v10i1e64125_app4.docx]

References

1. Panagioti M, Khan K, Keers RN, Abuzour A, Phipps D, Kontopantelis E, et al. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. BMJ 2019 Jul 17;366:l4185 [FREE Full text] [doi: 10.1136/bmj.l4185] [Medline: 31315828]
2. Martens J, Van Gerven E, Lannoy K, Panella M, Euwema M, Sermeus W, et al. Serious reportable events within the inpatient mental health care: impact on physicians and nurses. Rev Calid Asist 2016 Jul;31 Suppl 2:26-33. [doi: 10.1016/j.cali.2016.04.004] [Medline: 27318766]

XSL•FO
RenderX

3. Rinaldi C, Ratti M, Russotto S, Seys D, Vanhaecht K, Panella M. Healthcare students and medical residents as second victims: a cross-sectional study. Int J Environ Res Public Health 2022 Sep 26;19(19):12218 [FREE Full text] [doi: 10.3390/ijerph191912218] [Medline: 36231520]
4. Busch IM, Moretti F, Purgato M, Barbui C, Wu AW, Rimondini M. Psychological and psychosomatic symptoms of second victims of adverse events: a systematic review and meta-analysis. J Patient Saf 2020 Jun;16(2):e61-e74 [FREE Full text] [doi: 10.1097/PTS.0000000000000589] [Medline: 30921046]
5. Wu AW. Medical error: the second victim. The doctor who makes the mistake needs help too. BMJ 2000 Mar 18;320(7237):726-727 [FREE Full text] [doi: 10.1136/bmj.320.7237.726] [Medline: 10720336]
6. Schiess C, Schwappach D, Schwendimann R, Vanhaecht K, Burgstaller M, Senn B. A transactional "second-victim" model-experiences of affected healthcare professionals in acute-somatic inpatient settings: a qualitative metasynthesis. J Patient Saf 2021 Dec 01;17(8):e1001-e1018. [doi: 10.1097/PTS.0000000000000461] [Medline: 29384831]
7. Global patient safety action plan 2021-2030: towards eliminating avoidable harm in health care. World Health Organization (WHO). URL: https://iris.who.int/bitstream/handle/10665/343477/9789240032705-eng.pdf?sequence=1 [accessed 2024-04-29]
8. Hollnagel E, Wears RL, Braithwaite J. From safety-I to safety-II: a white paper. National Health Service.: The Resilient Health Care Net URL: https://www.england.nhs.uk/signuptosafety/wp-content/uploads/sites/16/2015/10/safety-1-safety-2-whte-papr.pdf [accessed 2024-09-03]
9. Berry JC, Davis JT, Bartman T, Hafer CC, Lieb LM, Khan N, et al. Improved safety culture and teamwork climate are associated with decreases in patient harm and hospital mortality across a hospital system. J Patient Saf 2020 Jun;16(2):130-136. [doi: 10.1097/PTS.0000000000000251] [Medline: 26741790]
10. Mira JJ, Carrillo I, Lorenzo S, Ferrús L, Silvestre C, Pérez-Pérez P, et al. The aftermath of adverse events in Spanish primary care and hospital health professionals. BMC Health Serv Res 2015 Apr 09;15:151 [FREE Full text] [doi: 10.1186/s12913-015-0790-7] [Medline: 25886369]
11. Garcia CL, Abreu LC, Ramos JL, Castro CF, Smiderle FR, Santos JA, et al. Influence of burnout on patient safety: systematic review and meta-analysis. Medicina (Kaunas) 2019 Aug 30;55(9):553 [FREE Full text] [doi: 10.3390/medicina55090553] [Medline: 31480365]
12. Van Slambrouck L, Verschueren R, Seys D, Bruyneel L, Panella M, Vanhaecht K. Second victims among baccalaureate nursing students in the aftermath of a patient safety incident: an exploratory cross-sectional study. J Prof Nurs 2021;37(4):765-770. [doi: 10.1016/j.profnurs.2021.04.010] [Medline: 34187676]
13. Rodriquez J, Scott SD. When clinicians drop out and start over after adverse events. Jt Comm J Qual Patient Saf 2018 Mar;44(3):137-145. [doi: 10.1016/j.jcjq.2017.08.008] [Medline: 29499810]
14. Vanhaecht K, Seys D, Russotto S, Strametz R, Mira J, Sigurgeirsdóttir S, et al. An evidence and consensus-based definition of second victim: a strategic topic in healthcare quality, patient safety, person-centeredness and human resource management. Int J Environ Res Public Health 2022 Dec 15;19(24):16869 [FREE Full text] [doi: 10.3390/ijerph192416869] [Medline: 36554750]
15. Reason J. Human error: models and management. BMJ 2000 Mar 18;320(7237):768-770 [FREE Full text] [Medline: 10720363]
16. Colla JB, Bracken AC, Kinney LM, Weeks WB. Measuring patient safety climate: a review of surveys. Qual Saf Health Care 2005 Oct;14(5):364-366 [FREE Full text] [doi: 10.1136/qshc.2005.014217] [Medline: 16195571]
17. Williams ES, Manwell LB, Konrad TR, Linzer M. The relationship of organizational culture, stress, satisfaction, and burnout with physician-reported error and suboptimal patient care: results from the MEMO study. Health Care Manage Rev 2007;32(3):203-212. [doi: 10.1097/01.HMR.0000281626.28363.59] [Medline: 17666991]
18. Barkell NP, Snyder SS. Just culture in healthcare: an integrative review. Nurs Forum 2021 Jan;56(1):103-111. [doi: 10.1111/nuf.12525] [Medline: 33231884]
19. Marx D. Patient safety and the just culture. Obstet Gynecol Clin North Am 2019 Jun;46(2):239-245. [doi: 10.1016/j.ogc.2019.01.003] [Medline: 31056126]
20. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q 1999 Jun;44(2):350-383. [doi: 10.2307/2666999]
21. Appelbaum NP, Dow A, Mazmanian PE, Jundt DK, Appelbaum EN. The effects of power, leadership and psychological safety on resident event reporting. Med Educ 2016 Mar;50(3):343-350. [doi: 10.1111/medu.12947] [Medline: 26896019]
22. Etchegaray JM, Ottosen MJ, Dancsak T, Thomas EJ. Barriers to speaking up about patient safety concerns. J Patient Saf 2020 Dec;16(4):e230-e234. [doi: 10.1097/PTS.0000000000000334] [Medline: 29112033]
23. Khan MF, Sewell MD, Alrawi A, Taif S, Divani K. Can a culture of team psychological safety and MDT proforma improve team performance and patient outcomes in spinal MDTs? Br J Neurosurg 2024 Jun;38(3):726-730. [doi: 10.1080/02688697.2021.1967288] [Medline: 35135402]
24. Mira JJ. Errores Honestos y Segundas Víctimas: Hacia una Cultura Justa para la Seguridad del Paciente. J Healthc Qual Res 2023;38(5):259-261. [doi: 10.1016/j.jhqr.2023.08.001] [Medline: 37657855]

XSL•FO
RenderX

25. Dietl JE, Derksen C, Keller FM, Lippke S. Interdisciplinary and interprofessional communication intervention: how psychological safety fosters communication and increases patient safety. Front Psychol 2023 Jun 15;14:1164288 [FREE Full text] [doi: 10.3389/fpsyg.2023.1164288] [Medline: 37397302]
26. Freeman M, Morrow LA, Cameron M, McCullough K. Implementing a just culture: perceptions of nurse managers of required knowledge, skills and attitudes. Nurs Leadersh (Tor Ont) 2016;29(4):35-45. [doi: 10.12927/cjnl.2016.24985] [Medline: 28281449]
27. Leotsakos A, Ardolino A, Cheung R, Zheng H, Barraclough B, Walton M. Educating future leaders in patient safety. J Multidiscip Healthc 2014 Sep 19;7:381-388 [FREE Full text] [doi: 10.2147/JMDH.S53792] [Medline: 25285012]
28. Tocco Tussardi I, Benoni R, Moretti F, Tardivo S, Poli A, Wu AW, et al. Patient safety in the eyes of aspiring healthcare professionals: a systematic review of their attitudes. Int J Environ Res Public Health 2021 Jul 15;18(14):7524 [FREE Full text] [doi: 10.3390/ijerph18147524] [Medline: 34299975]
29. Patient safety curriculum guide: multi-professional edition. World Health Organization (WHO). 2011. URL: https://iris.who.int/bitstream/handle/10665/44641/9789241501958_eng.pdf?sequence=1 [accessed 2024-04-29]
30. Lee EH, Pitts S, Pignataro S, Newman LR, D'Angelo EJ. Establishing psychological safety in clinical supervision: multi-professional perspectives. Clin Teach 2022 Apr;19(2):71-78. [doi: 10.1111/tct.13451] [Medline: 35001537]
31. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002 Jan 09;287(2):226-235. [Medline: 11779266]
32. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain. New York, NY: David McKay; 1956.
33. Ablah E, Biberman DA, Weist EM, Buekens P, Bentley ME, Burke D, et al. Improving global health education: development of a global health competency model. Am J Trop Med Hyg 2014 Mar;90(3):560-565 [FREE Full text] [doi: 10.4269/ajtmh.13-0537] [Medline: 24445206]
34. Schwappach D, Sendlhofer G, Kamolz LP, Köle W, Brunner G. Speaking up culture of medical students within an academic teaching hospital: need of faculty working in patient safety. PLoS One 2019 Sep 12;14(9):e0222461 [FREE Full text] [doi: 10.1371/journal.pone.0222461] [Medline: 31514203]
35. Carrillo I, Serpa P, Landa-Ramírez E, Guilabert M, Gómez-Ayala Y, López-Pineda A, et al. Speaking up about patient safety, withholding voice and safety climate in clinical settings: a cross-sectional study among Ibero-American healthcare students. Int J Public Health 2024 Jul 01;69:1607406 [FREE Full text] [doi: 10.3389/ijph.2024.1607406] [Medline: 39011389]
36. Sánchez-García A, Saurín-Morán PJ, Carrillo I, Tella S, Põlluste K, Srulovici E, et al. Patient safety topics, especially the second victim phenomenon, are neglected in undergraduate medical and nursing curricula in Europe: an online observational study. BMC Nurs 2023 Aug 24;22(1):283 [FREE Full text] [doi: 10.1186/s12912-023-01448-w] [Medline: 37620803]
37. Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet E-surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
38. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357 [FREE Full text] [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
39. Home page. The European Researchers' Network Working on Second Victims. URL: https://cost-ernst.eu/ [accessed 2024-04-29]
40. Richard A, Pfeiffer Y, Schwappach DD. Development and psychometric evaluation of the speaking up about patient safety questionnaire. J Patient Saf 2021 Oct 01;17(7):e599-e606. [doi: 10.1097/PTS.0000000000000415] [Medline: 28858000]
41. Lee NJ, An JY, Song TM, Jang H, Park SY. Psychometric evaluation of a patient safety competency self-evaluation tool for nursing students. J Nurs Educ 2014 Oct;53(10):550-562. [doi: 10.3928/01484834-20140922-01] [Medline: 25275988]
42. Schnall R, Stone P, Currie L, Desjardins K, John RM, Bakken S. Development of a self-report instrument to measure patient safety attitudes, skills, and knowledge. J Nurs Scholarsh 2008;40(4):391-394. [doi: 10.1111/j.1547-5069.2008.00256.x] [Medline: 19094156]
43. Erlingsson C, Brysiewicz P. A hands-on guide to doing content analysis. Afr J Emerg Med 2017 Sep;7(3):93-99 [FREE Full text] [doi: 10.1016/j.afjem.2017.08.001] [Medline: 30456117]
44. Vivas Maiques C, Moreno Prat M, Vilariño Cerdá B, García Serra H. Balint group experience during the COVID pandemic [Article in Spanish]. Aten Primaria 2021 Dec;53(10):102177 [FREE Full text] [doi: 10.1016/j.aprim.2021.102177] [Medline: 34562662]
45. Scott SD, Hirschinger LE, Cox KR, McCoig M, Hahn-Cover K, Epperly KM, et al. Caring for our own: deploying a systemwide second victim rapid response team. Jt Comm J Qual Patient Saf 2010 May;36(5):233-240. [doi: 10.1016/s1553-7250(10)36038-7] [Medline: 20480757]
46. Haviland C, Green J, Dzara K, Hardiman WO, Petrusa ER, Park YS, et al. Psychological safety between pediatric residents and nurses and the impact of an interdisciplinary simulation curriculum. BMC Med Educ 2022 Aug 29;22(1):649 [FREE Full text] [doi: 10.1186/s12909-022-03709-9] [Medline: 36038868]
47. Pfeifer L, Vessey J, Cazzell M, Ponte PR, Geyer D. Relationships among psychological safety, the principles of high reliability, and safety reporting intentions in pediatric nursing. J Pediatr Nurs 2023;73:130-136. [doi: 10.1016/j.pedn.2023.09.001] [Medline: 37683304]

XSL•FO
RenderX

48. Cho H, Steege LM, Arsenault Knudsen É. Psychological safety, communication openness, nurse job outcomes, and patient safety in hospital nurses. Res Nurs Health 2023 Aug;46(4):445-453. [doi: 10.1002/nur.22327] [Medline: 37370217]
49. O'Donovan R, McAuliffe E. A systematic review of factors that enable psychological safety in healthcare teams. Int J Qual Health Care 2020 Jun 04;32(4):240-250. [doi: 10.1093/intqhc/mzaa025] [Medline: 32232323]
50. Ulmer F, Krings R, Häberli C, Bally R, Schuchmann M, Huwendiek S, et al. Patient safety 4.0: "failure of the week" it's all about role modelling! [Article in German]. Dtsch Med Wochenschr 2023 Aug;148(15):e87-e97 [FREE Full text] [doi: 10.1055/a-2061-1554] [Medline: 37308082]
51. Deis JN, Smith KM, Warren MD, Throop PG, Hickson GB, Joers BJ, et al. Transforming the morbidity and mortality conference into an instrument for systemwide improvement. In: Henriksen K, Battles JB, Keyes MA, editors. Advances in Patient Safety: New Directions and Alternative Approaches. Volume 2. Rockville, MD: Agency for Healthcare Research and Quality (US); 2008.
52. Steel EJ, Janda M, Jamali S, Winning M, Dai B, Sellwood K. Systematic review of morbidity and mortality meeting standardization: does it lead to improved professional development, system improvements, clinician engagement, and enhanced patient safety culture? J Patient Saf 2024 Mar 01;20(2):125-130. [doi: 10.1097/PTS.0000000000001184] [Medline: 38038688]
53. Schwappach DL, Häsler L, Strodtmann L, Siggelkow A. Morbidity and mortality conferences in lower saxony: implementation status and further development needs [Article in German]. Z Evid Fortbild Qual Gesundhwes 2018 Sep;135-136:34-40 [FREE Full text] [doi: 10.1016/j.zefq.2018.06.004] [Medline: 30007770]
54. Vallé B, Gasq C, Dehours E, Van Tricht M, Bounes V, Lauque D, et al. Morbidity and mortality conferences in emergency departments: the French National Survey. Eur J Emerg Med 2013 Oct;20(5):364-366. [doi: 10.1097/MEJ.0b013e32835ad5a8] [Medline: 23117420]
55. de Vos MS, Marang-van de Mheen PJ, Smith AD, Mou D, Whang EE, Hamming JF. Toward best practices for surgical morbidity and mortality conferences: a mixed methods study. J Surg Educ 2018;75(1):33-42. [doi: 10.1016/j.jsurg.2017.07.002] [Medline: 28720425]
56. McVeigh TP, Waters PS, Murphy R, O'Donoghue GT, McLaughlin R, Kerin MJ. Increasing reporting of adverse events to improve the educational value of the morbidity and mortality conference. J Am Coll Surg 2013 Jan;216(1):50-56. [doi: 10.1016/j.jamcollsurg.2012.09.010] [Medline: 23127791]
57. de Vos MS, Verhagen MJ, Hamming JF. The morbidity and mortality conference: a century-old practice with ongoing potential for future improvement. Eur J Pediatr Surg 2023 Apr;33(2):114-119 [FREE Full text] [doi: 10.1055/s-0043-1760836] [Medline: 36720246]
58. Seys D, Panella M, Russotto S, Strametz R, Joaquín Mira J, Van Wilder A, et al. In search of an international multidimensional action plan for second victim support: a narrative review. BMC Health Serv Res 2023 Jul 31;23(1):816 [FREE Full text] [doi: 10.1186/s12913-023-09637-8] [Medline: 37525127]
59. Lyman B, Gunn MM, Mendon CR. New graduate registered nurses' experiences with psychological safety. J Nurs Manag 2020 May;28(4):831-839. [doi: 10.1111/jonm.13006] [Medline: 32173958]
60. McClintock AH, Fainstad T. Growth, engagement, and belonging in the clinical learning environment: the role of psychological safety and the work ahead. J Gen Intern Med 2022 Jul;37(9):2291-2296 [FREE Full text] [doi: 10.1007/s11606-022-07493-6] [Medline: 35710656]
61. O'Donovan R, McAuliffe E. A systematic review exploring the content and outcomes of interventions to improve psychological safety, speaking up and voice behaviour. BMC Health Serv Res 2020 Feb 10;20(1):101 [FREE Full text] [doi: 10.1186/s12913-020-4931-2] [Medline: 32041595]
62. Minehart RD, Rudolph J, Pian-Smith MC, Raemer DB. Improving faculty feedback to resident trainees during a simulated case: a randomized, controlled trial of an educational intervention. Anesthesiology 2014 Jan;120(1):160-171 [FREE Full text] [doi: 10.1097/ALN.0000000000000058] [Medline: 24398734]
63. Torralba KD, Jose D, Byrne J. Psychological safety, the hidden curriculum, and ambiguity in medicine. Clin Rheumatol 2020 Mar;39(3):667-671. [doi: 10.1007/s10067-019-04889-4] [Medline: 31902031]
64. The safety competencies: enhancing patient safety across the health professions. 2nd edition. Canadian Patient Safety Institute. URL: https://www.healthcareexcellence.ca/media/115mbc4z/cpsi-safetycompetencies_en_digital-final-ua.pdf [accessed 2024-04-29]
65. Certified Professional in Patient Safety (CPPS): overview. Institute for Healthcare Improvement. URL: https://www.ihi.org/education/certified-professional-patient-safety-cpps [accessed 2024-04-29]
66. Majority of health jobs held by women. Eurostat. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210308-1 [accessed 2021-03-08]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
CV: coefficient of variability
ERNST: European Researchers' Network Working on Second Victims

XSL•FO
RenderX

KAS: knowledge, attitudes, and skills
M&M: morbidity and mortality conference

Edited by B Lesselroth; submitted 10.07.24; peer-reviewed by P Lachman, H Lee; comments to author 02.08.24; revised version received 15.08.24; accepted 14.09.24; published 07.10.24.

Please cite as:
Carrillo I, Skoumalová I, Bruus I, Klemm V, Guerra-Paiva S, Knežević B, Jankauskiene A, Jocic D, Tella S, Buttigieg SC, Srulovici E, Madarasová Gecková A, Põlluste K, Strametz R, Sousa P, Odalovic M, Mira JJ
Psychological Safety Competency Training During the Clinical Internship From the Perspective of Health Care Trainee Mentors in 11 Pan-European Countries: Mixed Methods Observational Study
JMIR Med Educ 2024;10:e64125
URL: https://mededu.jmir.org/2024/1/e64125
doi:10.2196/64125
PMID:39374073

©Irene Carrillo, Ivana Skoumalová, Ireen Bruus, Victoria Klemm, Sofia Guerra-Paiva, Bojana Knežević, Augustina Jankauskiene, Dragana Jocic, Susanna Tella, Sandra C Buttigieg, Einav Srulovici, Andrea Madarasová Gecková, Kaja Põlluste, Reinhard Strametz, Paulo Sousa, Marina Odalovic, José Joaquín Mira. Originally published in JMIR Medical Education (https://mededu.jmir.org), 07.10.2024. 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.

XSL•FO
RenderX

Navigating Nephrology's Decline Through a GPT-4 Analysis of Internal Medicine Specialties in the United States: Qualitative Study

Jing Miao[*], MD, PhD; Charat Thongprayoon[*], MD; Oscar Garcia Valencia, MD; Iasmina M Craici, MD; Wisit Cheungpasitporn, MD

Division of Nephrology and Hypertension, Department of Medicine, Mayo Clinic, 200 1st st sw, Rochester, MN, United States

[*]these authors contributed equally

Corresponding Author:
Jing Miao, MD, PhD

Abstract

Background: The 2024 Nephrology fellowship match data show the declining interest in nephrology in the United States, with an 11% drop in candidates and a mere 66% (321/488) of positions filled.

Objective: The study aims to discern the factors influencing this trend using ChatGPT, a leading chatbot model, for insights into the comparative appeal of nephrology versus other internal medicine specialties.

Methods: Using the GPT-4 model, the study compared nephrology with 13 other internal medicine specialties, evaluating each on 7 criteria including intellectual complexity, work-life balance, procedural involvement, research opportunities, patient relationships, career demand, and financial compensation. Each criterion was assigned scores from 1 to 10, with the cumulative score determining the ranking. The approach included counteracting potential bias by instructing GPT-4 to favor other specialties over nephrology in reverse scenarios.

Results: GPT-4 ranked nephrology only above sleep medicine. While nephrology scored higher than hospice and palliative medicine, it fell short in key criteria such as work-life balance, patient relationships, and career demand. When examining the percentage of filled positions in the 2024 appointment year match, nephrology's filled rate was 66%, only higher than the 45% (155/348) filled rate of geriatric medicine. Nephrology's score decreased by 4% - 14% in 5 criteria including intellectual challenge and complexity, procedural involvement, career opportunity and demand, research and academic opportunities, and financial compensation.

Conclusions: ChatGPT does not favor nephrology over most internal medicine specialties, highlighting its diminishing appeal as a career choice. This trend raises significant concerns, especially considering the overall physician shortage, and prompts a reevaluation of factors affecting specialty choice among medical residents.

(JMIR Med Educ 2024;10:e57157) doi:10.2196/57157

KEYWORDS

artificial intelligence; ChatGPT; nephrology fellowship training; fellowship matching; medical education; AI; nephrology; fellowship; United States; factor; chatbots; intellectual; complexity; work-life balance; procedural involvement; opportunity; career demand; financial compensation

Introduction

The National Resident Matching Program released the 2024 Nephrology fellowship match data on November 29, 2023 [1], revealing a significant downturn in the specialty's appeal. Only 321 candidates secured nephrology positions, marking an 11% decrease from the prior year, leaving just more than half of the 180 nephrology programs filled. The trend is more obvious when considering that of the 488 spots available, a mere 66% (321/488) were taken [2], underscoring a persistent wane in the candidate-to-position ratio from 1.3 in 2011 to around 0.6 in recent years [3]. Alarmingly, only a small fraction of these roles were filled by US MD graduates, ranging from 15% to 26% between 2019 and 2024 [1,4].

This disinterest in nephrology is particularly concerning given the escalating shortage of nephrologists worldwide [5] and the burgeoning prevalence of chronic kidney conditions [6]. It is predicted that the United States alone may face a deficit of more than 139,000 physicians by 2030 [7], a scenario that casts a long shadow over the future of nephrology care and its sustainability. The publication of annual match data consistently amplifies these worries, leading to persistent debates [3,8-11]. Nonetheless, the underlying causes of this critical issue are still largely unexamined.

XSL•FO
RenderX

In this context, there is a growing curiosity about the role of advanced artificial intelligence (AI) tools such as ChatGPT in reshaping medical education and practice [12-15]. This study uses ChatGPT to analyze and juxtapose nephrology with other internal medicine specialties, aiming to illuminate the influences shaping medical career choices today and provide insights into decision-making in the evolving landscape of medical career planning.

Methods

Specialties Examined in This Study

Within the realm of internal medicine, there are 17 fellowship specialties other than Nephrology [4]. There are 4 advanced fellowships such as Adult Congenital Heart Disease, Advanced Heart Failure & Transplant Cardiology, Clinical Cardiac Electrophysiology, and Interventional Pulmonology, which are typically not options for Internal Medicine residents or internists who might consider a nephrology fellowship. Hence, these 4 advanced specialties were not included in our study.

Study Design

GPT-4, a sophisticated iteration of ChatGPT, was prompted to provide insights into choosing between nephrology and other 13 internal medicine specialties. To examine that the ChatGPT's response did not depend on the sequence of fellowships presented in the query, we also asked ChatGPT to choose between other specialties and nephrology in reverse scenarios.

The prompts used in this study have been provided in (Multimedia Appendix 1) and are presented in screenshot format. Specifically, we asked:

> *If you need to choose nephrology or [insert specialty name] fellowship, which one do you choose, you can describe but at the end you need choose one; each aspect comparisons may choose scores of 1 - 10.*

For the reverse scenarios, we used:

> *If you need to choose [insert specialty name] or nephrology fellowship, which one do you choose, you can describe but at the end you need choose one; each aspect comparisons may choose scores of 1 - 10.*

ChatGPT's responses are also presented in screenshot format. To prevent our content from being used to train the models, we disabled the "Data controls—Improve the model for everyone" option in the setting of ChatGPT. To minimize potential biases from the AI's memory of prior interactions and ensure the independence of each prompt and response, we started each query in a new chat session.

Evaluation

In evaluating the decline in interest in Nephrology, we used 7 criteria identified independently by ChatGPT. These criteria include (1) intellectual challenge and complexity, (2) work-life balance, (3) procedural involvement, (4) research and academic opportunities, (5) patient relationships and continuity of care, (6) career opportunity and demand, and (7) financial compensation. These factors were chosen based on their relevance and applicability to fellowship selection in the real world. While these criteria are consistent with those used in the analysis of other specialties, it is important to note that they were not established by the authors themselves.

Each criterion was rated on a scale from 1 to 10. We did not train ChatGPT on how to score each criterion, such as defining what constitutes a 1/10 or a 9/10. We did not use any weighting anchors in the ChatGPT scoring process.

We calculated a cumulative score for each specialty based on the 7 criteria, resulting in a maximum possible score of 70 per specialty. The comparative ratio of nephrology's score in each criterion over other specialties was calculated. Nephrology's score in each criterion was also compared with the average score of all other specialties.

Ethical Considerations

This study does not include human participants (no human subjects experimentation or intervention was conducted) and so does not require institutional review board approval.

Results

ChatGPT favored only nephrology over a single specialty, sleep medicine (Table 1). Despite accruing a total score surpassing that of hospice and palliative medicine, ChatGPT opted for palliative medicine instead (Figure 1). Analysis of the 2024 appointment year match data revealed that 66% of nephrology positions were filled, a rate that exceeded only that of geriatric medicine, which stood at 45% (155/348) (Figure 1).

Upon examining specific parameters, nephrology ranked comparatively lower in terms of career demand, research opportunities, and financial remuneration than most other specialties (Figure 2). Specifically, nephrology experienced a decline ranging from 4% to 14% in 5 principal domains: intellectual challenge, procedural involvement, career demand, research prospects, and financial compensation. Nonetheless, nephrology exhibited a relative improvement, with a 7% increase noted in both the aspects of work-life balance and the development of patient relationships (Figure 3).

The same scores and choices were observed when we asked GPT-4 to evaluate other specialties over nephrology in reverse scenarios (Multimedia Appendix 1).

XSL•FO
RenderX

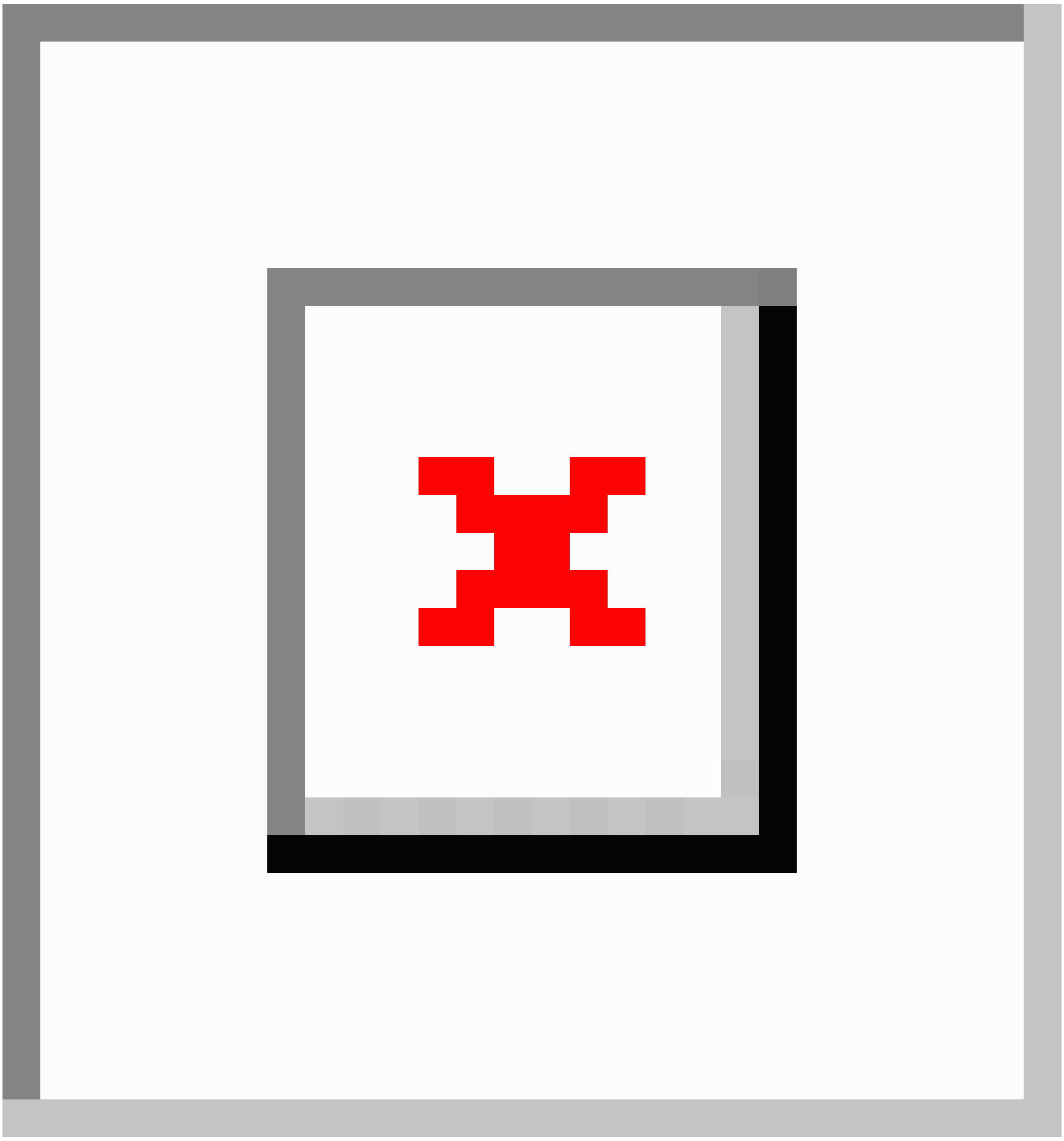
Table . Scale of nephrology and 13 other specialties on 7 criteria.

	Intellectual complexity	Work-life balance	Procedural involvement	Research and academic opportunities	Patient relationships and continuity of care	Career opportunity and demand	Financial compensation	Total score	ChatGPT's choice over Nephrology[a]
Nephrology	8	8	6	7	8	7	6	50	N/A[b]
Geriatric medicine	8	9	4	7	9	9	6	52	Yes
Infectious disease	9	8	4	9	7	8	6	51	Yes
Hospice and palliative medicine	7	9	4	6	9	8	6	49	Yes
Sleep medicine	7	9	4	6	7	6	6	45	No
Endocrinology, diabetes, and metabolism	8	9	4	8	8	8	6	51	Yes
Pulmonary Disease	8	7	8	8	7	8	7	53	Yes
Critical care medicine	9	6	9	8	5	9	8	54	Yes
Pulmonary disease and critical care medicine	9	6	9	8	7	8	8	55	Yes
Rheumatology	8	9	5	8	8	8	6	52	Yes
Hematology and oncology	9	6	7	9	8	8	8	55	Yes
Gastroenterology	9	7	9	8	7	8	8	56	Yes
Cardiovascular disease	9	6	9	9	7	8	9	57	Yes
Oncology	9	6	7	9	8	8	8	55	Yes
Average score[c]	8.5	7.2	6.8	8.0	7.5	7.9	7.2	52.7	N/A

[a]ChatGPT's preference for Nephrology compared with other Internal Medicine specialties when it comes to fellowship selection.
[b]N/A: not applicable.
[c]The average score of individual criterion across all other 13 specialties.

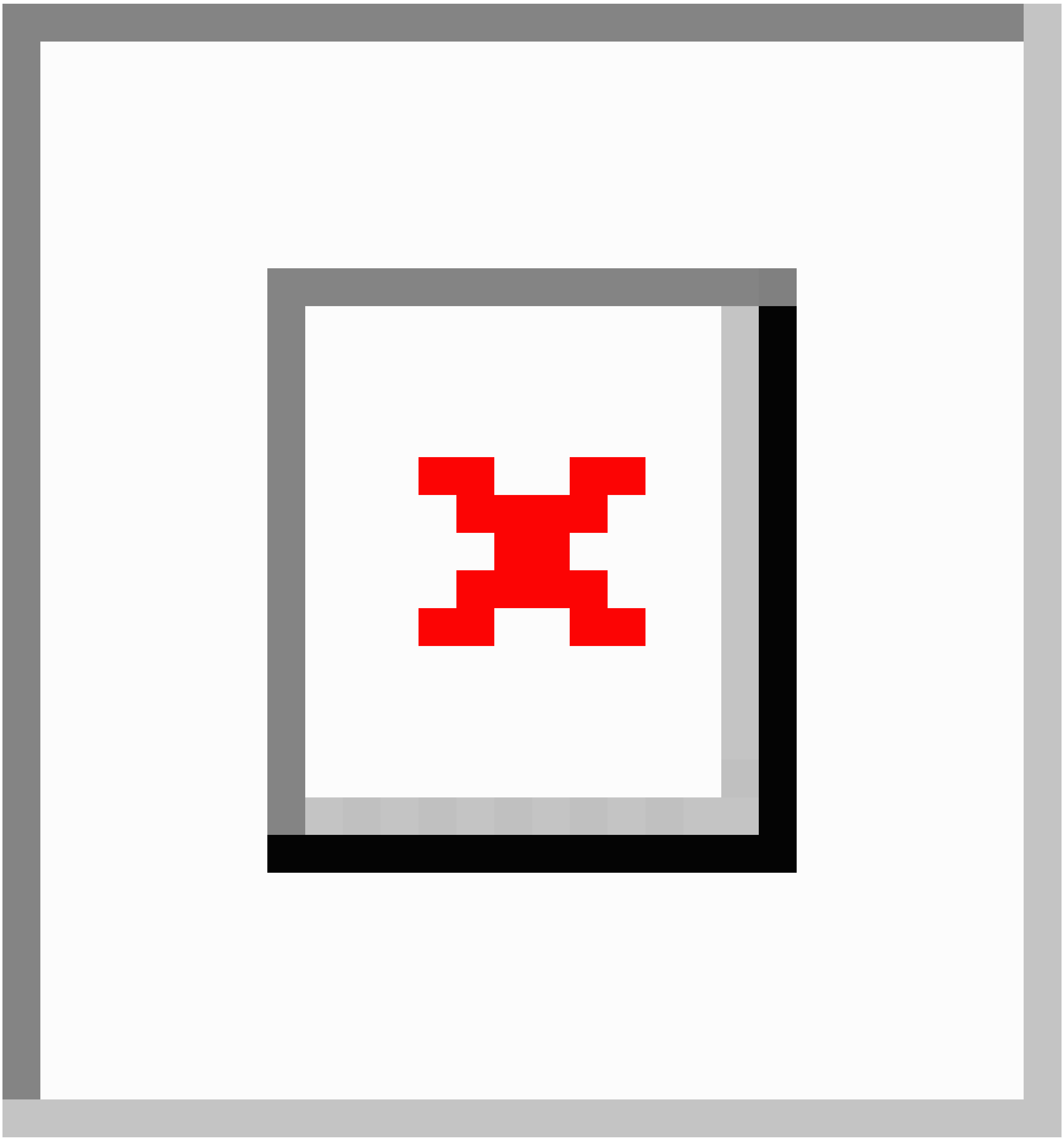
XSL•FO
RenderX

Figure 1. ChatGPT's score and the fellowship position fill rates. Seven criteria for each specialty were assessed by ChatGPT. Each criterion was scored on a scale from 1 to 10, resulting in a maximum possible score of 70. The total score assigned by ChatGPT to each specialty, along with its fellowship recommendations, is presented using diamonds. Nephrology's score (red diamond) surpassed only those of sleep medicine and hospice and palliative medicine (green diamond). ChatGPT recommended nephrology as a fellowship option only when compared with sleep medicine. ChatGPT's score and choice mainly align with the rank of positions filled in 2024 reported by the National Resident Matching Program (gray bar). The fill rate for nephrology fellowships (321/488, 65.8%) was only higher than that of geriatric medicine (155/348, 44.5%). AY: appointment year.



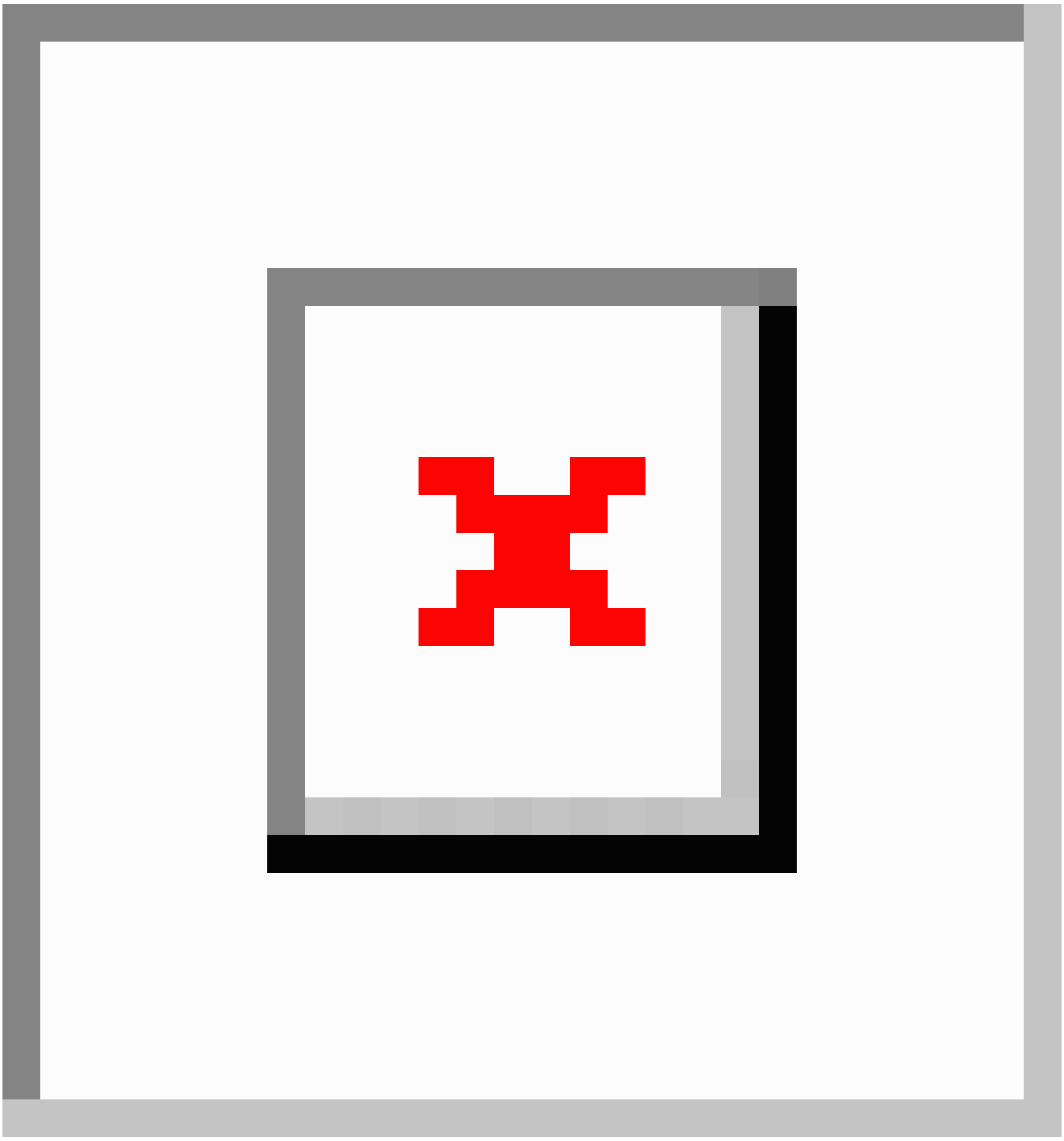
XSL•FO
RenderX

Figure 2. Comparisons of nephrology with other specialties in each criterion. The comparative ratio of nephrology to other specialties in terms of 7 criteria. color label for the ratio of nephrology to other specialties: black <1, white=1, and gray>1.



XSL•FO
RenderX

Figure 3. Comparisons of nephrology with the average score of other specialties in each criterion. Comparisons of the scale of various criteria in nephrology against the average scale for the other 13 specialties. Change in nephrology relative to other specialties shows the reduction in the percentage for nephrology relative to the average scale of the other 13 specialties.



Discussion

Principal Findings

The results underscore a diminished appeal in selecting nephrology as a career path, evident even in the preferences of sophisticated AI models such as ChatGPT when simulating fellowship choices. The factors contributing to this waning interest are likely diverse and personal. However, the significance of this trend cannot be overlooked, particularly in the context of the prevailing physician shortage.

Notably, the nephrology fellowship achieved a total score of 50, marginally surpassing hospice and palliative medicine, which scored 49, and sleep medicine at 45, among the 13 internal medicine specialties. However, ChatGPT recommended only nephrology over sleep medicine. Despite nephrology's higher overall score, ChatGPT favored hospice and palliative medicine due to its superior work-life balance, patient relationships, and career demand—especially pertinent given the aging population and increasing need for quality end-of-life care (Textbox 1). Nevertheless, ChatGPT noted that nephrology might be more suitable for those with a preference for technical aspects such as fluid and electrolyte management and renal pathophysiology. The lower score in intellectual complexity

XSL•FO
RenderX

and procedural involvement for hospice and palliative medicine may have also influenced ChatGPT's decision.

In fact, the reality for nephrologists is increasingly challenging. They face a growing workload and a diminishing control over their schedules, a situation exacerbated by the rising incidence of kidney diseases, especially chronic kidney disease and end-stage kidney disease [6]. The demands of the profession are extensive, involving long outpatient waitlists, demanding inpatient services, unpredictable night calls, and frequent visits to multiple dialysis units. These responsibilities, particularly the travel between units, consume significant time and effort [16]. In addition to their clinical duties, the 2023 Medscape Nephrologist Compensation Report states that nephrologists need to devote an average of 18.1 hours per week to support tasks such as paperwork and administration [17].

In terms of intellectual rigor, nephrology is on par with other specialties (8 vs 8.5). However, a national survey among internal medicine residents revealed that the field's broad scope and the complexity of kidney-related pathologies and physiologies deter many potential entrants [18-20]. The patients under nephrological care often present some of the most medically complex cases, marked by a plethora of comorbidities, intricate medication regimes, and a higher mortality risk. Despite these challenges, some find the diverse clinical conditions and the vast scope of practice in nephrology appealing [21,22]. Studies indicate that intellectual curiosity about kidney-related issues is a primary motivator for some choosing this career path [22]. Factors influencing this choice include a passion for the subject, a favorable work-life balance, mentorship availability, and exposure to the field [23]. However, exposure to nephrology during medical training is limited, with only a minority experiencing a rotation in this specialty during their clinical years, compared with a higher percentage during residency [18]. This limited exposure might be due to the complex nature of renal care, a relative scarcity of hands-on procedures compared with other specialties, and a lack of visible role models or mentors [24]. These factors contribute to the lower proportion of US MD graduates pursuing nephrology [4], highlighting a significant gap in early medical education and potential areas for enhancement in the field's approach to attracting and nurturing future talent.

Nephrology lags behind other medical specialties in financial reward, demand, and research prospects, impacting its appeal. It shows a 16% lower preference score in financial compensation, underscoring concerns highlighted in the 2023 Medscape Nephrologist Compensation Report. Nephrologists' average annual income falls below the median for all specialties [17]. In addition, the report indicates that nephrologists are in the bottom third of all specialties regarding how often they feel fairly compensated for their talents and time. In last year's report, nephrologists were ranked in the bottom spot. Furthermore, a survey among internal medicine residents reveals that the main obstacles deterring them from nephrology also include perceived financial inadequacy, intellectual rigor, work-life balance, and the potential to positively influence patient outcomes [18]. We recognize that financial considerations are multifaceted and significantly influenced by regional differences. Factors such as salary potential, cost of living, educational costs, and regional demand play crucial roles in deciding to choose nephrology as a specialty. Our study aimed to provide a general perspective, but we acknowledge the critical impact of regional financial factors on this decision-making process.

Recently, the perception of limited advancements and new therapeutic developments in nephrology has been recognized as a significant deterrent for choosing nephrology among internal medicine residents in the United States [25]. Despite ChatGPT scoring nephrology's career demand lower than the average for other specialties, there is, in reality, an increasing demand in the field of nephrology. This rise is observed not just in terms of patient needs but also in career opportunities within the specialty. This increase is driven by several factors, including the rising prevalence of chronic kidney disease, aging populations, and the associated complexities in managing these conditions [6]. As the number of individuals requiring specialized kidney care escalates, so does the need for skilled nephrologists to provide comprehensive and effective treatment. This surge in patient demand is creating more career opportunities within nephrology, indicating a promising future for those entering the field.

To address nephrology fellowship underfill rates, the American Society of Nephrology implemented measures such as the All-In policy, STARS (Students and Residents), and TREKS (Tutored Research Education for Kidney Scholars) programs [3,10]. ChatGPT suggests promoting nephrology's significance; enhancing training; fostering research; incorporating technology; encouraging collaboration, mentorship, and career growth; advocating work-life balance; increasing awareness; ensuring competitive pay; and broadening subspecialty choices.

Our study has certain constraints. It is important to acknowledge that the criteria for ranking established by ChatGPT are not fully transparent. While the criteria seem reasonable and applicable to real-world fellowship selection, the exact methodology and rationale behind their selection remain partially opaque. The overall score differences between nephrology and the 13 other specialties are relatively small. This limitation should be considered when interpreting the findings of our study. In addition, we did not use weighting anchors in the ChatGPT scoring process, recognizing that the importance of each criterion may vary. To mitigate hallucinations, a major concern in ChatGPT's responses, we implemented several measures such as asking the same questions in reverse scenarios, preventing the content from being used to train the models, and starting each query in a new chat session to minimize potential biases from the AI's memory of prior interactions. While these measures reduce the likelihood of hallucinations, we cannot completely exclude their possibility. It remains uncertain whether ChatGPT's choices truly reflect residents' sentiments. Addressing whether this issue originates from biases in perception or from broader systemic problems is crucial. Moreover, ChatGPT's scores and choices were consistent when the same questions were presented in reverse scenarios. However, previous studies, including our own [26], have shown varying levels of repeatability in ChatGPT's responses, indicating that repeatability might depend on the nature and type of the question. Finally, although we obtained

XSL•FO
RenderX

consistent results using another ChatGPT account with the same prompts, we cannot entirely exclude the possibility that different ChatGPT accounts might produce varying results due to differences in settings, usage habits, repeatability, and other unknown factors. Further research is needed to address this issue.

Textbox 1. ChatGPT's conclusion on deciding to pursue a fellowship in Hospice and Palliative Medicine.

Conclusion Based on this analysis, I would recommend choosing a Hospice and Palliative Medicine fellowship. It scores higher in work-life balance, patient relationships, and career demand, particularly relevant in the context of an aging population and the increasing need for quality end-of-life care. This choice should align with personal interests and career goals, especially if one is drawn to patient-centered, holistic care. If a preference lies in more technical aspects like fluid and electrolyte management and renal pathophysiology, then Nephrology might be more suitable.

Conclusions

ChatGPT, as an AI model, shows no bias toward nephrology over other internal medicine branches in fellowship choices, highlighting a broader decline in interest for this specialty. This trend is driven by factors such as financial incentives, career demands, and opportunities for research, which significantly influence specialty decisions. Moreover, intellectual stimulation and work-life balance are key factors. This issue, whether due to perceived or real barriers, demands immediate action in light of the physician shortage. Addressing these deterrents is essential to boost nephrology's attractiveness and fulfill the increasing demand for nephrologists, thereby maintaining exemplary health care standards.

Conflicts of Interest

None declared.

Multimedia Appendix 1

ChatGPT's responses to fellowship selection between nephrology and other 13 internal medicine specialties as well as between other specialties and nephrology in reverse scenarios.

[PDF File, 10019 KB - mededu_v10i1e57157_app1.pdf]

References

1. National Resident Matching Program. 2023 medicine and pediatric specialties match results report. 2023. URL: https://www.nrmp.org/wp-content/uploads/2023/11/2023-MPSM-Match-Results-Statistics-Report.pdf [accessed 2023-11-29]
2. Pivert KA. First look: AY 2024 match. : ASNDATA; 2023. URL: https://data.asn-online.org/posts/ay_2024_match/ [accessed 2023-11-29]
3. Parker MG, Sozio SM. The future nephrology workforce: there will be one. Clin J Am Soc Nephrol 2021 Nov;16(11):1752-1754. [doi: 10.2215/CJN.05040421] [Medline: 34281982]
4. National Resident Matching Program. Results and data specialties matching service 2023: appointment year. 2022. URL: https://www.nrmp.org/wp-content/uploads/2023/04/2023-SMS-Results-and-Data-Book.pdf [accessed 2024-10-01]
5. International Society of Nephrology. ISN global kidney health atlas. 2019. URL: https://www.theisn.org/wp-content/uploads/2023/10/GKHAtlas_2019_WebFIle_rev.pdf [accessed 2023-12-04]
6. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl (2011) 2022 Apr;12(1):7-11. [doi: 10.1016/j.kisu.2021.11.003] [Medline: 35529086]
7. Zhang X, Lin D, Pforsich H, Lin VW. Physician workforce in the United States of America: forecasting nationwide shortages. Hum Resour Health 2020 Feb 6;18(1):8. [doi: 10.1186/s12960-020-0448-3] [Medline: 32029001]
8. Shah HH, Fishbane S, Ross DW, Jhaveri KD, Sachdeva M. Subspecialty focus tracks during nephrology fellowship training. Am J Kidney Dis 2023 Dec;82(6):639-643. [doi: 10.1053/j.ajkd.2023.05.006] [Medline: 37516298]
9. Palleti SK. Fellowship in the United States as an exceptionally qualified applicant. Postgrad Med J 2023 Sep 21;99(1176):1115-1119. [doi: 10.1093/postmj/qgad036] [Medline: 37286194]
10. Cheng SC, Pivert KA, Sozio SM. "Make me a match": all-in and other trends in the nephrology match. Clin J Am Soc Nephrol 2022 Nov;17(11):1691-1693. [doi: 10.2215/CJN.04450422] [Medline: 35853729]
11. Adams ND. Attracting more residents into nephrology. Clin J Am Soc Nephrol 2012 Sep;7(9):1382-1384. [doi: 10.2215/CJN.07600712] [Medline: 22917705]
12. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
13. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
14. Garcia Valencia OA, Thongprayoon C, Jadlowiec CC, Mao SA, Miao J, Cheungpasitporn W. Enhancing kidney transplant care through the integration of chatbot. Healthcare (Basel) 2023 Sep 12;11(18):37761715. [doi: 10.3390/healthcare11182518] [Medline: 37761715]

XSL•FO
RenderX

15. Suppadungsuk S, Thongprayoon C, Krisanapan P, et al. Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. J Clin Med 2023 Aug 25;12(17):5550. [doi: 10.3390/jcm12175550] [Medline: 37685617]
16. Roberts JK. Burnout in nephrology: implications on recruitment and the workforce. Clin J Am Soc Nephrol 2018 Feb 7;13(2):328-330. [doi: 10.2215/CJN.09870917] [Medline: 29326310]
17. Koval ML. Medscape nephrologist compensation report 2023. : Medscape; 2023 URL: https://www.medscape.com/slideshow/2023-compensation-nephrologist-6016364 [accessed 2024-10-01]
18. Nakhoul GN, Mehdi A, Taliercio JJ, et al. "What do you think about nephrology?" A national survey of internal medicine residents. BMC Nephrol 2021 May 21;22(1):190. [doi: 10.1186/s12882-021-02397-9] [Medline: 34020598]
19. Jhaveri KD, Sparks MA, Shah HH, et al. Why not nephrology? A survey of US internal medicine subspecialty fellows. Am J Kidney Dis 2013 Apr;61(4):540-546. [doi: 10.1053/j.ajkd.2012.10.025] [Medline: 23332603]
20. Daniels MN, Maynard S, Porter I, Kincaid H, Jain D, Aslam N. Career interest and perceptions of nephrology: a repeated cross-sectional survey of internal medicine residents. PLoS One 2017;12(2):e0172167. [doi: 10.1371/journal.pone.0172167] [Medline: 28207893]
21. Beckwith H, Kingsbury M, Horsburgh J. Why do people choose nephrology? Identifying positive motivators to aid recruitment and retention. Clin Kidney J 2018 Oct;11(5):599-604. [doi: 10.1093/ckj/sfy076] [Medline: 30288258]
22. McMahon GM, Thomas L, Tucker JK, Lin J. Factors in career choice among US nephrologists. Clin J Am Soc Nephrol 2012 Nov;7(11):1786-1792. [doi: 10.2215/CJN.03250312] [Medline: 22956263]
23. Nair D, Pivert KA, Baudy A, Thakar CV. Perceptions of nephrology among medical students and internal medicine residents: a national survey among institutions with nephrology exposure. BMC Nephrol 2019 Apr 29;20(1):146. [doi: 10.1186/s12882-019-1289-y] [Medline: 31035944]
24. Moura-Neto JA. "To be, or not to be" a nephrologist: students' dilemma and a strategy for the field. Blood Purif 2021;50(4-5):696-701. [doi: 10.1159/000513155] [Medline: 33503624]
25. Beck N, Furgeson S, Chonchol M, Kendrick J. Internal medicine residents' perceptions of nephrology as a career: a focus group study. Kidney360 2020 Oct 29;1(10):1052-1059. [doi: 10.34067/KID.0003652020] [Medline: 35368786]
26. Miao J, Thongprayoon C, Garcia Valencia OA, et al. Performance of ChatGPT on nephrology test questions. Clin J Am Soc Nephrol 2024 Jan 1;19(1):35-43. [doi: 10.2215/CJN.0000000000000330] [Medline: 37851468]

Abbreviations

AI: artificial intelligence

Edited by B Lesselroth; submitted 06.02.24; peer-reviewed by FH Leung, MH Drazner, R Yin; revised version received 22.05.24; accepted 15.08.24; published 10.10.24.

Please cite as:
Miao J, Thongprayoon C, Garcia Valencia O, Craici IM, Cheungpasitporn W
Navigating Nephrology's Decline Through a GPT-4 Analysis of Internal Medicine Specialties in the United States: Qualitative Study
JMIR Med Educ 2024;10:e57157
URL: https://mededu.jmir.org/2024/1/e57157
doi:10.2196/57157

© Jing Miao, Charat Thongprayoon, Oscar Garcia Valencia, Iasmina M Craici, Wisit Cheungpasitporn. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.10.2024. 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.

XSL•FO
RenderX

Health Care Workers' Motivations for Enrolling in Massive Open Online Courses During a Public Health Emergency: Descriptive Analysis

Jennifer Jones[1], MBBS; Jamie Sewan Johnston[2], PhD; Ngouille Yabsa Ndiaye[3], MPH; Anna Tokar[3], PhD; Saumya Singla[2], MA; Nadine Ann Skinner[2], PhD; Matthew Strehlow[1,2], MD; Heini Utunen[3], PhD

1

2

3

Corresponding Author:
Jennifer Jones, MBBS

Abstract

Background: Massive open online courses (MOOCs) are increasingly used to educate health care workers during public health emergencies. In early 2020, the World Health Organization (WHO) developed a series of MOOCs for COVID-19, introducing the disease and strategies to control its outbreak, with 6 courses specifically targeting health care workers as learners. In 2020, Stanford University also launched a MOOC designed to deliver accurate and timely education on COVID-19, equipping health care workers across the globe to provide health care safely and effectively to patients with the novel infectious disease. Although the use of MOOCs for just-in-time training has expanded during the pandemic, evidence is limited regarding the factors motivating health care workers to enroll in and complete courses, particularly in low-income countries (LICs) and lower-middle–income countries (LMICs).

Objective: This study seeks to gain insights on the characteristics and motivations of learners turning to MOOCs for just-in-time training, to provide evidence that can better inform MOOC design to meet the needs of health care workers. We examine data from learners in 1 Stanford University and 6 WHO COVID-19 courses to identify (1) the characteristics of health care workers completing the courses and (2) the factors motivating them to enroll.

Methods: We analyze (1) course registration data of the 49,098 health care workers who completed the 7 focal courses and (2) survey responses from 6272 course completers. The survey asked respondents to rank their motivations for enrollment and share feedback about their learning experience. We use descriptive statistics to compare responses by health care profession and by World Bank country income classification.

Results: Health care workers completed the focal courses from all regions of the world, with nearly one-third (14,159/49,098, 28.84%) practicing in LICs and LMICs. Survey data revealed a diverse range of professional roles among the learners, including physicians (2171/6272, 34.61%); nurses (1599/6272, 25.49%); and other health care professionals such as allied health professionals, community health workers, paramedics, and pharmacists (2502/6272, 39.89%). Across all health care professions, the primary motivation to enroll was for personal learning to improve clinical practice. Continuing education credit was also an important motivator, particularly for nonphysicians and learners in LICs and LMICs. Course cost (3423/6272, 54.58%) and certification (4238/6272, 67.57%) were also important to a majority of learners.

Conclusions: Our results demonstrate that a diverse range of health care professionals accessed MOOCs for just-in-time training during a public health emergency. Although all health care workers were motivated to improve their clinical practice, different factors were influential across professions and locations. These factors should be considered in MOOC design to meet the needs of health care workers, particularly those in lower-resource settings where alternative avenues for training may be limited.

(JMIR Med Educ 2024;10:e51915) doi:10.2196/51915

KEYWORDS

massive open online course; MOOC; online learning; online courses; online course; health care education; medical education; education; training; professional development; continuing education; COVID-19 training; infectious disease outbreak response; emergency; public health; crisis; crises; outbreak; pandemic; COVID-19; SARS-CoV-2; coronavirus; humanitarian emergency response; health care workers; nurse; nurses; practitioner; practitioners; clinician; clinicians; health care worker; medic; low-income; lower-middle income; LIC; LMIC; developing country; developing countries; developing nation; developing nations; case study;

XSL•FO
RenderX

survey; surveys; descriptive analysis; descriptive analyses; motivation; motivations; lower-middle–income country; low-income country

Introduction

During the COVID-19 pandemic, massive open online courses (MOOCs) emerged as an invaluable source of training for health care workers globally [1-4]. Studies have demonstrated MOOCs' effectiveness in facilitating learning among practicing health care professionals [5,6], and their capability to deliver content rapidly and flexibly has established e-learning as a preferred method for transferring clinical skills and knowledge [6]. Their broad applicability, accessibility, and cost-effectiveness make MOOCs particularly appealing for continuing education (CE) requirements, also known as continuing medical education [5,7,8]. Consequently, MOOCs have been used for skill development and retention, competency assessment, and lifelong learning [9]. In low-income countries (LICs) and lower-middle–income countries (LMICs), MOOCs potentially increase access to essential health education content and reduce training costs for health care professionals [5,10,11].

Despite the increasing data on general MOOC enrollee motivations [12-15], there remains a significant gap concerning the specific factors motivating practicing health care professionals. Understanding the motivations of health care workers in LICs and LMICs to enroll in and complete health care–related MOOCs is crucial, as engagement and completion rates among this group are notably low [16-18]. By identifying what drives their participation, we can enhance MOOC design and dissemination, particularly for just-in-time learning initiatives during health emergencies—a time when organizations such as the World Health Organization (WHO) and national governments increasingly rely on MOOCs to rapidly disseminate critical information to health care workers.

This study aims to uncover the characteristics and motivations of health care professionals who enrolled in health care–related MOOCs during the COVID-19 pandemic—a period marked by an urgent need to rapidly disseminate critical health care information. Research indicates several potential reasons for enrolling in MOOCs. As a teaching model, MOOCs support adult learning principles targeting self-directed learners [17]. The self-directed learning model allows individuals to guide their learning process, establish their learning objectives, engage in individualized learning strategies, and manage their time based on their interests while still receiving access to curated content [17]. It can be presumed that learner motivations for engaging in MOOCs differ from those in traditional brick-and-mortar educational venues [19]. Prior studies suggest that primary intrinsic motivations for MOOC enrollment include personal interest and knowledge acquisition [12], whereas extrinsic motivations often involve certification and professional development opportunities [17]. However, the specific motivations driving health care workers, particularly those in LICs and LMICs, remain underexplored.

Although recent studies, such as Garrido et al [20] and a scoping review on MOOCs for health care worker education in low- and middle-income countries [21], have begun to explore the use of MOOCs for professional and workforce development, these insights predominantly focus on broad educational outcomes and employment advancements. Such research underscores the potential of MOOCs to enhance skill sets and career opportunities, highlighting the alignment of MOOC coursework with job market needs and professional certifications. However, these studies generally do not delve deeply into the specific intrinsic motivations of health care workers in LICs and LMICs to enroll in MOOCs, especially during health emergencies. In fact, in 2023, the WHO commissioned 3 systematic reviews of the literature to support guidelines for building just-in-time training during public health emergencies, finding a gap in the literature regarding the motivations of learners enrolling in relevant online courses, particularly in LMICs (WHO, unpublished data, 2023). Our study seeks to fill this void by examining the unique motivations behind MOOC enrollment, particularly during the unprecedented global crisis triggered by the COVID-19 pandemic.

This study contributes uniquely to the literature by investigating the key motivations for health care workers to enroll in MOOCs, with a special emphasis on provider type and country income level during a global health crisis. These insights are vital as learners in LICs and LMICs face challenges such as linguistic and cultural barriers, limited access to digital technology, low-bandwidth connectivity, infrastructure constraints, and limited digital literacy [5,10]. By understanding what motivates learners in these settings, our study provides foundational knowledge that can inform more thoughtful and effective MOOC design and recruitment strategies, ultimately improving knowledge transmission, learning outcomes, and course completion rates in regions with critical needs for health care worker training. This broad impact underscores the potential of targeted online education strategies to significantly enhance global health responses.

Methods

Study Design

In this study, we present a descriptive analysis of MOOC learner data to identify the characteristics and motivations of health care workers enrolled in 7 MOOCs designed to serve as just-in-time education for clinically practicing health care workers during the COVID-19 pandemic. We examine two sources of data: (1) course enrollment data (n=49,098) collected during course registration and (2) follow-up survey data (n=6272) collected from course completers.

Course Descriptions

In Table 1, we detail the 7 focal courses examined in this study. We selected 6 courses developed by the WHO in early 2020 to respond to the growing COVID-19 crisis. These courses were launched on the OpenWHO online platform, which serves as the WHO's learning hub for health emergencies. These courses build on the WHO's initial introductory COVID-19 course, which had 232,890 enrollments across 13 published languages by the end of March 2020 and provided general information

XSL•FO
RenderX

about the disease for a broad audience [22]. The 6 WHO courses were selected out of all 43 COVID-19 courses offered on the OpenWHO platform due to their greater content relevance to practicing health care workers. The 6 MOOCs focused on introducing health care workers to the novel disease and providing them with strategies to control its outbreak. Three courses were designed to provide health care workers with the basic tools needed to combat the pandemic and protect themselves from infection when providing health care services. Another 3 courses were designed to provide health care workers with an overview of the COVID-19 disease and provide learners with specific clinical strategies to address the pandemic. The courses were initially published in English and then rapidly translated into over 19 languages in the subsequent 2 months.

XSL•FO
RenderX

Table . Course descriptions.

Source and course title		Description	Languages	Date launched	Course duration	Enrolled learners, n
Stanford University						
	COVID-19 Training for Healthcare Workers	This course is designed for health care professionals. It provides an evidence-based approach to life-saving techniques for treating critically ill patients with COVID-19.	English, Hindi, Portuguese, French, and Spanish	July 17, 2020	8 h	101,734
OpenWHO						
	Hand Hygiene	This course is designed to summarize the WHO[a] guidelines on hand hygiene, associated tools, and ideas for effective implementation. The WHO guidelines support hand hygiene promotion and improvement in health care facilities worldwide.	Arabic, Chinese, Dutch, English, French, Macedonian, Portuguese, Russian, Shqip, Sinhalese, Somali, Spanish, Tamil, Tetum, and Turkish	June 3, 2020	1 h	274,116
	Personal Protective Equipment	The course is a guide for health care workers involved in patient care activities in a health care setting. It aims to show the type of personal protective equipment needed to correctly protect oneself.	Albanian, Arabic, Chinese, Dutch, English, French, Kazakh, Macedonian, Portuguese, Russian, Sinhalese, Somali, Spanish, Tamil, Tetum, Thai, and Turkish	April 15, 2020	15 min	346,200
	Occupational Health and Safety	This course is for health workers, incident managers, supervisors, and administrators who make policies and protocols for their health facilities. The WHO recommends a combination of measures for infection prevention and control, occupational health and safety, and psychosocial support.	Dutch, English, Indonesian, Macedonian, Portuguese, Spanish, and Swahili	August 30, 2020	1 h	85,504

XSL•FO
RenderX

Source and course title		Description	Languages	Date launched	Course duration	Enrolled learners, n
	Clinical Management: Patient Rehabilitation	The course is devoted to the rehabilitation of patients with COVID-19 by addressing needs of patients recovering from COVID-19, including patients with cognitive impairment, physical deconditioning and weakness, respiratory impairment, swallow impairment, and communication impairment, as well as techniques for rehabilitation.	Chinese, English, French, Macedonian, Russian, and Shqip	January 13, 2021	3 h	22,704
	Clinical Management: General Considerations	This course gives background on the pandemic, discusses facility operations, and addresses COVID-19 pandemic preparedness at all levels of health care provision. It also discusses ethical issues arising during COVID-19 care.	English, Indonesian, Macedonian, and Shqip	October 22, 2020	3 h	31,972
	Clinical Management: Acutely Ill Patients	Designed to prepare and support health providers as they provide emergency care to seriously ill patients with COVID-19, including a systematic approach via the WHO and ICRC[b] Basic Emergency Care course content.	English, Somali, and Spanish	May 5, 2021	6 h	14,190

[a]WHO: World Health Organization.

[b]ICRC: International Committee of the Red Cross.

To broaden the reach of learners in the study, we also included a Stanford University MOOC launched in August 2020 to equip health care workers with timely in-service education, to improve their ability to safely and effectively treat patients with the novel disease [23]. The Stanford MOOC was launched on both the Coursera and edX platforms, 2 US-based MOOC providers founded in 2012 that routinely provide university-level courses on various topics including health. As of November 2020, nearly 900 health-related courses were available on the Coursera platform alone [24]. The Stanford course was first developed in English and then translated into 4 additional languages.

The courses were promoted via their respective institutional networks. No paid advertisements were published. The Stanford course was promoted starting in July 2020, with emails sent to over 100,000 Coursera listserve subscribers. The course was also promoted through a variety of Stanford-affiliated social media channels and online publications, YouTube's spotlight channel, and direct sharing with a network of health education collaborators throughout the world by Stanford team members. The WHO courses were promoted as each course launched on the WHO website, the OpenWHO platform, and through WHO newsletters and mailing lists.

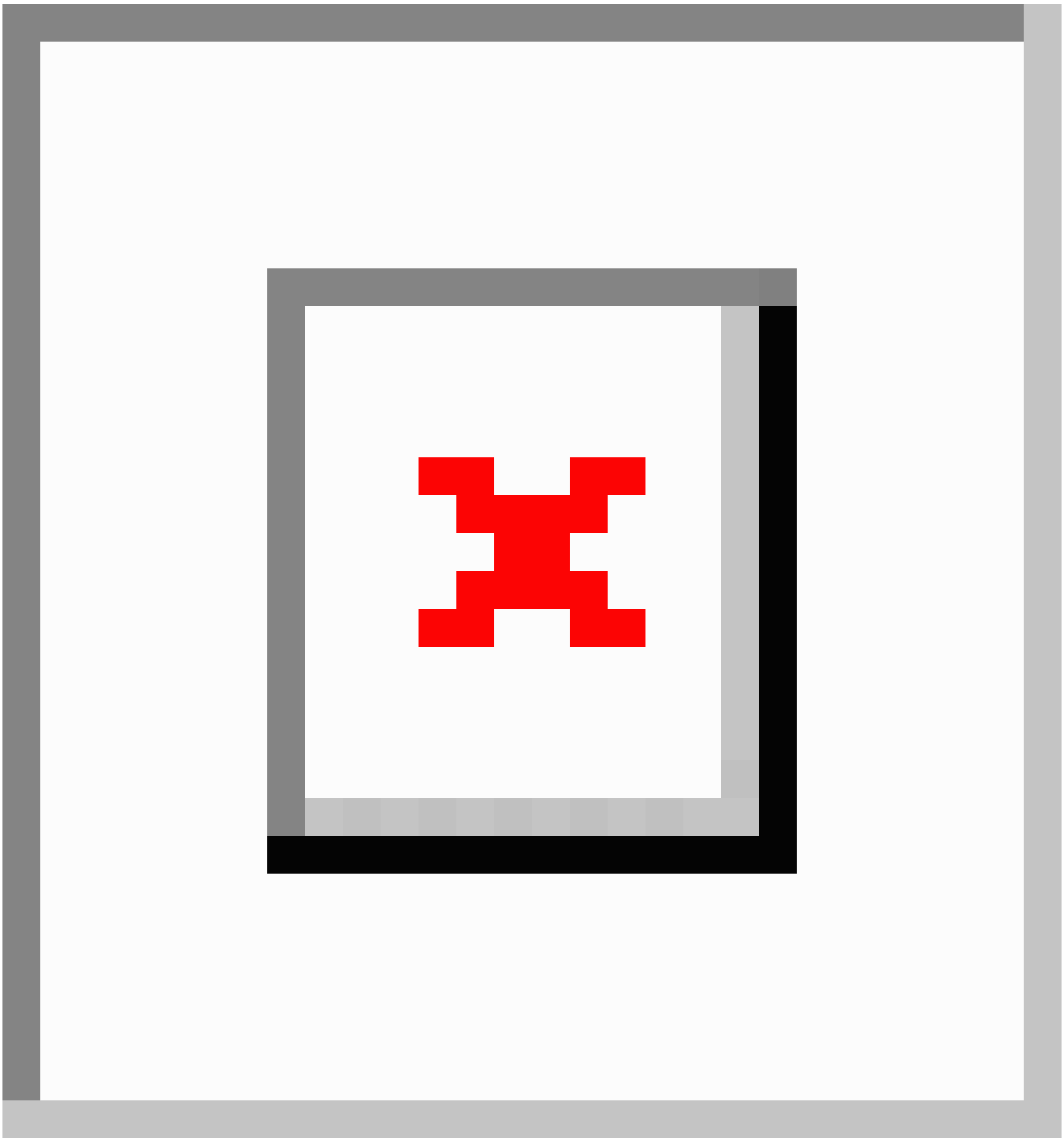
Data Collection

Figure 1 describes the flow diagram for study participation and data collection. We obtained data on all course enrollees via the

XSL•FO
RenderX

respective course platforms (OpenWHO for WHO courses and edX and Coursera for the Stanford course). Course completion was defined by course developers and identified through backend data available from the course platforms. Learner background data were collected via the respective platforms at the time of course registration and included the learners' age, gender, geographic location, and profession. The health care worker profession category included those identifying as being employed in the following professions: allopathic medicine (including physicians and physician assistants); traditional medicine; nursing (including nurses, nurse practitioners, nurse midwives, nursing instructors, and certified nursing assistants); allied health (including physical therapy, occupational therapy, speech pathology, medical assistants, and home health aides); community health; emergency medical services (including paramedics and emergency medical technicians); and pharmacy (including pharmacists and pharmacy technicians).

Figure 1. Flow diagram for study participation. The number of enrollees, course completers, and survey completers is shown for all learners and health care workers. The survey completer sample (shaded in gray) is the focal sample for this study. Health care workers included those who identified as being employed as health care professionals at enrollment and in the follow-up survey. Health care professions included the following: allied health; community health; nursing (including nurses, nurse practitioners, nurse midwives, nursing instructors, and certified nursing assistants); physician assistants; paramedics and emergency medical technicians; pharmacy; physicians; and traditional medicine.



XSL•FO
RenderX

We invited all enrollees who had completed the course they were enrolled in to complete an online survey (Multimedia Appendix 1) on the respective course platforms. To recruit WHO course learners, we sent 3 survey invitations to the email addresses provided by learners at the time of registration and through the OpenWHO automated course message. To recruit participants from the Stanford course, we sent 3 requests via Coursera and edX email announcements. The survey window was open from December 11, 2020, to September 28, 2021. The survey completion rate was 3.98% (12,170/305,849) among all course completers and 12.77% (6272/49,098) among health care workers completing the courses.

The 23-question survey collected information on learners' personal and professional demographics, information about their professional experience with COVID-19, and their ability to connect with physicians in their daily work. Respondents were asked to rank 6 possible motivations for course enrollment in the order of importance to them. Additionally, respondents were asked about their use of course certificates, including whether their employer required a certificate, if they planned to provide it to their employer, or if they planned to use it for CE credit. Finally, respondents were asked about the cost of MOOCs and how it impacted their decision to enroll in the course. All study authors were involved in the development of the survey. Questions were reviewed by all authors to include appropriate vocabulary, inclusive of globally used terminology. The survey was not adapted directly from any other source; however, the motivations included were drawn from anecdotal course feedback and the extant literature discussing motivations for MOOC enrollment.

Statistical Analysis

Because of the study focus, we limited our analytic sample to health care workers exclusively. To investigate the generalizability of our survey sample, we summarized the characteristics of all health care workers completing the courses (n=49,098) and health care workers completing the survey (n=6272) using descriptive statistics (mean, SD, and response rates). To compare the proportion of learners by characteristic between course completers and survey completers, we used the Pearson χ^2 test. To examine ranked enrollment motivators and compare across learner subgroups, we conducted multiple comparison tests using 1-way ANOVA, comparing the mean rank of motivations (dependent variable) by learner characteristics. The independent variables compared included differences by occupation (physicians vs nurses and physicians vs other health professionals) and country income classification (LICs and LMICs vs upper-middle–income countries [UMICs] and high-income countries [HICs]). All statistical analyses were conducted using Stata SE V15 (StataCorp).

Ethical Considerations

Informed consent was obtained from all learners. Participation was voluntary and no monetary compensation was provided to the participants. The collected data were anonymized. Approval for all aspects of this study design, including consent, outreach, data collection, surveying, and data analysis, was obtained from the Stanford University School of Medicine Institutional Review Board (protocol 57831).

Results

Learner Characteristics

As shown in Figure 1, as of September 2021, the 7 courses had 856,263 total enrollees, 90.47% (n=774,686) in WHO courses and 9.53% (n=81,577) in the Stanford course. In all, 13.3% (113,902/856,263) of enrollees and 16.05% (49,098/305,849) of course completers identified as practicing health care workers at course registration. The course completion rate was higher among health care workers (49,098/113,902, 43.1%) than overall enrollees (305,849/856,263, 35.72%).

Table 2 shows that nearly one-third (15,238/49,098, 31.04%) of the health care workers that completed a course were between the ages of 18 - 29 years, and 41.25% (20,252/49,098) identified as female. The region with the most health care workers that completed a course was Latin America and the Caribbean (10,665/49,098, 21.72%), followed by South Asia (7264/49,098, 14.79%), North America (7019/49,098, 14.3%), Europe and Central Asia (5365/49,098, 10.93%), East Asia and the Pacific (5278/49,098, 10.75%), Middle East and North Africa (3816/49,098, 7.77%), and sub-Saharan Africa (3502/49,098, 7.13%). Nearly one-third (14,159/49,098, 28.84%) of the health care workers who completed a course were from LICs (828/49,098, 1.69%) or LMICs (13,331/49,098, 27.15%).

XSL•FO
RenderX

Table . Health care worker characteristics, by course and survey completion. This table compares the characteristics of health care workers who completed the focal courses and follow-up survey. A higher proportion of course completers did not specify characteristics compared to survey completers. Because response options for age and gender were voluntary, a number of learners did not specify these characteristics. We show the numbers not specified for each. For course completion, geographic region was identified via course platform analytics; however, we were unable to identify a subset, shown as "not specified" in the table. For survey completion, geographic regions were identified primarily through survey self-reports. In 177 survey responses, location was not reported. For these cases, we used the survey response's IP address to identify the geographic region of the respondent. Percentages are shown for those for whom we have data on characteristics. Percentage for each categorical variable sum to 100.

Characteristics		Completed course (n=49,098), n (%)	Completed survey (n=6272), n (%)	*P* value
Course type				
	OpenWHO	38,837 (79.1)	2214 (35.3)	<.001
	Stanford University	10,261 (20.9)	4058 (64.7)	<.001
Age range (y)				
	18 - 29	15,238 (31.04)	2020 (32.21)	<.001
	30 - 39	9699 (19.75)	1560 (24.87)	.10
	40 - 49	4511 (9.19)	950 (15.15)	<.001
	50 - 59	2324 (4.73)	662 (10.55)	<.001
	60 - 69	691 (1.41)	232 (3.7)	<.001
	70+	233 (0.47)	35 (0.56)	.56
	Not specified	16,402 (33.41)	813 (12.96)	—[a]
Gender				
	Female	20,252 (41.25)	3057 (48.74)	<.001
	Male	12,758 (25.98)	2349 (37.45)	<.001
	Nonbinary or other	139 (2.83)	43 (0.69)	<.001
	Not specified	15,949 (32.48)	823 (13.12)	—
Geographic region				
	East Asia and Pacific	5278 (10.75)	894 (14.25)	<.001
	Europe and Central Asia	5365 (10.93)	666 (10.62)	<.001
	Latin America and Caribbean	10,665 (21.72)	1061 (16.92)	<.001
	Middle East and North Africa	3816 (7.78)	547 (8.72)	.66
	North America	7019 (14.3)	993 (15.83)	.29
	South Asia	7264 (14.79)	1393 (22.21)	<.001
	Sub-Saharan Africa	3502 (7.13)	718 (11.45)	<.001
	Not specified	6189 (12.61)	0 (0)	—
World Bank income classification				
	High income	14,157 (28.83)	1971 (31.43)	.01
	Upper-middle income	14,593 (29.72)	1611 (25.69)	<.001
	Lower-middle income	13,331 (27.15)	2468 (39.35)	<.001
	Low income	828 (1.69)	222 (3.54)	<.001
	Not specified	6189 (12.61)	0 (0)	—

[a]Not applicable.

Table 2 also compares the characteristics of health care workers completing the course, with the 12.77% (6272/49,098) completing the survey. We observe slight differences in the age and gender composition of survey completers with course completers, with the survey sample skewing older and more male. The survey sample includes a slightly larger share of participants from LICs (222/6272, 3.54%) and LMICs (2468/6272, 39.35%).

XSL•FO
RenderX

Table 3 describes the professions of the health care workers who completed the survey and their levels of physician supervision. Physicians represent 34.61% (2171/6272) of the survey sample, followed by nurses (1599/6272, 25.49%) and allied health professionals (1190/6272, 18.97%). This breakdown of professional roles is similar in LICs and LMICs and in UMICs and HICs. Of the nonphysician health care workers, more than a third (1315/3639, 36.14%) reported having access to a physician for consultation during less than 50% of their workday, although the majority (1989/2341, 84.96%) could contact a physician by phone if needed. Most health care workers either already cared for patients with COVID-19 (2793/6272, 44.53%) or anticipated caring for them (1940/6272, 30.93%) at the time of survey completion.

Table . Characteristics of the health care worker survey sample. Allied health included physical therapy, occupational therapy, speech pathology, medical assistants, and home health aides. Nursing included nurses, nurse midwives, nursing instructors, and certified nursing assistants. The question about the frequency of physicians being on site was asked of nonphysicians only. The question about physicians being available via phone was asked of nonphysicians who had indicated that physicians were not available on site 100% of the time. Across questions asking about the availability of physician and treating patients with COVID-19, survey respondents could indicate that the question was not applicable in their health care setting.

Characteristics		Total (n=6272), n (%)	HICs[a] and UMICs[b] (n=3582), n (%)	LMICs[c] and LICs[d] (n=2690), n (%)
Profession				
	Allied health	1190 (18.97)	663 (18.51)	527 (19.59)
	Community health worker	501 (7.99)	296 (8.26)	205 (7.62)
	Nursing	1599 (25.49)	1012 (28.25)	587 (21.82)
	Physician assistant or nurse practitioner	103 (1.64)	68 (1.9)	35 (1.3)
	Paramedic or emergency medical technician	272 (4.34)	159 (4.44)	113 (4.2)
	Pharmacist	330 (5.26)	106 (2.96)	224 (8.33)
	Physician	2171 (34.61)	1217 (33.98)	954 (35.46)
	Traditional medicine	106 (1.69)	61 (1.7)	45 (1.67)
Frequency of physicians being on site [e]				
	Always (100% of time)	1228 (33.75)	660 (30.88)	568 (37.82)
	Mostly (>50% of time)	1096 (30.12)	586 (27.42)	510 (33.95)
	Sometimes (<50% of time)	815 (22.4)	482 (22.55)	333 (22.17)
	Never (0% of time)	500 (13.74)	409 (19.14)	91 (6.06)
Physicians being available via phone [f]				
	Yes	1989 (82.5)	1180 (48.94)	809 (33.55)
	No	352 (14.6)	256 (10.62)	96 (3.98)
	Not specified	70 (2.9)	41 (1.7)	29 (1.2)
Treating patients with COVID-19 [g]				
	Currently treating	2793 (44.53)	1551 (43.3)	1242 (46.17)
	Anticipated in future	1940 (30.93)	1003 (28)	937 (34.83)
	Not anticipated	460 (7.33)	314 (8.77)	146 (5.43)
	Not specified	1079 (17.2)	714 (19.93)	365 (13.57)

[a]HIC: high-income country.

[b]UMIC: upper-middle–income country.

[c]LMIC: lower-middle–income country.

[d]LIC: low-income country.

[e]This survey question was only asked to nonphysician health care workers who work directly with physicians (n=3639). Percentages shown are out of applicable participants only.

[f]This survey question was only asked to nonphysician health care workers that work directly with physicians and do not have a physician on site 100% of the time (n=2411). Percentages shown are out of data provided with applicable respondents only. Not all applicable respondents responded to this question (n=70).

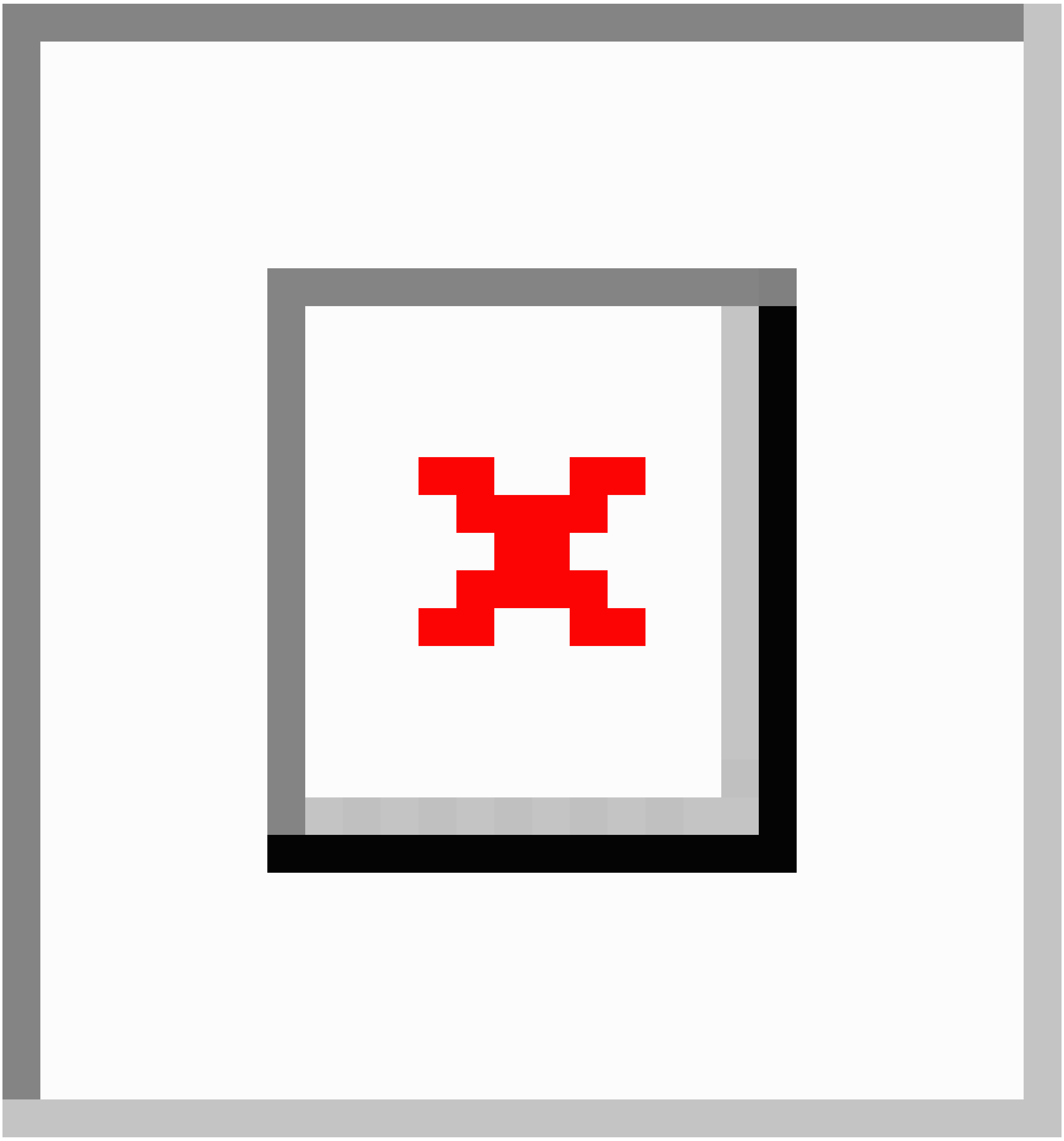
[g]Data on whether health care workers treat patients with COVID-19 were based on a voluntary question asked of patients at the time of course enrollment.

XSL•FO
RenderX

Learner Motivations

In the survey, health care workers were asked to rank in importance the following 6 potential motivating factors for course enrollment: to improve practice, to earn a certificate, CE, course brand, free cost of course, and employer recommendation. Figure 2 shows the ranking preferences across survey respondents. Among survey respondents ranking all factors (n=5518), the majority (n=3090, 56%) ranked “improve practice” as their top preference, with an additional 16% (n=883) ranking it as the second most important factor and 10% (n=552) ranking it as the third most important factor. The second and third most important factors were CE and to earn a certificate, with employer recommendation as the least most important factor ranked.

Figure 2. Percent of learners by motivation rank among health care providers (n=5518).



In Table 4, we show the ranking differences by the type of health care worker. Although the motivation of improving practice was ranked the highest across all subgroups, it was ranked higher by physicians, with a mean rank of 1.86, compared to nurses with a mean rank of 2.06 and other health care providers with a mean rank of 2.24. Nonphysicians ranked CE and employer recommendations higher than physicians. Certification also appears to matter more to nonphysicians, with 69.76% (2861/4101) choosing to obtain a certificate, 63.76% (2615/4101) providing a copy of the certificate to their employer, and 79.18% (3247/4101) using the certificate for a CE requirement. The course brand appears to be a more

XSL•FO
RenderX

important motivating factor to physicians compared to nonphysicians. Course cost did not appear to differentially influence course enrollment by the type of health care worker.

Table . Mean rank of motivation (1=highest rank, 6=lowest rank) and course perspectives by the type of health care worker. Physician is the reference category for comparisons. Nursing included nurses, nurses, midwives, and nursing assistants. Mean ranking does not include observations that skipped ranking altogether (n=745). Course perspectives include observations that skipped ranking but provided responses for these questions.

		Physician (n=2171), mean (SD)	Nursing (n=1599), mean (SD)	*P* value	Other (n=2502), mean (SD)	*P* value
Motivation (mean ranking)						
	Improve practice	1.86 (1.38)	2.06 (1.51)	<.001	2.24 (1.60)	<.001
	Earn certificate	3.52 (1.36)	3.53 (1.36)	.80	3.46 (1.42)	.16
	Continuing education requirement	3.63 (1.49)	3.31 (1.47)	<.001	3.46 (1.55)	<.001
	Course brand	3.58 (1.61)	4.17 (1.62)	<.001	3.92 (1.68)	<.001
	Course is free	3.83 (1.57)	3.77 (1.65)	.27	3.81 (1.61)	.68
	Employer recommended	4.66 (1.54)	4.47 (1.55)	.001	4.39 (1.61)	<.001
Course perspectives (proportion agreeing)						
	Would have taken course if it was not free	0.47 (0.50)	0.43 (0.50)	.01	0.46 (0.50)	.39
	Chose to obtain a certificate	0.63 (0.48)	0.71 (0.46)	<.001	0.69 (0.46)	<.001
	Gave a copy of the certificate to employer	0.55 (0.50)	0.65 (0.48)	<.001	0.63 (0.48)	<.001
	Will use the certificate for continuing education requirement	0.71 (0.45)	0.81 (0.39)	<.001	0.78 (0.41)	<.001

In Table 5, we show ranking differences by the location of health care workers, comparing differences in UMICs and HICs compared to LICs and LMICs. In LICs and LMICs, health care workers ranked CE and employer recommendation higher on average compared to learners in UMICs and HICs. Conversely, course brand appears to matter more for learners in UMICs and HICs. Certification was obtained by roughly the same proportion of learners in both subgroups, although learners in UMICs and HICs were more likely to give a copy of the certificate to their employer, whereas learners in LICs and LMICs were more likely to use the certificate for a CE requirement.

XSL•FO
RenderX

Table . Mean rank of motivation (1=highest rank, 6=lowest rank) and course perspectives by country classification. This table shows differences by World Bank income classifications: high-income country (HIC), upper-middle–income country (UMIC), lower-middle–income country (LMIC), and low-income country (LIC). Mean ranking does not include observations that skipped ranking altogether (n=745). Course perspectives include observations that skipped ranking but provided responses for these questions.

		HICs and UMICs (n=3582), mean (SD)	LICs and LMICs (n=2690), mean (SD)	*P* value
Motivation (mean ranking)				
	Improve practice	2.10 (1.52)	2.01 (1.49)	.04
	Earn certificate	3.45 (1.38)	3.57 (1.38)	.001
	Continuing education requirement	3.58 (1.54)	3.37 (1.48)	<.001
	Course brand	3.77 (1.65)	3.97 (1.66)	<.001
	Course is free	3.68 (1.59)	3.97 (1.61)	<.001
	Employer recommended	4.58 (1.58)	4.41 (1.57)	<.001
Course perspectives (proportion agreeing)				
	Would have taken course if it was not free	0.45 (0.50)	0.46 (0.50)	.61
	Chose to obtain a certificate	0.68 (0.47)	0.67 (0.47)	.22
	Gave a copy of the certificate to employer	0.65 (0.48)	0.57 (0.50)	<.001
	Will use the certificate for continuing education requirement	0.73 (0.44)	0.81 (0.39)	<.001

Generally, the fact that MOOCs were free was a lower-ranked motivator. Although interestingly, in the subgroup analysis, the course being free of cost was ranked lower in LICs and LMICs (mean 3.97, SD 1.61) than in UMICs and HICs (mean 3.68, SD 1.59; Table 5). However, when survey respondents were asked about their perspectives on the cost of MOOCs, more than half (3423/6272, 54.58%) of the health care workers indicated they would not have taken the course if there was an associated cost. This perspective was consistent across subgroup analyses of health care professional types and country-income levels.

Discussion

Principal Findings

Through a survey of 6272 health care workers worldwide who completed MOOCs for COVID-19 training across multiple platforms and organizations, our study provides unique insight into the factors motivating health care workers to enroll in and complete MOOCs during public health emergencies. We identified that the primary motivator for enrollment among health care workers was to improve their personal practice, followed by the pursuit of CE credit and certification. Course cost is an influential factor in the decision to enroll in an MOOC, with 54.58% (3423/6272) of respondents indicating that they would not have enrolled if the course had not been free. This first-of-its-kind analysis of health care worker motivations in just-in-time training MOOCs during a public health emergency fills an important gap in the existing literature, providing key insights for future course development and marketing.

Our findings highlight the widespread demand among health care workers for MOOC training during a public health crisis. Health care workers from over 200 countries and territories enrolled in and completed the COVID-19 MOOCs examined in this study, with a third (14,159/42,909, 33%) of course completers located in LICs and LMICs. Compared to the typical MOOC completion rates of under 10% [17,18], the 43.1% (49,098/113,902) completion rate among health care workers in the COVID-19 MOOCs in this study is notably high. Although the high rate of completion likely reflects the limited alternatives for training during the start of the COVID-19 pandemic, it may also indicate intrinsic motivation among health care workers, whose predominant reason for enrollment was to improve their personal practice.

We also observed that the COVID-19 MOOCs attracted a diverse range of health care providers globally. Although the majority (3770/6272, 60.11%) of respondents were nurses and physicians, 39.89% (2502/6272) reported working in other health care capacities including allied health, community health, emergency medical services, and pharmacy. Furthermore, we noted that motivations for enrollment varied by profession. Compared to physicians, nurses and other health care professionals were more motivated by CE credit, employer recommendations, and certification. Nurses and other health professionals were more likely to obtain certificates, provide a copy of the certificate to their employer, and use the certificates for CE requirements. Recognizing these differences in motivating factors across types of health care workers can inform the design of MOOCs that more effectively respond to the interests and needs of the targeted audience.

Despite these differences, the majority of all health care workers, including physicians, indicated their intention to use their certificates professionally, either by providing them to their

XSL•FO
RenderX

employers (3809/6272, 60.73%) or by earning CE credit (4788/6272, 76.34%). This finding underscores the potential for MOOCs to fill a gap in the CE arena, where traditional approaches often present barriers to completion. The common, traditional route for obtaining CE credits involves attendance at national or international medical conferences [7,8]; however, many such conferences were either canceled or transitioned to a web-based format during the pandemic. Given the time and travel requirements associated with conference attendance, MOOCs can serve as a viable and accessible alternative for learners. Interestingly, our study found that the use of course certificates for CE among learners in LICs and LMICs was higher than that in UMICs and HICs, which may reflect a lack of economically feasible options to earn CE credits in resource-limited geographies. Including certification in MOOC design may serve as an important motivator to increase enrollment and completion, particularly in LICs and LMICs, enhancing the attainment of timely health care education for the global health care workforce.

An additional benefit of online learning is the reduced cost for participants to obtain CE credits. Our study found that cost was a significant consideration for course participants, with 54.58% (3423/6272) of learners indicating they would not have taken the course if it had not been free. Although the course being free was slightly less important to learners in LICs and LMICs than those in UMICs and HICs, we speculate that in lower-income countries, learners with access to the technology required to participate in an online course may be relatively better off financially within their respective countries, and that those with lower incomes may not have the technology to enroll in the courses at all—only 3.54% (222/6272) of learners were from LICs. It is also possible that a single course participant may have shared access to the course with others.

Identifying the characteristics and motivations of specific groups of learners, such as those in LICs and LMICs, will aid in the design of future health care–related MOOCs to encourage participation and completion. Although many public health emergencies and disease outbreaks occur in LICs and LMICs with devastating impact, little data exist that examine the motivations of health care workers in these regions to enroll in just-in-time training MOOCs. Nevertheless, the WHO and various national health agencies frequently leverage MOOCs to disseminate critical health information during these emergencies. Future work should particularly investigate how to overcome barriers related to technology access and content accessibility with an eye toward equity, ensuring that the delivery of crucial health care worker training, particularly in times of emergency, is available to all. Likewise, future investigations should examine how online content is used and shared offline in contexts where the broader population has limited access to digital platforms, thereby enhancing the delivery of course materials through offline sharing.

Limitations

We recognize several methodological limitations inherent in our survey-based research. First, the potential for social desirability bias and selection bias due to voluntary participation limits the generalizability of our findings. To mitigate these biases, we deployed the survey across multiple learning platforms (Coursera, edX, and OpenWHO), each likely attracting different user demographics, and achieved a substantial sample size of 6272 respondents representing a diverse economic and geographic distribution. Additionally, we examined and reported only marginal differences between survey respondents and the overall course participants (as detailed in Table 2), although it remains a limitation that survey completers may not fully represent the broader learner population.

Second, the exclusive use of English for survey dissemination likely influenced the diversity of the respondents and further constrained the study's generalizability. Future studies could incorporate multiple language options to better capture a wider demographic.

Third, although the survey instrument was tailored to the specific contexts of the courses and discussed rigorously by experts across various fields—including educational assessment, emergency medicine, public health, and online learning—its lack of external validation presents a limitation. No prior studies identified during our review provided a validated instrument for assessing learner motivations in MOOCs, emphasizing the innovative aspect of our research while also necessitating a careful interpretation of our findings.

Fourth, our study's scope was restricted by the limitations in identifying patient-facing health care workers among enrollees, due to data collection methodologies on the OpenWHO platform until June 2020. This limitation hindered our capability to fully classify professions among participants. Future studies should aim to enhance the categorization of health care worker types and delve deeper into the differing motivations among these groups.

Finally, the dynamics of the COVID-19 pandemic—characterized by fluctuating case rates and mortality—suggest that motivations for enrolling in COVID-19–related MOOCs likely varied over time. Some health care workers might have enrolled early in anticipation of patient care needs, whereas others joined after gaining firsthand experience. This temporal variation in motivations, coupled with the evolving availability of other educational tools, presents a complex backdrop against which these motivations were formed. Future studies could benefit from aligning course enrollment data with local COVID-19 case trends to better understand these motivations.

Conclusion

Our study examined the motivations and characteristics of health care workers who engaged with MOOCs during the unprecedented COVID-19 health emergency. The analysis showed that the primary motivation for health care professionals was enhancing their personal practice. CE credit also proved to be a significant motivator, especially for those from LICs and LMICs. Additionally, the necessity of free access was clear, with more than half of the participants (3423/6272, 54.58%) indicating they would not have enrolled if fees were charged. These findings are important for the future development and deployment of MOOCs, ensuring that they not only are

XSL•FO
RenderX

accessible but also resonate with the intrinsic and extrinsic motivations of health care professionals from diverse geographic, training, and economic backgrounds. Future research should further investigate these motivations to see if they are consistent across different types and stages of health emergencies.

Acknowledgments

We would like to thank the World Health Organization (WHO) Health Emergencies Programme team, the Stanford Global Emergency Medicine (SEMI) team, and the Stanford Center for Health Education Digital Medic team for their work in developing and distributing the focal courses of this study. Additionally, we would like to thank all the learners who directly contributed to this research, especially the learners who participated in the survey and shared their experiences with our team.

Data Availability

Some data are available on reasonable request to the corresponding author.

Authors' Contributions

JJ and JSJ led the conceptualization and design of the study and oversaw all aspects of study implementation, writing, and editing. NYN and AT conducted the data collection. JSJ and SS conducted the quantitative analysis. NAS contributed to organizing and writing the manuscript. NYN, AT, NAS, HU, and MS contributed to the study's design, interpretation of findings, and revision of all drafts. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COVID-19 provider course: follow-up survey.
[DOCX File, 23 KB - mededu_v10i1e51915_app1.docx]

References

1. Utunen H, George R, Ndiaye N, Attias M, Piroux C, Gamhewage G. Responding to global learning needs during a pandemic: an analysis of the trends in platform use and incidence of COVID-19. Educ Sci 2020 Nov 22;10(11):345. [doi: 10.3390/educsci10110345]
2. Findyartini A, Greviana N, Hanum C, et al. Supporting newly graduated medical doctors in managing COVID-19: an evaluation of a massive open online course in a limited-resource setting. PLoS One 2021 Sep 10;16(9):e0257039. [doi: 10.1371/journal.pone.0257039] [Medline: 34506524]
3. Bendezu-Quispe G, Torres-Roman JS, Salinas-Ochoa B, Hernández-Vásquez A. Utility of massive open online courses (MOOCs) concerning outbreaks of emerging and reemerging diseases. . 2017 Sep 18 p. 1699. [doi: 10.5256/F1000RESEARCH.12639.D177854] [Medline: 29259764]
4. Bhattacharya S, Singh A, Hossain MM. Health system strengthening through massive open online courses (MOOCs) during the COVID-19 pandemic: an analysis from the available evidence. J Educ Health Promot 2020 Aug 31;9:195. [doi: 10.4103/jehp.jehp_377_20] [Medline: 33062728]
5. Liyanagunawardena TR, Aboshady OA. Massive open online courses: a resource for health education in developing countries. Glob Health Promot 2018 Sep;25(3):74-76. [doi: 10.1177/1757975916680970] [Medline: 28134014]
6. Regmi K, Jones L. A systematic review of the factors - enablers and barriers - affecting e-learning in health sciences education. BMC Med Educ 2020 Mar 30;20(1):91. [doi: 10.1186/s12909-020-02007-6] [Medline: 32228560]
7. Setia S, Tay JC, Chia YC, Subramaniam K. Massive open online courses (MOOCs) for continuing medical education - why and how? Adv Med Educ Pract 2019 Sep 11;10:805-812. [doi: 10.2147/AMEP.S219104] [Medline: 31572042]
8. Furtner D, Shinde SP, Singh M, Wong CH, Setia S. Digital transformation in medical affairs sparked by the pandemic: insights and learnings from COVID-19 era and beyond. Pharm Med 2022 Feb;36(1):1-10. [doi: 10.1007/s40290-021-00412-w] [Medline: 34970723]
9. Mahajan R, Gupta P, Singh T. Massive open online courses: concept and implications. Indian Pediatr 2019 Jun 15;56(6):489-495. [doi: 10.1007/s13312-019-1575-6] [Medline: 31278230]
10. King M, Pegrum M, Forsey M. MOOCs and OER in the Global South: problems and potential. The International Review of Research in Open and Distributed Learning 2018;19(5). [doi: 10.19173/irrodl.v19i5.3742]
11. Perryman LA, Hemmings-Buckler A, Seal T. Learning from TESS-India's approach to OER localisation across multiple Indian States. J Interact Media Educ 2014 Dec 23;2014(2):7. [doi: 10.5334/jime.af]

XSL•FO
RenderX

12. Bayeck RY. Exploratory study of MOOC learners' demographics and motivation: the case of students involved in groups. Open Praxis 2016 Jul 1;8(3):223-233. [doi: 10.5944/openpraxis.8.3.282]
13. Christensen G, Steinmetz A, Alcorn B, Bennett A, Woods D, Emanuel EJ. The MOOC phenomenon: who takes massive open online courses and why? SSRN. Preprint posted online on Apr 18, 2014. [doi: 10.2139/ssrn.2350964]
14. Zhong SH, Zhang QB, Li ZP, et al. Motivations and challenges in MOOCs with Eastern insights. Int J Inf Educ Technol 2016 Dec;6(12):954-960. [doi: 10.7763/IJIET.2016.V6.824]
15. Hew KF, Cheung WS. Students' and instructors' use of massive open online courses (MOOCs): motivations and challenges. Educ Res Rev 2014 Jun;12:45-58. [doi: 10.1016/j.edurev.2014.05.001]
16. Kizilcec RF, Piech C, Schneider E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Suthers D, Verbert K, Duval E, editors. LAK '13: Proceedings of the Third International Conference on Learning Analytics and Knowledge: Association for Computing Machinery; 2013:170-179. [doi: 10.1145/2460296.2460330]
17. Maya-Jariego I, Holgado D, González-Tinoco E, Castaño-Muñoz J, Punie Y. Typology of motivation and learning intentions of users in MOOCs: the MOOCKNOWLEDGE study. Educ Technol Res Dev 2020 Feb;68(1):203-224. [doi: 10.1007/s11423-019-09682-3]
18. Reich J, Ruipérez-Valiente JA. The MOOC pivot. Science 2019 Jan 11;363(6423):130-131. [doi: 10.1126/science.aav7958] [Medline: 30630920]
19. Kizilcec RF, Schneider E. Motivation as a lens to understand online learners: towards data-driven design with the OLEI scale. ACM Transac Comput Hum Int 2015 Mar 10;22(2):1-24. [doi: 10.1145/2699735]
20. Garrido M, Koepke L, Anderson S, Mena A, Macapagal M, Dalvit L. An examination of MOOC usage for professional workforce development outcomes in Colombia, the Philippines, & South Africa. Technology & Social Change Group, University of Washington Information School. 2016. URL: https://tascha.uw.edu/publications/an-examination-of-mooc-usage-for-professional-workforce-development-outcomes-in-colombia-the-philippines-south-africa/ [accessed 2024-05-30]
21. Nieder J, Nayna Schwerdtle P, Sauerborn R, Barteit S. Massive open online courses for health worker education in low- and middle-income countries: a scoping review. Front Public Health 2022 Jul 12;10:891987. [doi: 10.3389/fpubh.2022.891987] [Medline: 35903395]
22. Utunen H, Ndiaye N, Piroux C, George R, Attias M, Gamhewage G. Global reach of an online COVID-19 course in multiple languages on OpenWho in the first quarter of 2020: analysis of platform use data. J Med Internet Res 2020 Apr 27;22(4):e19076. [doi: 10.2196/19076] [Medline: 32293580]
23. COVID-19: training for healthcare workers. Stanford Online. URL: https://online.stanford.edu/courses/som-xche0007-covid-19-training-healthcare-workers [accessed 2021-06-15]
24. Top healthcare courses - learn healthcare online. Coursera. URL: https://www.coursera.org/search?query=healthcare [accessed 2020-11-22]

Abbreviations

CE: continuing education
HIC: high-income country
LIC: low-income country
LMIC: lower-middle–income country
MOOC: massive open-source online course
UMIC: upper-middle–income country
WHO: World Health Organization

Edited by B Lesselroth; submitted 16.08.23; peer-reviewed by A Bahattab, M Stoto, MN Shalaby, ZU Haq; revised version received 26.04.24; accepted 19.05.24; published 19.06.24.

Please cite as:
Jones J, Johnston JS, Ndiaye NY, Tokar A, Singla S, Skinner NA, Strehlow M, Utunen H
Health Care Workers' Motivations for Enrolling in Massive Open Online Courses During a Public Health Emergency: Descriptive Analysis
JMIR Med Educ 2024;10:e51915
URL: https://mededu.jmir.org/2024/1/e51915
doi:10.2196/51915

© Jennifer Jones, Jamie Sewan Johnston, Ngouille Yabsa Ndiaye, Anna Tokar, Saumya Singla, Nadine Ann Skinner, Matthew Strehlow, Heini Utunen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.6.2024. This is an

XSL•FO
RenderX

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.

XSL•FO
RenderX

The Use of Animations Depicting Cardiac Electrical Activity to Improve Confidence in Understanding of Cardiac Pathology and Electrocardiography Traces Among Final-Year Medical Students: Nonrandomized Controlled Trial

Alexandra M Cardoso Pinto[1,*], BSc; Daniella Soussi[1,*], BSc; Subaan Qasim[1], BSc; Aleksandra Dunin-Borkowska[1], BSc; Thiara Rupasinghe[2], BSc; Nicholas Ubhi[3], MBBS; Lasith Ranasinghe[4], BSc, MBBS

1

2

3

4

*these authors contributed equally

Corresponding Author:
Alexandra M Cardoso Pinto, BSc

Abstract

Background: Electrocardiography (ECG) interpretation is a fundamental skill for medical students and practicing medical professionals. Recognizing ECG pathologies promptly allows for quick intervention, especially in acute settings where urgent care is needed. However, many medical students find ECG interpretation and understanding of the underlying pathology challenging, with teaching methods varying greatly.

Objective: This study involved the development of novel animations demonstrating the passage of electrical activity for well-described cardiac pathologies and showcased them alongside the corresponding live ECG traces during a web-based tutorial for final-year medical students. We aimed to assess whether the animations improved medical students' confidence in visualizing cardiac electrical activity and ECG interpretation, compared to standard ECG teaching methods.

Methods: Final-year medical students at Imperial College London attended a web-based tutorial demonstrating the 7 animations depicting cardiac electrical activity and the corresponding ECG trace. Another tutorial without the animations was held to act as a control. Students completed a questionnaire assessing their confidence in interpreting ECGs and visualizing cardiovascular electrical transmission before and after the tutorial. Intervention-arm participants were also invited to a web-based focus group to explore their experiences of past ECG teaching and the tutorial, particularly on aspects they found helpful and what could be further improved in the tutorial and animations. Wilcoxon signed-rank tests and Mann-Whitney *U* tests were used to assess the statistical significance of any changes in confidence. Focus group transcripts were analyzed using inductive thematic analysis.

Results: Overall, 19 students attended the intervention arm, with 15 (79%) completing both the pre- and posttutorial questionnaires and 15 (79%) participating in focus groups, whereas 14 students attended the control arm, with 13 (93%) completing both questionnaires. Median confidence in interpreting ECGs in the intervention arm increased after the tutorial (2, IQR 1.5-3.0 vs 3, IQR 3-4.5; $P<.001$). Improvement was seen in both confidence in reviewing or diagnosing cardiac rhythms and the visualization of cardiac electrical activity. However, there was no significant difference between the intervention and control arms, for all pathologies (all $P>.05$). The main themes from the thematic analysis were that ECGs are a complex topic and past ECG teaching has focused on memorizing traces; the visualizations enabled deeper understanding of cardiac pathology; and ECG learning requires repetition, and clinical links remain essential.

Conclusions: This study highlights the value of providing concise explanations of the meaning and pathophysiology behind ECG traces, both visually and verbally. ECG teaching that incorporates relevant pathophysiology, alongside vignettes with discussions regarding investigations and management options, is likely more helpful to students than practices based solely on pattern recognition. Although the animations supported student learning, the key element was the tutor's explanations. These animations may be more helpful as a supplement to teaching, for instance, as open-access videos.

(JMIR Med Educ 2024;10:e46507) doi:10.2196/46507

XSL•FO
RenderX

KEYWORDS

medical education; cardiology; technology; clinical skills; cardiac; cardiac electrical activity; ECG; mixed methods study; students; education; medical professionals; development; web-based tutorial; teaching; cardiovascular; learning; electrocardiography

Introduction

Electrocardiography (ECG) interpretation is a fundamental skill necessary during medical school education and in the practice of clinical medicine and surgery. Recognizing pathologies such as an ST elevation or non-ST elevation myocardial infarction; bundle branch block (BBB); and arrhythmias, including atrial fibrillation (AF), supraventricular tachycardia, ventricular fibrillation, and ventricular tachycardia, allows for prompt intervention and improves patient care, especially in the acute setting where urgent interventions may be lifesaving.

Despite its importance, medical students struggle with interpreting ECGs, and teaching methods seem to vary greatly between systematic interpretation based on ECG segments and pattern recognition. This leads to a lack of confidence and inaccurate interpretation of ECGs, which could lead to adverse events including treatment delay or incorrect management of the pathology. Although automated computer interpretation may be available, this should not be used by itself to diagnose conditions, since clinical correlation is warranted and these algorithms may be inaccurate [1].

The major disruptions caused by the COVID-19, in which medical students had limited exposure to hospital wards and experienced most lectures and tutorials on the web rather than in person [2], served as a strong reminder of the need for investment in innovative teaching methods.

Literature already suggests that teaching should be focused on the understanding of lead placement, as well as the basics of electrophysiology and ECG, to better identify abnormalities [3]. Teaching should also be correlated with the clinical findings of a case, as this has been shown to lead to more accurate ECG diagnosis in practice. For instance, case-based learning (CBL) has been frequently used in recent years, which increases the practical knowledge and confidence of medical students and junior doctors, through clinically correlating various cardiac pathologies [4]. However, explaining the link between the underlying cardiac pathology and the traces demonstrated by the ECG is not common practice in medical school curricula.

Medical students repeatedly describe ECG interpretation as a challenging skill [5,6]. A study based in Israel reports that despite competence and confidence in ECG interpretation improving throughout medical school, levels remain low among final-year medical students [5]. Moreover, a study of Polish medical students highlighted students' lack of ability to recognize common and emergency cardiac pathologies [6]. Additionally, these results emphasized independent learning as the strongest predictor of competency, as opposed to attendance in formal teaching sessions [6]. This continues following medical school, with rates of accurate ECG interpretation being as low as 55.8% among trainee doctors [7]. These findings suggest the need for a review of ECG teaching methods.

Technology-enhanced learning has grown in popularity and has been trialed as a method to encourage active practice of ECG interpretation among medical students. Students in this cohort demonstrated better diagnostic accuracy, but rates of knowledge attrition 6 months after the study remained high [1]. These findings highlight that despite continued practice remaining important, current methods of teaching ECGs do not support students in gaining in-depth understanding, nor do they enable knowledge retention.

Methods of technology-enhanced education, including visualizations, have been trialed extensively for anatomy teaching [8], with reported increases in student engagement with the content [9]. Although greater engagement does not ensure improved understanding, it may be an important component in supporting effective teaching and learning [8]. Additionally, there is evidence to show that visualization tools are capable of supporting students' understanding of anatomy [10].

Therefore, this research team developed novel animations demonstrating the passage of electrical activity through the heart for different pathologies and showcased them alongside the corresponding live ECG traces during a web-based ECG tutorial for final-year medical students at Imperial College London. The aim of this study was to assess whether these animations are associated with the improvement of final-year medical students' confidence in both visualizing cardiac pathology and interpreting the corresponding ECGs, compared to standard ECG teaching methods that do not involve visual animations.

Methods

The study was designed as a nonrandomized controlled trial.

Recruitment

Year 6 Bachelor of Medicine, Bachelor of Surgery students from Imperial College London were invited to participate in the study. Messages were sent through student communication channels with the description of the study and tutorial. This included a link to the study information sheet as well as a sign-up link to register their interest in participating. The first 20 students to sign up were emailed by a member of the research team with the information sheet and focus group consent form attached. Students were asked to confirm their participation by returning the signed consent form via email. Students were given a week to confirm their participation, after which the space would be offered to others who registered interest until a total of 20 confirmations were reached. The process was repeated in the following year, with a new cohort of Year 6 Bachelor of Medicine, Bachelor of Surgery students at the same point in the academic year as the original cohort.

The sample size of 20 students per teaching session was agreed by the research team based on the tutor's preference, following their experience of what would be a feasible number of students to teach within the agreed timeframe. This decision was also

XSL•FO
RenderX

supplemented by evidence to suggest that cohorts of fewer than 30 students may enable better learning [11] and that cohorts of approximately 19 students may enable greater interaction [12].

Design and Delivery of the Tutorial

Prototypes for the ECG traces and animations were created on Microsoft PowerPoint by 2 junior doctors on the research team. A total of 7 ECG patterns (sinus rhythm, AF, atrial flutter, atrioventricular nodal re-entry tachycardia, atrioventricular re-entry tachycardia, right BBB, and left BBB) that are known to commonly arise in clinical practice and in exams were chosen, and the prototypes were converted into high production value animations using Adobe Illustrator and Adobe After Effects. The animations were produced by skilled members of the research team and took a collective total of 10 hours to produce. The final product consisted of a video animation of the electrical activity passing through the heart alongside an ECG rhythm strip (lead II) for the given abnormality—with the exception of BBBs, which were depicted alongside leads V1 and V6. The animation of electrical activity through the heart and the corresponding ECG trace were synchronized to demonstrate how each ECG deflection corresponds with the electrical activity within the heart. Depolarization was shown in yellow and repolarization was shown in green.

The tutorials were both delivered by a UK-based Academic Foundation doctor within the research team (LR) on Zoom (Zoom Video Communications) at a prespecified time on a weekday evening. LR has vast experience teaching medical students and designing medical educational material and had completed the Membership of the Royal Colleges of Physicians of the United Kingdom Part 1 Examination successfully at the time of delivering the sessions.

Participants logged on using their unique identifier code and kept their cameras turned off to maintain anonymity. The intervention tutorial involved going through each animation in turn and narrating the path of the electrical activity. To ensure the smooth running of the event, questions were reserved until the end of the session.

The control tutorial followed the same lesson plan as the intervention but involved the tutor narrating the path of electrical activity using an example 12-lead ECG without any animations. Participants were not explicitly told this tutorial would be the control arm but were instead invited to a standard ECG tutorial, following the same methods as the intervention. However, all participants would have read the study information sheet and known the aim of the study, which may have compromised the single-blinding process.

Questionnaires

An email was sent to participants 1 week prior to the tutorial with a link to an anonymous Qualtrics questionnaire to be completed before the tutorial (Multimedia Appendix 1). This questionnaire was composed of 5-point Likert-scale questions assessing participants' confidence in interpreting ECGs and visualizing cardiovascular electrical transmission in each of the cardiovascular pathologies covered in the tutorial. The questionnaire also included multiple-choice and free-text questions inquiring about previous formats of ECG teaching experienced by participants and their views on what could be improved about current ECG teaching generally.

A similar questionnaire was repeated at the end of the tutorial to assess change in confidence using the same 5-point Likert-scale questions, as well as free-text questions inquiring about participants' experience of the tutorial (Multimedia Appendix 1). A link and QR code to this questionnaire was shared at the end of the tutorial, prior to the start of the focus groups.

Participants were given a unique participant code, which they were asked to state at the start of each questionnaire. This enabled questionnaire responses to be paired while maintaining anonymity.

Focus Groups

Focus groups were only conducted at the end of the intervention tutorial. Participants who took part in the control arm were not invited for a focus group, as the primary purpose of this exercise was to understand participants' experience of the visual animations, which were not included in the control arm. Upon the completion of the tutorial, students were divided into 4 breakout rooms on Zoom, each designed to host 5 students and a single researcher. Participants were asked to unmute microphones to participate in the semistructured focus group and were invited to keep their cameras switched off if they wished to remain anonymous. The focus groups further explored participants' experiences of past ECG teaching and the current tutorial, with particular focus on aspects they found helpful and what could be further improved in the delivery of the tutorial and design of the visualizations (Multimedia Appendix 2).

Focus group questions were designed collaboratively by the research team, with feedback from an expert qualitative researcher (see the *Acknowledgments* section), to ensure that the questions were adequate in informing the study's aims and gave participants the opportunity to share their experiences of ECG learning openly. These were also reviewed by the ethics committee (see the *Ethical Considerations* section).

Focus groups were audio and video recorded on the platform. Recordings were deleted upon transcription, which took place within 2 weeks of the tutorial. Participants were asked if they wished to receive a copy of the transcription to review their statements (anonymized using their unique participant codes); those who asked for the transcription were sent the transcript by email and given 1 week to inform the research team of any redactions they wished to make.

Data Analysis

Questionnaire data were analyzed using descriptive statistics on Microsoft Excel. The Shapiro-Wilk test was used to determine the distribution of data. As this showed that the data were nonparametric, Wilcoxon signed-rank tests were used to assess the statistical significance of any reported changes in confidence between pre- and posttutorial questionnaire responses, for each of the intervention and control arms. The Mann-Whitney *U* test was used to compare differences in pre- and posttutorial confidence between the intervention and control arms of the study.

XSL•FO
RenderX

Focus group transcripts were analyzed using inductive thematic analysis, following Braun and Clarke's [13] stages of thematic analysis as guidance. This was done on NVivo 12.0 (Lumivero) by 2 researchers cooperatively. Themes were reviewed by a third researcher. Free-text questions from the questionnaires were analyzed following similar methods on Microsoft Excel.

Ethical Considerations

This study was approved by the Imperial College Education Ethics Review committee (EERP2122-086). Participation in questionnaires and focus groups was voluntary, with participants given the option to withdraw from the study at any point, up until 2 weeks following the completion of the postintervention questionnaire. All participants were provided with a study information sheet prior to confirming their consent for participation in the study. There was no financial compensation involved in this study. Information sheets explained the aim of the study, methods of data storage, and outputs. Participants were also provided with a unique identifier code generated by the research team to be placed at the start of the questionnaires, enabling data to remain paired while ensuring anonymity.

Results

Questionnaire Results

The first 20 students who signed up to participate in each tutorial were allocated a slot.

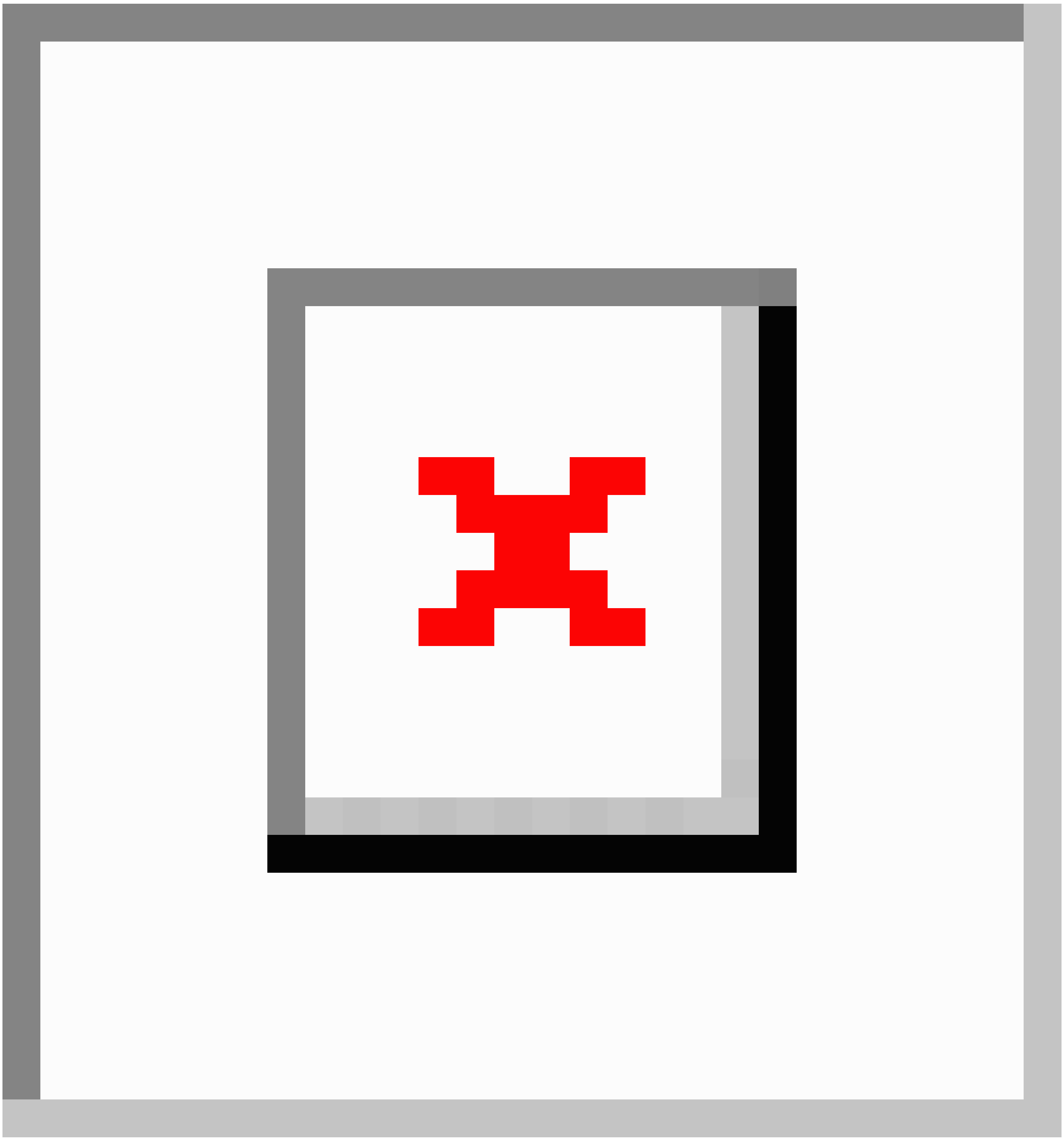
In the intervention tutorial, a total of 19 students attended. Of these, 15 (79%) completed both the pre- and posttutorial questionnaires.

All participants confirmed at least 1 prior method of ECG teaching, including didactic lectures (13/15, 87%), case-based tutorials (11/15, 73%), memorization of ECG features (9/15, 60%), animations (2/15, 13.3%), and practical sessions (1/15, 7%).

Overall, in the intervention group, median results for confidence in interpreting ECGs increased from the pretutorial scores (2, IQR 1.5-3) to the posttutorial scores (3, IQR 3-4.5; $P<.001$). Improvement was seen in both confidence in reviewing and diagnosing cardiac rhythms (Figure 1) and in visualizing electrical activity throughout the heart (Figure 2), across most of the pathologies illustrated.

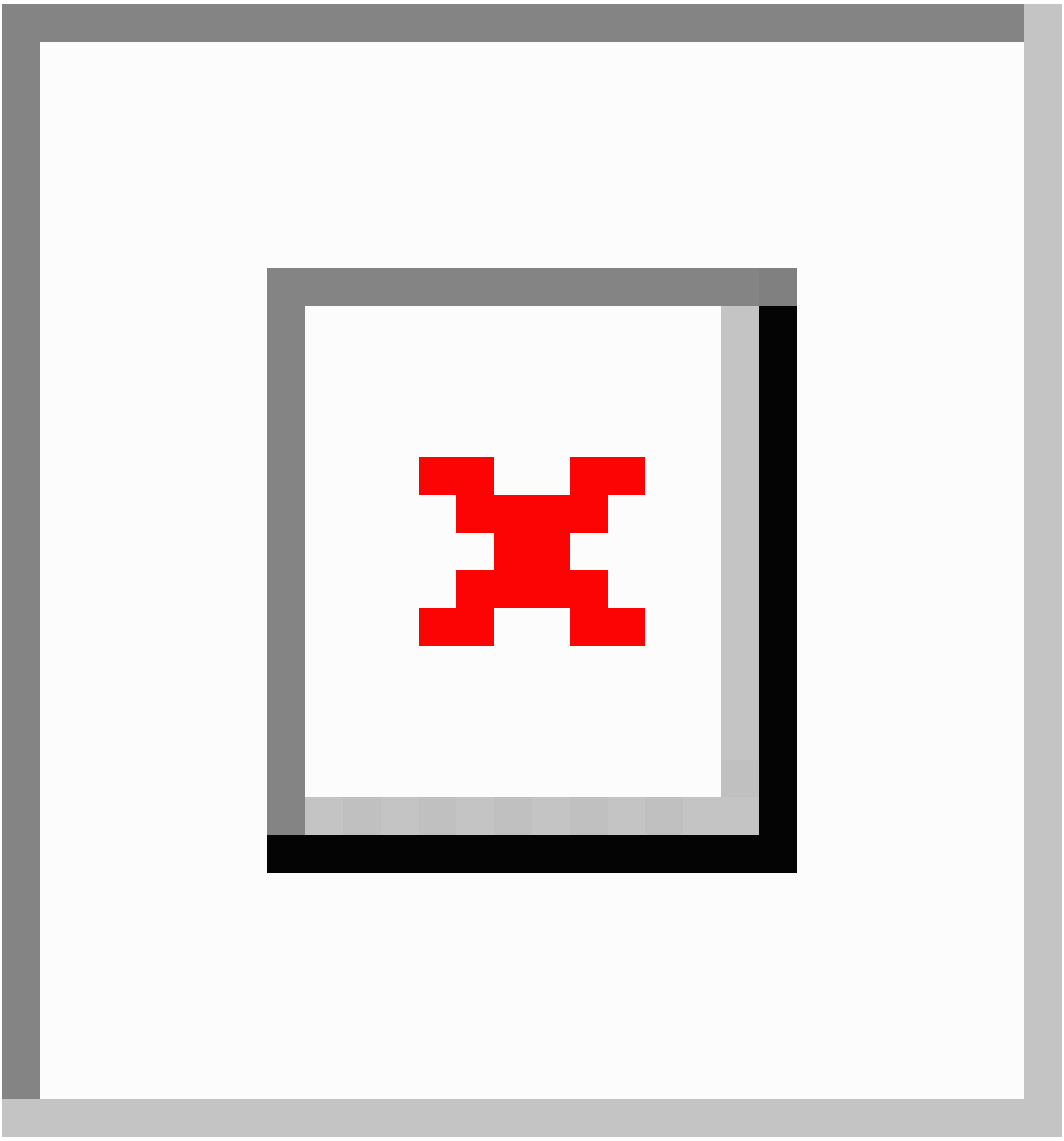
XSL•FO
RenderX

Figure 1. Confidence in reviewing and diagnosing cardiac rhythms and pathology on an ECG (median and IQR score on a Likert scale, from a 1=not confident at all to 5=extremely confident) for the intervention group (n=15). Wilcoxon signed-rank test results: *$P \leq .05$, **$P \leq .01$, and ***$P \leq .001$. AVNRT: atrioventricular nodal re-entry tachycardia; AVRT: atrioventricular re-entry tachycardia; ECG: electrocardiography; LBBB: left bundle branch block; RBBB: right bundle branch block.



XSL•FO
RenderX

Figure 2. Confidence in visualizing electrical activity through the heart in different ECG pathologies (median and IQR score on a Likert scale, from a 1=not confident at all to 5=extremely confident) for the intervention group (n=15). Wilcoxon signed-rank test results: **P*≤.05, ** *P*≤.01, and *** *P*≤.001. AVNRT: atrioventricular nodal re-entry tachycardia; AVRT: atrioventricular re-entry tachycardia; ECG: electrocardiography; LBBB: left bundle branch block; RBBB: right bundle branch block.



Participants showed the least confidence in reviewing and diagnosing ventricular pathologies compared to atrial pathologies, with atrioventricular nodal re-entry tachycardia and atrioventricular re-entry tachycardia scoring the lowest median confidence scores before and after the tutorial but also showing the greatest levels of improvement following the tutorial (Figure 1; Table S1 in Multimedia Appendix 3).

A similar pattern is observed for median scores in visualizing cardiac electrical activity (Figure 2; Table S2 in Multimedia Appendix 3). However, median confidence levels before the tutorial in this category were, overall, lower than the same measurement of confidence for reviewing and diagnosing cardiac rhythms. Nevertheless, the level improvement in confidence in visualizing electrical activity in the heart was overall greater than that in confidence in reviewing and diagnosing cardiac rhythms, with median posttutorial confidence levels also achieving higher levels in most cardiac pathologies than those for reviewing and diagnosing cardiac rhythms.

Participants reported greater enjoyment of this tutorial (median 4, IQR 3-4.5) compared to past ECG teaching (median 3, IQR 1.5-3; *P*=.02).

XSL•FO
RenderX

For the control arm of the study, a total of 14 students attended the tutorial, of which 13 (93%) completed both the pre- and posttutorial questionnaires.

Prior methods of ECG teaching were similar to those of the intervention group, including didactic lectures (11/13, 85%), case-based tutorials (10/13, 77%), memorization of ECG features (6/13, 46%), practical sessions (3/13, 23%), animations (2/13, 15%), and a website with example ECGs for self-learning (1/13, 8%).

Overall confidence in interpreting ECGs showed only slight improvement in the control group, from a median of 3 (IQR 2-3) to 3 (IQR 3-4; *P*=.01).

Pretutorial confidence scores were similar in the control and intervention arms (Table 1). For the control group, pretutorial median confidence scores were also lower for ventricular pathologies compared to atrial pathologies, and overall confidence scores for reviewing and diagnosing cardiac pathologies were higher than visualizing cardiac activity, which is similar to the pattern seen in the intervention group (Multimedia Appendix 4).

Table . Pretutorial median (IQR) confidence scores for control and intervention groups with *P* values (Mann-Whitney *U* Test).

Scores and pathologies		Intervention, median (IQR)	Control, median (IQR)	*P* value
Confidence in reviewing and diagnosing cardiac rhythms and pathology on an ECG[a]				
	Sinus rhythm	4.0 (3.4-4.0)	5.0 (3.9-5.0)	.07
	Atrial flutter	3.0 (1.8-3.8)	3.3 (3.9-5.0)	.35
	Atrial fibrillation	3.7 (2.4-4.1)	4.0 (3.0-4.0)	.19
	AVNRT[b]	1.5 (1.1-2.1)	2.0 (1.3-2.5)	.32
	AVRT[c]	1.5 (1.1-2.3)	2.0 (1.5-2.5)	.39
	RBBB[d]	2.2 (1.4-2.9)	2.5 (2.0-3.5)	.28
	LBBB[e]	2.2 (1.4-2.9)	2.5 (2.0-3.5)	.33
Confidence in visualizing electrical activity on an ECG				
	Sinus rhythm	3.0 (2.5-4.5)	4.0 (2.0-4.0)	.59
	Atrial flutter	2.0 (1.0-3.0)	2.0 (1.0-2.0)	>.99
	Atrial fibrillation	3.0 (1.5-3.0)	2.0 (1.0-4.0)	.98
	AVNRT	2.0 (1.0-3.0)	2.0 (1.0-3.0)	.94
	AVRT	2.0 (1.0-3.0)	2.0 (1.0-2.0)	.68
	RBBB	1.0 (1.0-2.0)	2.0 (1.0-4.0)	.18
	LBBB	1.0 (1.0-2.0)	2.0 (1.0-4.0)	.18
Overall confidence in interpreting ECGs		2.0 (1.0-3.0)	3.0 (2.0-3.0)	.50

[a]ECG: electrocardiography.

[b]AVNRT: atrioventricular nodal re-entry tachycardia.

[c]AVRT: atrioventricular re-entry tachycardia.

[d]RBBB: right bundle branch block.

[e]LBBB: left bundle branch block.

There was no statistically significant difference between the enjoyment of this tutorial (median 4, IQR 4-5) compared to past ECG teaching (median 4, IQR 3-4; *P*=.052).

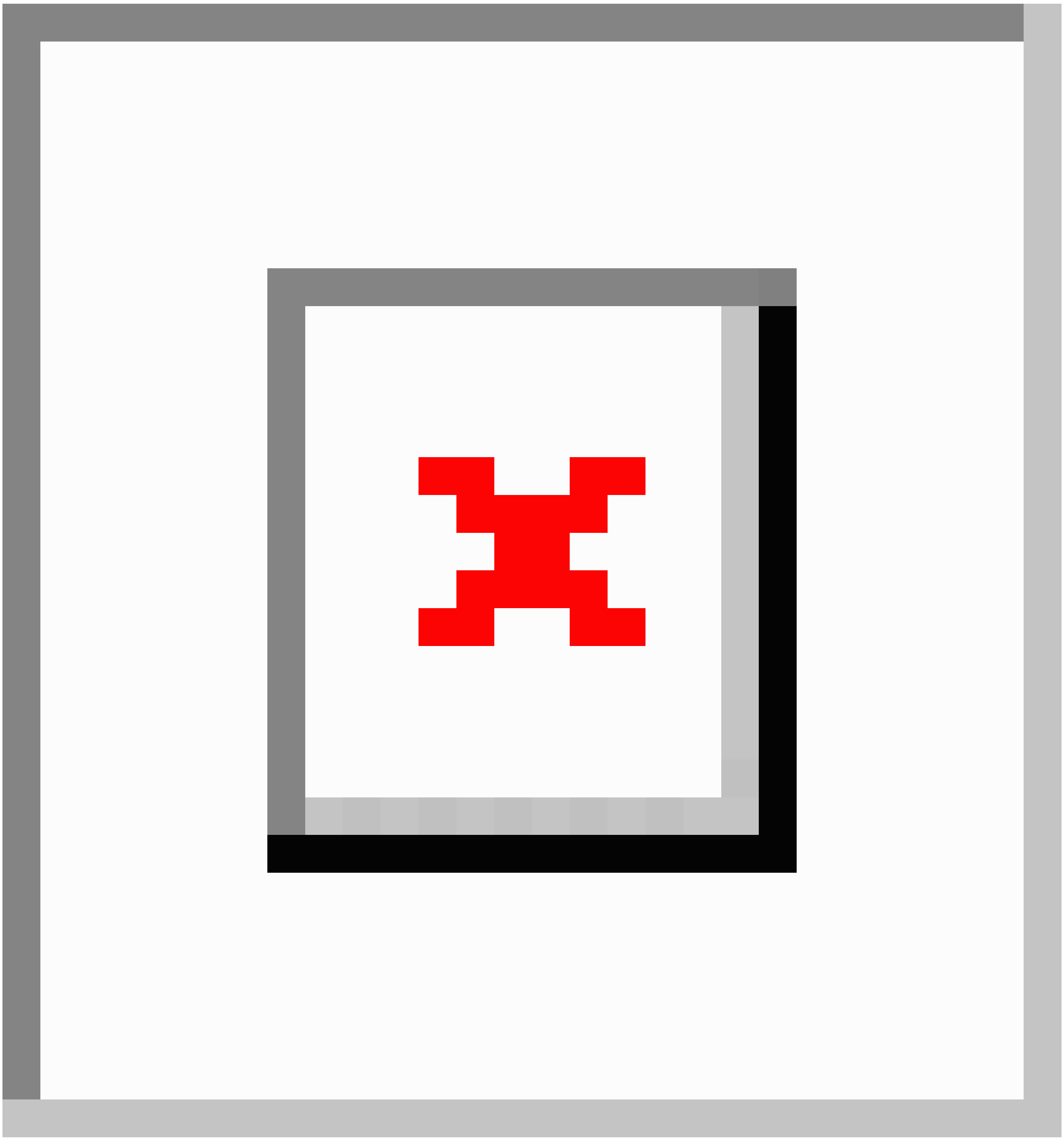
When comparing the change in confidence between the control and intervention groups for both reviewing and diagnosing pathology and visualizing electrical activity, no statistically significant difference was seen across all pathologies (all *P*>.05).

Data for confidence in reviewing and diagnosing cardiac rhythms and pathology showed greater improvements in the intervention group across most pathologies, except for AF. The greatest absolute difference between the intervention and control groups was seen for left BBB, although this was still statistically nonsignificant (*P*=.89; Figure 3). Data for confidence in visualizing cardiac electrical activity showed similar median changes in confidence across most pathologies, apart from right and left BBBs, where the intervention group showed greater improvement, although not statistically significant (*P*=.15 and *P*=.12, respectively; Figure 4).

There was also no statistically significant difference in median scores for the enjoyment of the tutorial when comparing control and intervention groups (*P*=.37).

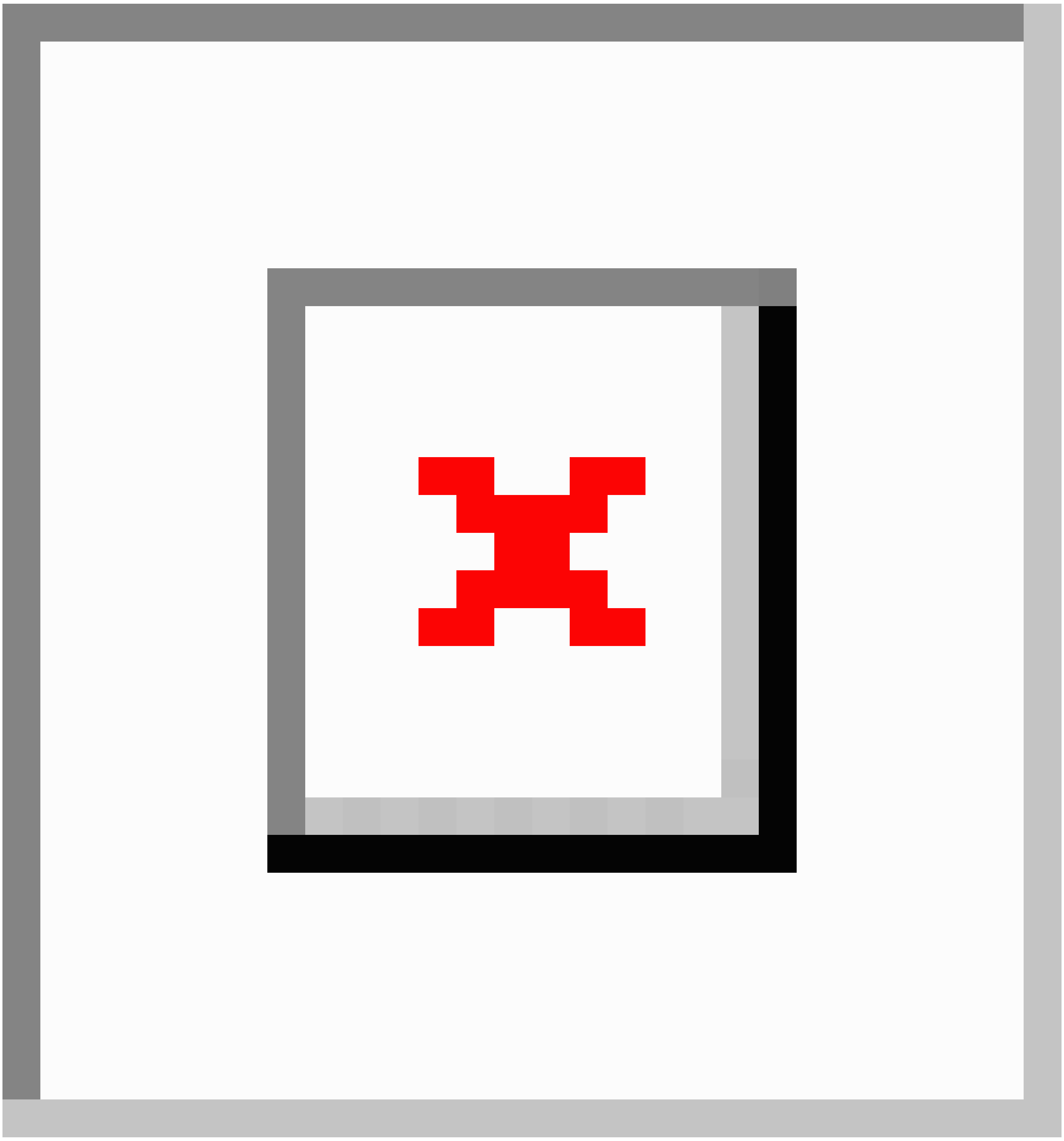
XSL•FO
RenderX

Figure 3. Median (IQR) scores for absolute difference in confidence in reviewing and diagnosing cardiac rhythms and pathology on an ECG for control (n=13) and intervention (n=15) groups. AVNRT: atrioventricular nodal re-entry tachycardia; AVRT: atrioventricular re-entry tachycardia; ECG: electrocardiography; LBBB: left bundle branch block; RBBB: right bundle branch block.



XSL•FO
RenderX

Figure 4. Median (IQR) scores for absolute difference in confidence in visualizing electrical activity through the heart in different ECG pathologies for control (n=13) and intervention (n=15) groups. AVNRT: atrioventricular nodal re-entry tachycardia; AVRT: atrioventricular re-entry tachycardia; ECG: electrocardiography; LBBB: left bundle branch block; RBBB: right bundle branch block.



Focus Group Results

Overview

A total of 15 (79%) out of 19 participants who attended the intervention tutorial took part in the focus groups. These were preallocated at random into 4 separate groups, which contained between 2 to 5 students (1 with 5 students, 2 with 4 students, and 1 with 2 students). Three key themes emerged from the analysis of focus group transcripts.

Past ECG Learning Has Been Centered on the Clinical Context and Memorizing Traces

All participants noted varied past ECG teaching, including formal lectures and tutorials focused on the principles of ECG interpretation throughout medical school but also informal teaching while on placement. However, there was an agreement that ECGs remained a challenging concept to learn. For example, one participant noted that they “found [ECGs] hard to understand and engage [with]” (participant 8), whereas another explained how “an ECG can kind of be a different language almost” (participant 7) and hence may take more time and effort to understand.

XSL•FO
RenderX

Past teaching experienced by participants, particularly informal teaching on placement, also focused on pattern recognition and correlating ECG signs to diagnoses.

> *I think at least with the ones I went through with the doctors and stuff, it was very much like a tick box or like oh the saw tooth pattern is this, this is this…* [Participant 12]

Although this format of learning was concise and focused on key knowledge required to be a Foundation Year 1 doctor, it did not promote deeper understanding of ECGs that could be applied to any ECG pattern.

> *I'd kind of leave knowing that if that exact ECG comes up, that was helpful, but otherwise I don't know really what or why it is that and then some of the actual understanding came from doing work outside of firms.* [Participant 30]

Participants also highlighted that "the key thing is kind of just repetition" (participant 7) when learning to interpret ECGs, and that "you also have to dedicate time yourself to go through it, if you really want to properly understand it" (participant 33).

Being taught systematic methods for ECG interpretation and presentation was reported to be useful; namely, it was "more relevant to us and our exams and practicals" (participant 15) when going through clinical cases alongside ECGs, which help provide clinical context to the ECG and "also gets you used to different subtleties, because between patients an ECG of the same condition can look slightly different" (participant 6).

The Animations and Associated Explanations Promoted a Deeper Understanding of Cardiac Electrical Activity

Overall, participants found the animations and accompanying explanations during the tutorial to be a helpful tool. The depiction of an ECG trace and heart animation simultaneously helped them understand the correlation between the 2, and hence, as a participant stated, "the first time I've properly understood what is exactly is going on [in the heart]" (participant 1) for the pathologies illustrated. One participant highlighted that "breaking it down into basics…and how it's reflected in the heart as well as it's corresponding trace…feels less like I'm trying to memorise something and more like I'm actually trying to figure it out" (participant 17).

The visual nature of these animations enabled participants to "clearly see how the electricity is conducted in the heart" (participant 2) and was noted to be an effective method to "consolidate what I know about the conditions that we went through" (participant 2).

This more thorough level of understanding was noted to be "quite useful" to participants, as "when going on the wards and I see an ECG, I can actually visualise how the heart is functioning" (participant 17). It was also perceived to be a helpful way of retaining their learning about heart pathologies and associated ECG traces in the longer-term, as "when you understand the reason why something is the way it is you are more likely to remember it" (participant 15).

Participants also stated the value of covering content that they considered relevant to their exams and starting work as Foundation Year 1 doctors: "I enjoyed the fact that we covered like a lot of a main conditions, so less of the more niche stuff" (participant 7). They also described this tool as more of a helpful "recap" (participant 2) of heart conditions and their associated ECG traces, as opposed to methods of ECG interpretation, which are often the focus in later years of medical school. One participant explained, "I just wish we were taught this way before [in earlier years of medical school]; it would make understanding a lot easier later" (participant 12).

Implementing This 1-Hour Tutorial Is Not Enough: ECG Learning Requires Repetition and Clinical Links Remain Essential

The key differentiating component of this tutorial was its animations: "I'm a visual learner, so I need to see it to understand it. So that's what's been a gamechanger for me, to actually see the animation" (participant 32). However, participants suggested that "having [the animation playing] even slower" (participant 8) or the opportunity to independently "use the scroller to advance" (participant 32) through the animation would be helpful to visualize more carefully "what is happening step-by-step in the heart and on the ECG" (participant 32). One participant also suggested potential value in "3D animations, that would be useful so you can turn the heart around and see all the fibres and all the [conduction activity]" (participant 17).

Despite the value of the tutorial in supporting students' understanding of cardiac pathologies, participants highlighted that there are additional factors that are important in contributing to in-depth learning. For instance, the need for repetition was widely acknowledged. Participants therefore asked that animations be made available for them to view independently. Additionally, the fast-paced nature of the tutorial, which covered multiple pathologies, means that some participants "didn't really have much of a time to get an understanding again, of like the condition" (participant 7), which might be resolved through independent revision with the animations or delivering the content through multiple teaching sessions.

Finally, participants noted the value of greater interaction with the audience, including the implementation of quizzes to test understanding and the integration of clinical cases for stronger clinical correlations.

Discussion

Overall, results for the intervention cohort demonstrate a statistically significant improvement in confidence when identifying abnormalities in ECG traces and visualizing cardiac electrical activity, compared to prior to attending this tutorial. However, a similar improvement was seen in the control group, with no statistically significant differences in improvement in confidence between the control and intervention groups. Although the focus groups highlighted a possible value in the use of animations demonstrating cardiac electrical activity synchronized to the corresponding ECG trace, the overall results suggest that perhaps this tool may be more adequate as a supplement to teaching.

Focus group transcripts provided fruitful data on how students have previously been taught how to interpret ECGs, how their

XSL•FO
RenderX

previous learning compared to how this tutorial was delivered, and what they thought of the animations used to support the tutorial delivery. Moreover, information on how to improve the session was also collected. The main themes that arose were that ECGs are regarded as a complex topic among students and that past ECG learning used CBL and involved the memorization of traces. Other main themes include that the animations and associated explanations promoted a deeper understanding of cardiac electrical activity (compared to past teaching) and that ECG learning requires repetition and clinical links remain essential. Students noted that their most helpful past teaching involved cases and clinical contextualization, which should therefore be considered in any form of teaching implemented to final-year students, as clinical context appears to be their learning priority.

The key commonality between the control and intervention groups was the provision of a concise explanation of cardiac electrical activity in the heart for each section of the ECG trace. Therefore, future studies may benefit from investigating ways of delivering this content most effectively, for example through CBL or team-based learning [14-16], or similar methods of enabling greater interaction between students, but with a focus on understanding the pathology as opposed to focusing on pattern recognition.

This study and its teaching session do not come without some limitations. For reasons described in the *Methods* section, this study was limited to up to 20 participants in each arm and was based in a single study-year group and university. Therefore, it is not possible to confirm that these results are generalizable. No data were collected on the demographics of participants, which would also be helpful in determining the generalizability of the findings. Furthermore, this study also did not directly assess knowledge; instead, it assessed confidence in knowledge. Confidence has greater subjectivity than knowledge-based assessments and is not a reliable alternative to assessing student learning. Therefore, future evaluations of these animations would benefit from a validated assessment of students before and after the tutorial.

The teaching session itself would have benefited from greater interactivity, which has been showed to be an important element to teaching [14-16]. The session was delivered in a more didactic way, compared to CBL or team-based learning methods, which may have compromised student engagement and therefore learning. None of these elements were incorporated in the teaching session mainly due to time constraints but also to maintain the focus of the session on evaluating the value of the animations in improving student confidence. For instance, the inclusion of cases would act as a confounding variable as students may be able to understand the pathology from the case rather than from interpreting the ECG, whereas other students might not have engaged as much with the animations when in a team compared to when working individually. It is important to note that clinical context is important, as supplementing teaching materials with a patient case helps students to better diagnose, investigate, and manage cardiac conditions, thus improving their clinical reasoning skills [16].

The animations are likely to be even more valuable if used alongside other helpful learning tools, including the design of more interactive tutorials by involving quizzes throughout and gamification, which is a concept recently discussed in the literature, wherein game design elements are used in nongame contexts to promote users' engagement [17]. Moreover, future teaching sessions would benefit from including the aforementioned clinical scenarios prior to demonstrating each pathology, intertwining the learning of relevant pathophysiology with clinical knowledge. Although the latter would allow greater contextualization and demonstrate the relevance of the learning to clinical practice, the former would provide the required background knowledge to understand the clinical manifestations, and management, of disease.

In addition, it is important to acknowledge that for students to confidently be able to interpret ECGs, they need to apply the concepts of spaced learning and repetition. Future teaching could be accompanied by resources such as a recording of the session, the slides and animations used, as well as single–best answer questions to enable students to consolidate and test their learning. A more appropriate method of using these animations may therefore be to provide these to students as an independent learning resource. When doing so, it would enable students to scroll through the animation and independently control its speed to match their learning needs and understanding. Additionally, students suggested to make the animation 3D and to demonstrate the full 12-lead ECG alongside the animation as opposed to a single lead only.

In conclusion, this study suggests that although incorporating visual animations to demonstrate the electrical activity of different pathologies in ECG teaching may be beneficial in improving students' confidence in interpreting ECGs and understanding the underlying pathology, it is not the only way that this can be achieved. Students benefited equally from verbal explanations, suggesting that the most essential part of future ECG teaching is providing emphasis on the relevant pathophysiology, presented alongside clinical vignettes in which discussions regarding investigations and management options can be made. Interactivity within teaching sessions using quizzes and spaced practice is also recommended, in which students can access the resources, including the animation used in the session, later, to help consolidate their learning. Nevertheless, the development of animations was a low-cost intervention enjoyed by students and was reported to support their learning and understanding of cardiac pathophysiology and interpretation of ECG traces. Therefore, it is hoped that making these animations available to students as a revision resource can supplement their current ECG teaching and individual study practices.

XSL•FO
RenderX

Acknowledgments

We would like to thank Dr Ana Baptista for the support and guidance in preparing the ethics application and reviewing the study design. Open-access publication fees for this article were covered by the Imperial College Open Access Fund.

Data Availability

Questionnaire data sets have been made available in Multimedia Appendix 5.

Conflicts of Interest

AMCP, DS, AD-B, and TR are volunteers at and LR is the founder of a medical charity, Make a Medic, which may consider developing and implementing the educational tool piloted in this study. However, the authors hold no financial or other similar benefits from this work or its outcomes.

Multimedia Appendix 1
Questionnaires.
[DOCX File, 23 KB - mededu_v10i1e46507_app1.docx]

Multimedia Appendix 2
Focus group questions.
[DOCX File, 8 KB - mededu_v10i1e46507_app2.docx]

Multimedia Appendix 3
Intervention data.
[DOCX File, 29 KB - mededu_v10i1e46507_app3.docx]

Multimedia Appendix 4
Control data and figures.
[DOCX File, 149 KB - mededu_v10i1e46507_app4.docx]

Multimedia Appendix 5
Raw data.
[XLSX File, 23 KB - mededu_v10i1e46507_app5.xlsx]

References

1. Viljoen CA, Millar RS, Manning K, Burch VC. Effectiveness of blended learning versus lectures alone on ECG analysis and interpretation by medical students. BMC Med Educ 2020 Dec 3;20(1):488. [doi: 10.1186/s12909-020-02403-y] [Medline: 33272253]
2. Papapanou M, Routsi E, Tsamakis K, et al. Medical education challenges and innovations during COVID-19 pandemic. Postgrad Med J 2022 May;98(1159):321-327. [doi: 10.1136/postgradmedj-2021-140032] [Medline: 33782202]
3. Viljoen CA, Millar RS, Manning K, Burch VC. Determining electrocardiography training priorities for medical students using a modified Delphi method. BMC Med Educ 2020 Nov 16;20(1):431. [doi: 10.1186/s12909-020-02354-4] [Medline: 33198726]
4. Williams B. Case based learning--a review of the literature: is there scope for this educational paradigm in prehospital education? Emerg Med J 2005 Aug;22(8):577-581. [doi: 10.1136/emj.2004.022707] [Medline: 16046764]
5. Vishnevsky G, Cohen T, Elitzur Y, Reis S. Competency and confidence in ECG interpretation among medical students. Int J Med Educ 2022 Nov 30;13:315-321. [doi: 10.5116/ijme.6372.2a55] [Medline: 36463574]
6. Kopeć G, Magoń W, Hołda M, Podolec P. Competency in ECG interpretation among medical students. Med Sci Monit 2015 Nov 6;21:3386-3394. [doi: 10.12659/msm.895129] [Medline: 26541993]
7. Cook DA, Oh SY, Pusic MV. Accuracy of physicians' electrocardiogram interpretations: a systematic review and meta-analysis. JAMA Intern Med 2020 Nov 1;180(11):1461-1471. [doi: 10.1001/jamainternmed.2020.3989] [Medline: 32986084]
8. Keenan ID, Ben Awadh A. Integrating 3D visualisation technologies in undergraduate anatomy education. Adv Exp Med Biol 2019;1120:39-53. [doi: 10.1007/978-3-030-06070-1_4] [Medline: 30919293]
9. Clunie L, Morris NP, Joynes VCT, Pickering JD. How comprehensive are research studies investigating the efficacy of technology-enhanced learning resources in anatomy education? a systematic review. Anat Sci Educ 2018 May 6;11(3):303-319. [doi: 10.1002/ase.1762] [Medline: 29236354]

XSL•FO
RenderX

10. de Barros N, Rodrigues CJ, Rodrigues AJ Jr, de Negri Germano MA, Cerri GG. The value of teaching sectional anatomy to improve CT scan interpretation. Clin Anat 2001 Jan;14(1):36-41. [doi: 10.1002/1098-2353(200101)14:1<36::AID-CA1006>3.0.CO;2-G] [Medline: 11135396]
11. Thomas R, Ditto Stritto ME. Student outcomes in online courses: when does class size matter? The Northwest eLearning Journal 2021 May 31;1(1). [doi: 10.5399/osu/nwelearn.1.1.5608]
12. Orellana A. Class size and interaction in online courses. Q Rev Distance Educ 2006 Jan;7(3):229-248 [FREE Full text]
13. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
14. James M, Baptista AMT, Barnabas D, et al. Collaborative case-based learning with programmatic team-based assessment: a novel methodology for developing advanced skills in early-years medical students. BMC Med Educ 2022 Feb 7;22(1):81. [doi: 10.1186/s12909-022-03111-5] [Medline: 35125094]
15. Burgess A, van Diggele C, Roberts C, Mellis C. Team-based learning: design, facilitation and participation. BMC Med Educ 2020 Dec 3;20(Suppl 2):461. [doi: 10.1186/s12909-020-02287-y] [Medline: 33272267]
16. Burgess A, Matar E, Roberts C, et al. Scaffolding medical student knowledge and skills: team-based learning (TBL) and case-based learning (CBL). BMC Med Educ 2021 Apr 26;21(1):238. [doi: 10.1186/s12909-021-02638-3] [Medline: 33902576]
17. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: defining "gamification". In: MindTrek ’11: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments: Association for Computing Machinery; 2011:9-15. [doi: 10.1145/2181037.2181040]

Abbreviations

AF: atrial fibrillation
BBB: bundle branch block
CBL: case-based learning
ECG: electrocardiography

Edited by TDA Cardoso; submitted 14.02.23; peer-reviewed by K Gupta, V Podder; revised version received 17.03.24; accepted 22.03.24; published 23.04.24.

Please cite as:
Cardoso Pinto AM, Soussi D, Qasim S, Dunin-Borkowska A, Rupasinghe T, Ubhi N, Ranasinghe L
The Use of Animations Depicting Cardiac Electrical Activity to Improve Confidence in Understanding of Cardiac Pathology and Electrocardiography Traces Among Final-Year Medical Students: Nonrandomized Controlled Trial
JMIR Med Educ 2024;10:e46507
URL: https://mededu.jmir.org/2024/1/e46507
doi:10.2196/46507

© Alexandra M Cardoso Pinto, Daniella Soussi, Subaan Qasim, Aleksandra Dunin-Borkowska, Thiara Rupasinghe, Nicholas Ubhi, Lasith Ranasinghe. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.4.2024. 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.

XSL•FO
RenderX

Original Paper

Nursing Students' Attitudes Toward Technology: Multicenter Cross-Sectional Study

Ana Luiza Dallora[1], MSc, PhD; Ewa Kazimiera Andersson[2], PhD; Bruna Gregory Palm[3], MSc, PhD; Doris Bohman[1,4], PhD; Gunilla Björling[5,6,7], PhD; Ludmiła Marcinowicz[8], PhD; Louise Stjernberg[9,10], PhD; Peter Anderberg[1,11], PhD

[1]Department of Health, Blekinge Institute of Technology, Karlskrona, Sweden
[2]Department of Health and Caring Sciences, Linnaeus University, Växjö, Sweden
[3]Department of Mathematics and Natural Sciences, Blekinge Institute of Technology, Karlskrona, Sweden
[4]Optentia Research Unit, North West University, Vanderbijlpark, South Africa
[5]School of Health and Wellfare, Jönköping University, Jönköping, Sweden
[6]Faculty of Nursing, Kilimanjaro Christian Medical University College, Moshi, United Republic of Tanzania
[7]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden
[8]Faculty of Health Sciences, Medical University of Bialystok, Białystok, Poland
[9]Department of Care Science, Malmö University, Malmö, Sweden
[10]Swedish Red Cross University, Huddinge, Sweden
[11]School of Health Sciences, University of Skövde, Skövde, Sweden

Corresponding Author:
Ana Luiza Dallora, MSc, PhD
Department of Health
Blekinge Institute of Technology
Valhallavägen 1
Karlskrona, 371 41
Sweden
Phone: 46 073 422 3667
Email: ana.luiza.moraes@bth.se

Abstract

Background: The growing presence of digital technologies in health care requires the health workforce to have proficiency in subjects such as informatics. This has implications in the education of nursing students, as their preparedness to use these technologies in clinical situations is something that course administrators need to consider. Thus, students' attitudes toward technology could be investigated to assess their needs regarding this proficiency.

Objective: This study aims to investigate attitudes (enthusiasm and anxiety) toward technology among nursing students and to identify factors associated with those attitudes.

Methods: Nursing students at 2 universities in Sweden and 1 university in Poland were invited to answer a questionnaire. Data about attitudes (anxiety and enthusiasm) toward technology, eHealth literacy, electronic device skills, and frequency of using electronic devices and sociodemographic data were collected. Descriptive statistics were used to characterize the data. The Spearman rank correlation coefficient and Mann-Whitney *U* test were used for statistical inferences.

Results: In total, 646 students answered the questionnaire—342 (52.9%) from the Swedish sites and 304 (47.1%) from the Polish site. It was observed that the students' technology enthusiasm (techEnthusiasm) was on the higher end of the Technophilia instrument (score range 1-5): 3.83 (SD 0.90), 3.62 (SD 0.94), and 4.04 (SD 0.78) for the whole sample, Swedish students, and Polish students, respectively. Technology anxiety (techAnxiety) was on the midrange of the Technophilia instrument: 2.48 (SD 0.96), 2.37 (SD 1), and 2.60 (SD 0.89) for the whole sample, Swedish students, and Polish students, respectively. Regarding techEnthusiasm among the nursing students, a negative correlation with age was found for the Swedish sample ($P<.001$; $\rho_{Swedish}=-0.201$) who were generally older than the Polish sample, and positive correlations with the eHealth Literacy Scale score ($P<.001$; $\rho_{all}=0.265$; $\rho_{Swedish}=0.190$; $\rho_{Polish}=0.352$) and with the perceived skill in using computer devices ($P<.001$; $\rho_{all}=0.360$; $\rho_{Swedish}=0.341$; $\rho_{Polish}=0.309$) were found for the Swedish, Polish, and total samples. Regarding techAnxiety among the nursing students, a positive correlation with age was found in the Swedish sample ($P<.001$; $\rho_{Swedish}=0.184$), and negative correlations

XSL•FO
RenderX

with eHealth Literacy Scale score (P<.001; ρ_{all}=−0.196; $\rho_{Swedish}$=−0.262; ρ_{Polish}=−0.133) and with the perceived skill in using computer devices (P<.001; ρ_{all}=−0.209; $\rho_{Swedish}$=−0.347; ρ_{Polish}=−0.134) were found for the Swedish, Polish, and total samples and with the semester only for the Swedish sample (P<.001; $\rho_{Swedish}$=−0.124). Gender differences were found regarding techAnxiety in the Swedish sample, with women exhibiting a higher mean score than men (2.451, SD 1.014 and 1.987, SD 0.854, respectively).

Conclusions: This study highlights nursing students' techEnthusiasm and techAnxiety, emphasizing correlations with various factors. With health care's increasing reliance on technology, integrating health technology–related topics into education is crucial for future professionals to address health care challenges effectively.

International Registered Report Identifier (IRRID): RR2-10.2196/14643

(JMIR Med Educ 2024;10:e50297) doi:10.2196/50297

KEYWORDS

nursing education; technophilia; eHealth; technology anxiety; technology enthusiasm; mobile phone

Introduction

Background

Health care costs have been growing faster than the economy for the past 17 years [1]. This upward trend is due to multifactorial causes related to the growth and aging of the population, increased prevalence of lifestyle-related noncommunicable diseases, increased prices of health services and pharmaceuticals, and the risk of global pandemics [2-4]. All these factors put high pressure on the health care systems, which have to deal with many challenges related to efficiency and productivity. The digitalization of the health care sector is strongly influencing the efforts to address health care challenges and involves the use of technologies such as information and communication technologies in health settings, which was later termed as *eHealth* [5].

The integration of eHealth in the health care sector points to greater use of technology to access health data, manage eHealth records, and engage in telehealth platforms, among others [6]. This is such an important topic that the European Commission issued the Digital Decade Policy Program targeting Europe's digital transformation by 2030 [7]. This policy envisions, among other goals, the achievement of a digitally skilled population, highlighting the importance of a highly digitally skilled health care workforce and inspiring initiatives in different European countries. In the United States, a similar government initiative promotes the use of health technologies to improve the quality, safety, and efficiency of and reduce disparities in health care delivery [8]. The merging of health care workforce and digital technologies became so evident that informatics is outlined as one of the core competencies in the nursing profession: "use information and technology to communicate, manage knowledge, mitigate error and support decision making" [9]. Accordingly, it is also increasingly important for registered nurses to become proficient in this aspect.

Incorporation of health technologies into nursing education and the preparedness of the new students to use these in clinical scenarios and practice are highly important and a growing concern for program administrators, educators, researchers, policy makers, and employers [10]. This concern is valid because despite many students having grown up with technology ingrained in their everyday life, they still report low confidence, difficulties, and not-so-positive views about applying digital skills in clinical contexts [11-15]. Therefore, it is important to investigate the nursing students' attitudes toward technologies, so that appropriate decisions can be made for educational purposes that might affect future patient care.

Many models assess user interaction with technology according to factors such as acceptance, motivation, adoption, adaptivity, and usability, which are known to play a role in technology use [16]. However, it is argued that both cognitive and emotional effectiveness affect behavior, and these are underlying factors that precede the specific, planned, and reasoned actions directed toward technology [17,18]. The concept of technophilia is a personality trait and a psychological construct that is related to a person's enthusiasm or positive feelings toward technology use and the absence of anxiety or fears and doubts regarding technology [19], and it is a general quality that could potentially influence a wide range of aspects of technology use. Contrary to models tailored to specific organizational tasks, the investigation of technophilia could provide a better picture of the students' needs regarding this proficiency.

Objectives

This study comprises a multicenter, cross-sectional investigation of technophilia among nursing students that aimed to (1) establish the levels of technophilia among nursing students of 3 educational institutions in Sweden and Poland regarding their enthusiasm and anxiety and (2) identify factors that could be associated with the students' technology enthusiasm (techEnthusiasm) and technology anxiety (techAnxiety).

Methods

Study Design

This study used a multicenter, cross-sectional design based on questionnaire data collected from nursing students in 3 different universities, in Sweden and Poland, in different stages in their education. The protocol for this study has been described previously [20]. This study adhered to the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) guideline for cross-sectional studies (Multimedia Appendix 1).

Setting

We collected data in the period between December 2019 and April 2020, using questionnaires administered to students

XSL•FO
RenderX

enrolled in the nurse education programs of 3 universities: 2 in Sweden (Blekinge Institute of Technology [BTH] and Swedish Red Cross University [SRCU]) and 1 in Poland (Medical University of Bialystok [MUB]). The undergraduate nursing education of both countries adheres to the European Union requirements, which comprises 180 European Credit Transfer and Accumulation System (ECTS) credits at the university level [21,22]. The educational programs in both countries result in a professional degree (ie, a diploma) and an academic degree (ie, a bachelor's degree), qualifying for a license as a registered nurse. At the time the study, the Swedish nursing education consisted of both theoretical and clinical practice courses—60% and 40% of the total curriculum, respectively. At the Polish institution, MUB, the nursing program consisted of 52% theoretical courses and 48% clinical practice courses. The students' exposure to eHealth or health technology courses at the time of the data collection was as follows:

- At BTH, eHealth is covered in nursing subjects during the whole program and in two dedicated courses in the curriculum:

1. An eHealth introductory course is offered in the third semester to all students (4.5 ECTS), which was completed at the time of the data collection.
2. An optional course on digitalization and eHealth was offered in the fifth semester (7.5 ECTS). It was chosen by approximately one-third of the fifth-semester students and was ongoing at the time of the data collection.

- At SRCU, eHealth was also incorporated into nursing subjects during the whole program and 1 optional course (7.5 ECTS) in medical technology, digitalization, and eHealth was offered in the fifth semester. However, this course started 5 weeks after this study's data collection.
- At MUB, eHealth was incorporated into nursing courses during the whole program.

Participants and Data Collection Procedures

A convenience sample of undergraduate nursing students, enrolled at the bachelor of nursing program at BTH, SRCU, and MUB, was used in this study. Students from the first, third, and fifth semesters were eligible to participate in this study. These semesters were chosen to obtain a sample incorporating the beginning, middle, and end of nursing education, which comprises 6 semesters.

Data were collected using a paper-based questionnaire administered to all undergraduate students from the first, third, and fifth semesters of the participating universities by research members who had no educational connections to the students. This was done to minimize response bias.

Questionnaire

The questionnaire was used to collect data about the participants' sociodemographics, self-reported attitudes toward technology, eHealth literacy, perceived skills in using electronic devices, and frequency of using electronic devices.

Data on Attitudes Toward Technology (Technophilia Instrument)

The outcome measures of this study are the self-reported data on attitudes toward technology scored by the Technophilia instrument (TechPH) [19]. The TechPH comprises 6 questions to capture behaviors related to adaptation and use of a new technology, which were derived from the content analysis of relevant technophilia measures. It results in 2 numeric scores ranging from 1 (low) to 5 (high): techEnthusiasm and techAnxiety. The TechPH was originally developed for measuring older adults' attitudes toward technology; however, published studies have already applied it on younger individuals, physicians, and dementia caretakers aged 18 to 44 years [23,24]. In this study, techEnthusiasm and techAnxiety have Cronbach α of 1 and 0.925, respectively, showing excellent internal consistency.

Sociodemographic Data

Sociodemographic data consisted of the participants' age; gender; focus of high school studies (health or social care, technology, or other); and previous work experience (health or social care, technology, or other).

eHealth Literacy Data (eHealth Literacy Scale)

The eHealth literacy was scored using the eHealth Literacy Scale (eHEALS) instrument [25]. The eHEALS is a self-report tool consisting of 8 questions and has already been validated in many languages and diverse populations including undergraduate health professionals [25,26]. The eHEALS produces a score ranging from 1 (low eHealth literacy) to 5 (high eHealth literacy).

Data on Perceived Skill in and Frequency of Using Technological Devices

Perceived skills in using electronic devices, namely, computers or laptops, tablets, and smartphones, were rated using a Likert-type scale ranging from 1 (not knowledgeable at all) to 5 (very knowledgeable). The frequencies of using electronic devices were rated using a Likert-type scale ranging from 1 (several times daily) to 5 (never).

Data Analysis

The descriptive statistics, namely, frequency, mean, and SD, were used to analyze the collected data. The Shapiro-Wilk test was used to assess data distribution. As the data were not normally distributed, nonparametric tests were used in the statistical analyses. Spearman rank correlation coefficient was used to measure the association among age, semester, perceived skills in using computers or laptops, and frequency of using electronic devices via the self-reported TechPH components—techEnthusiasm and techAnxiety. CIs were calculated to analyze the stability of the results. Mann-Whitney *U* test was used to assess gender differences regarding students' enthusiasm and anxiety toward technology. Sensitivity analyses were performed by removing the outliers and revealed that the interpretations were unperturbed, showing that extreme data points did not impact the study outcomes. For all the analyzes, a significance level of .05 was used. Stratification was used; therefore, results are presented for the whole sample, Swedish

XSL•FO
RenderX

students, and Polish students separately, to control for confounding. Entries with missing data were omitted from the analysis. The analyses were performed using R (version 1.4.1717; RStudio).

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki [27]. Participation in the study was voluntary. All participants were briefed about the study aims; that they could choose to not submit the questionnaire or submit it blank; and that by submitting the questionnaire, they would consent to participate in the study. All collected data were anonymous.

Permission to conduct the study was obtained from heads of the departments at all participating universities. In Poland, ethics approval was obtained from the ethics committee of Medical University of Bialystok (register number R-I-002/148/2017). In Sweden, the study did not require ethics approval according to the requirements of the Swedish Ethical Review Act 2003:460, 3-4§ [28], as the study did not explore sensitive personal data (eg, health, religion, political views, and ethnic heritage) or data relating to criminal offenses, did not involve physical intervention on the participants, and did not aim to affect the participants in any way or involve biological material.

Results

Sample Characteristics

In total, 646 students answered the questionnaire—342 (52.9%) from the Swedish sites and 304 (47.1%) from the Polish site. The response rates were 70.2% (646/920) for the whole sample, 63.1% (342/542) for the Swedish students, and 80.4% (304/378) for the Polish students. Nonresponders include students who decided not to submit the questionnaire or to submit it blank. None of the variables used in the analyses contained >5% of missing values.

The descriptive statistics are shown in Table 1, for the whole sample and for the Swedish and Polish students separately. Multimedia Appendix 2 shows the descriptive statistics along with the means and SDs for the techAnxiety and techEnthusiasm for each grouping shown in Table 1—for the whole sample, Swedish students, and Polish students separately. The mean age of the sample is 23.9 (SD 6.39) years, with the Swedish students being generally older and having a higher age variance (mean 27, SD 7.34 years) compared with the Polish students (mean 20.4, SD 1.72 years), as shown in Figure 1. While the Polish sample has a distribution that is more concentrated around the mean, the Swedish sample has a flatter distribution of ages. The sample was majorly composed of women students (555/646, 85.9%). Very few students had a high school focus on or previous work experience with technology before their nursing education. Overall, 50.3% (153/304) of the Polish students had a health and social care focus in high school, while this number was 23.9% (82/342) for the Swedish students. In terms of perceived skill in using electronic devices, most participants perceive themselves “knowledgeable” or “very knowledgeable” in all 3 categories: computers (479/646, 74.1%), smartphones (574/646, 88.9%), and tablets (388/646, 60.1%). Furthermore, 48.6% (314/646) of the participants answered that they use computers or laptops “several times daily” or “daily,” while this number reached 98.8% (638/646) for smartphones and 11.2% (72/646) for tablets. The students showed an overall high eHealth literacy, with 93.3% (603/646) scoring ≥3 points. The mean eHEALS scores for the overall, Polish, and Swedish samples were 3.95 (SD 0.75), 3.96 (SD 0.78), and 3.95 (SD 0.73), respectively, constituting high scores and showing an overall high perceived eHealth literacy.

XSL•FO
RenderX

Table 1. Frequency for the variables in the study for the whole, Swedish, and Polish samples.

	All students (N=646), n (%)	Swedish students (n_{Sweden}=342), n (%)	Polish students (n_{Poland}=304), n (%)
Age (years)			
18-25	478 (73.9)	179 (52.3)	299 (98.4)
>25	168 (26)	163 (47.7)	5 (1.6)
Gender			
Women	555 (85.9)	284 (83)	271 (89.1)
Men	89 (13.8)	56 (16.4)	33 (10.9)
Semester			
1	289 (44.7)	158 (46.2)	131 (43.1)
3	208 (32.2)	101 (29.5)	107 (35.2)
5	149 (23.1)	83 (24.2)	66 (21.7)
eHEALS[a] score			
<3	43 (6.7)	23 (6.7)	24 (7.9)
≥3	603 (93.3)	319 (93.3)	284 (93.4)
High school focus			
Health and social care	235 (36.4)	82 (23.9)	153 (50.3)
Technology	25 (3.9)	11 (3.2)	14 (4.6)
Other	374 (57.9)	242 (70.8)	132 (43.4)
Previous work experience			
Health and social care	211 (32.7)	188 (54.9)	23 (7.6)
Technology	12 (1.9)	7 (2)	5 (1.6)
Other	332 (51.4)	118 (34.5)	214 (70.4)
Skills: computer			
1: not knowledgeable at all	1 (0.2)	0 (0)	1 (0.3)
2	19 (2.9)	14 (4)	5 (1.6)
3	134 (20.7)	93 (27.2)	41 (13.5)
4	173 (26.8)	106 (30.9)	67 (22)
5: very knowledgeable	306 (47.4)	116 (33.9)	190 (62.5)
Skills: smartphone			
1: not knowledgeable at all	4 (0.6)	3 (0.9)	1 (0.3)
2	5 (0.8)	3 (0.9)	2 (0.7)
3	6 (0.9)	41 (11.9)	22 (7.2)
4	127 (19.7)	88 (25.7)	39 (12.9)
5: very knowledgeable	447 (69.2)	207 (60.5)	240 (78.9)
Skills: tablets			
1: not knowledgeable at all	66 (10.2)	25 (7.3)	41 (13.5)
2	69 (10.7)	47 (13.7)	22 (7.2)
3	107 (16.6)	66 (19.3)	41 (13.5)
4	149 (23.1)	88 (25.7)	61 (20)
5: very knowledgeable	239 (36.9)	100 (29.2)	139 (45.7)
Frequency: computer			
Several times daily	131 (20.3)	69 (20.2)	62 (20.4)
Daily	183 (28.3)	80 (23.4)	103 (33.9)

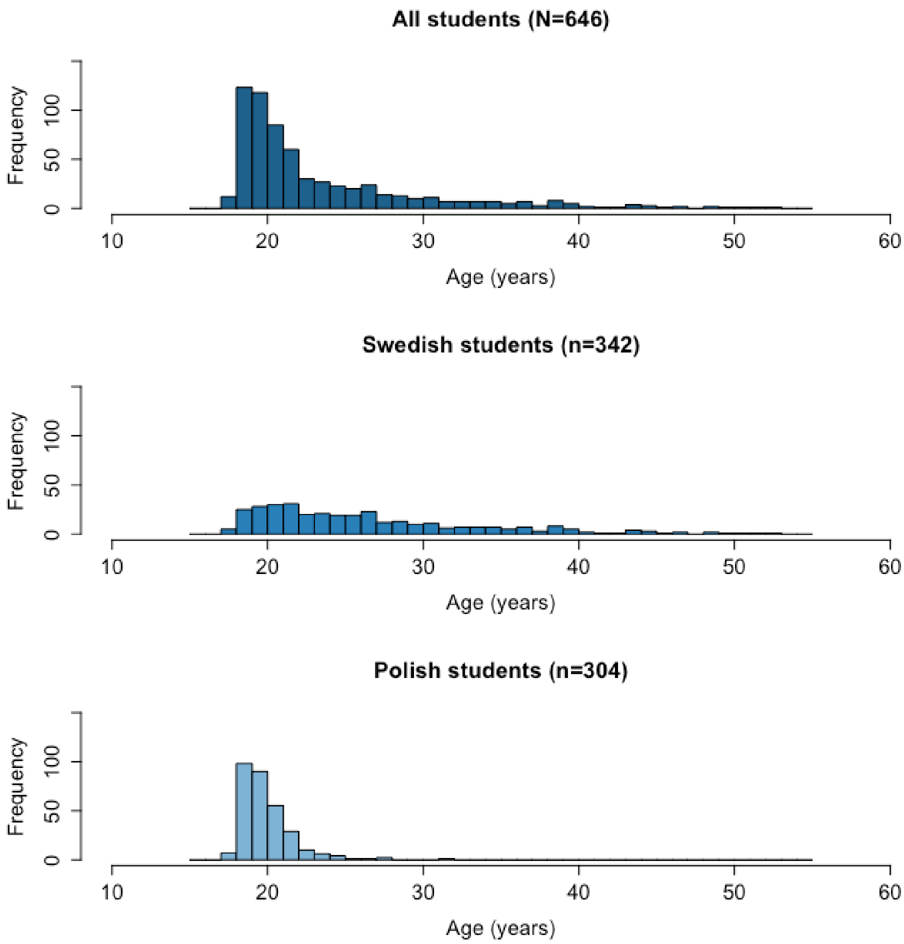
XSL•FO
RenderX

	All students (N=646), n (%)	Swedish students (n_{Sweden}=342), n (%)	Polish students (n_{Poland}=304), n (%)
Every week	152 (23.5)	85 (24.9)	67 (22)
Every month	39 (6)	29 (8.5)	10 (3.3)
Sometimes	98 (15.2)	42 (12.3)	56 (18.4)
Never	11 (1.7)	5 (1.5)	6 (1.9)
Frequency: smartphone			
Several times daily	576 (89.2)	302 (88.3)	274 (90.1)
Daily	62 (9.6)	34 (9.9)	28 (9.2)
Every week	3 (0.5)	2 (0.6)	1 (0.3)
Every month	0 (0)	0 (0)	0 (0)
Sometimes	2 (0.3)	1 (0.3)	1 (0.3)
Never	1 (0.2)	1 (0.3)	0 (0)
Frequency: tablet			
Several times daily	29 (4.5)	23 (6.7)	6 (1.9)
Daily	43 (6.7)	26 (7.6)	17 (5.6)
Every week	51 (7.9)	38 (11.1)	13 (4.3)
Every month	23 (3.6)	14 (4.1)	9 (2.9)
Sometimes	138 (21.4)	76 (22.2)	62 (20.4)
Never	332 (51.4)	135 (39.5)	197 (64.8)

[a]eHEALS: eHealth Literacy Scale.

XSL•FO
RenderX

Figure 1. Distribution of ages in the whole sample, Swedish students, and Polish students.



The mean and SD values for the self-reported techEnthusiasm for the whole sample, Swedish students, and Polish students were 3.83 (SD 0.90), 3.62 (SD 0.94), and 4.04 (SD 0.78), respectively, which constitutes a high overall technophilia. On the other hand, the mean and SD values for techAnxiety for the whole sample, Swedish students, and Polish students were 2.48 (SD 0.96), 2.37 (SD 1), and 2.60 (SD 0.89), respectively, displaying midrange values regarding the negative feelings toward technology.

Multimedia Appendix 2 shows the mean and SD values for both techEnthusiasm and techAnxiety according to different levels of socioeconomic, eHEALS, perceived skill, and frequency variables. The association of these variables with techEnthusiasm and techAnxiety is investigated in the following sections.

Factors Associated With TechEnthusiasm

The Spearman rank correlation coefficient was used to investigate the association of techEnthusiasm and the nonparametric variables of ordinal scale in the study, namely, age, semester, eHEALS score, perceived skill, and frequency of using electronic devices. These results are shown in terms of the Swedish, Polish, and overall samples in Table 2.

XSL•FO
RenderX

Table 2. Spearman rank correlation coefficient calculated for the whole sample, Swedish students, and Polish students separately, regarding technology enthusiasm.

	All students		Swedish students		Polish students	
	P value	ρ (95% CI)	*P* value	ρ (95% CI)	*P* value	ρ (95% CI)
Age	<.001	−0.238 (−0.310 to −0.163)	<.001	−0.201 (−0.302 to −0.096)	.65	−0.027 (−0.139 to 0.086)
Semester	.96	0.002 (−0.075 to 0.079)	.53	−0.034 (−0.140 to 0.073)	.44	0.044 (−0.068 to 0.156)
eHEALS[a] score	<.001	0.265 (0.190 to 0.336)	<.001	0.190 (0.084 to 0.291)	<.001	0.352 (0.246 to 0.449)
Skill: computer	<.001	0.360 (0.288 to 0.428)	<.001	0.341 (0.238 to 0.436)	<.001	0.309 (0.201 to 0.410)
Skill: smartphone	<.001	0.385 (0.315 to 0.452)	<.001	0.352 (0.253 to 0.445)	<.001	0.364 (0.258 to 0.460)
Skill: tablet	<.001	0.269 (0.194 to 0.342)	<.001	0.309 (0.204 to 0.406)	<.001	0.204 (0.092 to 0.310)
Frequency: computer	<.001	−0.153 (−0.230 to −0.074)	<.001	−0.176 (−0.283 to −0.065)	.01	−0.146 (−0.255 to −0.034)
Frequency: smartphone	.95	0.002 (−0.075 to 0.080)	.97	−0.002 (−0.109 to 0.105)	.87	0.009 (−0.103 to 0.122)
Frequency: tablet	.92	0.004 (−0.075 to 0.083)	.16	−0.079 (−0.189 to 0.032)	.59	−0.031 (−0.143 to 0.081)

[a]eHEALS: eHealth Literacy Scale.

A negative correlation was found between age and techEnthusiasm for the Swedish sample and overall sample, indicating that greater the age, lesser the techEnthusiasm score ($P<.001$; $\rho_{all}=-0.238$; $\rho_{Swedish}=-0.201$). This association might not have been significant for the Polish sample due to the lack of age variance observed in the Swedish sample (refer to Figure 1—the Polish students' age distribution presents a heavier tail compared to the Swedish ones). A positive correlation was found between eHealth literacy and techEnthusiasm, indicating that greater the eHEALS score, greater the techEnthusiasm score ($P<.001$; $\rho_{all}=0.265$; $\rho_{Swedish}=0.190$; $\rho_{Polish}=0.352$). A positive correlation was found between perceived skill in all investigated electronic devices and techEnthusiasm, indicating that greater the perceived skill, greater the techEnthusiasm (Table 2). In terms of frequency of use, a negative correlation was found between the use of computers and techEnthusiasm ($P<.001$ $\rho_{all}=-0.153$; $\rho_{Swedish}=-0.176$; $\rho_{Polish}=-0.146$). The negative values of ρ are due to the inverted Likert scale used for the question, that is, from "several times daily" to "never." Thus, the techEnthusiasm score increases with higher frequencies of use. It is noteworthy that even with low ρ values, the significant associations found are still relevant due to the large sample size. The narrow 95% CIs indicate low variability and stability of the results.

The Mann-Whitney *U* test was used to assess gender differences regarding the students' reported techEnthusiasm. No significant differences were found for Swedish, Polish, or overall samples ($P_{Swedish}=.45$, $P_{Polish}=.38$, and $P_{all}=.68$).

Factors Associated With TechAnxiety

Analogous statistical analyses were performed on the techAnxiety scores for the Swedish, Polish, and overall samples. Table 3 shows the Spearman rank correlation coefficients calculated for the same variables as for techEnthusiasm. A positive correlation was found between age and techAnxiety in the Swedish sample, indicating that greater the age, greater the techAnxiety score ($P<.05$; $\rho_{Swedish}=0.184$). Similar to techEnthusiasm, this association might not have been significant for the Polish sample due to the lack of age variance (Figure 1). A negative correlation was found between higher semesters and techAnxiety in the Swedish and overall samples, indicating that higher the students were in their education, lesser the techAnxiety score ($P<.05$; $\rho_{all}=-0.101$; $\rho_{Swedish}=-0.124$).

XSL•FO
RenderX

Table 3. Spearman rank correlation coefficient calculated for the whole sample, Swedish students, and Polish students separately, regarding technology anxiety.

Feature	All students		Swedish students		Polish students	
	P value	ρ (95% CI)	*P* value	ρ (95% CI)	*P* value	ρ (95% CI)
Age	.54	0.024 (−0.101 to 0.053)	<.001	0.184 (0.078 to 0.286)	.45	−0.043 (−0.155 to 0.070)
Semester	.01	−0.101 (−0.178 to −0.024)	.02	−0.124 (−0.229 to −0.018)	.19	−0.075 (−0.186 to 0.038)
eHEALS[a] score	<.001	−0.196 (−0.270 to −0.120)	<.001	−0.262 (−0.360 to −0.158)	.02	−0.133 (−0.242 to −0.020)
Skill: computer	<.001	−0.209 (−0.283 to −0.132)	<.001	−0.347 (−0.442 to −0.245)	.02	−0.134 (−0.244 to −0.022)
Skill: smartphone	<.001	−0.165 (−0.240 to −0.088)	<.001	−0.245 (−0.345 to −0.141)	.046	−0.114 (−0.224 to −0.002)
Skill: tablet	<.001	−0.251 (−0.324 to −0.175)	<.001	−0.347 (−0.442 to −0.244)	<.001	−0.191 (−0.298 to −0.080)
Frequency: computer	.495	0.028 (−0.052 to 0.107)	.16	0.080 (−0.033 to 0.191)	.61	−0.029 (−0.142 to 0.083)
Frequency: smartphone	.17	0.055 (−0.023 to 0.132)	.27	0.060 (−0.047 to 0.166)	.37	0.052 (−0.061 to 0.164)
Frequency: tablet	.19	0.053 (−0.026 to 0.132)	.77	0.017 (−0.095 to 0.128)	.50	0.039 (−0.074 to 0.151)

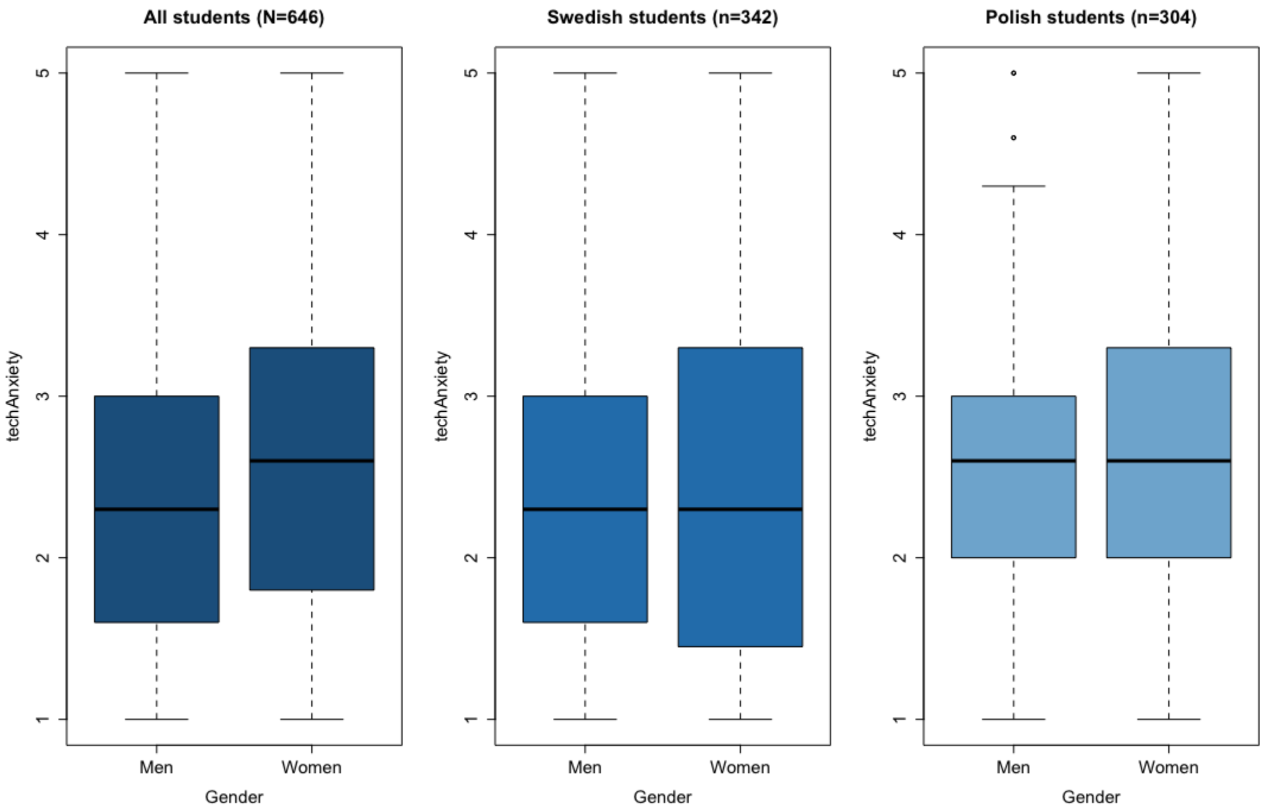
[a]eHEALS: eHealth Literacy Scale.

A negative correlation was found between eHealth literacy and techAnxiety (P<.001; ρ_{all}=−0.196; $\rho_{Swedish}$=−0.262; ρ_{Polish}=−0.133). A negative correlation was found between the perceived skill in all investigated devices and techAnxiety, indicating that greater the perceived skill, lesser the techAnxiety score (Table 3). Similar to techEnthusiasm, the low ρ values still show relevant associations due to the sample size. In addition, similar to techEnthusiasm, the narrow 95% CIs indicate low variability and stability of the results.

Gender differences regarding the students' reported techAnxiety were observed through the Mann-Whitney *U* test for the whole sample and the Swedish students ($P_{Swedish}$=.002; P_{Polish}=.69; P_{all}=.01). A considerable difference in the mean scores of the reported techAnxiety can be observed between men (1.987, SD 0.854) and women (2.451, SD 1.014) in the Swedish sample of students. This was not observed for the whole sample, with means of 2.240 (SD 0.90) and 2.521 (SD 0.963) for men and women students, respectively. However, upon closer inspection of the boxplots shown in Figure 2, the attributed gender differences can be observed when the distribution is analyzed. The Swedish women students present a higher dispersion of techAnxiety scores, whereas men present a heavier tail distribution, which in turn increases the distance between these groups.

XSL•FO
RenderX

Figure 2. Box plots for the reported technology anxiety in the whole, Swedish, and Polish samples.



Discussion

Principal Findings

Overview

This cross-sectional, multicenter study aimed to determine Swedish and Polish students' attitudes toward technology, specifically directed to enthusiasm and anxiety, and factors associated with those attitudes. The principal findings of this study are as follows: (1) in the Swedish sample (mean age 27, SD 7.34 years), the older the students were, the more anxious and less enthusiastic they were about technology; (2) the higher the students' eHealth literacy score was, the more enthusiastic and less anxious they were regarding technology (both Swedish and Polish samples); (3) the higher the perceived skill in using electronic devices was, the more enthusiastic and less anxious about technology the students were (both Swedish and Polish samples); (4) in the Swedish sample, the more senior the students were in their education (higher semesters), the less anxious they were toward technology they were; and (5) gender differences were found in the Swedish sample regarding anxiety toward technology. These will be further discussed in the following sections.

Attitudes Toward Technology and Age

The positive correlation between age and techAnxiety and negative correlation between age and techEnthusiasm, meaning that greater the age, lesser the enthusiasm and higher the anxiety toward technology, is an interesting finding, as many Swedish students (163/342, 47.6%) fall into the mature student category. This concept does not have a definition, but published literature usually considers the individuals who enter higher education at the age of 26 to 30 as mature students who are believed to be different from their younger colleagues [29]. The fact that many universities have a changing cohort with a higher rate of accepted mature students, meanwhile adopting more and more technologies as teaching enhancements [30], raises concerns about how the students' attitudes toward technology could affect their learning. Technology-enhanced learning methods in the classroom can promote high-order thinking, that is, rationalizing on a level higher than memorizing or telling facts as told [31]. These teaching approaches affect the attainment of the subject being taught and decrease subject anxieties [29,32,33]. In the specific case of nursing, a systematic review of literature by Labrague et al [34] shows positive results in using high-fidelity simulations for enhancing the self-confidence of nursing students in managing their duties. Identifying the students' needs could be important in these scenarios, so that learning could be efficiently delivered. A recent study investigated mature students' attitudes toward technology and found no significant differences from younger students. However, the attitudes considered in the study instrument were confidence and a sense of utility [29]. The hypothesis suggests that the observed phenomenon related to age may not be applicable to the Polish sample due to its younger ages and lower variability (mean 20.4, SD 1.72 years).

Attitudes Toward Technology and eHealth Literacy

The positive correlation between eHealth literacy score and enthusiasm toward technology is not surprising because computer literacy is one of the domains assessed in the eHEALS instrument. The negative correlation between eHealth literacy

XSL•FO
RenderX

score and anxiety toward technology is also important to be considered. Even with different approaches to eHealth in nursing education, it seems important for the nursing students' attitudes toward technology, which would later influence their use of technology in clinical scenarios. Registered nurses commonly rely on advice from their colleagues as their primary information source to inform daily clinical practice [35-37]. However, this information channel has an inherent risk of diverging from the best evidence available in published literature, which could impact the quality of patient care [38]. The best clinical practices can be readily accessed through reference materials and web-based publications in nursing journals. Thus, if anxiety toward technology is a factor that is identified as a barrier to pursuing such information, this means that it should be addressed in their education.

Attitudes Toward Technology and Perceived Skills in Using Electronic Devices

Another study finding suggests that students exhibiting higher perceived skill in using electronic devices (computers or laptops, smartphones, and tablets) also demonstrated more enthusiastic and less anxious attitudes toward technology. Only few studies have investigated electronic device use in nursing education. However, investigating this topic is important because despite nursing students reporting proficiency in computer skills, a lack of exposure to new devices can still lead to hesitancy in their use [39]. In the recent years, mobile apps have been trialed and shown to support the education and practice training of nursing students [40]. This technology facilitates access to patient care resources, fostering self-directed learning and problem-solving [41]. A study by Kenny et al [42] investigated the impact of using smartphones and tablets with a QR code scanning app linking to educational information on the nursing students' anxiety levels while performing psychomotor skills in the patient care setting. The study found that providing students with access to these tools helped to reduce anxiety by offering quick access to reputable patient care information [42]. Previous studies of bank employees also found an inverse relationship between techAnxiety and computer skills [43]. The study did not find any significant associations with the frequency of use, which is consistent with this study, with the exception of techEnthusiasm related to computer or notebook. However, it can be argued that this could simply be a direct result of being enthusiastic and wanting to engage with it daily. While no studies in the literature approached the topic of techEnthusiasm and skill, a study by Revilla Muñoz et al [44] reported lower levels of techAnxiety in high school teachers after information and communications technology training.

Attitudes Toward Technology and Semester

In the Swedish sample, students exhibited lower levels of anxiety toward technology the further in their education they were. This could be related to how health technology topics are being addressed in specific courses given in higher semesters, which was not the case for the Polish university at the time of the study. This could indicate that having specific courses with eHealth and health technology curricula could be useful to address techAnxiety in students. A scoping review by Nes et al [45] highlights that the current state of nursing education indicates a prevalent lack of focus on technology and technological literacy, favoring teaching over engaging with technological advancements in the clinical field, resulting in limited exposure to such developments. This holds significance because practitioners are likely to navigate ongoing technological advancements throughout their careers. Therefore, nursing education should be viewed as a platform that fosters lifelong learning, placing emphasis on proactive engagement and critical thinking in response to technological progress [46].

Attitudes Toward Technology and Gender

This study also found gender differences regarding techAnxiety in the Swedish students (mature student sample). There is sparse published literature about gender and computer anxiety, and findings do not seem to provide a conclusion [47-50]. Sparse literature has been published in the area of techAnxiety and even less so in the techEnthusiasm domain; this may compromise the credibility of the findings from comparing these studies. It can be argued that the findings of studies conducted more than a decade ago are difficult to interpret without the context of the time they were published in, because with the rapid technological advancements of the past years, the relationship between users and technology has changed drastically.

Implications to Practice

Understanding the factors that influence techEnthusiasm and techAnxiety holds important practical implications, particularly in the context of health care innovation and access to care.

TechAnxiety and techEnthusiasm can impact the technology acceptance level of a new health care solution. Low levels of acceptance are related to implementation delays and even complete system failures [51]. According to the systematic literature review by AlQudah et al [52], which included 142 studies, the key factors associated with health technology acceptance are its ease of use and perceived usefulness, which are measured using the widespread Technology Acceptance Model instrument. In addition, anxiety and computer self-efficacy are the next extensively studied factors related to health care technology acceptance, which aligns with the focus of this study.

A qualitative study conducted with nurses who have lower levels of digital literacy [53] explored factors related to health IT acceptance in this population. The results portrayed that these nurses show little enthusiasm toward technology and even considered the use of such technological tools as "bad patient-centered care." Addressing those attitudes toward technology is a challenge and should be tailored to special needs, as these individuals also reported that the training sessions are conducted in large groups and that the pace is too fast for them.

Telemedicine, eHealth records, health IT systems, and mobile apps emerge as important health technologies that are directed to improve productivity and effectiveness of the health care sector. During the COVID-19 pandemic, digital health strategies, which include such systems, were imperative for providing continuity of care; economic, social, geographical, time, and cultural accessibility; and coordination of care, among others [54]. However, during those difficult times, several health

XSL•FO
RenderX

professionals were unprepared to use such technologies [54]. Having health personnel that is trained to use different health technologies proved to be imperative to build preparedness for unusual health emergency situations. Strategies to address the problem of accessibility of health care in remote or rural areas could also use such technologies [52]. Hence, it is important to understand students' digital savviness to devise strategies to address health technology topics accordingly in the curricula of health-related undergraduate programs.

Limitations

As this was a self-reported survey study, care must be taken when extrapolating the results shown in this paper. Response bias was mitigated by involving researchers with no educational connections with the surveyed participants. An earlier study of self-reported technology use presented only marginal errors to the respondents' true use [55]. Another important limitation of this study is the disproportionate number of women and men participants, with the former consisting of 85.9% (555/646) of the whole sample. Although the statistical tests used in the analyses are robust against data imbalance, the magnitude of such imbalance could have affected the results. In addition, the use of a convenience sample can limit how the findings of this study can be generalized. However, in this study, the involvement of different universities from different countries as data sources helped to reduce this risk. Finally, the instrument used in this study was initially crafted and validated for use with older adults. Although published evidence exists for its use in younger populations [23,24], the results should be interpreted with caution, recognizing the potential for age-related bias.

Conclusions

This cross-sectional, multicenter study emphasized the importance of nursing students' enthusiasm and anxiety toward technology and highlighted the factors associated with these attitudes. As health care increasingly relies on technologies such as telemedicine, eHealth records, health IT systems, and mobile apps, the integration of health technology topics into educational curricula becomes imperative, taking the students' attitudes toward technology into consideration, so that in the future, these professionals are prepared to address future health care challenges. Future qualitative studies should investigate nursing students who portray high anxiety and low enthusiasm toward technology to further validate the results presented in this paper and understand their points of view, so that pedagogical strategies can be developed to incorporate health technology topics in the curricula.

Acknowledgments

The authors are especially grateful to the study participants for their time and interest in participating in the study. This study is part of the eHealth in Nursing Education (eNurseEd) study and was supported financially by the participating universities as an educational improvement effort. The funding source was not involved in the review design, analysis, interpretation of findings, writing of the paper, or the decision to submit the paper for publication.

Data Availability

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

Authors' Contributions

ALD, DB, and PA conceived the study design. EKA, DB, LS, GB, and LM performed the data collection. ALD and BGP performed the data analysis. ALD, EKA, BGP, and PA drafted the paper. All authors have contributed to the authorship and approved the final version of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

STROBE (Strengthening the Reporting of Observational studies in Epidemiology) statement—checklist of items that should be included in the reports of cross-sectional studies.

[DOCX File , 32 KB - mededu_v10i1e50297_app1.docx]

Multimedia Appendix 2

Descriptive statistics in terms of sociodemographic variables, technology anxiety, and technology enthusiasm. The first panel shows the frequency and percentages, and the second and third panels show the results in terms of the technology anxiety and technology enthusiasm scores.

[PDF File (Adobe PDF File), 102 KB - mededu_v10i1e50297_app2.pdf]

References

XSL•FO
RenderX

1. Global spending on health: a world in transition. World Health Organization. 2019 Dec 12. URL: https://www.who.int/publications/i/item/WHO-HIS-HGF-HFWorkingPaper-19.4 [accessed 2024-04-18]
2. Muka T, Imo D, Jaspers L, Colpani V, Chaker L, van der Lee SJ, et al. The global impact of non-communicable diseases on healthcare spending and national income: a systematic review. Eur J Epidemiol 2015 Apr 18;30(4):251-277. [doi: 10.1007/s10654-014-9984-2] [Medline: 25595318]
3. Dieleman JL, Squires E, Bui AL, Campbell M, Chapin A, Hamavid H, et al. Factors associated with increases in US health care spending, 1996-2013. JAMA 2017 Nov 07;318(17):1668-1678 [FREE Full text] [doi: 10.1001/jama.2017.15927] [Medline: 29114831]
4. Yeganeh H. An analysis of emerging trends and transformations in global healthcare. Int J Health Gov 2019 May 22;24(2):169-180. [doi: 10.1108/ijhg-02-2019-0012]
5. Tortorella GL, Fogliatto FS, Mac Cawley Vergara A, Vassolo R, Sawhney R. Healthcare 4.0: trends, challenges and research directions. Prod Plan Control 2019 Dec 17;31(15):1245-1260. [doi: 10.1080/09537287.2019.1702226]
6. Brown J, Pope N, Bosco AM, Mason J, Morgan A. Issues affecting nurses' capability to use digital technology at work: an integrative review. J Clin Nurs 2020 Aug;29(15-16):2801-2819. [doi: 10.1111/jocn.15321] [Medline: 32416029]
7. Europe's digital decade. European Commission. URL: https://digital-strategy.ec.europa.eu/en/policies/europes-digital-decade [accessed 2024-04-18]
8. Kim HN. A conceptual framework for interdisciplinary education in engineering and nursing health informatics. Nurse Educ Today 2019 Mar;74:91-93. [doi: 10.1016/j.nedt.2018.12.010] [Medline: 30639937]
9. Cronenwett L, Sherwood G, Barnsteiner J, Disch J, Johnson J, Mitchell P, et al. Quality and safety education for nurses. Nurs Outlook 2007 May;55(3):122-131. [doi: 10.1016/j.outlook.2007.02.006] [Medline: 17524799]
10. Kleib M, Nagle LM, Furlong KE, Paul P, Duarte Wisnesky U, Ali S. Are future nurses ready for digital health?: informatics competency baseline assessment. Nurse Educ 2022 Mar 25;47(5):E98-104 [FREE Full text] [doi: 10.1097/NNE.0000000000001199] [Medline: 35324499]
11. Edirippulige S, Samanta M, Armfield NR. Assessment of self-perceived knowledge in e-health among undergraduate students. Telemed J E Health 2018 Feb;24(2):139-144. [doi: 10.1089/tmj.2017.0056] [Medline: 28708457]
12. van Houwelingen CT, Ettema RG, Kort HS, ten Cate O. Internet-generation nursing students' view of technology-based health care. J Nurs Educ 2017 Dec 01;56(12):717-724 [FREE Full text] [doi: 10.3928/01484834-20171120-03] [Medline: 29206261]
13. Brown J, Morgan A, Mason J, Pope N, Bosco AM. Student nurses' digital literacy levels: lessons for curricula. Comput Inform Nurs 2020 Mar 13;38(9):451-458. [doi: 10.1097/CIN.0000000000000615] [Medline: 33955370]
14. Elder BL, Koehn ML. Assessment tool for nursing student computer competencies. Nurs Educ Perspect 2009;30(3):148-152. [Medline: 19606656]
15. Miller LA, Stimely ME, Matheny PM, Pope MF, McAtee RE, Miller KA. Novice nurse preparedness to effectively use electronic health records in acute care settings: critical informatics knowledge and skill gaps. Online J Nurs Inform 2014;18(2).
16. Taherdoost H. A review of technology acceptance and adoption models and theories. Procedia Manuf 2018;22:960-967. [doi: 10.1016/j.promfg.2018.03.137]
17. Perlusz S. Emotions and technology acceptance: development and validation of a technology affect scale. In: Proceedings of the IEEE International Engineering Management Conference. 2004 Presented at: IEMC 2004; October 18-21, 2004; Singapore, Singapore. [doi: 10.1109/iemc.2004.1407500]
18. Edison SW, Geissler GL. Measuring attitudes towards general technology: antecedents, hypotheses and scale development. J Target Meas Anal Mark 2003 Nov 1;12(2):137-156. [doi: 10.1057/palgrave.jt.5740104]
19. Anderberg P, Eivazzadeh S, Berglund JS. A novel instrument for measuring older people's attitudes toward technology (TechPH): development and validation. J Med Internet Res 2019 May 23;21(5):e13951 [FREE Full text] [doi: 10.2196/13951] [Medline: 31124467]
20. Anderberg P, Björling G, Stjernberg L, Bohman D. Analyzing nursing students' relation to electronic health and technology as individuals and students and in their future career (the eNursEd study): protocol for a longitudinal study. JMIR Res Protoc 2019 Oct 01;8(10):e14643 [FREE Full text] [doi: 10.2196/14643] [Medline: 31573945]
21. Högskoleförordning (1993:100). The Swedish Ministry of Education and Research. URL: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/hogskoleforordning-1993100_sfs-1993-100/ [accessed 2024-04-17]
22. Regulation of the Minister of Science and Higher Education of May 9, 2012 on education standards for the following fields of study: medicine, dentistry, pharmacy, nursing and midwifery. Sejm of the Republic of Poland. 2012. URL: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20120000631 [accessed 2024-04-18]
23. Eastwood KW, May R, Andreou P, Abidi S, Abidi SS, Loubani OM. Needs and expectations for artificial intelligence in emergency medicine according to Canadian physicians. BMC Health Serv Res 2023 Jul 25;23(1):798 [FREE Full text] [doi: 10.1186/s12913-023-09740-w] [Medline: 37491228]
24. Lee AR, McDermott O, Orrell M. Understanding barriers and facilitators to online and app activities for people living with dementia and their supporters. J Geriatr Psychiatry Neurol 2023 Sep 04;36(5):366-375 [FREE Full text] [doi: 10.1177/08919887221149139] [Medline: 36597870]

XSL•FO
RenderX

25. Lee J, Lee EH, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021 Nov 15;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
26. Park H, Lee E. Self-reported eHealth literacy among undergraduate nursing students in South Korea: a pilot study. Nurse Educ Today 2015 Feb;35(2):408-413. [doi: 10.1016/j.nedt.2014.10.022] [Medline: 25466791]
27. WMA declaration of Helsinki – ethical principles for medical research involving human subjects. World Medical Association. 2022 Sep 6. URL: https://tinyurl.com/3kakpnaw [accessed 2023-11-30]
28. Lag (2003:460) om etikprövning av forskning som avser människor. Utbildningsdepartementet (The Ministry of Education). 2003. URL: https://tinyurl.com/w7xf39pj [accessed 2024-04-10]
29. Staddon RV. Bringing technology to the mature classroom: age differences in use and attitudes. Int J Educ Technol High Educ 2020 Mar 23;17:11. [doi: 10.1186/s41239-020-00184-4]
30. Shelton C. “Virtually mandatory”: a survey of how discipline and institutional commitment shape university lecturers’ perceptions of technology. Brit J Educ Tech 2013 May 09;45(4):748-759. [doi: 10.1111/bjet.12051]
31. Thomas A, Thorne G. How to increase higher order thinking. Center for Development and Learning. 2009. URL: http://www.thekeytotorah.com/uploads/2/5/5/8/25587179/how_to_increase_higher_order_thinking0001.pdf [accessed 2024-04-18]
32. Charles-Ogan G, Williams C. Flipped classroom versus a conventional classroom in the learning of mathematics. Br J Educ 2015 Jun;3(6):71-77 [FREE Full text]
33. Marshall EM, Staddon RV, Wilson DA, Mann VE. Addressing maths anxiety and engaging students with maths within the curriculum. MSOR Connections. 2017. URL: https://tinyurl.com/22szdmy7 [accessed 2024-04-18]
34. Labrague LJ, McEnroe-Petitte DM, Bowling AM, Nwafor CE, Tsaras K. High-fidelity simulation and nursing students' anxiety and self-confidence: a systematic review. Nurs Forum 2019 Jul 10;54(3):358-368. [doi: 10.1111/nuf.12337] [Medline: 30852844]
35. Fossum M, Opsal A, Ehrenberg A. Nurses' sources of information to inform clinical practice: an integrative review to guide evidence-based practice. Worldviews Evid Based Nurs 2022 Oct 04;19(5):372-379 [FREE Full text] [doi: 10.1111/wvn.12569] [Medline: 35244324]
36. Alving BE, Christensen JB, Thrysøe L. Hospital nurses' information retrieval behaviours in relation to evidence based nursing: a literature review. Health Info Libr J 2018 Mar;35(1):3-23 [FREE Full text] [doi: 10.1111/hir.12204] [Medline: 29327483]
37. Ebenezer C. Nurses’ and midwives’ information behaviour: a review of literature from 1998 to 2014. New Library World 2015;116(3/4):155-172 [FREE Full text] [doi: 10.1108/NLW-07-2014-0085]
38. Schaafsma F, Verbeek J, Hulshof C, van Dijk F. Caution required when relying on a colleague's advice; a comparison between professional advice and evidence from the literature. BMC Health Serv Res 2005 Aug 31;5:59 [FREE Full text] [doi: 10.1186/1472-6963-5-59] [Medline: 16131405]
39. Day-Black C, Merrill EB. Using mobile devices in nursing education. ABNF J 2015;26(4):78-84. [Medline: 26665501]
40. O'Connor S, Andrews T. Smartphones and mobile applications (apps) in clinical nursing education: a student perspective. Nurse Educ Today 2018 Oct;69:172-178. [doi: 10.1016/j.nedt.2018.07.013] [Medline: 30096510]
41. Gambo JM, Bahreman NT, Watties-Daniels D, Neal M, Swoboda SM. Can mobile technology enhance learning and change educational practice? Comput Inform Nurs 2017 Aug;35(8):375-380. [doi: 10.1097/CIN.0000000000000380] [Medline: 28796667]
42. Kenny LA, Gaston T, Powers K, Isaac-Dockery A. Anxiety in nursing students: the impact of using mobile technology with quick response codes. Nurse Educ Today 2020 Jun;89:104382. [doi: 10.1016/j.nedt.2020.104382] [Medline: 32200133]
43. Shah MM, Hassan R, Embi R. Technology acceptance and computer anxiety. In: Proceedings of the International Conference on Innovation Management and Technology Research. 2012 Presented at: ICIMTR 2012; May 21-22, 2012; Malacca, Malaysia. [doi: 10.1109/ICIMTR.2012.6236408]
44. Revilla Muñoz O, Alpiste Penalba F, Fernández Sánchez J, Santos OC. Reducing techno-anxiety in high school teachers by improving their ICT problem-solving skills. Behav Inf Technol 2016 Sep 07;36(3):255-268. [doi: 10.1080/0144929x.2016.1221462]
45. Nes AA, Steindal SA, Larsen MH, Heer HC, Lærum-Onsager E, Gjevjon ER. Technological literacy in nursing education: a scoping review. J Prof Nurs 2021 Mar;37(2):320-334 [FREE Full text] [doi: 10.1016/j.profnurs.2021.01.008] [Medline: 33867086]
46. Ralph N, Birks M, Chapman Y, Francis K. Future-proofing nursing education: an Australian perspective. SAGE Open 2014 Nov 11;4(4). [doi: 10.1177/2158244014556633]
47. Heinssen RK, Glass CR, Knight LA. Assessing computer anxiety: development and validation of the Computer Anxiety Rating Scale. Comput Hum Behav 1987;3(1):49-59. [doi: 10.1016/0747-5632(87)90010-0]
48. Cohen BA, Waugh GW. Assessing computer anxiety. Psychol Rep 1989 Dec 31;65(3 Pt 1):735-738. [doi: 10.2466/pr0.1989.65.3.735] [Medline: 2608829]
49. Parasuraman S, Igbaria M. An examination of gender differences in the determinants of computer anxiety and attitudes toward microcomputers among managers. Int J Man Mach Stud 1990 Mar;32(3):327-340. [doi: 10.1016/s0020-7373(08)80006-5]

XSL•FO
RenderX

50. He J, Freeman LA. Are men more technology-oriented than women? The role of gender on the development of general computer self-efficacy of college students. J Inf Syst Educ 2010;21(2):203.
51. Ketikidis P, Dimitrovski T, Lazuras L, Bath PA. Acceptance of health information technology in health professionals: an application of the revised technology acceptance model. Health Informatics J 2012 Jun 24;18(2):124-134 [FREE Full text] [doi: 10.1177/1460458211435425] [Medline: 22733680]
52. AlQudah AA, Al-Emran M, Shaalan K. Technology acceptance in healthcare: a systematic review. Appl Sci 2021 Nov 09;11(22):10537. [doi: 10.3390/app112210537]
53. De Leeuw JA, Woltjer H, Kool RB. Identification of factors influencing the adoption of health information technology by nurses who are digitally lagging: in-depth interview study. J Med Internet Res 2020 Aug 14;22(8):e15630 [FREE Full text] [doi: 10.2196/15630] [Medline: 32663142]
54. Silva CR, Lopes RH, de Goes Bay O, Martiniano CS, Fuentealba-Torres M, Arcêncio RA, et al. Digital health opportunities to improve primary health care in the context of COVID-19: scoping review. JMIR Hum Factors 2022 May 31;9(2):e35380 [FREE Full text] [doi: 10.2196/35380] [Medline: 35319466]
55. Deane FP, Podd J, Henderson RD. Relationship between self-report and log data estimates of information system usage. Comput Hum Behav 1998 Dec;14(4):621-636. [doi: 10.1016/S0747-5632(98)00027-2]

Abbreviations

BTH: Blekinge Institute of Technology
ECTS: European Credit Transfer and Accumulation System
eHEALS: eHealth Literacy Scale
MUB: Medical University of Bialystok
SRCU: Swedish Red Cross University
STROBE: Strengthening the Reporting of Observational studies in Epidemiology
TechAnxiety: technology anxiety
TechEnthusiasm: technology enthusiasm
TechPH: Technophilia instrument

Edited by T de Azevedo Cardoso; submitted 26.06.23; peer-reviewed by O Navarro, T Xu, R Staddon; comments to author 24.10.23; revised version received 15.12.23; accepted 22.03.24; published 29.04.24.

Please cite as:
Dallora AL, Andersson EK, Gregory Palm B, Bohman D, Björling G, Marcinowicz L, Stjernberg L, Anderberg P
Nursing Students' Attitudes Toward Technology: Multicenter Cross-Sectional Study
JMIR Med Educ 2024;10:e50297
URL: https://mededu.jmir.org/2024/1/e50297
doi:10.2196/50297
PMID:38683660

©Ana Luiza Dallora, Ewa Kazimiera Andersson, Bruna Gregory Palm, Doris Bohman, Gunilla Björling, Ludmiła Marcinowicz, Louise Stjernberg, Peter Anderberg. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.04.2024. 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.

XSL•FO
RenderX

Original Paper

Exploring HTML5 Package Interactive Content in Supporting Learning Through Self-Paced Massive Open Online Courses on Healthy Aging: Mixed Methods Study

Pratiwi Rahadiani[1], BEd, MT, MSc; Aria Kekalih[1,2], MD, MTI, PhD; Diantha Soemantri[3], MD, MMEdEd, PhD; Desak Gede Budi Krisnamurti[1,4], BPharm, MSc, PhD

[1]Center of e-Learning Cluster, Indonesia Medical Education and Research Institute, Faculty of Medicine, Universitas Indonesia, Central Jakarta, Indonesia

[2]Department of Community Medicine, Faculty of Medicine, Universitas Indonesia, Central Jakarta, Indonesia

[3]Department of Medical Education, Faculty of Medicine, Universitas Indonesia, Central Jakarta, Indonesia

[4]Department of Medical Pharmacy, Faculty of Medicine, Universitas Indonesia, Central Jakarta, Indonesia

Corresponding Author:
Desak Gede Budi Krisnamurti, BPharm, MSc, PhD
Center of e-Learning Cluster, Indonesia Medical Education and Research Institute
Faculty of Medicine
Universitas Indonesia
Jl. Salemba Raya No. 6
Central Jakarta, 10430
Indonesia
Phone: 62 2129189160
Email: desak.gede@ui.ac.id

Abstract

Background: The rapidly aging population and the growth of geriatric medicine in the field of internal medicine are not supported by sufficient gerontological training in many health care disciplines. There is rising awareness about the education and training needed to adequately prepare health care professionals to address the needs of the older adult population. Massive open online courses (MOOCs) might be the best alternative method of learning delivery in this context. However, the diversity of MOOC participants poses a challenge for MOOC providers to innovate in developing learning content that suits the needs and characters of participants.

Objective: The primary outcome of this study was to explore students' perceptions and acceptance of HTML5 package (H5P) interactive content in self-paced MOOCs and its association with students' characteristics and experience in using MOOCs.

Methods: This study used a cross-sectional design, combining qualitative and quantitative approaches. Participants, predominantly general practitioners from various regions of Indonesia with diverse educational backgrounds and age groups, completed pretests, engaged with H5P interactive content, and participated in forum discussions and posttests. Data were retrieved from the online questionnaire attached to a selected MOOC course. Students' perceptions and acceptance of H5P interactive content were rated on a 6-point Likert scale from 1 (strongly disagree) to 6 (strongly agree). Data were analyzed using SPSS (IBM Corp) to examine demographics, computer literacy, acceptance, and perceptions of H5P interactive content. Quantitative analysis explored correlations, while qualitative analysis identified recurring themes from open-ended survey responses to determine students' perceptions.

Results: In total, 184 MOOC participants agreed to participate in the study. Students demonstrated positive perceptions and a high level of acceptance of integrating H5P interactive content within the self-paced MOOC. Analysis of mean (SD) value across all responses consistently revealed favorable scores (greater than 5), ranging from 5.18 (SD 0.861) to 5.45 (SD 0.659) and 5.28 (SD 0.728) to 5.52 (SD 0.627), respectively. This finding underscores widespread satisfaction and robust acceptance of H5P interactive content. Students found the H5P interactive content more satisfying and fun, easier to understand, more effective, and more helpful in improving learning outcomes than material in the form of common documents and learning videos. There is a significant correlation between computer literacy, students' acceptance, and students' perceptions.

Conclusions: Students from various backgrounds showed a high level of acceptance and positive perceptions of leveraging H5P interactive content in the self-paced MOOC. The findings suggest potential new uses of H5P interactive content in MOOCs,

XSL•FO
RenderX

such as interactive videos with pop-up questions, to substitute for synchronous learning. The study underscores the significance of tailored educational strategies in supporting the professional development of health care professionals.

(JMIR Med Educ 2024;10:e45468) doi:10.2196/45468

KEYWORDS

HTML5 package; H5P; students' perspectives; students' acceptance; massive open online courses; MOOCs; healthy aging; self-paced MOOC; student; perception; acceptance; opinion; attitude; MOOC; self-paced; self-guided; online course; online learning; geriatric; gerontology; gerontological; learning

Introduction

Indonesia is one of the most populous countries in the world, with a population of about 268 million people. In 2021, there were 10.8% or around 29.3 million older adults in the population. By 2045, the ratio of older adults is expected to increase to 19.9% [1]. The rising number of older adults poses potential challenges and vulnerability in medical, psychological, economic, and social domains within this population. The aging process is unavoidable and can potentially bring various challenges in terms of health and quality of life [2]. Furthermore, it contributes to increasing mortality from noncommunicable diseases among older adults [3].

In response to this problem, the Indonesian government launched the 2016-2019 National Action Plan for Elderly Health to improve the quality of life of older adults through health service programs. However, this program has not been run optimally. The implementation of geriatric services in hospitals is still not realized. The policy exists and the implementation of the program is quite good, but the instrument for measuring it has not been understood and used by service providers [4]. According to the Central Bureau of Statistics data, Indonesia had 173,779 doctors and 2,287,142 health workers spread throughout the country in 2021 [5,6]. However, data from the Health Human Resources Information System in 2018 reported that only 23% of health community service centers had qualified health care professionals and 15% had none [7]. This indicates that the coverage of health services for older adults has not met expectations and is unevenly distributed [8]. Doctors and health care professionals need to be equipped with the knowledge and skills to ensure a healthy and independent older adult population.

Moreover, while geriatric medicine is a rapidly growing field within internal medicine, gerontological training is still limited across many health care disciplines today [9,10]. This highlights the importance of raising awareness about education and training in geriatric health care, ensuring health professionals are adequately prepared to address the needs of the older adult population. Therefore, education for health professionals and continuing education for practitioners is required to reframe medical care to meet the needs and personal quality of life goals of older adults [11].

Many health care professionals experience difficulties traveling to attend face-to-face continuing training due to their busy work schedules and their location in remote rural areas. However, time and travel restrictions during the COVID-19 pandemic catalyzed a shift from face-to-face to online education [12]. As a form of e-learning, massive open online courses (MOOCs) have emerged as an invaluable tool to address some of the training challenges experienced in developing countries by supporting content delivery that is efficient, cost-effective, and accessible [13]. MOOCs can provide training for health care professionals and increase the dissemination of information about public health issues to the public [14]. Skinner et al [15] found that health care professionals need MOOC content that is easy to adapt and share. Thus, MOOCs can effectively deliver learning materials in health care education and continuing education for practitioners.

Despite the many benefits of MOOCs, such as their open nature and enabling teachers to reach a large and diverse group of participants, there is a perception of social isolation regarding the lack of interaction between teachers and students and between students [16].

Furthermore, the diversity of participants makes it difficult to engage students, which further adds to the complexity of the students' interactions with the course content [17], even though such interaction is the most common interaction form in MOOCs [18]. This challenges MOOC providers to develop innovative learning content that suits the needs and characters of users to increase student engagement and mastery of learning. The existing MOOC platforms use videos as the main information delivery method. The videos are presented in a 1-way format where students are passive viewers, making learning feel monotonous [18,19].

The use of the HTML5 package (H5P) to develop interactive learning content can make the class more interactive and fun and encourage self-directed learning. Moreover, H5P allows students' learning outcomes to be recorded so that they can be used as evaluation material for instructors [20]. H5P is a simple and easy-to-use open-source technology without the need for plugins or Shareable Content Object Reference Model (SCORM) standards. The interactive learning content can be developed in formats such as interactive videos, course presentations, image hotspots, and branching scenarios [20,21].

Previous studies have demonstrated how H5P interactive content can support online learning environments, including blended learning, flipped classrooms, active learning, and virtual simulations. However, few studies have focused specifically on the use of H5P within Moodle-based MOOC platforms, especially in health profession education and continuing education [22-24]. Moreover, limited research exists on Indonesian students' perspectives regarding their participation in self-paced MOOCs with interactive content H5P, especially in the context of healthy aging. This study explored students'

XSL•FO
RenderX

perceptions and levels of acceptance of H5P interactive content in a self-paced MOOC.

Methods

Design and Settings

The study used a combination of qualitative and quantitative approaches. The participants enrolled in a course comprised of 8 topics on the MOOC platform (Table 1). The course included a pretest, learning material, forum discussions, and a posttest on every topic. At the end of the course, we asked participants to complete a self-reflection task regarding their experience during the course. The course was developed using a self-paced learning method so participants could progress at their own pace. Learning material provided in the H5P took the form of interactive books and interactive videos. There were between 3 and 8 videos per topic in the module, with an average duration of around 12 minutes for each video, and each interactive video included pop-up questions. A survey was attached at the end of the course for students to answer questions on their acceptance of H5P interactive content and their views on whether it supports self-paced learning.

Table 1. Tittle and estimated duration to complete for each topic.

Topic	Tittle	Duration (hours)
1	The Role of Healthy Aging and Risk Factors That Can Lead to Non-Communicable Disease	2
2	Elderly and its Problems	2
3	Metabolic and Hormonal Aspects of Aging	2
4	Risk Factors and Cardiovascular Disease in the Elderly	4
5	Neurodegenerative Diseases in the Elderly and Prevention	2
6	The Role of Physical Activity as Prevention of Non-Communicable Diseases in the Elderly	3
7	The Role of Nutrition as Prevention of Non-Communicable Diseases in the Elderly	8
8	The Role of Social Support and Medical Funding in Overcoming Diseases in the Elderly	4

Participants

Due to their scattered locations throughout Indonesia, the participants were recruited via email and WhatsApp (Meta Platforms). We identified potential participants through various educational networks and organizations dedicated to supporting students across Indonesia. Specifically, we leveraged our existing contacts within these networks and used databases of educational institutions. Additionally, participants were identified based on recommendations from local contacts familiar with the target demographics. Furthermore, to ensure the relevance of participants, we specifically targeted general practitioners practicing in diverse locations throughout Indonesia. The participants were enrolled in an online course on the MOOC platform from August to September 2021. There were 796 participants in the course. This study used purposive sampling as its sampling method. The sample was selected based on participants' completion of the course. A total of 184 participants completed the course and were selected to participate and analyze in this study.

Ethical Considerations

The ethics committee of the Faculty of Medicine, Universitas Indonesia-Cipto Mangunkusumo Hospital, has approved this study (KET-511/UNI2.F1/ETIK/PPM.00.02/2021). Informed consent was obtained from all participants involved in the study. Confidentiality and anonymity were preserved during data collection and processing.

Data Collection

We retrieved data from the online questionnaire attached to the MOOC course. The questionnaires were adapted and modified from a variety of existing instruments [25-27]. The questionnaire was written in Bahasa Indonesia and included items on the demographic characteristics of students (age range, sex, education, and experience of MOOCs), their computer literacy rated on a 5-point Likert scale from 1 (bad) to 5 (excellent), students' perceptions and acceptance of H5P interactive content rated on a 6-point Likert scale from 1 (strongly disagree) to 6 (strongly agree). The student perception questionnaire was adapted and modified from the study by Muthuprasad et al [27]. The open-ended question was used to emphasize findings from the students' perception questionnaire. The researcher with experts was involved in reviewing the questionnaire items and validated it using content validation methods. Meanwhile, students' acceptance was measured using the Technology Acceptance Model (TAM) adopted and modified from the study by Masrom [26], which consists of 2 variables—perceived ease of use and perceived usefulness. According to Davis [28], these 2 variables of TAM are the 2 main factors that influence behavioral intentions toward new technologies and they affect the actual use of the technologies [28]. The validity and reliability of the instrument were tested in the previous study. The result indicates that the validity test of students' perceived ease of use variables ranged from 0.783 to 1.000, and perceived usefulness ranged from 0.731 to 0.977. Cronbach α values of perceived ease of use and perceived usefulness were 0.975 and 0.973, respectively. Scores greater than 0.7 indicate reliability and good internal consistency [29].

Quantitative Analysis

We analyzed the data using SPSS (version 25.0; IBM Corp). Descriptive analysis was used to analyze students' characteristics, perceptions, and acceptance of the H5P content. A bivariate correlation (Pearson correlation) was conducted to determine the relationships among predictor factors such as

XSL•FO
RenderX

students' computer literacy and their perceptions and acceptance of H5P interactive content in terms of perceived ease of use and perceived usefulness based on the TAM. Conditions for multicollinearity and homoscedasticity were met. No multicollinearity between independent and dependent variables was found in the preliminary study. The variance inflation factor values were all below 5. We investigated perceived ease of use, perceived usefulness, and students' perception as dependent variables, and age, sex, education, computer literacy, and MOOC experience as independent variables.

Qualitative Analysis

In the qualitative analysis, we identified themes related to students' perceptions and opinions from the open-ended survey questions using a constant comparison process [30] and inductive analysis. First, the researchers read each response to gain a grasp of the information. Second, they broke the data into smaller units, coded and labeled the units according to the content they contained. Finally, categories and overarching themes were defined. During the analysis, the data were constantly compared back and forth between the current data and previous data that had been coded. All data analysis was completed manually using Microsoft Excel.

Results

Student Characteristics

Table 2 presents all the detailed characteristics of students. The participants ranged in age from 36 to 45 years old, with 70.7% (n=130) female participants. Most participants live on Java Island and the rest are scattered throughout Indonesia (Figure 1 [31]).

The educational backgrounds of students were dominated by participants who held bachelor's (40.2%, n=74) and master's (42.9%, n=79) degrees. Most of the participants were doctors (66.8%, n=123). In terms of MOOC experience, most participants had no previous experience of taking MOOCs.

XSL•FO
RenderX

Table 2. Students' characteristics (N=184).

Characteristics	Values, n (%)
Age range (years)	
<25	33 (17.9)
26-35	49 (26.6)
36-45	63 (34.2)
46-55	26 (14.1)
>56	13 (7.1)
Sex	
Male	54 (29.3)
Female	130 (70.7)
Domicile	
Nanggroe Aceh Darussalam	1 (0.5)
Sumatera Barat	4 (2.2)
Riau	1 (0.5)
Kepulauan Riau	1 (0.5)
Jambi	3 (1.6)
Bengkulu	1 (0.5)
Sumatera Selatan	3 (1.6)
Lampung	7 (3.8)
Banten	10 (5.4)
Daerah Khusus Ibukota Jakarta	57 (31.0)
Jawa Barat	37 (20.1)
Jawa Tengah	12 (6.5)
Jawa Timur	25 (13.6)
Daerah Istimewa Yogyakarta	5 (2.7)
Bali	4 (2.2)
Nusa Tenggara Timur	1 (0.5)
Kalimantan Barat	2 (1.1)
Kalimantan Selatan	1 (0.5)
Kalimantan Timur	1 (0.5)
Sulawesi Selatan	5 (2.7)
Sulawesi Tenggara	2 (1.1)
Overseas	1 (0.5)
Education	
High school	5 (2.7)
Bachelor's degree	74 (40.2)
Master's degree	79 (42.9)
Doctoral degree	26 (14.1)
Occupation	
Student	11 (6.0)
Doctor	123 (66.8)
Health worker	2 (1.1)
Lecturer	48 (26.1)

XSL•FO
RenderX

Characteristics	Values, n (%)
MOOC[a] experience	
Yes	63 (34.2)
No	121 (65.8)

[a]MOOC: massive open online courses.

Figure 1. Characteristics of students based on the domicile area (adapted from Vemaps [31], which is published under Creative Commons Attribution (CC-BY) [32].

Computer Literacy (Self-Efficacy)

Computer literacy or computer self-efficacy of students who participated in the course are presented in Table 3. The mean value of computer literacy responses ranged from 3.92 to 4.28, and the SD ranged from 0.725 to 0.871. Overall, students' computer literacy was good. The item on computer maintenance ability (installing programs, applications, or software) had the lowest mean value of 3.92 (SD 0.871), and the item related to browsing skills had the highest mean value of 4.28 (SD 0.751).

Table 3. Computer literacy of participants (N=184).

Statements	Values, mean (SD)
CL1—Ability to use computer	4.21 (0.725)
CL2—Basic computer operation skills (typing accuracy and speed, moving cursor)	4.24 (0.773)
CL3—Running computer programs, software, or applications	4.12 (0.759)
CL4—Computer maintenance ability (installing programs or applications)	3.92 (0.871)
CL5—Browsing skills	4.28 (0.751)

Perception and Opinion of the Participant Toward H5P Interactive Content Usage

Table 4 shows the results of a descriptive analysis of students' perceptions of H5P interactive content. The mean value of all responses falls between 5.18 and 5.45, and the SDs range from 0.605 to 0.861, indicating that students generally have a positive perception toward H5P interactive content leveraging. The results indicate that students perceived H5P interactive content as fun, easier, and more helpful in promoting understanding and improving learning outcomes, and more satisfying than material in the form of common documents and learning videos. Thus, students' perceptions of how interactive content can replace synchronous learning had the lowest mean value of 5.18.

XSL•FO
RenderX

Table 4. Students' perception toward H5P[a] interactive content (N=184).

Statement	Values, mean (SD)	Range
SP1—I prefer the presentation of material in interactive form (interactive video and interactive book) than in the form of common documents or learning videos	5.45 (0.659)	3-6
SP2—Presentation of material in interactive form (interactive video and interactive book) provides a more enjoyable learning experience than presentation in the form of common documents or learning videos	5.43 (0.633)	3-6
SP3—Presentation of material in interactive form (interactive video and interactive book) makes it easier for me to understand learning compared to presentation in the form of common documents or learning videos	5.43 (0.605)	3-6
SP4—Presentation of material in interactive form (interactive video and interactive book) is more effective than presentation in the form of common documents or learning videos	5.38 (0.675)	1-6
SP5—Presentation of material in interactive form (interactive video and interactive book) is more effective than compared to presentation in the form of common documents or learning videos	5.36 (0.687)	1-6
SP6—Presentation of material in interactive form (interactive videos and interactive books) makes me more focused than presenting in the form of common documents or learning videos	5.38 (0.691)	1-6
SP7—I am more satisfied with the presentation of material in interactive form (interactive videos and interactive books) than in the form of common documents or learning videos	5.34 (0.786)	1-6
SP8—Presentation of material in interactive form (interactive video and interactive book) can replace synchronous learning	5.18 (0.861)	1-6

[a]H5P: HTML5 package.

The students' perceptions and opinions about H5P interactive content were also explored in the open-ended questions supporting the findings from the perception assessment. According to the responses, students believe that H5P interactive content could increase focus and memory retention, as illustrated by the following quotes.

> *Overall, the learning content is very good. If possible, learning videos with pop-up questions are applied to all materials because they can increase focus and memory retention on related materials.* [Female, 27 years old]

> *…Giving interactive book material in approximately 15 minutes can maintain concentration. Moreover, having a quiz in the middle of a video can help me to stay focused.* [Female, 45 years old]

> *…Videos with pop-up questions really help to increase my attention when taking lessons…* [Female, 28 years old]

Furthermore, students also noted the advantages of interactive videos with pop-up questions. They reported that they provide a different and better learning experience than synchronous learning through Zoom (Zoom Video Communications), are good and interactive, and can cross-check their understanding during the learning process. So, students suggest adding more questions during the video, as highlighted in the quotes.

> *Increasing number of pop-up questions in the middle of learning materials…* [Female, 30 years old]

> *I suggest applying pop-up questions in the midst of all interactive videos, so that it will provide a different and better learning experience than learning through Zoom.* [Male, 24 years old]

> *I think interactive videos with pop-up questions are good and interactive. If possible, every video should have an interaction (pop-up questions) so we can cross-check our understanding during the learning process.* [Male, 31 years old]

Students' Acceptance of H5P Interactive Content

The findings presented in Table 5 indicate that students had a high level of acceptance of H5P interactive content. All 5 items measuring perceived ease of use and the 5 items measuring perceived usefulness recorded mean values above 5 points. This indicates that respondents found the H5P interactive content easy to use and effective for helping them learn on the MOOC.

XSL•FO
RenderX

Table 5. Students' acceptance to H5P[a] interactive content (N=184).

Statements	Values, mean (SD)	Range
Perceived ease of use		
PEU1—I can easily access the interactive learning content available in this module	5.36 (0.663)	3-6
PEU2—The navigation provided in the interactive learning content in this module can be easily understood	5.28 (0.728)	3-6
PEU3—I can easily understand the instructions given on the interactive learning content in this module	5.38 (0.683)	3-6
PEU4—I can easily operate the interactive learning content in this module	5.37 (0.742)	2-6
PEU5—Overall interactive learning content is easy to use	5.45 (0.651)	3-6
Perceived usefulness		
PU1—The questions that arise during the learning video playback can increase my attention (attention) to the material presented	5.52 (0.627)	3-6
PU2—The interactive video presented was able to trigger my curiosity further about the material	5.41 (0.679)	3-6
PU3—The interactive video presented was able to make it easier for me to understand the material	5.40 (0.602)	4-6
PU4—The use of interactive learning content (H5P) allows me to better manage my study time	5.43 (0.658)	3-6
PU5—Overall, I find it helpful to have H5P interactive learning content	5.51 (0.572)	4-6

[a]H5P: HTML5 package.

H5P Interactive Content Acceptance and Student Performance in the MOOC

This study examined several factors, including age range, sex, education, MOOC experience, computer literacy, perceived ease of use, perceived usefulness, and students' perception. The correlations between these factors are presented in Table 6. The results reveal a significantly negative correlation between computer literacy and both age and sex. In contrast, there is a significantly positive correlation between computer literacy and MOOC experience. Computer literacy had a positive impact on perceived ease of use, perceived usefulness, and students' perception.

Table 6. Correlation among variables.

	1	2	3	4	5	6	7
Age range (years)	1	—[a]	—	—	—	—	—
Sex	0.237[b]	1	—	—	—	—	—
Education	0.587[b]	0.182[c]	1	—	—	—	—
MOOC[d] experience	–0.226[b]	–0.315[b]	–0.034	1	—	—	—
Computer literacy	–0.441[b]	–0.172[c]	–0.135	0.277[b]	1	—	—
Perceived ease of use	–0.124	–0.050	–0.112	0.065	0.395[b]	1	—
Perceived usefulness	–0.079	0.054	–0.154[c]	0.024	0.275[b]	0.795[b]	1
Students' perception	–0.140	0.007	–0.119	0.037	0.339[b]	0.662[b]	0.752[b]

[a]Not applicable.
[b]Correlation is significant at 0.01 level (2-tailed).
[c]Correlation is significant at 0.05 level (2-tailed).
[d]MOOC: massive open online courses.

Discussion

This study explored students' perceptions and acceptance of H5P interactive content. We found that students have positive perceptions and acceptance, particularly of interactive books and videos with pop-up questions. The lack of instructor presence and student-instructor interactions in a self-paced MOOC potentially hampers students' commitment and intention to commit to further learning. Therefore, student-content interaction needs to be maximized [33]. Student-content interaction may include reading information, watching videos, completing assignments, interacting with computer-based multimedia, using simulations, searching for information, and working on projects [34]. According to Kuo et al [35], students reported that interaction with the course content increased their satisfaction with the course. MOOCs offer openly accessible online participation, meaning participants can be from diverse backgrounds. The data on the characteristics of participants in

XSL•FO
RenderX

this study show various backgrounds in terms of age, sex, education, job and employment status, computer literacy, previous knowledge of the material or topic, and experience with MOOCs. This indicates that such learning content should be designed to meet the diverse needs of students.

Our findings indicate that students from diverse backgrounds showed a high level of acceptance of H5P interactive content in the self-paced MOOC. They preferred H5P interactive content over traditional teaching methods because it is more fun, easier, more effective, more helpful in facilitating learning and improving learning outcomes, and more satisfying. Moreover, students also found that H5P interactive content (interactive video) provided a different and better learning experience than synchronous learning. Previous research suggests that more satisfying experiences will lead to better learning outcomes [36].

Another interesting finding emerged from the open-ended question analysis. We found that students perceived that pop-up questions during a video increased their focus and memory retention. They even suggested incorporating more questions during the video. This feedback has led to another research question regarding the optimum number of questions needed to make learning more effective and less monotonous. While previous studies have explored the relationship between the length of the intervals between questions and the rates of correct answers [37], the ideal number of questions has yet to be investigated.

Al-Adwan [38] defined computer self-efficacy as individuals' beliefs about their computer abilities and skills. Research indicates that computer skills are one of the critical factors in successful e-learning. Moreover, many empirical studies have demonstrated that computer self-efficacy is a key predictor of perceived ease of use and usefulness in e-learning [34,35]. This aligns with the findings of this study, which indicated computer self-efficacy had a significantly positive impact on participants' perception and acceptance of the perceived ease of use and usefulness of H5P interactive content in the self-paced MOOC. Students with high levels of computer self-efficacy are less likely to be discouraged by difficulties [39]. In contrast, students with low computer self-efficacy are more likely to give up when faced with challenges. To improve the learning experience, we suggest providing a step-by-step written or video tutorial to familiarize people with lower computer literacy. Students' prior experience of MOOCs positively correlated with computer literacy but had no significant impact on their perception of ease of use and usefulness of H5P interactive content. This contrasts with the findings of Wang et al [40], which demonstrated that students' prior experiences were positively associated with their satisfaction with their current learning experience.

This study has some limitations. First, the sample size is small and the participants share individual opinions. We used open-ended questions at the end of the questionnaire to provide more detailed responses to the survey responses; however, we did not follow up with the participants to confirm their answers. This survey is subject to individual variations among participants; their personal circumstances, work, and educational routines might have influenced their answers. Moreover, this study is also limited to the type of H5P interactive content used, namely H5P interactive video. Nevertheless, these study findings can form the basis for a pilot project to further analyze the potential of H5P interactive content to improve the interaction and engagement of students in self-paced MOOCs. We found that H5P interactive content videos with pop-up questions can substitute for synchronous learning; however, further study is necessary to examine its impact on learning outcomes.

In conclusion, this research suggests the use of H5P interactive content, especially interactive books with pop-up questions, can potentially substitute for synchronous learning in the context of self-paced MOOCs. Positive perceptions and high-level acceptance by students toward the use of H5P interactive content suggest that it is suitable for diverse participants of MOOCs from various regions of Indonesia, with diverse educational backgrounds and age groups. Future research should compare students' learning performance in self-paced MOOCs with and without H5P interactive content.

Acknowledgments

The authors would like to acknowledge the support given by Indonesia Medical Education and Research Institute (IMERI), Faculty of Medicine, Universitas Indonesia and Healthy Aging Module Team who made this research work. This research was funded by PT. Nutricia Indonesia Sejahtera and also supported by Universitas Indonesia through Publikasi Terindeks Internasional (PUTI) grants 2022 with contract (NKB-583/UN2.RST/HKP.05.00/2022).

Data Availability

All data generated or analyzed during this study are included in this published paper.

Authors' Contributions

PR and DGBK performed the conceptualization. PR, AK, and DS contributed to the methodology. PR and AK contributed to the software. DS, DGBK, and AK performed the validation. PR and DS did the formal analysis. PR did the investigation. PR contributed to the resources. PR contributed to the data curation. PR contributed to writing—original draft preparation. DS, AK, and DGBK contributed to writing—review and editing. PR and AK performed the visualization. DGBK did the supervision. PR did the project administration. PR and DGBK did the funding acquisition. All authors have read and agreed to the published version of the paper.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Statistik Penduduk Lanjut Usia 2021 [2021 statistics on the older adult population]. Badan Pusat Statistik (BPS). Jakarta: © Badan Pusat Statistik URL: https://www.bps.go.id/publication/2021/12/21/c3fd9f27372f6ddcf7462006/statistik-penduduk-lanjut-usia-2021.html [accessed 2022-08-11]
2. Basrowi RW, Rahayu EM, Khoe LC, Wasito E, Sundjaya T. The road to healthy ageing: what has Indonesia achieved so far? Nutrients 2021;13(10):3441 [FREE Full text] [doi: 10.3390/nu13103441] [Medline: 34684441]
3. Tham TY, Tran TL, Prueksaritanond S, Isidro J, Setia S, Welluppillai V. Integrated health care systems in Asia: an urgent necessity. Clin Interv Aging 2018;13:2527-2538 [FREE Full text] [doi: 10.2147/CIA.S185048] [Medline: 30587945]
4. Policy paper Analisis Kebijakan Mewujudkan Lanjut Usia Sehat Menuju Lanjut Usia Aktif (Active Ageing) [Policy paper: policy analysis to realise healthy elderly towards active elderly (active ageing)]. Center for Analysis of Health Determinant Ministry of Health Republic of Indonesia.: Ministry of Health Republic of Indonesia; 2019. URL: http://www.padk.kemkes.go.id/uploads/download/Analisis_Lansia.pdf [accessed 2022-09-16]
5. Mahmudan A. Indonesia miliki 173.779 dokter pada 2021 [Indonesia will have 173,779 doctors in 2021]. Data Indonesia. 2022 May 20. URL: https://dataindonesia.id/ragam/detail/indonesia-miliki-173779-dokter-pada-2021 [accessed 2022-09-15]
6. Mahmudan A. Tenaga Kesehatan Indonesia Didominasi Perawat pada 2021 [Indonesian health workers will be dominated by nurses in 2021]. Data Indonesia. 2022 May 12. URL: https://dataindonesia.id/ragam/detail/tenaga-kesehatan-indonesia-didominasi-perawat-pada-2021 [accessed 2022-09-15]
7. Rencana AKSI program 2020-2024 [Action plan for the 2020-2024 program]. Badan Penelitian dan Pengembangan Kesehatan Kementerian Kesehatan Republik Indonesia. 2020. URL: https://e-renggar.kemkes.go.id/file_performance/1-416151-01-3tahunan-835.pdf [accessed 2022-09-16]
8. PMK No. 25 Tentang Rencana Aksi Nasional Kesehatan Lanjut Usia Tahun 2016-2019 [Regulation of the Minister of Health of the Republic of Indonesia no. 25 regarding the national action plan for elderly health 2016-2019]. Ministry of Health Republic of Indonesia. 2016. URL: http://hukor.kemkes.go.id/uploads/produk_hukum/PMK_No._25_ttg_Rencana_Aksi_Nasional_Kesehatan_Lanjut_Usia_Tahun_2016-2019_.pdf [accessed 2022-09-16]
9. Nair BKR. Geriatric Medicine: A Problem-Based Approach. Germany: Springer; 2018.
10. Bardach SH, Rowles GD. Geriatric education in the health professions: are we making progress? Gerontologist 2012;52(5):607-618 [FREE Full text] [doi: 10.1093/geront/gns006] [Medline: 22394495]
11. Gugliucci MR, Weaver SA. Array of opportunities in health professions education programs to advance older adult health care. Gerontol Geriatr Educ 2020;41(1):1-3. [doi: 10.1080/02701960.2020.1725262] [Medline: 32009568]
12. Schulte TL, Gröning T, Ramsauer B, Weimann J, Pin M, Jerusalem K, et al. Impact of COVID-19 on continuing medical education-results of an online survey among users of a non-profit multi-specialty live online education platform. Front Med (Lausanne) 2021;8:773806 [FREE Full text] [doi: 10.3389/fmed.2021.773806] [Medline: 34869493]
13. Liyanagunawardena TR, Aboshady OA. Massive open online courses: a resource for health education in developing countries. Glob Health Promot 2018;25(3):74-76. [doi: 10.1177/1757975916680970] [Medline: 28134014]
14. DiBartolo MC, Halick JA. Massive open online courses: a global educational strategy to disseminate geriatric content. J Gerontol Nurs 2021;47(4):3-4 [FREE Full text] [doi: 10.3928/00989134-20210309-04] [Medline: 34038255]
15. Skinner NA, Job N, Krause J, Frankel A, Ward V, Johnston JS. The use of open-source online course content for training in public health emergencies: mixed methods case study of a COVID-19 course series for health professionals. JMIR Med Educ 2023;9:e42412 [FREE Full text] [doi: 10.2196/42412] [Medline: 36735834]
16. Gitlin L, Hodgson N. Online training-can it prepare an eldercare workforce? Generations 2016;40(1):71-81.
17. Al Mamun MA, Lawrie G, Wright T. Exploration of learner-content interactions and learning approaches: the role of guided inquiry in the self-directed online environments. Comput Educ 2022;178:104398. [doi: 10.1016/j.compedu.2021.104398]
18. Julia K, Peter VR, Marco K. Educational scalability in MOOCs: analysing instructional designs to find best practices. Comput Educ 2021;161:104054. [doi: 10.1016/j.compedu.2020.104054]
19. Nawrot I, Doucet A. Building engagement for MOOC students: introducing support for time management on online learning platforms. 2014 Presented at: Proceedings of the 23rd International Conference on World Wide Web; 2014 April 07; Seoul, Korea p. 1077-1082.
20. Brame CJ. Effective educational videos. Vanderbilt University. URL: https://cft.vanderbilt.edu/guides-sub-pages/effective-educational-videos/ [accessed 2020-10-20]
21. Create, share and reuse interactive HTML5 content in your browser. H5P. URL: https://h5p.org/ [accessed 2022-09-16]
22. Magro J. H5P. J Med Libr Assoc 2021;109(2):351-354. [doi: 10.5195/jmla.2021.1204]
23. Sinnayah P, Salcedo A, Rekhari S. Reimagining physiology education with interactive content developed in H5P. Adv Physiol Educ 2021;45(1):71-76 [FREE Full text] [doi: 10.1152/advan.00021.2020] [Medline: 33529142]
24. Killam LA, Luctkar-Flude M. Virtual simulations to replace clinical hours in a family assessment course: development using H5P, gamification, and student co-creation. Clin Simul Nurs 2021;57:59-65. [doi: 10.1016/j.ecns.2021.02.008]

XSL•FO
RenderX

25. Tsironis A, Katsanos C, Xenos M. Comparative usability evaluation of three popular MOOC platforms. 2016 Presented at: IEEE Global Engineering Education Conference (EDUCON); 2016 April 10-13; Abu Dhabi, United Arab Emirates. [doi: 10.1109/educon.2016.7474613]
26. Masrom M. Technology acceptance model and E-learning. In; 2007 Presented at: 12th International Conference on Education, Sultan Hassanal Bolkiah Institute of Education Universiti in Brunei Darussalam; 2007 May 21-24; Brunei.
27. Muthuprasad T, Aiswarya S, Aditya KS, Jha GK. Students' perception and preference for online education in India during COVID -19 pandemic. Soc Sci Humanit Open 2021;3(1):100101 [FREE Full text] [doi: 10.1016/j.ssaho.2020.100101] [Medline: 34173507]
28. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319. [doi: 10.2307/249008]
29. Rahadiani P, Kekalih A, Krisnamurti DGB. Use of H5P interactive learning content in a self-paced MOOC for learning activity preferences and acceptance in an Indonesian medical elective module. Afr J Sci Technol Innov Dev 2023;15(7):844-851. [doi: 10.1080/20421338.2023.2209482]
30. Anderson C. Presenting and evaluating qualitative research. Am J Pharm Educ 2010;74(8):141 [FREE Full text] [doi: 10.5688/aj7408141] [Medline: 21179252]
31. Multicolor map of Indonesia with provinces. Vemaps. 2019 Jul 25. URL: https://vemaps.com/indonesia/id-07 [accessed 2023-11-01]
32. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/
33. Kim D, Jung E, Yoon M, Chang Y, Park S, Kim D, et al. Exploring the structural relationships between course design factors, learner commitment, self-directed learning, and intentions for further learning in a self-paced MOOC. Comput Educ 2021;166:104171. [doi: 10.1016/j.compedu.2021.104171]
34. Bernard RM, Abrami PC, Borokhovski E, Wade CA, Tamim RM, Surkes MA, et al. A meta-analysis of three types of interaction treatments in distance education. Rev Educ Res 2009;79(3):1243-1289. [doi: 10.3102/0034654309333844]
35. Kuo Y, Walker AE, Belland BR, Schroder KEE. A predictive study of student satisfaction in online education programs. Int Rev Res Open Distrib Learn 2013;14(1):16-39.
36. Shih P, Muñoz D, Sánchez F. The effect of previous experience with information and communication technologies on performance in a web-based learning program. Comput Human Behav 2006;22(6):962-970. [doi: 10.1016/j.chb.2004.03.016]
37. Wachtler J, Hubmann M, Zöhrer H, Ebner M. An analysis of the use and effect of questions in interactive learning-videos. Smart Learn Environ 2016;3(1):13. [doi: 10.1186/s40561-016-0033-3]
38. Al-Adwan AS. Investigating the drivers and barriers to MOOCs adoption: the perspective of TAM. Educ Inf Technol 2020;25(6):5771-5795. [doi: 10.1007/s10639-020-10250-z]
39. Salloum SA, Al-Emran M, Shaalan K, Tarhini A. Factors affecting the E-learning acceptance: a case study from UAE. Educ Inf Technol 2018;24(1):509-530. [doi: 10.1007/s10639-018-9786-3]
40. Wang C, Xie A, Wang W, Wu H. Association between medical students' prior experiences and perceptions of formal online education developed in response to COVID-19: a cross-sectional study in China. BMJ Open 2020;10(10):e041886 [FREE Full text] [doi: 10.1136/bmjopen-2020-041886] [Medline: 33122327]

Abbreviations

H5P: HTML5 Package
MOOC: Massive Open Online Course
SCORM: Shareable Content Object Reference Model
TAM: Technology Acceptance Model

Edited by T de Azevedo Cardoso; submitted 03.01.23; peer-reviewed by T Staubitz, H Alshawaf, M Piano; comments to author 19.07.23; revised version received 01.11.23; accepted 22.07.24; published 22.08.24.

Please cite as:
Rahadiani P, Kekalih A, Soemantri D, Krisnamurti DGB
Exploring HTML5 Package Interactive Content in Supporting Learning Through Self-Paced Massive Open Online Courses on Healthy Aging: Mixed Methods Study
JMIR Med Educ 2024;10:e45468
URL: https://mededu.jmir.org/2024/1/e45468
doi:10.2196/45468
PMID:39049507

XSL•FO
RenderX

©Pratiwi Rahadiani, Aria Kekalih, Diantha Soemantri, Desak Gede Budi Krisnamurti. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.08.2024. 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.

XSL•FO
RenderX

Using ChatGPT in Psychiatry to Design Script Concordance Tests in Undergraduate Medical Education: Mixed Methods Study

Alexandre Hudon[1], BEng, MD; Barnabé Kiepura[1], MD; Myriam Pelletier[2]; Véronique Phan[3], MSc, MD

1

2

3

Corresponding Author:
Alexandre Hudon, BEng, MD

Abstract

Background: Undergraduate medical studies represent a wide range of learning opportunities served in the form of various teaching-learning modalities for medical learners. A clinical scenario is frequently used as a modality, followed by multiple-choice and open-ended questions among other learning and teaching methods. As such, script concordance tests (SCTs) can be used to promote a higher level of clinical reasoning. Recent technological developments have made generative artificial intelligence (AI)–based systems such as ChatGPT (OpenAI) available to assist clinician-educators in creating instructional materials.

Objective: The main objective of this project is to explore how SCTs generated by ChatGPT compared to SCTs produced by clinical experts on 3 major elements: the scenario (stem), clinical questions, and expert opinion.

Methods: This mixed method study evaluated 3 ChatGPT-generated SCTs with 3 expert-created SCTs using a predefined framework. Clinician-educators as well as resident doctors in psychiatry involved in undergraduate medical education in Quebec, Canada, evaluated via a web-based survey the 6 SCTs on 3 criteria: the scenario, clinical questions, and expert opinion. They were also asked to describe the strengths and weaknesses of the SCTs.

Results: A total of 102 respondents assessed the SCTs. There were no significant distinctions between the 2 types of SCTs concerning the scenario (P=.84), clinical questions (P=.99), and expert opinion (P=.07), as interpretated by the respondents. Indeed, respondents struggled to differentiate between ChatGPT- and expert-generated SCTs. ChatGPT showcased promise in expediting SCT design, aligning well with *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* criteria, albeit with a tendency toward caricatured scenarios and simplistic content.

Conclusions: This study is the first to concentrate on the design of SCTs supported by AI in a period where medicine is changing swiftly and where technologies generated from AI are expanding much faster. This study suggests that ChatGPT can be a valuable tool in creating educational materials, and further validation is essential to ensure educational efficacy and accuracy.

(JMIR Med Educ 2024;10:e54067) doi:10.2196/54067

KEYWORDS

psychiatry; artificial intelligence; medical education; concordance scripts; machine learning; ChatGPT; evaluation; education; medical learners; learning; teaching; design; support; tool; validation; educational; accuracy; clinical questions; educators

Introduction

Undergraduate Medical Education

Undergraduate medical studies offer a wide range of learning opportunities through various teaching methods for medical students [1]. The competencies required are partly dictated by the Medical Council of Canada, and these skills are regularly assessed throughout the undergraduate medical education (UGME) program. Training programs must incorporate clinical reasoning instruction to aid students in developing this crucial competency [2]. The Bloom taxonomy is a useful tool for clearly identifying the cognitive level targeted by different teaching methods [3]. The taxonomy helps determine the appropriate methods for teaching and evaluating students based on the desired level of competency. Although various teaching methods are used, clinical situations followed by multiple-choice questions, as well as open-ended questions, are commonly used initially [4]. However, these types of questions have limitations when it comes to assessing a student's analysis and clinical reasoning [5]. To address this, script concordance tests (SCTs) can be used to enhance the development of higher-level clinical reasoning skills [6].

The Use of SCTs

Methods such as SCTs are grounded in clinical cases designed to mirror real-life clinical scenarios, where information may be incomplete or unclear. The process involves presenting an initial vignette with some preliminary hypotheses, followed by additional information given to the student. SCTs assess how

XSL•FO
RenderX

this new information influences the likelihood of the initial hypotheses being considered as correct or relevant [6]. Students express the impact on the initial hypothesis using a 5-level Likert scale ranging from “much less likely” to “much more likely.” This process serves as a proxy for clinical reasoning, aiming to replicate decision-making in actual clinical practice. Typically, specialists in the subject develop the cases, and a robust SCT should comprise a minimum of 60 questions for strong internal validity [7-9]. The student’s responses are then compared to those of an expert panel, ideally consisting of at least 10 experts. Research suggests that 15 experts are necessary for high-impact testing, with minimal added benefit beyond 20 experts [10]. A notable limitation of SCTs is acceptability; a study on SCT acceptability with surgical residents revealed that experts tend to be more satisfied than students. Experts found the questions to be representative of real-life clinical settings [11]. However, SCTs may potentially provide a more precise assessment of students’ clinical reasoning compared to multiple-choice questions [12]. In psychiatry, the use of SCTs is emerging. Early data indicate good internal validity, with a correlation between learners’ education level, test scores, and improvement in evaluations tested before and after a psychiatry rotation [13].

The creation of SCTs demands a substantial investment of human resources [14]. Moreover, the questions are influenced by the designers’ inherent biases, necessitating multiple rounds of refinement with field experts [15]. This iterative process can lead to delays in developing educational materials. In a time when efficiency is crucial—such as during the COVID-19 pandemic or in situations with limited teaching resources—swift adaptations and improvements in the effectiveness of certain teaching methods may be imperative to uphold the quality of medical training [16,17].

Large Language Models and Their Uses in SCT Design

For clinician-educators seeking assistance in crafting educational materials, recent advancements include the availability of generative artificial intelligence (AI) tools, including large language models (LLM) such as ChatGPT (OpenAI) [18,19]. Originally designed for the public, these tools are currently under scrutiny by various companies and educational institutions to assess their limitations and advantages [20]. Numerous studies highlight the tool’s utility in developing clinical vignettes within medical studies and other health science domains [21]. However, to date, there is no study demonstrating the educational quality of SCT vignettes produced using ChatGPT. Before integrating tools such as ChatGPT into the design of educational materials, it is crucial to evaluate the quality of scenarios, questions, and related expertise generated by ChatGPT, as well as its ability to assess clinical reasoning. It is equally important to consider the potential limitations in using such tools for medical education material design. Although these generative models can be beneficial, they may also introduce errors that limit their usefulness [18]. As for medical students’ attitude toward AI, a recent study on the subject reported that medical students viewed AI in medicine as reliable, trustworthy, and technically competent, although they expressed limited confidence in its capabilities. While acknowledging AI’s intelligence, they did not consider it to be anthropomorphic. The consensus was that fundamental AI knowledge, covering its operation, ethics, applications, reliability, and potential risks, should be integrated into medical education [22].

Objective and Hypotheses

The primary goal of this project is to investigate how SCTs generated by ChatGPT compare to those produced by clinical experts in 3 key aspects: the scenario (stem), clinical questions, and expert opinion. A secondary objective is to assess whether blind evaluators can distinguish between an SCT generated by ChatGPT and one crafted by experts. Additionally, another subobjective aims to identify the advantages and limitations of the clinical vignettes under examination. Our hypothesis posits that the clinical SCTs created by ChatGPT will likely be considered acceptable by the medical community in terms of scenarios and clinical questions. However, we anticipate that their use with learners may necessitate supervision from clinical experts. Preliminary studies have indicated that AI is a promising tool to aid clinician-educators in designing clinical scenarios. Still, given that the underlying algorithms rely on potentially erroneous data, it is crucial to validate and fine-tune the content before using them as educational materials for learners.

Methods

Ethical Considerations

This study received the approval of the ethics of research committee of the Université de Montréal (approval 2023-4906). Participants were given a description of the study in the letter they received and were asked for their consent for their data to be used. Data were anonymized. The participants received no compensation for this study.

Recruitment

The project was aimed at residents and clinician-educators in the field of psychiatry since SCTs are already used in UGME programs. To be included in the study, participants needed to be either clinician-educators in the field of psychiatry or medical residents in psychiatry affiliated with 1 of Québec’s 4 universities that offer UGME programs (McGill University, Université de Montréal, Université de Sherbrooke, and Université Laval). Psychiatrists not involved in an UGME program were excluded. A total of 100 participants were anticipated for this study, according to similar studies to determine whether there were significant differences between clinical vignettes developed by ChatGPT or those developed by experts [23,24]. Convenience sampling was conducted with the help of the departments of psychiatry of the 4 universities listed above, and a letter was sent out by email that includes a link to a survey that contained all the questions from this study.

Data Collection

A web-based survey, hosted on LimeSurvey (LimeSurvey GmbH), featured 3 SCTs generated by ChatGPT and 3 SCTs previously crafted by experts in the field, currently used in the digital learning environment at the Université de Montréal. The experts consisted of experienced psychiatrists and primary care physicians who underwent training in SCT concepts. As the primary language for the participants is French, the survey was

XSL•FO
RenderX

conducted in French. The original, comprehensive survey in French is available in Multimedia Appendix 1, with an English translation provided in Multimedia Appendix 2. Participants assessed the SCTs based on their respective roles. Due to the anonymous nature of the survey and the inclusion criteria requiring respondents to be either psychiatry residents or physicians, additional demographic data were not collected. The study did, however, document information on the participants' level of training (resident doctors vs clinician-educators) and their level of clinical experience (0-5, 6-10, or ≥10 y).

Each SCT was evaluated by the participants using the conceptual framework developed by Fournier et al [9] for creating SCTs. This conceptual framework provides a general guideline for SCTs. The SCTs involve real-life medical situations, each describing as a short scenario with some uncertainty. To solve the problem presented in each scenario, there are multiple relevant options available for the medical student. Each scenario, along with its questions, is considered an item. The questions are divided into 3 parts. The first part provides a relevant diagnostic or management option. The second part introduces a new clinical finding, such as a physical sign or test result. The third part uses a 5-point Likert scale for examinees to express their decision on how the new finding affects the option, considering direction (positive, negative, or neutral) and intensity. Examinees are tasked with determining the impact of the new information, and the Likert scale is used to capture their decisions, as script theory suggests that clinical reasoning involves qualitative judgments.

Three components are evaluated by this framework when constructing SCTs: the scenario, clinical questions, and expert opinion. The scenario refers to the stem presented by the SCTs. The clinical questions are the individual questions adding a key element to the stem to stimulate clinical reasoning. The expert opinion refers to the opinion of an expert in the field giving a subjective appreciation as to the ability of the SCT to generate clinical reasoning. The elements of this framework are presented in Table 1. A common SCT template was used for both SCTs generated by ChatGPT and the experts in the field to ensure that the presentation of the SCTs does not create bias.

Table . The script concordance test (SCT) components with their relevant questions as per the framework by Fournier et al [9] for the evaluation and conception of SCTs.

SCT components and questions		Potential answers
Scenarios		
	S1. Describes a challenging circumstance, even for experts	Yes or no
	S2. Describes an appropriate situation for test takers	Yes or no
	S3. The scenario is necessary to understand the question and to set the context	Yes or no
	S4. The clinical presentation is typical	Yes or no
	S5. The scenario is well written	Yes or no
Clinical questions		
	Q1. The questions are developed using a key element approach	Yes or no
	Q2. In the opinion of experts, the options are relevant	Yes or no
	Q3. The same option is not found in 2 consecutive questions	Yes or no
	Q4. The new information (second column) makes it possible to test the link between the new information and the option (first column) in the context described	Yes or no
	Q5. Likert-scale anchors are clearly defined and unambiguous	Yes or no
	Q6. Questions are expanded to distribute responses equally across all Likert-scale values	Yes or no
	Q7. Questions are designed to provide a balance between low and high variability	Yes or no

Expert Opinion

The participants needed to state if the SCT was generated (or not) by ChatGPT (single-blinded mode), give their main hypothesis as to the main diagnosis studied in the SCT, and state in free-text style the strengths and weaknesses of each SCT.

XSL•FO
RenderX

Creating SCTs With ChatGPT

The ChatGPT tool operates through commands or prompts to enhance its performance. These prompts must offer a context of use, an expertise level, and a specific task. Following the typical steps involved in creating SCTs, we designed the prompts based on the approach outlined in Fournier et al [9]. In this initial study on the subject, we did not explore different sets of prompts, and the generated SCTs were used without modification.

The following commands were entered into ChatGPT to create the SCTs:

> *1. Act as an expert in university pedagogy of health sciences, in the field of psychiatry.*
>
> *2. Also acts as an expert in designing thumbnails by script matching.*
>
> *3. Generates a script matching vignette that includes three questions for the following diagnosis: (diagnosis name), according to DSM-5.*
>
> *4. Create questions linked to the vignette which start with if you think of "a diagnostic hypothesis" and you find "a sign or a symptom", this hypothesis is probable or not (from −2 to 2, using a Likert scale)*

Choosing the ChatGPT 3.5 algorithm as the main LLM for this task made sense for a few key reasons. This algorithm has a vast knowledge base covering a wide array of medical topics, making it an adequate tool for instructors crafting medical questions for medical students [25]. Its natural language comprehension, used in various medical fields, aids in question development [26]. The model's flexibility allows educators to create different types of questions to suit various learning styles and assessment methods. Notably, ChatGPT 3.5 supports multiple languages, including French, making it accessible for instructors in French-speaking regions. The model's ability to grasp context enables the creation of questions that build on existing knowledge, providing a more cohesive learning experience [27]. Educators can save time with the model's human-like text generation based on specific prompts or instructions. It is also crucial to highlight that this algorithm is open access and free, a substantial consideration when cost is a factor in choosing educational tools. Additionally, it is noteworthy that generating an SCT takes less than a minute on average with this tool.

Selecting Existing Expert-Created SCTs

Three SCTs were chosen at random from the 10 SCTs currently available to learners on the digital learning platform for the clinical psychiatry clerkship rotation at Université de Montréal. As stated above, a total of 3 ChatGPT-generated SCTs and 3 expert-created SCTs were chosen to limit the possibility that chance alone would identify the SCTs generated by ChatGPT from those produced by experts.

Statistical Analysis

A combined mixed method analysis was conducted with qualitative and quantitative components.

Qualitative Analysis

We conducted a content analysis by examining participants' open responses regarding the advantages and drawbacks of the presented SCTs. The objective was to pinpoint the primary types of benefits and limitations for emphasis. After receiving the open-ended survey responses, we individually extracted emergent themes from respondents using the grounded theory design framework [28]. Subsequently, AH and MP created an initial classification scheme based on these emerging themes. They applied this scheme to annotate the open-ended responses using the Qualitative Data Analysis Miner program (Provalis Research). Any discrepancies in annotations among responders were deliberated upon until a consensus was reached.

Quantitative Analysis

We conducted a descriptive statistical analysis to showcase the proportion of participants accurately identifying SCTs generated by ChatGPT compared to those crafted by experts. This same approach was applied to diagnostic hypotheses.

Additionally, we performed a descriptive statistical analysis to compare SCT scores based on the domains of the scenario and clinical questions, following the conceptual framework by Fournier et al [9]. Using a χ^2 test, we assessed the average results within each domain for the SCTs generated by ChatGPT and those by the experts. This allowed us to observe any statistical differences in the responses (yes or no) for various criteria within the scenario and clinical questions domains. We established a statistical significance threshold of $P<.05$ to identify noteworthy observations between the 2 types of SCTs.

Results

Participants Characteristics

A total of 102 participants completed the survey. Considering that there are an estimated 400 teaching clinicians in psychiatry in Quebec (about a third of the 1200 practicing psychiatrists), as well as 235 medical residents in psychiatry, this represents 16.1% (102/635) of the pool of potential responders. From the 102 participants, 45 (44.1%) identified as medical residents in psychiatry, 2 (2%) identified as teaching psychiatrists with less than 5 years of experience, 16 (15.7%) identified as teaching psychiatrists with between 6 and 10 years of experience, and 39 (38.2%) identified as teaching psychiatrists with more than 10 years of experience.

SCT Evaluation

The pooled averages of evaluations of the SCTs for each domain of assessment, stratified by the respondent categories, are shown in Table 2. A complete table reporting the evaluations of the respondents for each individual component of the domains of assessment is available in Multimedia Appendix 3. SCTs 2, 3 and 4 were generated by ChatGPT. It can be observed that there was no significant distinction between the pooled results for the SCTs generated by ChatGPT as compared to those generated by experts in the field. The questions related to the scenario component of the SCTs received better approval from the participants as compared to the clinical questions component.

XSL•FO
RenderX

Table . Responses for every component of the script concordance test (SCT) evaluations for the 6 SCTs, stratified by respondent categories. “Yes” indicates that the respondents agreed that the domain was elaborated appropriately.

SCT and evaluated component		Medical residents (n=45), n (%)	Teaching physicians (≤5 y; n=2), n (%)	Teaching physicians (6-10 y; n=16), n (%)	Teaching physicians (≥10 y; n=39), n (%)	Pooled average (N=102), n (%)
SCT 1						
	Scenario (yes)	30 (67)	2 (100)	12 (75)	31 (79)	75 (74)
	Clinical questions (yes)	29 (64)	2 (100)	13 (81)	28 (72)	72 (71)
	Is it a ChatGPT-generated scenario? (correct answers)	25 (44)	1 (50)	6 (38)	18 (54)	50 (49)
SCT 2 [a]						
	Scenario (yes)	29 (64)	2 (100)	13 (81)	25 (64)	69 (68)
	Clinical questions (yes)	30 (67)	2 (100)	14 (88)	25 (64)	71 (70)
	Is it a ChatGPT-generated scenario? (correct answers)	22 (49)	0 (0)	6 (38)	18 (46)	46 (45)
SCT 3 [a]						
	Scenario (yes)	28 (62)	2 (100)	12 (75)	26 (67)	68 (67)
	Clinical questions (yes)	28 (62)	2 (100)	13 (81)	25 (64)	68 (67)
	Is it a ChatGPT-generated scenario? (correct answers)	16 (36)	0 (0)	4 (25)	16 (41)	36 (35)
SCT 4 [a]						
	Scenario (yes)	28 (62)	2 (100)	11 (69)	26 (67)	67 (66)
	Clinical questions (yes)	25 (56)	2 (100)	14 (88)	28 (72)	69 (68)
	Is it a ChatGPT-generated scenario? (correct answers)	19 (42)	1 (50)	6 (38)	12 (31)	38 (37)
SCT 5						
	Scenario (yes)	26 (58)	2 (100)	11 (69)	26 (67)	65 (64)
	Clinical questions (yes)	27 (60)	2 (100)	13 (81)	28 (72)	70 (69)
	Is it a ChatGPT-generated scenario? (correct answers)	21 (53)	2 (100)	8 (50)	23 (59)	54 (53)
SCT 6						
	Scenario (yes)	27 (60)	2 (100)	12 (75)	26 (67)	67 (66)
	Clinical questions (yes)	24 (53)	2 (100)	13 (81)	27 (69)	66 (65)
	Is it a ChatGPT-generated scenario? (correct answers)	21 (53)	1 (50)	8 (50)	18 (46)	48 (47)

[a]Script concordance tests created by ChatGPT.

Participants could not identify which SCT was created by ChatGPT from those created by experts in the field, as observed in Table 2. Teaching clinicians with more than 10 years of experience tended to better recognize SCTs generated by ChatGPT than their peers with less experience and medical residents, except for SCT 4.

XSL•FO
RenderX

Comparisons Between ChatGPT- and Expert-Generated SCTs

When using the pooled observations for the scenario and clinical questions domains across the SCTs generated by ChatGPT and those generated by experts, no statistically significant distinctions were observed when comparing both types of SCTs (all *P*>.05), as seen in Table 3.

Table . Comparisons of the script concordance tests (SCTs) generated by ChatGPT as opposed to those generated by experts in the field.

Components	SCTs 1, 5, and 6 (experts), average score (%)	SCTs 2, 3, and 4 (ChatGPT), average score (%)	*P* value (ChatGPT-generated vs expert-generated SCTs)
Scenario	66.40	67.27	.84
Clinical questions	70.05	68.86	.99
Identifying if generated by AI[a]	54	40	.07

[a]AI: artificial intelligence.

Reported Strengths and Weaknesses of the SCTs

Overview

Only 39 (38.2%) of the 102 participants wrote at least 1 comment on the strengths or weaknesses for each of individual SCT. The strengths and weaknesses of the SCTs generated by ChatGPT were similarly reported across all the respondents and resembled those identified for the SCTs generated by experts in the field. Respondents reported that SCTs generated by ChatGPT were well aligned with the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*) but were also too caricatural.

Strengths of the SCTs Generated by Experts in the Field

Overall, 3 (8%) of the 39 respondents indicated for 1 or more SCTs generated by experts in the field that the scenario represented typical clinical challenges. Most of the respondents (27/39, 69%) reported that the SCTs used clear prompts to test clinical reasoning. Sample responses included the following:

> *This concordance test was easy to follow as because the scenarios were concise and the prompts were clear. [Respondent 1]*

> *In terms of clarity, the prompts were well written and it was very simple to see how they could elicit clinical reasoning. [Respondent 9]*

Strengths of the SCTs Generated by ChatGPT

Almost all respondents (32/39, 82%) mentioned that the SCTs were using typical clinical signs and symptoms reported in the *DSM-5*. Some (5/39, 13%) indicated that the SCTs were very well nuanced. Sample responses included the following:

> *This scenario corresponds to the textbook's description of the presented diagnosis. [Respondent 4]*

> *I see that these prompts do not try to derive too much from the differential diagnoses intended for the suggested clinical presentation. They offered a degree of flexibility to enable the student to use their clinical reasoning. [Respondent 71]*

Limitations of the SCTs Generated by Experts in the Field

In all, 2 (5%) of the 39 respondents mentioned that they found the SCTs straightforward and unchallenging. There were no other comments regarding the limitations of the SCTs generated by experts in the field. Sample responses included the following:

> *This scenario is too easy. I find little value as it is clear for the student that we are looking at the specific diagnosis. [Respondent 1]*

> *I don't see how this is challenging for the medical student who is going to take this test. [Respondent 80]*

Limitations of the SCTs Generated by ChatGPT

Most respondents (29/39, 74%) reported the SCTs generated by ChatGPT as caricatural or stereotypical clinical presentations as observed in textbooks with little regard to atypical presentations. A total of 7 (18%) respondents indicated that the SCTs generated by ChatGPT were too simple, as they tended to include additional information that were too trivial when attempting to challenge the responder's clinical reasoning. Sample responses included the following:

> *This is very trivial. I mean, it is not very difficult to find out what are the answers to these prompts as they clearly hint towards the same diagnosis. [Respondent 3]*

> *It would be interesting to add more challenging prompts as they tend to be very simplistic and poorly represent complex clinical cases as they are very stereotypical to what is found in the DSM-5. [Respondent 4]*

Discussion

Principal Findings

The aim of this study was to compare SCTs created by ChatGPT to SCTs produced by clinical specialists on the scenario (stem), clinical questions, and expert opinions. There were no significant distinctions between the SCTs generated by ChatGPT as compared to those developed by experts in the field for the evaluated components. The strengths and weaknesses were similar across the 2 types of SCT. Respondents reported that

XSL•FO
RenderX

the SCTs generated by ChatGPT were well aligned with the *DSM-5* but were also too caricatural.

Comparison With Prior Work

Since the creation of ChatGPT, it has been used in various areas of medical education such as digital teaching assistants and personalized education [29]. As a recent exploration study on the role of LLMs such as ChatGPT demonstrated, these models can provide interactive cases in a medical education context [30]. Considering these previous studies of ChatGPT in the development of medical education tools, it is possible that the inability to recognize a SCT generated by ChatGPT from one developed by experts in the field can be explained by the generative nature of this LLM. As such, a recent review on the use of ChatGPT in health care has identified that this form of AI can be used for problem-based learning and critical thinking in health care education [31]. However, it is mentioned in the literature that although the quality of the scenarios (or information) generated by ChatGPT might appear impressive, there is a need for an expert to assess the content generated, as it might be an amalgamation of erroneous information [32].

Although a few comments were provided regarding the strengths and limitations of both types of SCTs, they align with what is commonly reported in the literature for similar tasks. Some respondents noted caricature-like scenarios, possibly attributed to the robotic and dehumanized nature often associated with vignettes produced by LLMs [33]. It is plausible that more intricate prompts could have resulted in more nuanced scenarios. Therefore, the mentioned strengths of the scenarios and clinical questions, particularly their clinical alignment with the *DSM-5*, may be tied to the fact that this was one of the prompts used when conceptualizing interactions with ChatGPT during the creation of the SCTs.

In the field of psychiatry, applications of ChatGPT to medical education are limited. Among the limited available evidence, a novel study tested the knowledge of ChatGPT by exposing it to 100 clinical cases vignettes, and it performed extremely well [34]. Another similar use of ChatGPT was as an aid to answer clinical questions. A recent study evaluated the performance of users (psychiatrist and medical residents in the Netherlands) using ChatGPT as compared to nonusers for answering several questions in psychiatry, and it was observed that the users had better and faster responses as compared to nonusers [35]. Although these applications differ from this study, they might hint that ChatGPT currently has a database that holds relevant data in the field of psychiatry, which might explain the realism of scenarios and prompts observed for SCTs 2, 3, and 4.

There are substantial ethical considerations that must be accounted for when using such tool to assist medical educators. As an example, it is important to consider that ChatGPT (and other LLMs) are bound to the data they have been trained with along with their inherent biases [36]. Cross-validation of the generated information is often necessary to ensure that learners are not exposed to false information [37].

Limitations

Although web-based surveys offer convenience in distribution, they struggle with the challenge of accurately identifying the characteristics of the assessed population [38]. In our survey, we did not differentiate between those formally trained in SCTs and those who merely encountered them during their medical training, thus introducing potential limitations in generalizing the results. It is plausible that clinicians more experienced with SCTs were more likely to participate in the survey, but our recruitment from psychiatry departments exclusively helps mitigate this bias. Interpretation biases may also be present, as not all participants might be familiar with the framework used in this study. We did not explore acceptability regarding the use of generative AI in SCT creation, marking another limitation. Additionally, we did not compare different prompts, and it is conceivable that alternative sets of prompts could have produced better results for the SCTs generated by ChatGPT. Opting for a different language model might have yielded varied performances, and it is plausible that alternative models could outperform ChatGPT in this context.

Conclusions

In an era of rapidly evolving medicine and where technologies derived from AI are growing even more quickly, this study is the first to focus on the design of SCTs assisted by AI. The primary goal of this study highlighted that no statistical differences were found between the SCTs generated by ChatGPT and those created by clinical experts in the field of psychiatry for the elaboration of a scenario and the clinical questions presented in the SCTs. On average, the respondents incorrectly identified which SCTs were created with the help of AI. The major strength of SCTs generated by ChatGPT was that they were consistent with the *DSM-5*, whereas the caricatural quality or triviality of the SCTs generated by ChatGPT were the main weaknesses reported by the respondents. A possible way to mitigate this effect would be to provide more complex prompts to the generative AI or editing some details of the vignette. This study opens the door to larger-scale studies in this area to assess the impact of such aid on the academic success of medical students and how it can be used to improve efficiencies.

Acknowledgments

This study did not receive any financial support.

Authors' Contributions

AH, BK, MP, and VP contributed to the study conceptualization and writing of the original manuscript. All authors participated in the investigation and validation process. All authors edited the manuscript draft and reviewed the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original survey in French.
[PDF File, 502 KB - mededu_v10i1e54067_app1.pdf]

Multimedia Appendix 2
Translated survey in English.
[PDF File, 949 KB - mededu_v10i1e54067_app2.pdf]

Multimedia Appendix 3
Responses for every component of the script concordance test (SCT) evaluations for the 6 SCTs, stratified by the category of respondents.
[DOCX File, 27 KB - mededu_v10i1e54067_app3.docx]

References

1. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010 Aug;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
2. Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Acad Med 2020 Aug;95(8):1166-1171. [doi: 10.1097/ACM.0000000000003027] [Medline: 31577583]
3. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc 2015 Jul;103(3):152-153. [doi: 10.3163/1536-5050.103.3.010] [Medline: 26213509]
4. Heist BS, Gonzalo JD, Durning S, Torre D, Elnicki DM. Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: a mixed methods study. J Grad Med Educ 2014 Dec;6(4):709-714. [doi: 10.4300/JGME-D-14-00176.1] [Medline: 26140123]
5. Butler AC. Multiple-choice testing in education: are the best practices for assessment also good for learning? J Appl Res Mem Cogn 2018 Jul;7(3):323-331. [doi: 10.1016/j.jarmac.2018.07.002]
6. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12(4):189-195. [doi: 10.1207/S15328015TLM1204_5] [Medline: 11273368]
7. Giet D, Massart V, Gagnon R, Charlin B. Le test de concordance de script en 20 questions. Twenty questions on script concordance tests [Article in French]. Pédagogie Médicale 2013 Feb 4;14(1):39-48. [doi: 10.1051/pmed/2012026]
8. Petrucci AM, Nouh T, Boutros M, Gagnon R, Meterissian SH. Assessing clinical judgment using the script concordance test: the importance of using specialty-specific experts to develop the scoring key. Am J Surg 2013 Feb;205(2):137-140. [doi: 10.1016/j.amjsurg.2012.09.002] [Medline: 23246286]
9. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Inform Decis Mak 2008 May 6;8:18. [doi: 10.1186/1472-6947-8-18] [Medline: 18460199]
10. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Med Educ 2012 Jun;46(6):552-563. [doi: 10.1111/j.1365-2923.2011.04211.x] [Medline: 22626047]
11. Leclerc AA, Nguyen LHP, Charlin B, Lubarsky S, Ayad T. Assessing the acceptability of script concordance testing: a nationwide study in otolaryngology. Can J Surg 2021 May 26;64(3):E317-E323. [doi: 10.1503/cjs.014919] [Medline: 34038060]
12. See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. Med Educ 2014 Nov;48(11):1069-1077. [doi: 10.1111/medu.12514] [Medline: 25307634]
13. Kazour F, Richa S, Zoghbi M, El-Hage W, Haddad FG. Using the script concordance test to evaluate clinical reasoning skills in psychiatry. Acad Psychiatry 2017 Feb;41(1):86-90. [doi: 10.1007/s40596-016-0539-6] [Medline: 27178278]
14. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. Teach Learn Med 2010 Jul;22(3):180-186. [doi: 10.1080/10401334.2010.488197] [Medline: 20563937]
15. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ 2013 Dec;47(12):1175-1183. [doi: 10.1111/medu.12283] [Medline: 24206151]
16. Walters M, Alonge T, Zeller M. Impact of COVID-19 on medical education: perspectives from students. Acad Med 2022 Mar 1;97(3S):S40-S48. [doi: 10.1097/ACM.0000000000004525] [Medline: 34789656]
17. Saeki S, Okada R, Shane PY. Medical education during the COVID-19: a review of guidelines and policies adapted during the 2020 pandemic. Healthcare (Basel) 2023 Mar 16;11(6):867. [doi: 10.3390/healthcare11060867] [Medline: 36981524]
18. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607. [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
19. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-03-20]

XSL•FO
RenderX

20. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
21. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 9;11:e48808. [doi: 10.2196/48808] [Medline: 37812468]
22. Kimmerle J, Timm J, Festl-Wietek T, Cress U, Herrmann-Werner A. Medical students' attitudes toward AI in medicine and their expectations for medical education. J Med Educ Curric Dev 2023 Dec 6;10:23821205231219346. [doi: 10.1177/23821205231219346] [Medline: 38075443]
23. Martínez-Mesa J, González-Chica DA, Bastos JL, Bonamigo RR, Duquia RP. Sample size: how many participants do I need in my research? An Bras Dermatol 2014;89(4):609-615. [doi: 10.1590/abd1806-4841.20143705] [Medline: 25054748]
24. Asiamah N, Mensah H, Oteng-Abayie EF. Do larger samples really lead to more precise estimates? a simulation study. Am J Educ Res 2017 Jan;5(1):9-17. [doi: 10.12691/education-5-1-2]
25. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
26. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
27. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci 2023 Aug 23;10(8):230658. [doi: 10.1098/rsos.230658] [Medline: 37621662]
28. Chun Tie Y, Birks M, Francis K. Grounded theory research: a design framework for novice researchers. SAGE Open Med 2019 Jan 2;7:2050312118822927. [doi: 10.1177/2050312118822927] [Medline: 30637106]
29. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2023 Mar 10. [doi: 10.1002/ase.2270] [Medline: 36916887]
30. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023 Aug 14;9:e50945. [doi: 10.2196/50945] [Medline: 37578830]
31. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
32. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J 2023 Feb 28;64(1):1-3. [doi: 10.3325/cmj.2023.64.1] [Medline: 36864812]
33. Ashraf H, Ashfaq H. The role of ChatGPT in medical research: progress and limitations. Ann Biomed Eng 2024 Mar;52(3):458-461. [doi: 10.1007/s10439-023-03311-0] [Medline: 37452215]
34. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr 2023 Nov;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
35. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry 2023 Oct;22(3):479-480. [doi: 10.1002/wps.21145] [Medline: 37713576]
36. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 6;9:e48163. [doi: 10.2196/48163] [Medline: 37279048]
37. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol 2023 Sep 20;13(4):170-178. [doi: 10.5662/wjm.v13.i4.170] [Medline: 37771867]
38. Andrade C. The limitations of online surveys. Indian J Psychol Med 2020 Oct 13;42(6):575-576. [doi: 10.1177/0253717620957496] [Medline: 33354086]

Abbreviations

AI: artificial intelligence
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
LLM: large language model
SCT: script concordance test
UGME: undergraduate medical education

XSL•FO
RenderX

Edited by G Eysenbach, SR Mogali, TDA Cardoso; submitted 28.10.23; peer-reviewed by I Mlakar, J Kimmerle; revised version received 06.03.24; accepted 07.03.24; published 04.04.24.

Please cite as:
Hudon A, Kiepura B, Pelletier M, Phan V
Using ChatGPT in Psychiatry to Design Script Concordance Tests in Undergraduate Medical Education: Mixed Methods Study
JMIR Med Educ 2024;10:e54067
URL: https://mededu.jmir.org/2024/1/e54067
doi:10.2196/54067

© Alexandre Hudon, Barnabé Kiepura, Myriam Pelletier, Véronique Phan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.4.2024. 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.

XSL•FO
RenderX

Exploring the Performance of ChatGPT-4 in the Taiwan Audiologist Qualification Examination: Preliminary Observational Study Highlighting the Potential of AI Chatbots in Hearing Care

Shangqiguo Wang[1], PhD; Changgeng Mo[2], PhD; Yuan Chen[3], PhD; Xiaolu Dai[4], PhD; Huiyi Wang[5], MSc; Xiaoli Shen[6], MSc

1
2
3
4
5
6

Corresponding Author:
Yuan Chen, PhD

Abstract

Background: Artificial intelligence (AI) chatbots, such as ChatGPT-4, have shown immense potential for application across various aspects of medicine, including medical education, clinical practice, and research.

Objective: This study aimed to evaluate the performance of ChatGPT-4 in the 2023 Taiwan Audiologist Qualification Examination, thereby preliminarily exploring the potential utility of AI chatbots in the fields of audiology and hearing care services.

Methods: ChatGPT-4 was tasked to provide answers and reasoning for the 2023 Taiwan Audiologist Qualification Examination. The examination encompassed six subjects: (1) basic auditory science, (2) behavioral audiology, (3) electrophysiological audiology, (4) principles and practice of hearing devices, (5) health and rehabilitation of the auditory and balance systems, and (6) auditory and speech communication disorders (including professional ethics). Each subject included 50 multiple-choice questions, with the exception of behavioral audiology, which had 49 questions, amounting to a total of 299 questions.

Results: The correct answer rates across the 6 subjects were as follows: 88% for basic auditory science, 63% for behavioral audiology, 58% for electrophysiological audiology, 72% for principles and practice of hearing devices, 80% for health and rehabilitation of the auditory and balance systems, and 86% for auditory and speech communication disorders (including professional ethics). The overall accuracy rate for the 299 questions was 75%, which surpasses the examination's passing criteria of an average 60% accuracy rate across all subjects. A comprehensive review of ChatGPT-4's responses indicated that incorrect answers were predominantly due to information errors.

Conclusions: ChatGPT-4 demonstrated a robust performance in the Taiwan Audiologist Qualification Examination, showcasing effective logical reasoning skills. Our results suggest that with enhanced information accuracy, ChatGPT-4's performance could be further improved. This study indicates significant potential for the application of AI chatbots in audiology and hearing care services.

(JMIR Med Educ 2024;10:e55595) doi:10.2196/55595

KEYWORDS

ChatGPT; medical education; artificial intelligence; AI; audiology; hearing care; natural language processing; large language model; Taiwan; hearing; hearing specialist; audiologist; examination; information accuracy; educational technology; healthcare services; chatbot; health care services

Introduction

In recent years, the rapid advancement of large language models (LLMs) has significantly expanded their usage in various domains. Among the leading artificial intelligence (AI) chatbots—such as Bard, Bing, and ChatGPT—there has been a notable increase in diverse applications in everyday life. Prominently, ChatGPT, launched by OpenAI in November 2022 [1], stands out in the realm of AI chatbots. This model, known for its proficiency in generating and comprehending human-like text, showcases remarkable natural language processing skills. It has the capability to grasp complex queries, furnish insightful responses, and participate in meaningful conversations, thus

XSL•FO
RenderX

broadening the scope of AI's practicality in everyday scenarios [2,3].

ChatGPT represents a significant advancement in the field of natural language processing, exemplifying the latest developments in LLMs, particularly within the subset of autoregressive language models. Such generative LLMs, including ChatGPT, are predominantly trained on extensive text corpora. They use the decoder element of a transformer model, a groundbreaking architecture introduced by Vaswani et al [4] in 2017. This model is adept at predicting subsequent tokens in text sequences, a capability that has been progressively refined in subsequent research [5,6]. The transformer model, upon which ChatGPT is built, has revolutionized natural language processing. Its core strength lies in its ability to process text sequences efficiently, facilitating tasks such as language translation, question answering, and text summarization. One of the key features of this architecture is the self-attention mechanism, which allows it to understand long-range dependencies between words in a sentence without the need for sequential processing. This feature not only enhances efficiency compared to older recurrent neural network architectures but also offers improved interpretability, linking the semantic and syntactic structures of language inputs more effectively [4]. In addition to these capabilities, ChatGPT has evolved to incorporate real-time and knowledge-based information through various plug-ins. The introduction of GPT-4 by OpenAI in 2023 has expanded ChatGPT's proficiency to include processing both image and text inputs [1], marking a new milestone in the versatility and applicability of AI in diverse contexts.

ChatGPT has received considerable attention and exploration in its application within health care. The integration of ChatGPT into health care demonstrates its significant potential in enhancing patient education and handling general inquiries, marking it as a vital informational and supportive tool [7]. The broad applicability of AI chatbots in health care extends beyond patient interaction, serving clinicians, researchers, and students, with ChatGPT showing effectiveness in personalizing patient interactions and providing consumer health education [8,9]. This trend aligns with the overarching aim in health care AI to increase accessibility to medical knowledge and make care more affordable. Chatbots offer continuous health advice and support, potentially improving patient outcomes by reducing the need for in-person consultations. Additionally, they provide health care professionals with valuable insights for more informed patient care decision-making, though concerns regarding data transparency have been noted [10]. ChatGPT is capable of generating empathetic, high-quality responses to health-related queries, often comparable to those of physicians, and shows promise in producing emotionally aware responses with potential for continuous improvement [11,12]. In low- and middle-income countries, ChatGPT has great potential as a pivotal tool in public health efforts. Its advantages span various domains such as health literacy, screening, triage, remote support, mental health, multilingual communication, medical training, and professional support, addressing numerous challenges in these health care systems [13]. Furthermore, ChatGPT's role as a supplementary educational tool in areas requiring aptitude, problem-solving, critical thinking, and reading comprehension has been highlighted. The ChatGPT-4 version, in particular, shows potential in applications such as discharge summarization and group learning, enhancing human-computer interaction through verbal fluency [14,15]. However, the need for embracing these advancements while ensuring patient safety and recognizing the limitations of AI in intricate clinical cases is emphasized [16].

The evolution of computational sciences in hearing care services and research has given rise to the field of computational audiology. This approach combines algorithms, machine learning, and data-driven modeling for audiological diagnosis, treatment, and rehabilitation, using biological, clinical, and behavioral theories to augment care for patients and professionals [17]. The rapid development of AI technologies, especially LLMs such as ChatGPT, has significantly contributed to this field's growth. ChatGPT's advanced capabilities position it as a potential tool for patient interaction, education, aural rehabilitation program, and preliminary diagnostics in audiology [18,19]. However, it is crucial to recognize its current limitations. While it can handle complex interactions, it is not a substitute for human expertise in specialized areas such as audiology and is limited in interpreting nuanced medical information or performing physical diagnostics. AI chatbots have shown immense potential in hearing health care, aiding patients, clinicians, and researchers. Their applications range from initial screenings, educational support, and teleaudiology services for patients, to data analysis and decision support for clinicians and researchers [19]. In countries with vast geographical areas and imbalanced hearing care resources, AI chatbots could significantly enhance the development of hearing care services. Very recently, explorations into the use of AI chatbots for answering questions pertaining audiological knowledge have shown that AI chatbots can serve as a tool to access basic audiological information [20]. However, the accuracy and reliability of information provided by these tools remain a concern [19].

Despite the significant potential of AI chatbots to enhance hearing care services, research in this area remains sparse. AI chatbots' ability to understand questions and provide logical responses based on available information is crucial. This capability suggests promising applications in hearing care, including educational support, patient assistance in clinical settings, and aid for clinical staff. By engaging with AI chatbots, students, teachers, patients, and clinical personnel could significantly improve learning outcomes, patient care, and clinical practice efficiency. Therefore, this study starts from the most fundamental aspects to explore the performance of the current commercial version of ChatGPT-4 in taking an audiologist qualification examination (ie, the Taiwan Audiologist Qualification Examination). This investigation not only assesses the accuracy of responses to test questions but also explores the ability of the current AI chatbot to comprehend and logically respond to examination questions. These capabilities form the cornerstone for future integration of AI chatbots into educational support or clinical service assistance.

XSL•FO
RenderX

Methods

Materials

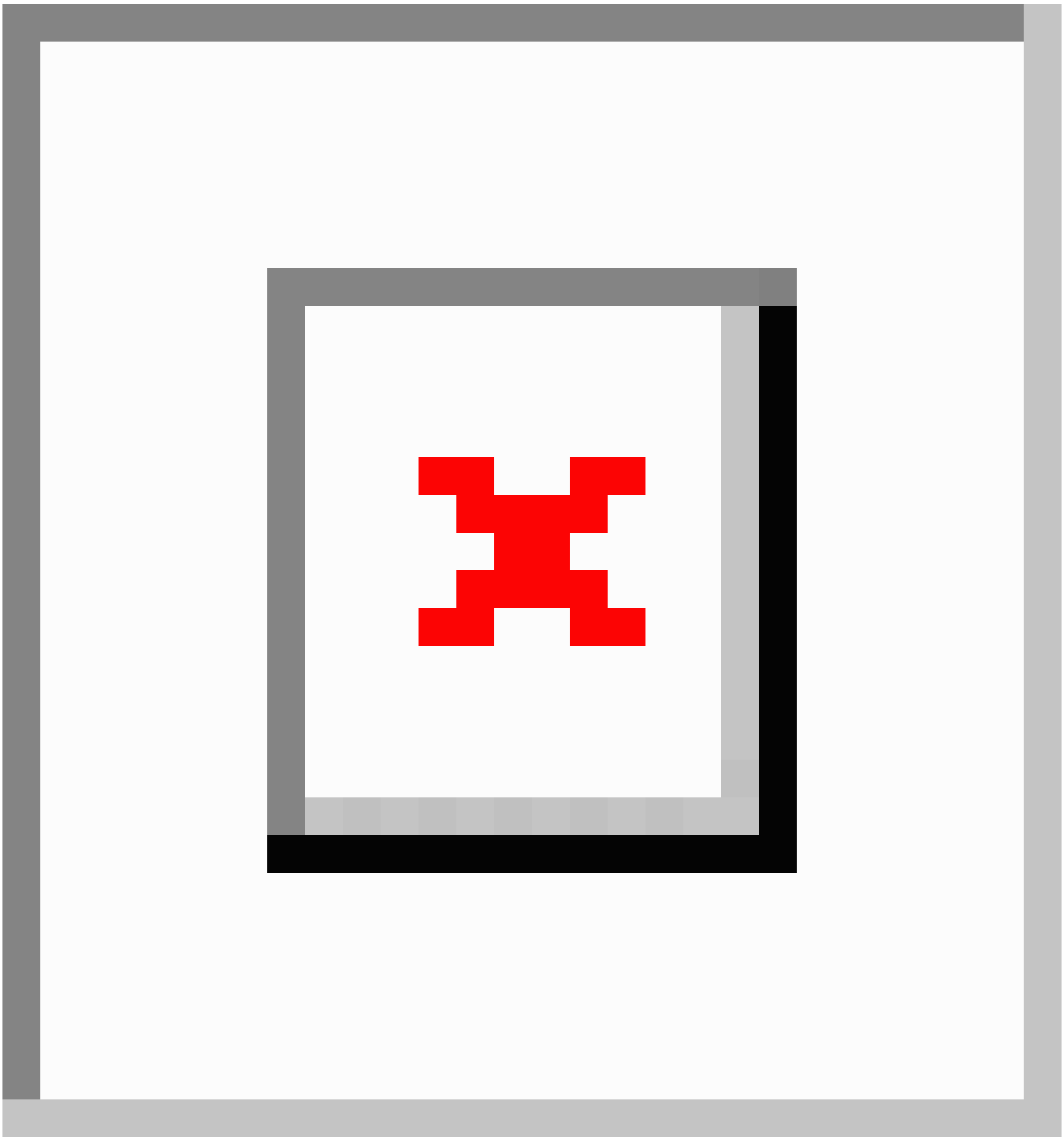
This study used the 2023 Taiwan Audiologist Qualification Examination [21]—a professional licensing examination for audiologists in Taiwan. Candidates of this examination are required to have a bachelor's or masters's degree in audiology and at least 6 months or 375 hours of clinical practice. The examination comprises six subjects: (1) basic auditory science, (2) behavioral audiology, (3) electrophysiological audiology, (4) principles and practice of hearing devices, (5) health and rehabilitation of the auditory and balance systems, and (6) auditory and speech communication disorders (including professional ethics). Each subject consists of 50 multiple-choice questions, except for behavioral audiology, which has 49 questions, totaling 299 questions in all. The examination papers featured 7 images, pivotal for answering 13 of the questions. Notably, these images were embedded directly within the PDF version of the examination rather than being provided as separate attachments. However, it is important to highlight that the images' resolution was relatively low, and they were presented without color. When extracted and saved in JPEG format, the images ranged in size from 12.7 to 27.2 KB and had resolutions spanning from 82 to 150 DPI. All related PDFs are accessible for download from the official source [21].

Prompt Engineering

Recognizing the significant influence of prompt engineering (where "prompt" refers to the input text provided by the user, which the model responds to) on the outputs of generative LLMs, a standardized prompt format was used in this research: "Please answer the following multiple-choice question as a hearing care professional, providing reasoning and explanation." This format was chosen to assess the logic and reasoning behind ChatGPT's responses. The original examination questions, a blend of Traditional Chinese and English, often used bilingual terminology for specialized concepts. For this study, ChatGPT was instructed to respond in English. Additionally, ChatGPT was not informed that these questions were from the Taiwan Audiologist Qualification Examination to prevent it from sourcing specific information to increase the accuracy of its responses. An example of a prompt and response is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. A template of a question posed to ChatGPT-4 and its corresponding responses. The example provided is the first question from the basic auditory science examination. The italicized English translations in the question stem are solely for reader comprehension and are not included in the actual prompt.



Procedure

All questions and correct answers were downloaded from the official website of the Taiwan Ministry of Examination in PDF format [21]. Subsequently, all questions were pre-edited in a Word (Microsoft Corp) document to avoid formatting issues. ChatGPT-4, enhanced with DALL-E (Decoder-Only Autoregressive Language and Image Synthesis), browsing, and analysis capabilities, was used for the test from December 10 to 12, 2023. A separate chat was used for each subject of the examination. Despite being within the same subject, the questions essentially had no overlapping content. For questions presented as images, both the image and the text format of the question were provided to leverage ChatGPT-4's image recognition capabilities. It is important to note that the resolution of the images supplied in the test was relatively poor, which could have potentially impacted the accuracy of image recognition. Following the approach of Gilson et al [9], the reasons for errors in incorrectly answered questions were categorized as follows: (1) logical errors: the response correctly identifies relevant information but fails to translate this information into an appropriate answer; (2) information errors: ChatGPT either overlooks a key piece of information, whether present in the question stem or from external sources, or shows a lack of expected knowledge; and (3) statistical errors: the error is due to a miscalculation, including explicit arithmetic errors

XSL•FO
RenderX

or incorrect estimations of statistical data. Authors SW and CM, both having a PhD in audiology, reviewed the original questions in Chinese and the GPT's responses in English, and then compared ChatGPT-4's responses to the official correct answers provided for the examination (all multiple-choice questions) to determine whether each question was answered correctly. They then performed a cross-check to ensure the accuracy of this step. Subsequently, SW and CM classified the incorrect answers into the 3 aforementioned categories and compared their classification results. In case of any discrepancies, they consulted with HW, who has a master's degree in public health, to reach a consensus and make a final decision together.

Ethical Considerations

This research did not involve human participants or private data and was therefore exempt from ethics approval by the ethics committee of Ningbo College of Health Sciences.

Data Analysis

The data analysis for this study was straightforward and conducted using Excel (Microsoft Corp). Our primary objective was to calculate the accuracy rate of ChatGPT-4 when tasked with taking the Taiwan Audiologist Qualification Examination.

Results

Overall Performance

ChatGPT-4 demonstrated commendable performance in the Taiwan Audiologist Qualification Examination. The accuracy rates for the 6 subjects were as follows: 88% for basic auditory science, 63% for behavioral audiology, 58% for electrophysiological audiology, 72% for principles and practice of hearing devices, 80% for health and rehabilitation of the auditory and balance systems, and 86% for auditory and speech communication disorders (including professional ethics). The overall accuracy rate for the 299 questions was 75% (see Table 1). The examination's passing criteria include an average accuracy rate of 60% across all subjects. Thus, ChatGPT-4 successfully passed this examination. Records of all ChatGPT-4's responses to the test questions can be found in the supplements (Multimedia Appendices 1-6). A detailed review of ChatGPT 4's responses revealed that errors were not caused by logical or statistical errors; instead, all incorrect answers resulted from information errors.

Table . Performance of ChatGPT-4 in the 2023 Taiwan Audiologist Qualification Examination.

	Questions, n	Correct responses, n	Accuracy rate, %
Basic auditory science	50	44	88
Behavioral audiology	49	31	63
Electrophysiological audiology	50	29	58
Principles and practice of hearing devices	50	36	72
Health and rehabilitation of the auditory and balance systems	50	40	80
Auditory and speech communication disorders (including professional ethics)	50	43	86
Total questions	299	233	75
Questions with images	13	8	62
Images	7	4	57

Information Errors Leading to Incorrect Answers

Lack of Correct Information Sources

Most incorrect answers were due to ChatGPT-4 relying on inaccurate information. For example, in the basic auditory science examination, question 43 involved identifying an incorrect statement about temporal masking among options A, B, C, and D. Option B stated that forward masking occurs when a signal appears after the masking noise, even with a 200-millisecond gap between the 2 stimuli (which is a misconception). The correct answer was that this statement is false, but ChatGPT-4's explanation suggested otherwise:

> *Forward Masking with 200 ms Gap (B): This is correct. Forward masking can occur when the signal appears after the masking noise, even with a gap as long as 200 ms.*

This response indicates that ChatGPT-4 did not have accurate information to answer this question. As a matter of fact, no masking occurs when the time interval between the masker and the signal equals or exceeds 200 milliseconds [22]. Similarly, for option D, which stated that forward masking is more effective than backward masking, ChatGPT-4's choice and explanation were as follows:

> *Forward Masking More Effective Than Backward Masking (D): This is typically incorrect. Backward masking is often found to be more effective than forward masking.*

While this appears correct at first glance, it lacks the clarification that backward masking is more effective than forward masking when the same time interval exists between the masker and signal, with more masking occurring when the masker follows the signal [23]. However, from the perspective of the

XSL•FO
RenderX

information available to ChatGPT-4, option D was the most reasonable choice.

Conflicting Information Sources

Multiple sources of information also led to incorrect responses from ChatGPT-4. For example, question 22 in the electrophysiological audiology examination asked about the most appropriate frequency for tympanometry in newborns, with the correct answer being 1000 Hz. Despite its simplicity and lack of complexity in reasoning, ChatGPT-4 provided an incorrect response:

> *(A) 112 Hz is typically used for infants and young children. Newborns and young children have a more compliant ear canal than adults, and lower frequencies such as 226 Hz can sometimes overestimate middle ear pressure and compliance. Therefore, a lower frequency probe tone is often recommended.*

Considering the straightforward nature of this question, the authors expected an accurate answer from ChatGPT-4. Therefore, after the conclusion of all tests, the question was posed to ChatGPT-4 twice more in separate new dialogues to avoid information cross-over. Two completely different responses were generated by ChatGPT-4:

> *226 Hz Standard: The 226 Hz probe tone is the standard frequency used in clinical tympanometry for all ages, including newborns. This frequency has been proven to provide reliable and accurate measurements of middle ear function in this age group. It is effective in identifying middle ear fluid, a common issue in newborns, and other conditions that might affect hearing. … In conclusion, 226 Hz (option B) is the most suitable probe tone frequency for tympanometry in newborns, due to its effectiveness in assessing middle ear function accurately in this specific age group.*

and

> *When performing tympanometry on a newborn, the most appropriate probe tone frequency is typically 1000 Hz (Option C). Let me explain why:*
>
> *Age-Related Differences: The middle ear mechanics in infants, especially newborns, differ significantly from those in older children and adults. These differences are mainly due to the size and composition of the middle ear and ear canal, as well as the compliance of the tympanic membrane.*
>
> *High-Frequency Probe Tones: A 1000 Hz probe tone is recommended for newborns and infants up to around six months of age. This higher frequency provides more accurate assessments of middle ear function in this age group, as it is less affected by the high compliance of the infant ear canal.*
>
> *Therefore, the most appropriate option for tympanometry in newborns is 1000 Hz (Option C), as it offers the most reliable assessment of middle ear function in this specific age group.*

This indicates that ChatGPT-4 may provide different answers each time based on varying sources of information, particularly when these sources have conflicts or inconsistencies.

Image Information Recognition

In the examination, 13 questions could be answered only through the recognition of images to extract information. ChatGPT-4 correctly answered 8 of these questions. Images 1 to 4 are from the behavioral audiology subject, images 5 and 6 are from the electrophysiological audiology subject, and image 7 is from the principles and practice of hearing devices subject. Out of the 7 images provided in total, ChatGPT-4 successfully recognized 4. The criterion for determining successful recognition was assessing the accuracy of ChatGPT-4's interpretation of image content and its ability to extract pertinent information for answering questions. Authors SW and CM independently evaluated this aspect and subsequently performed a cross-check of their assessments.

Discussion

Principal Findings

This study evaluated ChatGPT-4's performance in the 2023 Taiwan Audiologist Qualification Examination. The eligibility criteria for this examination are having a degree in audiology or a related field and a minimum of 6 months or 375 hours of clinical practice. The minimum required accuracy rate to pass the examination is set at 60%. In the 2023 examination, 88.5% of candidates achieved this accuracy rate or higher, effectively passing the examination. ChatGPT-4 achieved an overall accuracy rate of 75%, meeting the passing criterion necessary for candidates to obtain the basic qualification for practicing as clinical audiologists in Taiwan. It performed notably well in subjects that required more analytical reasoning and contextual decision-making, such as health and rehabilitation of the auditory and balance systems and auditory and speech communication disorders (including professional ethics). The proficiency of LLMs in integrating and interpreting information logically was evident in subjects demanding contextual knowledge. However, in fields such as electrophysiological audiology, which depend more on precise knowledge points, the accuracy of ChatGPT-4 was challenged when confronted with incorrect or insufficient information. In our study, the original questions were in both Chinese and English. We requested ChatGPT to provide responses in English, and the translation between the 2 languages did not negatively impact either the comprehension or the accuracy of the responses. In addition, although this research was a preliminary examination of ChatGPT-4's capabilities in image recognition within audiology examinations, it is important to note that the number of images used was limited, and their quality and resolution were suboptimal. Nevertheless, despite these constraints, ChatGPT-4 demonstrated a moderately acceptable level of image recognition performance, successfully identifying over half of the content within the images.

Comparative analysis with the existing literature indicates that LLMs such as ChatGPT have shown promising results in medical examinations [24-26], particularly GPT-4 [27]. The model's ability to pass examinations that are challenging for

XSL•FO
RenderX

many humans has been noted [9,28]. In our study, which involved multiple-choice questions, ChatGPT was tasked with not only selecting answers but also articulating the reasoning behind its choices. Notably, ChatGPT-4 has substantially reduced the incidence of logical and statistical errors that were more prevalent in its predecessors. Its accuracy rate in examinations based on multiple-choice questions has increased from 53.6% with GPT-3 and -3.5 to 75.1% with GPT-4 [27]. The absence of logical errors, in particular, suggests that ChatGPT-4 has an enhanced ability to understand questions accurately and make decisions that are logically coherent, using the information it has access to [9,27]. This advancement is especially relevant in the context of audiology examinations, especially in the multiple-choice question format, where statistical reasoning has traditionally been less emphasized. In this study, the primary challenge faced by ChatGPT-4 in accurately answering questions was identified as information errors, which primarily manifest in 2 distinct forms: a lack of correct information sources and the presence of conflicting information sources. The former issue directly impacts the ChatGPT's performance; for several questions, ChatGPT-4 lacked the necessary correct information to either directly answer or logically deduce the correct responses. Despite its training on an extensive database of information [29,30], it became evident that ChatGPT-4 does not possess a comprehensive knowledge base required to flawlessly address specialized queries within the field of audiology, a discipline that demands a high degree of professional expertise. Conversely, the presence of conflicting information sources contributed to erroneous responses from ChatGPT. The model, although equipped with a wealth of information, was not developed with a focus on audiological knowledge or audiology best practices. This abundance of data, however, presents a challenge in verifying the accuracy and reliability of the information, especially when multiple sources offer conflicting viewpoints on widely discussed topics. This was exemplified in this study's results, where a question on a fundamental concept in audiology—the use of 1000 Hz in tympanometry for children—resulted in ChatGPT-4 providing 3 distinct answers. Additionally, the lowest scores were observed in questions related to electrophysiological audiology. Apart from the issue of inaccurate sources of information, another possible reason is that information about electrophysiological audiology might be more specialized than that in other subjects, thereby restricting the amount of information that ChatGPT has access to. This is unlike the case with hearing aids or auditory rehabilitation, where a vast amount of information is readily available on the internet. Altogether, this highlights the necessity for further refinement of LLMs, emphasizing the integration of more precise and professionally relevant information. This approach will ensure that responses are derived from verified and accurate data sources, pointing to a crucial direction for future research in this area [31].

AI Chatbots and Audiology

The introduction of advanced AI models such as ChatGPT-4 has significant implications in hearing health care [18]. ChatGPT's capacity to process and analyze extensive data makes it a potentially useful tool for patients, clinicians, and researchers [19]. Our findings suggest that given training with reliable information, even the current iteration of ChatGPT-4 holds considerable promise for application in hearing care services. This potential is likely to increase alongside the continual advancements in LLMs. The results of this study show that when faced with professional-level audiology questions, AI chatbots can provide answers with a high accuracy rate and logical reasoning. Furthermore, their ability to mimic human-like responses suggests that they are capable of assisting in educational learning and hearing care awareness. However, this is contingent upon first building a fine-tuned model with accurate information sources.

In the context of patient care, AI chatbots could act as digital audiologists, providing answers to a range of hearing-related queries. Their easy accessibility may be beneficial for early hearing screenings and prompt intervention or medical attention. AI chatbots also have the potential to educate patients about hearing issues and offer psychological support for conditions such as tinnitus. They have been shown to have potential in managing mental health concerns and demonstrate a level of empathy that can surpass that of human physicians [11,32]. The development of AI chatbots as qualified audiologists could greatly enhance teleaudiology services. For clinicians, AI chatbots could serve as supportive tools, offering rapid references or recommendations based on current research and guidelines, aiding in diagnosis and treatment suggestions [33], and facilitating the creation of diagnostic or referral reports [34]. This is especially relevant in regions with limited hearing care resources [13], where AI chatbots could play a vital role in both the education of hearing care professionals and in augmenting clinical services. This enhancement could lead to improved overall quality and availability of hearing care services, ultimately benefiting individuals with hearing impairments. Similarly, researchers in auditory science could use AI chatbots to streamline their research processes. However, the effectiveness of these proposed applications depends on the thorough and complete validation of the chatbots' functionality and information accuracy.

Limitations

This study represents a preliminary exploration of an AI chatbot's performance in an audiologist qualification examination. However, several limitations must be acknowledged. First, the selected examination questions were exclusively multiple-choice, with a subset requiring integrated information for reasoning. This format lacks open-ended questions that typically mirror the complexity of real-world clinical scenarios in hearing care, where audiologists address diverse and intricate issues beyond isolated knowledge points. Future research could extend to evaluating AI chatbots in handling complex audiology cases. Second, while this study included an assessment of ChatGPT-4's image recognition capabilities, the quality of the images in the original test files was suboptimal. Additionally, the number of questions involving image information was limited, which constrained the ability of this study to draw substantial conclusions about this functionality.

XSL•FO
RenderX

Conclusions

In conclusion, the findings of this study show that ChatGPT 4 achieved a 75% accuracy rate in the 2023 Taiwan Audiologist Qualification Examination, thus successfully passing it. The primary reason for ChatGPT-4's incorrect responses was identified to be information errors, including both a lack of correct information sources and the presence of conflicting information sources. Therefore, a fine-tuned model containing accurate hearing care information sources has the potential to further enhance the feasibility of AI chatbot applications in hearing care services. However, passing the examination does not imply that ChatGPT-4 can become a qualified clinical audiologist in Taiwan; rather, it only indicates that ChatGPT-4 has some basic knowledge required for the audiology profession. Adequate clinical internship hours are also a crucial requirement for the actual practice of audiology in Taiwan, and its performance in handling real clinical cases remains unknown.

Acknowledgments

The authors wish to express gratitude for the rapid development of large language models, which brings hope for future improvement in hearing care globally. ChatGPT was used to proofread and correct grammatical issues in this article.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon request.

Authors' Contributions

SW led the study's design, managed the data, performed the analysis, and contributed to drafting the manuscript. CM, YC, and XD helped conceptualize the study. HW also participated in the analysis, while XS oversaw the study's administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT transcripts: basic auditory science.
[PDF File, 10148 KB - mededu_v10i1e55595_app1.pdf]

Multimedia Appendix 2
ChatGPT transcripts for behavioral audiology.
[PDF File, 9601 KB - mededu_v10i1e55595_app2.pdf]

Multimedia Appendix 3
ChatGPT transcripts for auditory and speech communication disorders (including professional ethics).
[PDF File, 9737 KB - mededu_v10i1e55595_app3.pdf]

Multimedia Appendix 4
ChatGPT transcripts for electrophysiological audiology.
[PDF File, 8874 KB - mededu_v10i1e55595_app4.pdf]

Multimedia Appendix 5
ChatGPT transcripts for principles and practice of hearing devices.
[PDF File, 10975 KB - mededu_v10i1e55595_app5.pdf]

Multimedia Appendix 6
ChatGPT transcripts for health and rehabilitation of the auditory and balance systems.
[PDF File, 9752 KB - mededu_v10i1e55595_app6.pdf]

References

1. ChatGPT. OpenAI. 2023. URL: https://openai.com/chatgpt [accessed 2024-04-16]
2. Haleem A, Javaid M, Singh RP. An era of ChatGPT as a significant futuristic support tool: a study on features, abilities, and challenges. BenchCouncil Trans Benchmarks Stand Eval 2022 Oct;2(4):100089. [doi: 10.1016/j.tbench.2023.100089]
3. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Physical Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]

XSL•FO
RenderX

4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4 to 9, 2017; Long Beach, CA p. 5999-6009 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [accessed 2024-04-23]
5. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: 34th Conference on Neural Information Processing Systems (NeurIPS 2020); Dec 6 to 12, 2020; Vancouver, BC (virtual) URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [accessed 2024-04-23]
6. Dai Z, Yang Z, Yang Y, Carbonell J, Le Q, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2019:2978-2988 URL: https://aclanthology.org/P19-1285.pdf [accessed 2024-04-16] [doi: 10.18653/v1/P19-1285]
7. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 4;47(1):33. [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
10. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Feb;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
11. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
12. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058. [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
13. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac 2023 Dec;41:100905. [doi: 10.1016/j.lanwpc.2023.100905] [Medline: 37731897]
14. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ 2023 Apr 26;9:e47737. [doi: 10.2196/47737] [Medline: 37099373]
15. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
16. Kleebayoon A, Wiwanitkit V. Issues for consideration about use of ChatGPT. Comment on 'Performance of ChatGPT on specialty certificate examination in dermatology multiple-choice questions'. Clin Exp Dermatol 2023 Jun 13:llad202. [doi: 10.1093/ced/llad202]
17. Wasmann JW, Lanting CP, Huinck WJ, et al. Computational audiology: new approaches to advance hearing health care in the digital age. Ear Hear 2021;42(6):1499-1507. [doi: 10.1097/AUD.0000000000001041] [Medline: 33675587]
18. Sooful P, Simpson A, Thornton M, Šarkic B. The AI revolution: rethinking assessment in audiology training programs. Hear J 2023 Nov;76(11):000. [doi: 10.1097/01.HJ.0000995264.80206.87]
19. Swanepoel DW, Manchaiah V, Wasmann JW. The rise of AI chatbots in hearing health care. Hear J 2023;76(4):26. [doi: 10.1097/01.HJ.0000927336.03567.3e]
20. Jedrzejczak WW, Kochanek K. Comparison of the audiological knowledge of three chatbots – ChatGPT, Bing Chat, and Bard. medRxiv. Preprint posted online on Nov 22, 2023. [doi: 10.1101/2023.11.22.23298893]
21. Post-examination question inquiry platform. Ministry of Examination ROC (Taiwan). 2023. URL: https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx [accessed 2024-04-16]
22. Durrant J, Lovrinic J. Introduction to psychoacoustics: temporal aspects of hearing. In: Durrant J, Lovrinic J, editors. Bases of Hearing Science: Lippincott Williams & Wilkins; 1995:294-299.
23. Elliott LL. Backward masking: monotic and dichotic conditions. J Acoust Soc Am 1962 Aug 1;34(8):1108-1115. [doi: 10.1121/1.1918253]
24. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023 Dec;28(1):2220920. [doi: 10.1080/10872981.2023.2220920] [Medline: 37307503]
25. Watari T, Takagi S, Sakaguchi K, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ 2023 Dec 6;9:e52202. [doi: 10.2196/52202] [Medline: 38055323]
26. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore) 2023 Aug 11;102(32):e34673. [doi: 10.1097/MD.0000000000034673] [Medline: 37565917]
27. Newton P, Xiromeriti M. ChatGPT performance on MCQ exams in higher education. A pragmatic scoping review. EdArXiv Preprints. Preprint posted online on Jun 18, 2024. [doi: 10.35542/osf.io/sytu3]

XSL•FO
RenderX

28. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation study. JMIR Nurs 2023 Jun 27;6:e47305. [doi: 10.2196/47305] [Medline: 37368470]
29. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
30. Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023 Aug;229(2):172. [doi: 10.1016/j.ajog.2023.04.020] [Medline: 37088277]
31. Vaid A, Landi I, Nadkarni G, Nabeel I. Using fine-tuned large language models to parse clinical notes in musculoskeletal pain disorders. Lancet Digit Health 2023 Dec;5(12):e855-e858. [doi: 10.1016/S2589-7500(23)00202-9]
32. Tal A, Elyoseph Z, Haber Y, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioeth 2023 Oct;23(10):74-77. [doi: 10.1080/15265161.2023.2250297] [Medline: 37812102]
33. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
34. Zhou Z. Evaluation of ChatGPT’s capabilities in medical report generation. Cureus 2023 Apr;15(4):e37589. [doi: 10.7759/cureus.37589] [Medline: 37197105]

Abbreviations

AI: artificial intelligence
DALL-E: Decoder-Only Autoregressive Language and Image Synthesis
LLM: large language model

Edited by G Eysenbach, TDA Cardoso; submitted 18.12.23; peer-reviewed by H Cullington, P Sooful, R Eikelboom; revised version received 09.03.24; accepted 22.03.24; published 26.04.24.

Please cite as:
Wang S, Mo C, Chen Y, Dai X, Wang H, Shen X
Exploring the Performance of ChatGPT-4 in the Taiwan Audiologist Qualification Examination: Preliminary Observational Study Highlighting the Potential of AI Chatbots in Hearing Care
JMIR Med Educ 2024;10:e55595
URL: https://mededu.jmir.org/2024/1/e55595
doi:10.2196/55595

© Shangqiguo Wang, Changgeng Mo, Yuan Chen, Xiaolu Dai, Huiyi Wang, Xiaoli Shen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.4.2024. 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.

XSL•FO
RenderX

Naloxone Coprescribing and the Prevention of Opioid Overdoses: Quasi-Experimental Metacognitive Assessment of a Novel Education Initiative

Michael Enich[1], MD, PhD; Cory Morton[2], MSW, PhD; Richard Jermyn[3], MD

1

2

3

Corresponding Author:
Cory Morton, MSW, PhD

Abstract

Background: Critical evaluation of naloxone coprescription academic detailing programs has been positive, but little research has focused on how participant thinking changes during academic detailing.

Objective: The dual purposes of this study were to (1) present a metacognitive evaluation of a naloxone coprescription academic detailing intervention and (2) describe the application of a metacognitive evaluation for future medical education interventions.

Methods: Data were obtained from a pre-post knowledge assessment of a web-based, self-paced intervention designed to increase knowledge of clinical and organizational best practices for the coprescription of naloxone. To assess metacognition, items were designed with confidence-weighted true-false scoring. Multiple metacognitive scores were calculated: 3 content knowledge scores and 5 confidence-weighted true-false scores. Statistical analysis examined whether there were significant differences in scores before and after intervention. Analysis of overall content knowledge showed significant improvement at posttest.

Results: There was a significant positive increase in absolute accuracy of participant confidence judgments, confidence in correct probability, and confidence in incorrect probability (all *P* values were <.05). Overall, results suggest an improvement in content knowledge scores after intervention and, metacognitively, suggest that individuals were more confident in their answer choices, regardless of correctness.

Conclusions: Implications include the potential application of metacognitive evaluations to assess nuances in learner performance during academic detailing interventions and as a feedback mechanism to reinforce learning and guide curricular design.

(JMIR Med Educ 2024;10:e54280) doi:10.2196/54280

KEYWORDS

naloxone; coprescribing; prescription; academic detailing; metacognition; metacognitive evaluation; pharmacotherapy; pharmaceutic; pharmaceutical; education; educational intervention; opioid; opioid overdose; harm reduction

Introduction

In 2020, of the 91,799 drug-related overdoses in the United States, 75% involved an opioid [1]. Naloxone is an invaluable tool to prevent opioid overdose [2], and coprescription initiatives (or programs to encourage providers to prescribe naloxone to patients receiving opioids) are a common, important intervention to reduce fatal overdoses. This is especially true in primary care settings, since eligible patients who meet the Centers for Disease Control and Prevention coprescription guidelines are underprescribed naloxone to take home [3].

Academic detailing programs are educational outreach approaches used to improve clinician decision-making, and they have consistently shown a positive effect on altering prescriber behavior and clinical decision-making [4]. Health systems have implemented naloxone coprescription academic detailing programs with demonstrated positive effects on the number of providers prescribing and patients receiving naloxone [2,5]. Critical evaluations of such programs have shown acceptability and feasibility of such interventions [3], including positive impact on increasing the number of prescriptions despite hesitancy around the logistics of implementation [6] and increasing the number of prescriptions after brief interventions [7].

Learners in the health professions are important allies for combatting overdose; interventions have been developed for health profession students to be trained in overdose education and naloxone distribution [8]. Results from student-focused overdose education and naloxone distribution interventions indicate increases in average participant knowledge around

XSL•FO
RenderX

identifying and responding to suspected overdoses [8]. To extend knowledge on changes in participant thinking during educational interventions, one area to consider is the effect on participant metacognition. Here, metacognition refers to the beliefs, attitudes, and confidence related to influencing a particular cognitive task, colloquially summarized as thinking about thinking. The measurement of participant metacognitive processes is especially important in health education because of the importance of the desired outcomes and the need for a life span approach to learning in the health professions. Improvements in metacognition in health education interventions have been linked to improved content knowledge acquisition, improved clinical reasoning, and decreased avoidable errors [9]. However, assessing metacognition is not often a focus of medical education evaluation, and those who wish to evaluate metacognition are often met with a lack of clarity on how to effectively measure it [10].

The dual purposes of this brief report are to (1) present a metacognitive evaluation of a naloxone coprescription academic detailing intervention for health professions students and practitioners and (2) describe the application of a metacognitive evaluation for future medical education interventions.

Methods

Data

Participants completed a self-paced, web-based academic detailing naloxone coprescription intervention implemented by Rowan University School of Osteopathic Medicine (RUSOM). This brief continuing medical education (CME)–eligible course provided a standardized, evidence-based curriculum to train RUSOM–affiliated health care providers, administrators, students, and executives across a variety of health care settings on how to implement and sustain naloxone coprescribing programs. Participants were recruited via email, and the only incentive was providing the CME credit at no cost. Consent was provided by agreeing to a question prompt to continue each survey after reading the informed consent documentation.

Data for this analysis came from a 20-item knowledge assessment based on the Centers for Disease Control and Prevention naloxone coprescription guidelines [11], previously validated overdose knowledge assessment instruments [12], as well as guidance based on best practices in implementation science [13]. It was delivered in a pre-post design, where participants completed the knowledge assessment before accessing the educational intervention and after course completion. The course was designed as a single module to be completed in 1 session. Participants were eligible for CME credits after completion if they achieved a passing score; multiple individuals had more than 1 posttest score as they attempted to meet that minimum score. To avoid a bias in results, knowledge scores that came chronologically first were used as the posttest score in all analyses.

To assess metacognition, items were designed with confidence-weighted true-false (CTF) scoring, which combines traditional true-false questions with each learner's rated confidence for each item (I am confident this is true; I believe this is true, but I am unsure; I believe this is false, but I am unsure; and I am confident this is false). CTF is a useful and simple means to measure both cognitive and metacognitive achievements [14].

Study Sample

The sample includes any individual who completed both pre- and posttest assessments for the naloxone coprescription educational intervention between dates April 2020 and July 2021. To access the intervention, participants had to register via a university web application (from which voluntary demographic data were derived) and then log in to their learning management system. While the intervention provided an opportunity for CME credits, any individual was able to register for and take the course, including nonprescribers and students.

Analysis

Descriptive statistics were calculated for individuals in the study sample. For both the pre- and posttest, 3 content knowledge scores were calculated: the summed CTF score (where confidently incorrect scores equal 0 points and confidently correct scores equal 4 points), the percent correct CTF score (based on maximum of 80), and the binary percent correct score (true/false [T/F]) (number correct regardless of confidence or number of items). In addition, metacognitive scores were calculated using the methods described by Dutke and Barenberg [14] and included absolute accuracy of confidence judgments (AC), bias of confidence judgments (BS), confidence correct probability (CCP), confidence incorrect probability (CIP), and discrimination between correct and incorrect decisions (DIS). AC reflects the overall match between participant confidence and the outcome of their choice. An increase in AC suggests that individuals are better able to gauge both when they are confident in correct answers and unconfident in incorrect answers. BS is similar to AC but gives an indication of the direction and severity of participant ability to correctly asses their level of confidence in an answer. Values close to 0 indicate an exact match between confidence or nonconfidence and correctness or incorrectness, positive values suggest overconfidence (more confident but less correct), and negative values indicate underconfidence (less confident but more correct). However, the BS does not indicate the relative contribution of confidence to correct or incorrect answers, and the CCP and CIP are used to discern the respective probabilities of being confident that the answer is correct (CCP) or confident that the answer is incorrect (CIP). A higher CCP score indicates higher confidence when the answer is correct. A lower CIP score indicates less confidence when the answer is incorrect. A high CCP and low CIP suggests improvement in metacognition. Finally, the DIS is the difference between the CCP and CIP probabilities and is used to indicate how reliably a participant discriminates between correct and incorrect answers, with higher values indicating appropriate participant metacognitive monitoring and the ability to discriminate between concepts that are known and those that need reinforcement [14]. To correct for a left-skewed distribution of assessment values, Wilcoxon signed rank analyses were applied to assess changes in individual scores between pre- and posttest assessments. Finally, Rosenthal correlations were calculated to determine

XSL•FO
RenderX

the effect size of the intervention on each metacognitive score. Item-level examinations of CTF distribution were completed to add context to the metacognitive outcomes and identify concepts in the naloxone coprescription framework that may need reinforcement. McNemar tests were used to determine whether there was a significant change in correctness from pre- to posttest for each item. Statistical analyses were completed using Stata 17 (StataCorp LLC).

Ethical Considerations

This study was approved by the Rutgers University Institutional Review Board (ID2019000275). Participants were provided informed consent at pretest and posttest, and data were deidentified prior to analysis. The course and CME credit were provided at no cost to participants, and no additional compensation was provided.

Results

Sample descriptive statistics are shown in Table 1; 307 individuals completed both pre- and posttests. As shown in Table 2, analysis of overall test scores showed a statistically significant improvement in content knowledge after completing the educational intervention, both in CTF score and binary correct-incorrect score. For both, the effect size of the intervention was moderate.

Significant differences in metacognitive scores suggest potential improvements in metacognitive monitoring occurred during the intervention. There is a statistically significant increase in absolute AC with a moderate effect size, suggesting that after intervention individuals are better able to gauge when they are *confident* in *correct* answers and *unconfident* in *incorrect* answers. For BS, median response values changed from negative to positive with a strong effect size, suggesting an overall change from being underconfident (negative values) in answer choices to appropriately confident (null or positive values) after intervention. Both CCP and CIP had a significant, positive change after intervention with strong effects. There was a significant decrease in DIS score after intervention with a very low effect size, which likely reflects an underlying increase in confidence in incorrect answers after academic detailing.

Table 3 shows the CTF and binary T/F frequencies for each item and an indication of significant change from pre- to posttest using McNemar test. Most items saw their binary correct answers increase at posttest; only 1 item (item 15) saw a significant decline in correct answers (t_{306}=−4.41; P=.04). This item was part of a conceptual group of questions (items 7, 12, and 15) on determining individual risk of overdose using the Risk Index for Overdose or Serious Opioid-Induced Respiratory Depression (RIOSORD) tool. From a metacognitive perspective, this group of questions also saw the frequency of confident incorrect answers increase at posttest between 117% and 350%.

Table . Demographic characteristics of participants (N=307).

Characteristics	Participants
Sex, n (%)	
Male	77 (25.1)
Female	106 (34.5)
Undisclosed	124 (40.4)
Race/ethnicity, n (%)	
White/non-Hispanic	88 (28.7)
Black/non-Hispanic	18 (5.9)
Hispanic	13 (4.2)
Native American	3 (1.0)
Asian/Pacific Islander	4 (1.3)
Undisclosed	181 (59.0)
Credentials, n (%)	
Health professions students	213 (69.4)
Prescribers (MD[a], DO[b], NP-C[c], or PA[d])	48 (15.6)
Pharmacists	3 (1.0)
Other health professional	4 (1.3)
Undisclosed	39 (12.7)
Age (years), mean (SD)	32 (11.6)

[a]MD: medical doctor.

[b]DO: doctor of osteopathic medicine.

[c]NP-C: nurse practitioner.

[d]PA: physician’s assistant.

XSL•FO
RenderX

Table . Naloxone coprescription program metacognitive scores (N=307).

	Preintervention			Postintervention			*df*	*z*	Effect size[a]
	Median	IQR	Range	Median	IQR	Range			
CTF[b] overall score	36	32 to 40	17 to 51	43	33 to 48	18 to 58	306	−9.41[c]	−0.54 (moderate)
Binary true/false score	60	53.3 to 66.7	28.3 to 85	71.7	55 to 80	30 to 97	306	−9.41[c]	−0.54 (moderate)
Absolute accuracy of confidence judgments	0.55	0.45 to 0.65	0.15 to 0.95	0.65	0.55 to 0.80	0.2 to 1	306	−9.42[c]	−0.54 (moderate)
Bias of the confidence judgments	−0.35	−0.50 to −0.10	−0.85 to 0.65	0.10	−0.15 to 0.25	−0.80 to 0.70	306	−13.08[c]	−0.75 (strong)
Confident correct probability	0.36	0.17 to 0.62	0 to 1	0.88	0.64 to 1	0 to 1	306	−13.59[c]	−0.78 (strong)
Confident incorrect probability	0.14	0 to 0.43	0 to 1	0.80	0.38 to 1	0 to 1	306	−12.82[c]	−0.73 (strong)
Discrimination between correct and incorrect decisions	0.11	0 to 0.26	−0.63 to 0.92	0	0 - 0.21	−0.54 to 1	306	2.85[d]	0.16 (weak)

[a]Rosenthal correlation (1991).
[b]CTF: confidence-weighted true-false.
[c]*P*<.001.
[d]*P*<.01 (for this entry: *P*=.004).

XSL•FO
RenderX

Table . Item-level frequency distribution of confidence-weighted true-false (CTF) and binary true (T)/false (F) choices at pre- and posttests (N=307).

	Pretest		Posttest		McNemar test on binary pre- and postperformance		Confident incorrect answers, % change
	CTF (%)	Binary choice (%)	CTF (%)	Binary choice (%)	Test statistic	*P* value	
Item 1: Naloxone coprescription efforts have been shown to increase access to naloxone for high-risk patients only in primary care settings. [Correct: F]					2.47	.16	100
Sure True	11	True: 38	22	True: 33			
Unsure True	27		11				
Unsure False	36	False: 62	11	False: 66			
Sure False	26		55				
Item 2: Higher doses of naloxone may be safely used if a person is suspected of overdosing from synthetic opioids such as Fentanyl. [Correct: T]					26.18	<.001	−28.5
Sure False	7	False: 27	5	False: 11			
Unsure False	19		6				
Unsure True	45	True: 73	22	True: 88			
Sure True	28		66				
Item 3: Clinicians can prescribe only naloxone to patients receiving opioid prescriptions. [Correct: F]					1.25	.26	64
Sure True	11	True: 28	18	True: 26			
Unsure True	17		8				
Unsure False	30	False: 72	12	False: 75			
Sure False	42		63				
Item 4: A person under the influence of an opioid can be arrested and charged for being under the influence of a controlled substance if he or she seeks medical assistance for himself or herself or someone else. [Correct: F]					0.91	.34	5
Sure True	6	True: 22	14	True: 21			
Unsure True	16		7				
Unsure False	21	False: 78	5	False: 80			
Sure False	57		75				
Item 5: The cheapest form of naloxone is the naloxone autoinjector made by Evzio. [Correct: F]					16.20	<.001	129
Sure True	7	True: 47	18	True: 33			
Unsure True	40		15				
Unsure False	37	False: 53	6	False: 67			
Sure False	16		61				
Item 6: Writing a prescription for Evzio, Narcan, or generic will each result in a patient receiving the same product. [Correct: F]					13.23	<.001	82
Sure True	17	True: 56	31	True: 44			
Unsure True	39		13				
Unsure False	29	False: 44	13	False: 55			
Sure False	15		42				
Item 7: Naloxone should be coprescribed to patients only with a RIOSORD [a] score of >18. [Correct: F]					0.81	.37	350

XSL•FO
RenderX

	Pretest		Posttest		McNemar test on binary pre- and postperformance		Confident incorrect answers, % change
	CTF (%)	Binary choice (%)	CTF (%)	Binary choice (%)	Test statistic	*P* value	
Sure True	8	True: 59	36	True: 56			
Unsure True	51		20				
Unsure False	30	False: 41	11	False: 44			
Sure False	11		33				
Item 8: Facilitators involved in leading the implementation of a naloxone coprescribing program are limited to clinical staff. [Correct: F]					19.76	<.001	−32
Sure True	9	True: 38	16	True: 26			
Unsure True	30		10				
Unsure False	31	False: 62	14	False: 74			
Sure False	30		60				
Item 9: Academic detailing is a service provided by academic professionals (ie, faculty at educational institutions) who provide clinicians with information on new clinical guidelines and how to implement them. [Correct: T]					0.02	.88	300
Sure False	1	False: 7	4	False: 7			
Unsure False	6		3				
Unsure True	58	True: 93	20	True: 93			
Sure True	35		73				
Item 10: A social marketing program for patients is likely to be more effective in a larger health system such as JerseyCare, as opposed to a smaller practice such as Johnson Family Practice. [Correct: F]					0.15	.70	4
Sure True	16	True: 54	35	True: 56			
Unsure True	38		21				
Unsure False	34	False: 46	18	False: 44			
Sure False	12		26				
Item 11: Tailoring aspects of the naloxone coprescription checklist to accommodate your practice is not recommended because it will limit the effectiveness of the naloxone coprescription program. [Correct: F]					0.38	.54	−6
Sure True	7	True: 35	21	True: 33			
Unsure True	27		12				
Unsure False	41	False: 65	17	False: 67			
Sure False	24		50				
Item 12: The RIOSORD tool calculates a patient's risk of overdose according to his or her mental health comorbidities. [Correct: F]					8.56	.003	117
Sure True	29	True: 87	63	True: 80			
Unsure True	58		17				
Unsure False	9	False: 13	5	False: 20			
Sure False	4		15				
Item 13: Developing a stakeholder analysis can be an effective way to both engage and motivate stakeholders as well as facilitate buy-in to your coprescribing program. [Correct: T]					18.67	<.001	−33

XSL•FO
RenderX

	Pretest		Posttest		McNemar test on binary pre- and postperformance		Confident incorrect answers, % change
	CTF (%)	Binary choice (%)	CTF (%)	Binary choice (%)	Test statistic	*P* value	
Sure False	3	False: 12	2	False: 4			
Unsure False	10		2				
Unsure True	56	True: 88	17	False: 96			
Sure True	32		79				
Item 14: Organizational Readiness Assessments allow facilitators to identify the likelihood that instituting a change in their practice will be successful. [Correct: T]					9.14	.003	−50
Sure False	2	False: 7	1	False: 3			
Unsure False	6		2				
Unsure True	57	True: 93	17	True: 98			
Sure True	36		81				
Item 15: Gap analyses reveal unmet gaps in naloxone coprescribing to patients with a RIOSORD score of >18. [Correct: F]					4.41	.04	191
Sure True	22	True: 84	64	True: 90			
Unsure True	63		26				
Unsure False	13	False: 16	6	False: 10			
Sure False	3		4				
Item 16: Studies show that patients prescribed naloxone are more likely to engage in risky opioid-related behaviors because of a decreased perception of risk. [Correct: F]					6.86	.009	45
Sure True	11	True: 34	16	True: 27			
Unsure True	23		11				
Unsure False	35	False: 66	12	False: 73			
Sure False	31		61				
Item 17: Provider stigma is a barrier to coprescribing naloxone. [Correct: T]					14.29	<.001	50
Sure False	2	False: 14	3	False: 5			
Unsure False	12		2				
Unsure True	29	True: 86	9	True: 96			
Sure True	57		87				
Item 18: The RE-AIM[b] framework is useful in structuring the evaluation and sustainability of your naloxone coprescription program. [Correct: T]					10.12	.002	0
Sure False	1	False: 9	1	False: 3			
Unsure False	8		2				
Unsure True	58	True: 91	21	True: 97			
Sure True	32		76				
Item 19: Providers in private practice with <10 staff members can implement the RE-AIM framework and naloxone coprescribing checklist effectively. [Correct: T]					2.06	.15	67
Sure False	3	False: 16	5	False: 12			
Unsure False	14		7				
Unsure True	58	True: 84	26	True: 89			
Sure True	26		63				
Item 20: In order to have the best results, implementation frameworks must be used in full and should not be combined. [Correct: F]					6.88	.009	73

XSL•FO
RenderX

	Pretest		Posttest		McNemar test on binary pre- and postperformance		Confident incorrect answers, % change
	CTF (%)	Binary choice (%)	CTF (%)	Binary choice (%)	Test statistic	*P* value	
Sure True	22	True: 64	38	True: 57			
Unsure True	42		19				
Unsure False	28	False: 36	19	False: 44			
Sure False	8		25				

[a]RIOSORD: Risk Index for Overdose or Serious Opioid-Induced Respiratory Depression.

[b]RE-AIM: reach, effectiveness, adoption, implementation, and maintenance.

Discussion

In summary, findings suggest that the naloxone coprescription academic detailing intervention was effective at delivering content area knowledge and stimulating metacognition about coprescription practices. From a knowledge gain perspective, the intervention saw increases in participant knowledge along the key objectives of a naloxone coprescription program. In addition, metacognitively, results suggest that individuals were more likely to be confident in their answer choices after the intervention. While the confidence gain was seen mostly among participants who chose correct answers, a small number of participants also became overconfident in their incorrect answers. This finding could support the development of refresher courses as a tactic to reexpose those who were overconfident to the material to correct any misunderstanding of course content [15], and for naloxone-prescribing programs, refreshers would be needed to account for the changing nature of the naloxone marketplace or clinical guidelines for overdose risk. The absolute AC significantly improved after intervention. Participants were better able to confidently discern correct and incorrect answers at posttest.

Across medical education settings, metacognitive evaluations have been implemented successfully, which has resulted in improvements in metacognition itself [16], the learning and retrieval of basic science information [17], and moderation of performance test anxiety in observed clinical examinations [18]. Even withstanding the complexity of metacognitive measurement concepts [10], CTF presents itself as a simple mechanism for metacognitive evaluation available to medical educators and evaluators, allowing them to assess potential areas of weakness in content delivery and specific areas where students may struggle with concepts [14]. Academic detailing programs applying metacognitive evaluative processes may be best served by developing feedback loops for learners and curriculum designers driven by the results CTF tests. In our results, learners were the most confident in incorrect answers for questions detailing the specifics of assessing individual risk of overdose. Feedback to learners could provide clarification on application of the RIOSORD tool through follow-up emails, refresher courses, or the development of learning communities to support implementation and adoption. Feedback to curriculum designers may prompt an evaluation of course content to identify what course updates were needed to ensure key concept delivery.

This specific intervention was self-paced and web-based, a common format available for CME. Electronic interventions have been shown to be no different for metacognition than in-person interventions, despite having no formal educator to guide the process [19]. This is important evidence to bolster the benefit of web-based continuing education [9], especially given the proliferation of web-based education that occurred during the COVID-19 pandemic [20]. Evidence suggests that if learners are going to engage in a self-paced curriculum, adding a metacognitive layer forces learners to critically think about their content knowledge acquisition [21]. The identified potential overconfidence observed in this study after receiving education is consistent with other metacognitive evaluations [22].

This study is not without limitations. The evaluation used a 1-group pretest, posttest design, which limits generalizability of the findings. While the course on best practices for coprescribing was brief and designed to be completed in 1 session, it is unknown how or whether other naloxone initiatives may have influenced participants. The academic detailing program's enrollment was open to RUSOM and its affiliates; it is not possible to rule out selection bias as 1 factor influencing score improvements.

While metacognitive processing was shown to be important for behavior change, we do not have a long-term measure to determine whether the intervention resulted in increased naloxone prescription or even whether learners went on to implement coprescription initiatives in their practice settings. As referenced earlier, there are multiple ways to assess metacognition, of which CTF is one, and the validity of one accepted measure of metacognition has yet to be established. However, this particular method of assessing metacognition with multiple conceptual domains allows evaluators to use several diagnostic measures to understand the conditions under which knowledge gain is occurring in educational interventions. Future research could measure long-term changes in these particular scores, tracking metacognitive monitoring as skills are applied, and potentially correlate both cognitive and metacognitive changes with on-the-ground prescribing and implementation behaviors.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Drug overdose. Centers for Disease Control. 2022. URL: https://www.cdc.gov/drugoverdose/deaths/index.html [accessed 2022-06-14]
2. Coffin PO, Behar E, Rowe C, et al. Nonrandomized intervention study of naloxone coprescription for primary care patients receiving long-term opioid therapy for pain. Ann Intern Med 2016 Aug 16;165(4):245-252. [doi: 10.7326/M15-2771] [Medline: 27366987]
3. Wilson CG, Rodriguez F, Carrington AC, Fagan EB. Development of a targeted naloxone coprescribing program in a primary care practice. J Am Pharm Assoc (2003) 2017;57(2S):S130-S134. [doi: 10.1016/j.japh.2016.12.076] [Medline: 28189537]
4. Soumerai SB, Avorn J. Principles of educational outreach ('academic detailing') to improve clinical decision making. JAMA 1990 Jan 26;263(4):549-556. [Medline: 2104640]
5. Bounthavong M, Devine EB, Christopher MLD, Harvey MA, Veenstra DL, Basu A. Implementation evaluation of academic detailing on naloxone prescribing trends at the United States Veterans Health Administration. Health Serv Res 2019 Oct;54(5):1055-1064. [doi: 10.1111/1475-6773.13194] [Medline: 31313839]
6. Behar E, Rowe C, Santos GM, et al. Acceptability of naloxone co-prescription among primary care providers treating ppatients on long-term opioid therapy for pain. J Gen Intern Med 2017 Mar;32(3):291-295. [doi: 10.1007/s11606-016-3911-z] [Medline: 27815762]
7. Behar E, Rowe C, Santos GM, Santos N, Coffin PO. Academic detailing pilot for naloxone prescribing among primary care providers in San Francisco. Fam Med 2017 Feb;49(2):122-126. [Medline: 28218937]
8. Moses TE, Moreno JL, Greenwald MK, Waineo E. Developing and validating an opioid overdose prevention and response curriculum for undergraduate medical education. Subst Abuse 2022;43(1):309-318. [doi: 10.1080/08897077.2021.1941515]
9. Medina MS, Castleberry AN, Persky AM. Strategies for improving learner metacognition in health professional education. Am J Pharm Educ 2017 May;81(4):78. [doi: 10.5688/ajpe81478] [Medline: 28630519]
10. Akturk AO, Sahin I. Literature review on metacognition and its measurement. Proc Soc Behav Sci 2011;15:3731-3736. [doi: 10.1016/j.sbspro.2011.04.364]
11. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain—United States, 2016. JAMA 2016 Apr 19;315(15):1624-1645. [doi: 10.1001/jama.2016.1464] [Medline: 26977696]
12. Williams AV, Strang J, Marsden J. Development of Opioid Overdose Knowledge (OOKS) and Attitudes (OOAS) Scales for take-home naloxone training evaluation. Drug Alcohol Depend 2013 Sep 1;132(1-2):383-386. [doi: 10.1016/j.drugalcdep.2013.02.007] [Medline: 23453260]
13. Rycroft-Malone J, Bucknall T. Models and Frameworks for Implementing Evidence-Based Practice: Linking Evidence to Action: John Wiley & Sons; 2010:288.
14. Dutke S, Barenberg J. Easy and informative: using confidence-weighted true–false items for knowledge tests in psychology courses. Psychol Learn Teach 2015 Nov 1;14(3):250-259. [doi: 10.1177/1475725715605627]
15. Bushuveb S, Bansbach J, Bentele M, et al. Overconfidence effects and learning motivation refreshing BLS: an observational questionnaire study. Resusc Plus 2023;14:100369. [doi: 10.1016/j.resplu.2023.100369]
16. Hong WH, Vadivelu J, Daniel EGS, Sim JH. Thinking about thinking: changes in first-year medical students' metacognition and its relation to performance. Med Educ Online 2015;20(1):27561. [doi: 10.3402/meo.v20.27561] [Medline: 26314338]
17. Hennrikus EF, Skolka MP, Hennrikus N. Applying metacognition through patient encounters and illness scripts to create a conceptual framework for basic science integration, storage, and retrieval. J Med Educ Curric Dev 2018;5:2382120518777770. [doi: 10.1177/2382120518777770] [Medline: 29845119]
18. O'Carroll PJ, Fisher P. Metacognitions, worry and attentional control in predicting OSCE performance test anxiety. Med Educ (Chicago Ill) 2013;47(6):562-568. [doi: 10.1111/medu.12125]
19. Norman E. The relationship between metacognitive experiences and learning: is there a difference between digital and non-digital study media? Comput Hum Behav 2016;9:301-309. [doi: 10.1016/j.chb.2015.07.043]
20. Kansal AK, Gautam J, Chintalapudi N, Jain S, Battineni G. Google trend analysis and paradigm shift of online education platforms during the COVID-19 pandemic. Infect Dis Rep 2021;13(2):418-428. [doi: 10.3390/idr13020040]
21. Tuysuzoglu BB, Greene JA. An investigation of the role of contingent metacognitive behavior in self-regulated learning. Metacogn Learn 2015;10(1):77-98. [doi: 10.1007/s11409-014-9126-y]
22. von Hoyer JF, Kimmerle J, Holtz P. Acquisition of false certainty: learners increase their confidence in the correctness of incorrect answers after online information search. Comput Assist Learn 2022;38(3):833-844. [doi: 10.1111/jcal.12657]

Abbreviations

AC: accuracy of confidence judgments
BS: bias of confidence judgments

XSL•FO
RenderX

CCP: confidence correct probability
CIP: confidence incorrect probability
CME: continuing medical education
CTF: confidence-weighted true-false
DIS: discrimination between correct and incorrect decisions
RIOSORD: Risk Index for Overdose or Serious Opioid-Induced Respiratory Depression
RUSOM: Rowan University School of Osteopathic Medicine

Edited by B Lesselroth; submitted 03.11.23; peer-reviewed by I Zakrocka, N Li, S Linder; revised version received 31.07.24; accepted 19.08.24; published 28.10.24.

Please cite as:
Enich M, Morton C, Jermyn R
Naloxone Coprescribing and the Prevention of Opioid Overdoses: Quasi-Experimental Metacognitive Assessment of a Novel Education Initiative
JMIR Med Educ 2024;10:e54280
URL: https://mededu.jmir.org/2024/1/e54280
doi:10.2196/54280

© Michael Enich, Cory Morton, Richard Jermyn. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.10.2024. 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.

XSL•FO
RenderX

Impact of Ophthalmic Knowledge Assessment Program Scores and Surgical Volume on Subspecialty Fellowship Application in Ophthalmology Residency: Retrospective Cohort Study

Amanda Kay Hertel, BS; Radwan S Ajlan, MBBCh

University of Kansas School of Medicine, 7400 State Line Road, Prairie Village, KS, United States

Corresponding Author:
Radwan S Ajlan, MBBCh

Abstract

Background: Ophthalmology residents take the Ophthalmic Knowledge Assessment Program (OKAP) exam annually, which provides percentile rank for multiple categories and the total score. In addition, ophthalmology residency training programs have multiple subspecialty rotations with defined minimum procedure requirements. However, residents' surgical volumes vary, with some residents exceeding their peers in specific subspecialty rotations.

Objective: This study aims to identify if there is a difference in OKAP examination scores and surgical volume exposure during ophthalmology residency training between nonfellowship and fellowship applicants and among various subspecialties.

Methods: A retrospective review of OKAP scores and surgical procedure numbers of graduating residents in an accredited academic ophthalmology residency program in the Midwest United States was conducted. Data were collected from 2012 to 2022.

Results: A total of 31 residents were identified. Most residents decided to pursue fellowship training upon graduation (20/31, 65% residents), and the rest chose to practice comprehensive ophthalmology (11/31, 35% residents). A total of 18/31 residents had OKAP score reports available. The fellowship group outperformed the nonfellowship group in multiple subsections and the total exam (P=.04). Those pursuing fellowship training in glaucoma performed higher on the Glaucoma section (P=.004) and the total exam (P=.005). Residents pursuing cornea performed higher on nearly all subsections, including External Disease and Cornea (P=.02) and the total exam (P=.007). The majority of the surgical volume exposure was identical between fellowship and nonfellowship groups. Those who pursued glaucoma fellowship performed more glaucoma filtering and shunting procedures (P=.03). Residents going into pediatrics fellowship were primary surgeons in more strabismus cases (P=.01), assisted in fewer strabismus cases (P<.001), and had no difference in the total number of strabismus surgeries.

Conclusions: In our program, residents pursuing fellowship training had higher OKAP scores on multiple sections and the total exam. There was no significant difference in the overall surgical volume averages between fellowship and nonfellowship groups, but few differences existed in subspecialty procedures among fellowship applicants. Larger multicenter studies are needed to clarify the relationship between OKAP scores and ophthalmology fellowship decisions nationwide.

(JMIR Med Educ 2024;10:e60940) doi:10.2196/60940

KEYWORDS

residency; fellowship; ophthalmology; OKAP; surgical training; ophthalmology resident; ophthalmology residency program; examination; surgical volume exposure; fellowship training; surgical volume; exposure; Ophthalmic Knowledge Assessment Program

Introduction

Fellowship and Subspecialty Statistics

The number of ophthalmology residents who pursue further fellowship training has been increasing for more than a decade [1]. In 2005, around 64% of ophthalmology residents in the United States pursued subspecialty training, while 36% pursued comprehensive ophthalmology [2]. However, between 2012 and 2017, the percentage of ophthalmology residents in the United States pursuing subspecialty training increased to 70.3% [3]. Ophthalmology residency graduates in Canada have similar statistics, with 64% pursuing subspecialty training [4]. In the United States, vitreoretinal (36%), cornea (25%), glaucoma (13%), oculoplastic (10%), and pediatric (10%) fellowships are the most common [2]. Fewer ophthalmologists pursued fellowship training in anterior segment (2%), neuro-ophthalmology (0.7%), and uveitis (0.7%) [2]. These percentages are similar to the 2017 through 2022 San Francisco Match Data [5]. The match rate for all ophthalmology residents applying for fellowship is 73.7% [3]. Subspecialties with the greatest number of positions offered each year include retina,

XSL•FO
RenderX

cornea, glaucoma, pediatrics, and uveitis [3]. The match rate was highest for retina, followed by cornea, glaucoma, pediatrics, and uveitis [3].

Factors Influencing the Decision to Pursue a Fellowship

Based on a survey of ophthalmology residents, the top factors for deciding to pursue a fellowship include the desire for additional surgical training, additional clinical training, and increased job market competitiveness [6]. This is similar to a study stating that gaining more special skills and working with new technology were the top motivating factors for fellowship [7]. Another study found that acquiring special skills, perceived more favorable job market, and prestige were the top motivating factors for pursuing subspecialty training [2]. In contrast, those who pursued comprehensive ophthalmology were motivated by the anticipated work hours and geographical preference [2]. Other studies have also found that those going into comprehensive ophthalmology had significantly higher student loan amounts [7]. It is also noted that those deciding to do subspecialty training were more likely to intend to practice in an academic setting [2,7]. In addition, ophthalmology residents applying for fellowship had significantly more first-author publications than those going into comprehensive ophthalmology [7].

Studies have found that gender has no significant impact on the decision to pursue subspecialty training following ophthalmology residency [2,4,6,8]. However, there were gender differences among specific subspecialties. It was found that significantly more males pursued vitreoretinal surgery, while more females pursued strabismus and pediatric ophthalmology [4,8]. It has also been found that more males pursued oculoplastic [4] and anterior segment [8] fellowships. Females were also more likely to pursue neuro-ophthalmology [8]. Other studies have shown that age, ethnicity, marital status, presence of children, or level of educational debt had no statistically significant impact on the decision to pursue subspecialty training [2].

A survey done in 2005 on graduating residents found that the number of ocular procedures performed during residency did not significantly differ between residents who decided to do a fellowship and those who went into comprehensive ophthalmology [2]. However, there may be more variation when looking at specific subspecialties, as this study only evaluated fellowship versus no fellowship cohorts. Other factors, including elective time, career counseling, and the amount of dedicated time for research, also did not impact the decision to pursue a fellowship [2].

Factors Influencing the Decision to Pursue Specific Subspecialties

One study evaluated residents' decisions to pursue a fellowship in neuro-ophthalmology [9]. The top reasons graduating residents decided not to pursue this fellowship included stronger interests in other fields, types of patients seen, no intraocular surgery, and the assumption that it is a nonsurgical discipline. Factors influencing the decision to pursue neuro-ophthalmology included interest in clinical diseases and interaction with other specialty fields. There were no differences between the groups regarding the degree of exposure to neuro-ophthalmology in medical school, the presence of a dedicated neuro-ophthalmology rotation in residency, or the rotation timing [9].

Another study evaluating the decisions impacting glaucoma fellowship found that residents entering a glaucoma fellowship had performed more glaucoma filtering procedures, were less likely to publish a paper, and were less likely to have time allocated for research than residents who pursued different fellowship training. Those seeking glaucoma fellowship also found challenging diagnostic problems, types of patient problems, academic careers, and working with new technology as less important. Residents pursuing glaucoma fellowship also decided later than residents selecting other subspecialties [10].

A recent study published in 2023 evaluated ophthalmology residents' perceptions of pediatric ophthalmology, and it was found that the desire to work with children was the most significant factor in pursuing a pediatric ophthalmology residency. The majority of residents also did not believe pediatric ophthalmology to be a prestigious specialty [11], and some residents also found pediatric patients difficult to examine [12], both deterrents to pursuing pediatric ophthalmology. In addition, concerns of economic factors and compensation in this field have also been discussed [11-13].

Ophthalmic Knowledge Assessment Program

The Ophthalmic Knowledge Assessment Program (OKAP) is a multiple-choice nationwide examination used to assess the ophthalmic knowledge of ophthalmology residents compared with their peers at the same training stage (postgraduate years [PGY] 2, 3, and 4). OKAP exams are taken electronically in a supervised environment in March each year. Residents in their first postgraduate year do not take the OKAP exam because the first year is spent rotating through other specialties as an internship year with little time in ophthalmology. OKAP exam scores are frequently requested when applying for competitive fellowship programs. Essential factors to score well on this examination include increased time spent studying, the use of question banks, and the incorporation of OKAP materials into the residency program [14]. In a study surveying retina, glaucoma, and cornea fellowship programs, OKAP scores were ranked as moderately important by program directors for fellowship applications [15]. It is notable that not all fellowship programs evaluate OKAP scores and that OKAP scores are only one indicator of proficiency in ophthalmology. Despite the first national OKAP exam occurring in 1968 [16], no studies have evaluated the performance of OKAP scores on the decision to pursue fellowship training after an ophthalmology residency.

Significance

There is limited literature examining the factors that influence fellowship training selection, especially for specific subspecialties. No studies have evaluated the impact of OKAP scores and intraocular procedural volume on specific subspecialty of choice by residents. This study aims to better understand the factors leading to selecting a subspecialty after graduating from an ophthalmology residency program by evaluating the OKAP scores and the number of surgical

XSL•FO
RenderX

procedures from graduating residents. This research can then provide valuable information to ophthalmology residency program directors as programs can be better designed to graduate ophthalmologists in subspecialties of need.

Methods

Study Methods

A retrospective review of OKAP scores and intraocular procedure numbers of graduating residents in an accredited ophthalmology residency program in the Midwest United States was conducted. Data were collected for all ophthalmology residents graduating from the program between 2012 to 2022. Data collection included OKAP scores, gender, fellowship decision, and procedural volume for all prior residents meeting the inclusion criteria (graduation date from 2012 to 2022). OKAP score reports were available on the San Francisco Match website for residents graduating between 2017 and 2022. OKAP examinations are required and taken by ophthalmology residents during their second, third, and fourth residency training years. The exam is currently taken online while proctored in person. All OKAP scores were available, but only OKAP scores from each residents' final program year were used in data analysis. This is because the final year OKAP scores are the most representative of cumulative ophthalmic knowledge, and are closer to the final decision of practice or fellowship.

Statistical Analysis

Descriptive statistics were completed for the demographics, number of surgical cases, and OKAP scores among the various groups of interest. For comparing the various groups, *t* tests were used with the settings of two-sample unequal variance with a one-tailed distribution. Statistical analysis was performed using Microsoft Excel.

Ethical Considerations

Institutional review board (IRB) approval was obtained from the University of Kansas Medical Center (study ID: 00150405). This was a retrospective review with secondary analysis of data. The IRB approval covers secondary analysis performed without the need for additional consent. All data were deidentified immediately following initial collection. No compensation was provided.

Results

Overview

A total of 31 residents were identified. Most residents decided to pursue fellowship training upon graduation (20/31, 65% residents), and the rest decided to practice comprehensive ophthalmology (11/31, 35%residents). Overall, 20 separate fellowships were completed by residents in Glaucoma (7/20, 35% fellowships), Surgical Retina (5/20, 25% fellowships) and Medical Retina (2/20, 10% fellowships), Cornea (4/20, 20% fellowships), and Pediatrics (2/20, 10% fellowships). No residents completed fellowships in Oculoplastics, Uveitis, or Anterior Segment in this residency program.

Out of the 31 residents identified, 42% (13/31) were female, while 58% (18/31) were male (Table 1). Of those who pursued fellowship (n=20), 50% (10/20) were female, and 50% (10/20) were male, while those who did not pursue a fellowship were 27% (3/11) female and 73% (8/11) male. When looking at specific subspecialties, 29% (2/7) of residents who pursued a Glaucoma fellowship were female. In addition, 50% (2/4) of Cornea fellowships, 100% Pediatric fellowships (2/2), 0% (0/2) of Medical Retina fellowships, and 60% (3/5) of Surgical Retina fellowships were female.

Table . Resident demographics (N=31).

Demographics	Female residents, n/N (%)	Male residents, n/N (%)
Total	13/31 (42)	18/31 (58)
Pursued fellowship	10/20 (50)	10/20 (50)
No fellowship; comprehensive ophthalmology	3/11 (27)	8/11 (73)
Glaucoma	2/7 (29)	5/7 (71)
Cornea	2/4 (50)	2/4 (50)
Pediatrics	2/2 (100)	0/2 (0)
Medical Retina	0/2 (0)	2/2 (100)
Surgical Retina	3/5 (60)	2/5 (40)

Fellowship Versus No Fellowship

A total of 18 OKAP score reports since 2017 were available. Around half of these residents pursued a fellowship (56%, 10/18 residents), while the others pursued comprehensive ophthalmology (44%, 8/18 residents). On the OKAP scores, the fellowship group outperformed the nonfellowship group in General Medicine (*P*=.03), Ophthalmic Pathology and Intraocular Tumors (*P*=.04), Lens and Cataract (*P*=.04), and Total Exam (*P*=.04). The difference in OKAP percentile rank compared with the entire group average for the fellowship and no fellowship cohorts is detailed in Table 2.

XSL•FO
RenderX

Table . Difference in Ophthalmic Knowledge Assessment Program score average percentile rank compared with the whole cohort.

	General medicine	Fundamentals and principles of ophthalmology	Clinical optics	Ophthalmic pathology and intraocular tumors	Neuro-ophthalmology	Pediatric ophthalmology and strabismus	Orbit, eyelids, and lacrimal system	External disease and cornea	Intraocular inflammation and uveitis	Glaucoma	Lens and cataract	Retina and vitreous	Refractive surgery	Total exam
Fellowship (n=10)	13[a]	12	8	10[a]	−1	10	4	8	10	9	13[a]	8	6	12[a]
No fellowship (n=8)	−17[a]	−14	−10	−13[a]	1	−13	−5	−10	−13	−11	−16[a]	−10	−8	−15[a]
Glaucoma (n=6)	12	10	9	19[b]	9	13	16[a]	8	13	22[b]	9	8	8	21[b]
Cornea (n=2)	18[a]	17[a]	23[b]	19	1	27[b]	23[b]	14[a]	35[b]	24[b]	33[b]	11	35[b]	21[b]
Medical Retina (n=2)	11	12	−7	−24	−32[a]	−15[a]	−51[a]	2	−24	−47[b]	4	6	−26[a]	−23

[a]$P<.05$.

[b]$P<.01$

All residents had surgical case number reports available. There were minimal differences between the fellowship and no fellowship group in procedural volume. However, the fellowship cohort did assist in more oculoplastic and orbit cases (*P*=.02) and more eyelid laceration cases (*P*=.02). All other surgical volume categories had no statistical significance between the fellowship and no fellowship groups (Tables S1-S4 in Multimedia Appendix 1).

Subspecialty

The specialties represented among the OKAP score reports were Glaucoma (60%, 6/10 residents), Cornea (20%, 2/10 residents), and Medical Retina (20%, 2/10 residents). Analysis of OKAP scores was then performed on the subspecialty cohorts compared with all other residents. Those who ended up pursuing fellowship training in glaucoma performed higher on Ophthalmic Pathology and Intraocular Tumors (*P*=.004), Orbit, Eyelids, and Lacrimal System (*P*=.02), Glaucoma (*P*=.004), and on total exam (*P*=.005). Those pursuing cornea performed higher on nearly all subsections and ontotal exam (*P*=.007). The sections they scored higher on include General Medicine (*P*=.01), Fundamentals and Principles of Ophthalmology (*P*=.02), Clinical Optics (*P*=.002), Pediatric Ophthalmology and Strabismus (*P*<.001), Orbit, Eyelids, and Lacrimal System (*P*=.004), External Disease and Cornea (*P*=.02), Intraocular Inflammation and Uveitis (*P*<.001), Glaucoma (*P*=.004), Lens and Cataract (*P*=.002), and Refractive Surgery (*P*<.001). The other fellowship represented in the OKAP scores was Medical Retina. Those who ended up pursuing medical retina did perform lower on some subsections. However, their total exam score and their Retina and Vitreous section scores were no different than the rest of the graduates. Residents who ended up pursuing medical retina scored lower on Neuro-ophthalmology (*P*=.05), Pediatric Ophthalmology and Strabismus (*P*=.02), Orbit, Eyelids, and Lacrimal System (*P*=.02), Glaucoma (*P*<.001), and Refractive Surgery (*P*=.02).

The procedural volume completed by the various subspecialties was also analyzed. Those who pursued a fellowship in glaucoma performed fewer cases of panretinal laser photocoagulation (*P*=.01) and strabismus (*P*=.04) as primary. They, however, assisted on more eyelid laceration cases (*P*=.03) and had more total glaucoma filtering and shunting procedures (*P*=.03). Those seeking cornea had fewer assist cases for laser trabeculoplasties (*P*=.004), laser iridotomies (*P*=.01), and chalazion excisions (*P*=.002). The cornea fellowship group assisted more and had a higher total number of retinal vitreous cases (*P*=.002 and .007, respectively). They also performed more open globe trauma cases (*P*=.03). Residents who went into pediatric ophthalmology performed fewer intravitreal injections (*P*<.001), and eyelid laceration repair (*P*=.002), as primary. They also had fewer total keratorefractive surgeries (*P*=.003) and globe trauma cases (*P*=.009). However, those pursuing pediatrics were the primary surgeons for more strabismus cases (*P*=.01) and assisted in fewer strabismus cases (*P*<.001). There was no significant difference in the total number of strabismus surgeries performed by graduates who pursued pPediatric Ophthalmology compared with other residents. Those going into Medical Retina had fewer total keratoplasty cases (*P*=.007) and assisted in more ptosis and blepharoplasty cases (*P*=.01). Those going into surgical retina had fewer total case numbers for cataract, Yttrium Aluminum Garnet capsulotomy, laser trabeculoplasty, and glaucoma surgeries (*P*=.02, .002, <.001, <.001, respectively). The surgical retina group also had fewer total pterygium or conjunctival cases (*P*=.04) and keratorefractive surgeries (*P*=.001; Tables S1-S4 in Multimedia Appendix 1).

XSL•FO
RenderX

Discussion

Principal Findings

To the best of our knowledge, this is the first study to evaluate the effect of OKAP scores on fellowship choice. In this study, it was found that most graduates from this residency program pursued fellowship training, with the most common fellowships being Glaucoma, Surgical Retina, Cornea, Medical Retina, and Pediatrics, respectively. When analyzing the OKAP scores, it was found that residents pursuing fellowship training scored higher on the total OKAP exam and on multiple subsections. It was also found that residents applying for specific subspecialties scored higher on that subsection of the OKAP examination. In addition, the procedural volume did not significantly differ between the fellowship and nonfellowship groups with some variations when analyzing specific subspecialties.

OKAP Scores Difference Between Fellowship and Nonfellowship

In multiple categories and on total exam, ophthalmology residents who pursued fellowship training scored higher than those who did not. One possible explanation for this difference is that OKAP scores are sometimes included in a Fellowship Application and have been identified as a moderately important aspect of a fellowship application by some program directors [15]. Residents considering a fellowship may have prepared more for the exam to be competitive for fellowship applications. In addition, this finding could suggest that residents who scored high on the OKAP were later influenced to pursue fellowship training due to their more competitive applications. Since this is a retrospective review, it is unknown whether residents entered this program with a preexisting interest in fellowship. Therefore, higher OKAP scores in general may serve as a predictor for residency programs that a resident is more likely to pursue fellowship training.

Ophthalmic Knowledge Assessment Program Score Differences Among Subspecialties

Those who pursued glaucoma fellowship also scored higher on total exam and in multiple subsections, including Glaucoma, compared with the rest of the cohort. In addition, those who pursued cornea fellowship outperformed the fellowship and nonfellowship groups on total exam and in many subcategories, including External Disease and Cornea. This could indicate that residents who decided to apply for a specific fellowship in a certain specialty had increased interest and knowledge in the field, spent more time learning that content, and scored higher on the associated OKAP sections. This could be one way for residency programs to identify applicants more likely to pursue fellowship training in a specific area. Those pursuing medical retina scored lower on some OKAP subsections but not on total exam or the Retina and Vitreous section. One potential reason is that medical retina (and retina in general) has an overall higher match rate, and a greater number of positions offered, making it less competitive than some of the other specialties [3]. This may mean OKAP scores were perceived to have less impact on matching into medical retina programs. It is also noted that while there are variations in the OKAP scores, all residents have successfully graduated from the residency program and are board-certified.

Procedural Volume Between Fellowship and Nonfellowship

Interestingly, in our study, when comparing the procedural volume of residents who pursued fellowship to those who did not, the only statistically significant difference was in the oculoplastics and orbit and eyelid laceration categories, despite no residents seeking fellowship in oculoplastics. This agrees with other studies that have shown the number of ocular procedures did not significantly differ between residents going into fellowship programs versus comprehensive ophthalmology [2]. It is also noted that this study evaluates a greater variety of procedures due to various technological advancements since 2005. This shows that while students might have different career paths, the bulk of surgical training during ophthalmology residency is the same. This is also important for those going into comprehensive ophthalmology, as comprehensive ophthalmologists must be well-versed in many areas.

Procedural Volume Differences Among Subspecialties

In this study, specific subspecialties had some variations within procedural volumes. For example, those who pursued glaucoma fellowship had a greater total number of glaucoma filtering and shunting procedures. This is consistent with another study that found residents entering a glaucoma fellowship performed more glaucoma filtering procedures [10]. This may mean those residents sought more opportunities to get involved in glaucoma cases and pursued extra training. Residents who ended up in pediatrics were the surgeons for more strabismus cases but had no statistically significant difference in the total number of strabismus cases. Interestingly, residents who went into cornea had no differences in corneal surgery procedural volume, and those who went into medical or surgical retina had no differences in retinal vitreous and intraretinal injections. It is also noted that despite these variations, all residents met the minimum training requirements in each area.

Residency Demographics and Nonpursued Subspecialties

The percentage of program graduates deciding to pursue fellowship training (64.5%) is close to the approximately 70% of ophthalmology residency graduates in the United States who pursued fellowship training [3]. The lower percentage pursuing a fellowship in this specific program may be because this program has no associated fellowship programs. Those who selected this residency program may have been less motivated by fellowship programs initially. While retina, cornea, and pediatric percentages are similar to averages in the United States, significantly more residents from this program pursued glaucoma fellowship [2]. It is also noted that no residents pursued a fellowship in oculoplastics despite it being the fourth most common ophthalmology fellowship nationwide [2]. One potential explanation is that the Oculoplastics application is during PGY-3 year instead of PGY-4 like the others, meaning residents must make this decision early [17]. This is also considered to be one of the more competitive fellowship programs, which may deter residents from applying. In addition,

XSL•FO
RenderX

no residents pursued a fellowship in uveitis, neuro-ophthalmology, or anterior segment. However, in the United States, these fellowship programs account for smaller percentages out of all fellowships [2].

Gender

While the majority of residents in this study identified as male, half of the cohort who pursued fellowship training identified as female. Conversely, most residents who did not pursue fellowship training were male. This is consistent with other studies that have shown gender to have no significant impact on the decision to attain subspecialty training in ophthalmology [2,4,6,8]. At least 50% of residents in this program who pursued fellowships in pediatrics (100%, 2/2 residents), surgical retina (60%, 3/5 residents), and cornea (50%, 2/4 residents) identified as female. This is consistent with other studies that have found more female ophthalmologists complete fellowships in Pediatrics and Strabismus [4,8]. However, these studies also found most residents who pursued Surgical Retina fellowship identify as male [4,8]. Interestingly, in our cohort, the majority of residents who pursued Retina Surgery were female. This could be due to the small cohort size or program-specific variations like the presence of female surgical retina attendings.

Other Factors Affecting Decision to Pursue a Fellowship

This study assessed the impact of OKAP scores and procedural volume on fellowship decision. Other studies and surveys have found that other important factors in the decision to subspecialize in ophthalmology include wanting additional training (both clinical and surgical) [2,6], acquiring specialized skills [7], working with new technology [7], increased job market competitiveness [2,6], and prestige [2]. In addition, those who have fellowship training are more likely to practice in academic settings [2,7], and to have first-author publications [7]. Therefore, residents who want to work in research or academics may be more inclined to subspecialize.

Limitations

The limitations of this study include the small sample size of a single mid-western program. The limited sample size can limit the generalizability of this study in a larger program. In addition, this residency program did not have associated fellowship programs during the study period, which, may have influenced residents' decision to pursue fellowship training, or attracted residents who wanted to maximize their surgical experience without fellows to practice comprehensive ophthalmology after graduation. Another limitation is the retrospective nature of this study, which does not allow examining if residents decided on fellowship before or after OKAP examination scores. It is unknown at what point in training each resident decided to pursue a fellowship. In addition, no residents decided to pursue training in certain fellowships (eg, oculoplastics and neuro-ophthalmology), so there are no data on those specific specialties.

Strengths

This study was conducted in an ophthalmology residency program accredited by the Accreditation Council for Graduate Medical Education. All residents in the program passed their minimal required surgical procedure volumes. All graduates of this program are certified by the American Board of Ophthalmology or preparing to set the board exams if graduated in 2022 at the time of the paper submission.

Conclusion

OKAP performance showed there were differences between fellowship and nonfellowship graduates in our program. Overall, residents in this program who pursued a fellowship scored higher than those who did not pursue a fellowship on multiple sections and on total OKAP examination. There were no significant differences in the overall surgical volume averages between fellowship and nonfellowship groups, but a few differences existed in subspecialty procedures among fellowship applicants. Despite these variations, all residents exceeded the minimum training requirements. Larger multicenter studies are needed to better clarify OKAP score relation to fellowship and subspecialty application decisions nationwide.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data.
[DOCX File, 25 KB - mededu_v10i1e60940_app1.docx]

References

1. Tsou BC, Aguwa UT, Arsiwala LT, et al. Trends in cornea fellowship applications and appplicant characteristics: a San Francisco match analysis. J Acad Ophthalmol (2017) 2022 Jul;14(2):e216-e223. [doi: 10.1055/s-0042-1756199] [Medline: 37388181]
2. Gedde SJ, Budenz DL, Haft P, Tielsch JM, Lee Y, Quigley HA. Factors influencing career choices among graduating ophthalmology residents. Ophthalmology 2005 Jul;112(7):1247-1254. [doi: 10.1016/j.ophtha.2005.01.038] [Medline: 15921748]
3. Zafar S, Bressler NM, Golnik KC, et al. Fellowship match outcomes in the U.S. from 2010 to 2017: analysis of San Francisco match. Am J Ophthalmol 2020 Oct;218:261-267. [doi: 10.1016/j.ajo.2020.06.008] [Medline: 32574772]

XSL•FO
RenderX

4. Sivachandran N, Noble J, Dollin M, O'Connor MD, Gupta RR. Trends in subspecialty training by Canadian ophthalmology graduates. Can J Ophthalmol 2016 Jun;51(3):201-206. [doi: 10.1016/j.jcjo.2015.10.011] [Medline: 27316270]
5. Secondary ophthalmology fellowship statistics. SF Match. 2022. URL: https://sfmatch.org/specialty/ophthalmology-fellowship/Statistics [accessed 2024-10-21]
6. Czyz CN, Kashyap R, Wayman LL. Factors influencing fellowship training among ophthalmology residents: a pilot study. HCA Healthc J Med 2022;3(5):271-282. [doi: 10.36518/2689-0216.1382] [Medline: 37425250]
7. Chen X, Zafar S, Srikumaran D, et al. Factors influencing postgraduate career decisions of ophthalmology residents. J Acad Ophthalmol 2020 Jul;12(2):e124-e133. [doi: 10.1055/s-0040-1715808]
8. Al-Essa RS, Al-Otaibi MD, Al-Qahtani BS, Masuadi EM, Omair A, Alkatan HM. Future ophthalmology practice pattern: a survey of Saudi Board of Ophthalmology residents. Saudi J Ophthalmol 2019;33(1):1-6. [doi: 10.1016/j.sjopt.2019.01.005] [Medline: 30930655]
9. Solomon AM, Patel VR, Francis CE. Factors affecting ophthalmology resident choice to pursue neuro-ophthalmology fellowship training. J Neuroophthalmol 2022 Mar 1;42(1):56-61. [doi: 10.1097/WNO.0000000000001239] [Medline: 33770011]
10. Gedde SJ, Budenz DL, Haft P, Lee Y, Quigley HA. Factors affecting the decision to pursue glaucoma fellowship training. J Glaucoma 2007 Jan;16(1):81-87. [doi: 10.1097/01.ijg.0000243474.36213.08] [Medline: 17224755]
11. Lee KE, Sussberg JA, Nelson LB, Thuma TBT. In the setting of heightened economic and workforce issues, what are third-year (PGY-4) ophthalmology residents' perspectives of pediatric ophthalmology? J Pediatr Ophthalmol Strabismus 2023;60(2):95-100. [doi: 10.3928/01913913-20230111-02] [Medline: 36975113]
12. Hasan SJ, Castanes MS, Coats DK. A survey of ophthalmology residents' attitudes toward pediatric ophthalmology. J Pediatr Ophthalmol Strabismus 2009;46(1):25-29. [doi: 10.3928/01913913-20090101-09] [Medline: 19213273]
13. Bernstein BK, Nelson LB. Workforce issues in pediatric ophthalmology. J Pediatr Ophthalmol Strabismus 2020 Jan 1;57(1):9-11. [doi: 10.3928/01913913-20191101-01] [Medline: 31972034]
14. Zafar S, Wang X, Srikumaran D, et al. Resident and program characteristics that impact performance on the Ophthalmic Knowledge Assessment Program (OKAP). BMC Med Educ 2019 Jun 7;19(1):190. [doi: 10.1186/s12909-019-1637-4] [Medline: 31174525]
15. Kempton JE, Shields MB, Afshari NA, Dou W, Adelman RA. Fellow selection criteria. Ophthalmology 2009 May;116(5):1020-1020. [doi: 10.1016/j.ophtha.2008.12.047] [Medline: 19410971]
16. Rubin ML. The Ophthalmic Knowledge Assessment Program (OKAP): a personal view. Surv Ophthalmol 1988;32(4):282-287. [doi: 10.1016/0039-6257(88)90176-2] [Medline: 3347895]
17. Saleh GM, Athanasiadis I, Collin JRO. Training and oculoplastics: past, present and future. Orbit 2013 Apr;32(2):111-116. [doi: 10.3109/01676830.2013.764448] [Medline: 23514028]

Abbreviations

IRB: institutional review board
OKAP: Ophthalmic Knowledge Assessment Program

Edited by B Lesselroth; submitted 26.05.24; peer-reviewed by HA Serhan, J Syed; revised version received 07.09.24; accepted 24.09.24; published 13.11.24.

Please cite as:
Hertel AK, Ajlan RS
Impact of Ophthalmic Knowledge Assessment Program Scores and Surgical Volume on Subspecialty Fellowship Application in Ophthalmology Residency: Retrospective Cohort Study
JMIR Med Educ 2024;10:e60940
URL: https://mededu.jmir.org/2024/1/e60940
doi:10.2196/60940

© Amanda Kay Hertel, Radwan S Ajlan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.11.2024. 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.

XSL•FO
RenderX

Original Paper

Using Project Extension for Community Healthcare Outcomes to Enhance Substance Use Disorder Care in Primary Care: Mixed Methods Study

MacKenzie Koester[1], MPH; Rosemary Motz[1], MPH, MA, RDN; Ariel Porto[1], MPH; Nikita Reyes Nieves[1], MPH; Karen Ashley[1], EdD

Weitzman Institute, Moses Weitzman Health System, Washington, DC, United States

Corresponding Author:
MacKenzie Koester, MPH
Weitzman Institute
Moses Weitzman Health System
1575 I Street Northwest
Suite 300
Washington, DC, 20005
United States
Phone: 1 8603476971
Email: koestem@mwhs1.com

Abstract

Background: Substance use and overdose deaths make up a substantial portion of injury-related deaths in the United States, with the state of Ohio leading the nation in rates of diagnosed substance use disorder (SUD). Ohio's growing epidemic has indicated a need to improve SUD care in a primary care setting through the engagement of multidisciplinary providers and the use of a comprehensive approach to care.

Objective: The purpose of this study was to assess the ability of the Weitzman Extension for Community Healthcare Outcomes (ECHO): Comprehensive Substance Use Disorder Care program to both address and meet 7 series learning objectives and address substances by analyzing (1) the frequency of exposure to the learning objective topics and substance types during case discussions and (2) participants' change in knowledge, self-efficacy, attitudes, and skills related to the treatment of SUDs pre- to postseries. The 7 series learning objective themes included harm reduction, team-based care, behavioral techniques, medication-assisted treatment, trauma-informed care, co-occurring conditions, and social determinants of health.

Methods: We used a mixed methods approach using a conceptual content analysis based on series learning objectives and substances and a 2-tailed paired-samples *t* test of participants' self-reported learner outcomes. The content analysis gauged the frequency and dose of learning objective themes and illicit and nonillicit substances mentioned in participant case presentations and discussions, and the paired-samples *t* test compared participants' knowledge, self-efficacy, attitudes, and skills associated with learning objectives and medication management of substances from pre- to postseries.

Results: The results of the content analysis indicated that 3 learning objective themes—team-based care, harm reduction, and social determinants of health—resulted in the highest frequencies and dose, appearing in 100% (n=22) of case presentations and discussions. Alcohol had the highest frequency and dose among the illicit and nonillicit substances, appearing in 81% (n=18) of case presentations and discussions. The results of the paired-samples *t* test indicated statistically significant increases in knowledge domain statements related to polysubstance use (P=.02), understanding the approach other disciplines use in SUD care (P=.02), and medication management strategies for nicotine (P=.03) and opioid use disorder (P=.003). Statistically significant increases were observed for 2 self-efficacy domain statements regarding medication management for nicotine (P=.002) and alcohol use disorder (P=.02). Further, 1 statistically significant increase in the skill domain was observed regarding using the stages of change theory in interventions (P=.03).

Conclusions: These findings indicate that the ECHO program's content aligned with its stated learning objectives; met its learning objectives for the 3 themes where significant improvements were measured; and met its intent to address multiple substances in case presentations and discussions. These results demonstrate that Project ECHO is a potential tool to educate multidisciplinary providers in a comprehensive approach to SUD care.

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e48135) doi:10.2196/48135

KEYWORDS

continuing medical education; telementoring; substance use disorder treatment; substance use disorder; SUD; primary care; Extension for Community Healthcare Outcomes; Project ECHO

Introduction

Background

In the United States, overdose deaths continue to be a major cause of injury-related deaths. Since the onset of the COVID-19 pandemic, numbers have only accelerated, and the state of Ohio has led the nation in high substance use disorder (SUD) rates, including drug use and prescription drug use. The Centers for Disease Control and Prevention ranks the state among the top 5 across the United States with the highest rates of opioid overdose deaths [1]. While research has shown an increase in the number of people enrolled in substance use treatment in Ohio between 2015 and 2019 there was still a notable high increase in the annual average prevalence of past-year illicit drug use disorder in Ohio (3.6%) compared to the regional average (3%) and the national average (2.9%) [2]. In addition, past-month alcohol use disorder (9.3%), cannabis use disorder (5.8%), and tobacco use disorder (35.2%) were higher than the national average among young adults aged 18-25 years [2]. Ohio's growing epidemic has highlighted the need to improve SUD care in a primary care setting by training providers to better address differences in care and social determinants of health through the use of behavioral techniques, harm-reduction philosophy of care, medication management, and a team-based care approach.

Weitzman Extension for Community Healthcare Outcomes: Comprehensive Substance Use Disorder Care Program

Beginning in 2021, Buckeye Health Plan and Ohio University Heritage College of Osteopathic Medicine have partnered with the Weitzman Institute (WI), a national primary care research, policy, and education institute, to provide targeted support and education to Ohio primary care medical and behavioral health providers working with underserved patients, especially those in the rural, southeastern Appalachian region, using the evidence-based Project Extension for Community Healthcare Outcomes (ECHO) learning model. Project ECHO uses frequent videoconference sessions to connect a target audience of learners with subject matter experts for didactic and case-based instruction and engaged discussion [3]. Through regular attendance at these sessions, Project ECHO aims to equip learners with the knowledge, confidence, and skills to better manage complex cases.

WI has over 11 years of experience in developing and delivering Project ECHO programs to meet the needs of providers working in resource-limited settings. As an early adopter of the model in 2012, Weitzman ECHO programs have been offered in 22 topic areas to over 8000 health care professionals across all 50 states, Washington D.C., and Puerto Rico. Working in collaboration, Buckeye Health Plan and Ohio University aimed to leverage this expertise and offer multiple Project ECHO programs each year for providers in topics of the greatest need and interest.

As described, one of Ohio's most dire population health needs is to improve outcomes for patients experiencing addiction. Thus, SUD was selected as the second ECHO program developed through this partnership. More specifically, opioids are a heightened concern throughout both Ohio and the United States, and the opioid epidemic has spurred significant funding allocations, such as the Biden Administration's US $1.5 billion award to states and territories to end the epidemic [4]. However, there are many additional substances of concern, both illicit and nonillicit, such as alcohol, tobacco, cannabis, methamphetamine, and cocaine [5], which may receive less attention given the directed funding for opioids. For this reason, it was decided that the ECHO would address not only opioids, or any one substance, but rather be designed to provide techniques to help providers address SUD overall through a comprehensive, team-based lens and a harm reduction philosophy of care. Reflecting this broad topical approach, the program was titled the Weitzman ECHO: Comprehensive Substance Use Disorder Care (CSUDC ECHO) program.

CSUDC ECHO consisted of 24 twice-monthly sessions held between July 2021 and July 2022. Each 1-hour session included a 20- to 25-minute didactic presentation followed by 1 patient case submitted by a participant before the session and discussed live for the remaining 35-40 minutes. Textbox 1 outlines the didactic presentation topics for each session. A multidisciplinary core faculty facilitated each session and was comprised of 1 physician with dual board certification in family medicine and addiction medicine and experienced in treating SUDs at federally qualified health centers; 1 nurse practitioner who developed and leads a federally qualified health center medication-assisted treatment (MAT) program; 1 supervisory licensed counselor; and 1 population health expert. Together, the faculty built a 12-month curriculum covering diverse topics such as medication management, team-based care, trauma-informed care, stages of change and motivational interviewing, polysubstance use and co-occurring conditions, and coordinating levels of care.

XSL•FO
RenderX

Textbox 1. Weitzman Extension for Community Healthcare Outcomes: Comprehensive Substance Use Disorder Care didactic topic by session.

Session and didactic topic

1. Philosophy of care (no case presentation).
2. Harm reduction strategies.
3. Principals of medication management.
4. Team-based care: care provision partners.
5. Trauma-informed care: an overview.
6. Motivational interviewing.
7. Stages of change for addiction.
8. Assessing stages of change and stage-based interventions.
9. Medications for opioid use disorder basics.
10. Behavioral health and primary care coordination.
11. Transitions of care.
12. Polysubstance use.
13. Social determinants of health including barriers or challenges (no case presentation).
14. Adverse childhood experiences.
15. Legal factors and access.
16. Mental health crisis and coordination of care.
17. Medication-assisted treatment for alcohol and tobacco use disorders.
18. Self-determination and strength-based approaches.
19. Contingency management for substance use disorder.
20. HIV and hepatitis C virus in patients with substance use disorder.
21. Screening, brief intervention and referral to treatment into primary care.
22. Stimulant use disorder treatment and medication management.
23. Co-occurring mental health substance use disorder.
24. Tobacco cessation for polysubstance patients.

Participants were recruited by email blasts targeted to each partner's network of Ohio primary care providers and other members of the care team. A total of 109 participants attended at least one session, 16 participants attended between 7 and 11 sessions, and 23 participants attended over 12 (half) of the sessions. On average, there were 32 attendees at each session. Continuing education credits were offered to medical providers, behavioral health providers, and nurses.

Purpose of Study

The purpose of this study was to assess the ability of CSUDC ECHO to both address and meet 7 learning objectives (Textbox 2) and address multiple substances by analyzing (1) the frequency of exposure to the learning objective topics and substance types during case discussions and (2) participants' knowledge, self-efficacy, skills, and attitudes related to the treatment of SUDs pre- to postprogram.

Textbox 2. Weitzman Extension for Community Healthcare Outcomes: Comprehensive Substance Use Disorder Care learning objectives.

- Project a harm reduction philosophy of care into your treatment of patients experiencing substance use disorders and explain this concept to peers.
- Use the care team more effectively to improve the management of patients experiencing substance use disorders.
- Use motivational interviewing and other behavioral techniques to improve patient outcomes related to substance use disorders.
- Better differentiate and implement medication management strategies for patients experiencing substance use disorders.
- Illustrate trauma-informed practices in the screening, assessment, and treatment of patients experiencing substance use disorders.
- Describe and manage common co-occurring conditions and polysubstance use more effectively in patients experiencing substance use disorders.
- Distinguish and address factors related to social determinants of health faced by specific populations experiencing substance use disorders.

XSL•FO
RenderX

Methods

Study Design and Data Collection

This study used a mixed methods design, using a conceptual content analysis [6] analyzing ECHO participant-led case presentations, as well as a 2-tailed paired-samples *t* test of participant self-reported learner outcomes. All ECHO attendees who registered and attended the Project ECHO CSUDC sessions are included in the deductive content analysis. All ECHO attendees who registered before and through the first session of the series were invited to complete a preseries survey (n=106) via Qualtrics survey software (Qualtrics). The preseries survey remained open for 3 weeks from June 25, 2021, to July 18, 2021. A total of 79 responses were received (n=79) for a response rate of 75%. Upon completion of the ECHO series, active attendees (ie, those that were still active at the conclusion of the series and did not officially drop from the series, as well as those who enrolled throughout the series) were invited to complete a postseries survey via Qualtrics Survey Software (n=90). The postseries survey remained open for 4 weeks from July 7, 2022, to August 2, 2022. A total of 25 responses were received (n=25) for a response rate of 28%. A total of 16 consented participants completed both the preseries and postseries surveys (n=16) and are included in the paired-samples *t* tests statistical analysis.

Ethical Considerations

This study was approved by the Community Health Center, Inc, Institutional Review Board (IRB; 1190) on January 6, 2022. Informed consent was accounted for by the authors through the administration of a consent form on the postseries survey gathering participant consent to use their deidentified survey data for the paired-samples *t* test analysis. The deductive content analysis was considered a secondary analysis and was given exempt status. All data used in this study were deidentified, accounting for privacy and confidentially. No compensation for participation in this study was deemed necessary by the IRB.

Survey Tools

The preseries and postseries surveys were internally created and based on the Consolidated Framework for Implementation Research (CFIR) [7] and Moore's Model of Outcomes Assessment Framework [8]. The specific CFIR domains assessed for include intervention characteristics, outer setting, inner setting, characteristics of individuals, and process [7]. Additionally, the levels of Moore's Model of Outcomes Assessment Framework assessed for include level 2 (satisfaction), level 3a (declarative knowledge), level 3b (procedural knowledge), level 4 (competence), level 5 (performance), and level 6 (patient health) [8]. The surveys assessed changes in participants' self-reported knowledge, attitudes, self-efficacy, and skills through statements centered on the series' learning objectives. The preseries survey also collected participant characteristics including provider type and years of experience working with patients diagnosed with SUDs, as well as team-based care practices. Additionally, the postseries survey collected information on engagement and practice changes. The preseries survey instrument is presented in Multimedia Appendix 1 and the postseries survey instrument is presented in Multimedia Appendix 2.

While the preseries survey and postseries survey tools were based on CFIR [7] and Moore's Model of Outcomes Assessment Framework [8], both surveys were internally designed. The internal research and evaluation and CSUDC ECHO programmatic teams created the survey tools through several iterations of the internal review, which also consisted of selecting the appropriate domain (ie, knowledge, attitudes, self-efficacy, and skills) to assess each series' learning objective. Each domain used a 5-point Likert scale to assess responses. The surveys were then presented to the CSUDC ECHO series stakeholders and faculty for review and approval before administering the surveys to the ECHO attendees. See Multimedia Appendices 1 and 2 for the domain placement of learning objectives and the 5-point Likert scales.

Conceptual Content Analysis

To further evaluate Weitzman ECHO CSUDC aims, researchers conducted a conceptual content analysis [6] using a set of a priori themes extracted from the series' learning objectives. Series' learning objectives are detailed in Textbox 2. To establish a priori themes, researchers met before the launch of the ECHO to examine the series' 7 learning objectives and extracted 7 themes for the content analysis. The themes were: harm reduction, team-based care, behavioral techniques, MAT, trauma-informed care, co-occurring conditions, and social determinants of health. To assess the frequency to which multiple substances were discussed, the themes also included 5 illicit and nonillicit substances of concern: alcohol, stimulants, opioids, cannabis, tobacco, or nicotine, plus polysubstance use when any 2 or more of these substances were identified. A conceptual analysis approach was used to gauge the dose and frequency of all learning objective themes and selected illicit and nonillicit substances. The content analysis aimed to confirm the discussion of the series' learning objectives during case presentations and to determine to what extent multiple substances were able to be addressed.

Researchers evaluated all 22 participant-led ECHO case presentations and discussions for the presence of the selected themes in the prepared participant cases, faculty recommendations, and participant recommendations. Case presentations and discussions consisted of participants independently preparing a patient case to present and receive participant and faculty guidance for a patient treatment plan. Case presentations were recorded and transcribed using Zoom videoconferencing software (Zoom Video Communications, Inc). The transcriptions were then used for the conceptual content analysis.

To ensure coding accuracy, 4 researchers independently coded 27% (n=6) of the case presentations and met to reconcile discrepancies and better establish coding parameters. After reconciling discrepancies, 1 researcher coded the remaining 16 case presentations and discussion transcripts. The content analysis themes and descriptions are presented in Table 1.

XSL•FO
RenderX

Table 1. Conceptual content analysis themes and learning objectives.

Theme	Learning objective
Harm reduction	Project a person-centered philosophy of care into your treatment of patients experiencing substance use disorders and explain this concept to peers.
Team-based care	Use the care team more effectively to improve the management of patients experiencing substance use disorders.
Behavioral techniques	Use motivational interviewing and other behavioral techniques to improve patient outcomes related to substance use disorders.
Medication-assisted treatment	Differentiate and implement medication management strategies for patients experiencing substance use disorders.
Trauma-informed care	Illustrate trauma-informed practices in the screening, assessment, and treatment of patients experiencing substance use disorders.
Co-occurring conditions	Describe and manage common co-occurring conditions and polysubstance use more effectively in patients experiencing substance use disorders.
Social determinants of health	Distinguish and address factors related to social determinants of health faced by specific populations experiencing substance use disorders.

Paired-Samples t Test

To determine if Project ECHO CSUDC affected participant learner outcomes, researchers calculated mean scores reported on a Likert scale of 1 to 5 and conducted a paired-samples *t* test to compare pre- and postseries scores at a .05 significance level. The surveys consisted of matching statements assessing knowledge, self-efficacy, attitudes, and skills associated with the series' learning objectives. The data were assessed for normality and homogeneity of variance and the assumptions were met. The data analysis was conducted using SPSS Statistics for Windows (version 26.0; IBM Corp).

Results

Participant Characteristics

CSUDC ECHO participants were asked to report their role type on the preseries survey. Of the participants that responded to the survey items (n=79), a majority were other care team members (n=32; 41%) followed by behavioral health providers (n=30; 38%) and medical providers (n=16; 21%). Additionally, participants were asked to indicate their years of experience working with SUDs. Most participants had between 1 and 5 years of experience (n=23; 29%) followed by 6-10 years (n=15; 19%), 11-20 years (n=14; 18%), less than 1 year (n=13; 16%), 7 participants indicated they do not work directly with patients (n=7; 9%), 21-30 years (n=4; 5%), 31-40 years (n=2; 3%), and more than 40 years of experience (n=1; 1%). Full participant characteristics of the entire CSUDC ECHO attendees, excluding the paired-samples *t* test sample, the paired-samples *t* test sample only, and all combined CSUDC ECHO attendees are provided in Table 2.

The attendance data of participants included in the paired-samples *t* test analysis were analyzed. Further, 6 (n=6; 38%) of the paired-samples *t* test participants attended 1% (n=1) to 25% (n=6) of the 24 CSUDC ECHO sessions, 3 (n=3; 19%) attended 26% (n=7) to 49% (n=11) of the sessions, 4 (n=4; 25%) attended 50% (n=12) to 75% (n=18) of the sessions, and 3 (n=3; 19%) attended 76% (n=19) to 100% (n=24) of the sessions.

XSL•FO
RenderX

Table 2. Participant characteristics of all ECHO[a] participants and paired-samples *t* test analysis sample.

	CSUDC[b] ECHO attendees (excluding paired-samples *t* test participants; n=63)	Paired-samples *t* test participants (n=16)	All CSUDC ECHO attendees (n=79)
Role type, n (%)			
Medical providers	13 (21)	3 (19)	16 (20)
Behavioral health providers	22 (35)	8 (50)	30 (38)
Other care team members	27 (43)	5 (31)	32 (41)
Missing	1 (2)	0 (0)	1 (1)
Years of SUD[c] care experience, n (%)			
Less than 1	11 (17)	2 (13)	13 (16)
1-5	19 (30)	4 (25)	23 (29)
6-10	12 (19)	3 (19)	15 (19)
11-20	11 (17)	3 (19)	14 (18)
21-30	2 (3)	2 (13)	4 (5)
31-40	2 (3)	0 (0)	2 (3)
≥40	0 (0)	1 (6)	1 (1)
Does not work directly with patients	6 (10)	1 (6)	7 (9)

[a]ECHO: Extension for Community Healthcare Outcomes.

[b]CSUDC: Comprehensive Substance Use Disorder Care.

[c]SUD: substance use disorder.

Conceptual Content Analysis

The conceptual content analysis indicated that all of the a priori themes relating to the learning objectives resulted in high frequencies and doses, appearing in a majority of case presentations and discussions. Further, 3 themes appeared in 100% (n=22) of case presentations and discussions, including team-based care at a frequency of 156, followed by harm reduction at a frequency of 152, and social determinants of health at a frequency of 135. In total, 4 themes appeared in less than 100% (n=22) of case presentations and discussions, but above 81% (n=18), including co-occurring conditions with a frequency of 118 and appearing in 95% (n=21) of case presentations and discussions, followed by behavioral techniques at a frequency of 108 and appearing in 91% (n=20) of case presentations and discussions, MAT at a frequency of 89 and appearing in 86% (n=19) of case presentations and discussions, and trauma-informed care at a frequency of 79 and appearing in 82% (n=18) case presentations and discussions. Additionally, multiple substances were represented but at differing frequencies. The substance that resulted in the highest frequency and dose was alcohol at a frequency of 64 and appeared in 81% (n=18) of case presentations and discussions, followed by stimulants at a frequency of 55 and 77% (n=17) of case presentations and discussions, opioids at a frequency of 49 and 59% (n=13) of case presentations and discussions. Cannabis resulted with a frequency of 38 but appeared in 64% (n=14) of case presentations and discussions. Finally, tobacco and nicotine resulted in the lowest frequency at 11 and dose appearing in 27% (n=6) of case presentations and discussions. When evaluating polysubstance use, which was limited to the use of two or more of the listed substances, we found a dose of 95% (n=21) of case presentations and discussions. The frequency of polysubstance use was not included in the conceptual content analysis since it was not a learning objective theme and the emphasis of the conceptual content analysis was focused on the specific illicit and nonillicit substance types. The results of the conceptual content analysis are presented in Table 3.

XSL•FO
RenderX

Table 3. The results of frequency and percentage of case appearances (dose) of conceptual content analysis themes.

Theme	Frequency	Case appearances (dose), n (%)
Team-based care	156	22 (100)
Harm reduction	152	22 (100)
Social determinants of health	136	22 (100)
Co-occurring conditions	118	21 (95)
Behavioral techniques	108	20 (91)
MAT[a]	89	19 (86)
Trauma-informed care	79	18 (82)
Substance type: alcohol	64	18 (81)
Substance type: stimulant	55	17 (77)
Substance type: opioid	49	13 (59)
Substance type: cannabis	38	14 (64)
Substance type: tobacco and nicotine	11	6 (27)
Polysubstance use of substance types	—[b]	21 (95)

[a]MAT: medication-assisted treatment.

[b]—: not available.

Paired-Samples t Test

Knowledge

In total, 4 knowledge domain statements resulted in statistically significant increases: understanding polysubstance use in patients experiencing SUD (*P*=.02), understanding the approach colleagues in other disciplines use to address SUD (*P*=.02), knowledge of medication management strategies for nicotine use disorder (*P*=.03), and knowledge of medication management strategies for opioid use disorder (OUD; *P*=.003). Additionally, all knowledge domain statements resulted in an increased change in mean score from preseries to postseries. The results of the knowledge domain preseries and postseries scores are presented in Table 4.

Table 4. The results of the paired-samples *t* test for the knowledge domain.

Statement	Preseries mean score (SD; 1-5)	Postseries mean score (SD; 1-5)	Change in mean	*P* value
I understand polysubstance use in patients experiencing substance use disorders	3.63 (1.03)	4.25 (0.45)	+0.62	.02
I understand factors related to social determinants of health faced by specific populations experiencing substance use disorders	4.13 (0.89)	4.31 (0.60)	+0.18	.38
I understand the approach of my colleagues in other disciplines (ie, behavioral health if you are a medical provider) to substance use disorder care	3.69 (0.87)	4.25 (0.58)	+0.56	.02
Knowledge of the different medication management strategies for patients experiencing—nicotine use disorder	3.40 (1.12)	4.00 (0.76)	+0.60	.03
Knowledge of the different medication management strategies for patients experiencing—alcohol use disorder	3.53 (0.99)	4.00 (0.54)	+0.47	.07
Knowledge of the different medication management strategies for patients experiencing—stimulant use disorder	3.07 (0.10)	3.71 (0.83)	+0.64	.10
Knowledge of the different medication management strategies for patients experiencing—opioid use disorder	3.56 (0.96)	4.19 (0.83)	+0.63	.003

Attitudes

No attitudes domain statements resulted as statistically significant. All attitudes domain statements resulted in an increased change in mean score from preseries to postseries except the statement about a treatment plan for a patient experiencing an illicit SUD only being successful if abstinence is maintained, which resulted in a negative change in mean score. The negative change in mean score from preseries to postseries was the appropriate direction of change for alignment with promoting a harm reduction philosophy. The results of the attitudes domain preseries and postseries scores are presented in Table 5.

XSL•FO
RenderX

Table 5. The results of the paired-samples *t* test for the attitudes domain.

Statement	Preseries mean score (SD; 1-5)	Postseries mean score (SD; 1-5)	Change in mean	*P* value
It is important to practice a harm reduction philosophy when treating patients experiencing substance use disorders	4.60 (0.63)	4.73 (0.46)	+0.13	.43
Practicing a harm reduction philosophy in the treatment of patients experiencing substance use disorders leads to better patient outcomes	4.47 (0.83)	4.60 (0.63)	+0.13	.50
It is important to identify factors related to social determinants of health that patients experiencing substance use disorders may be facing	4.71 (0.61)	4.86 (0.36)	+0.15	.44
Addressing factors related to social determinants of health in the treatment of patients experiencing substance use disorders leads to better patient outcomes	4.69 (0.60)	4.81 (0.40)	+0.12	.50
A treatment plan for a patient experiencing an illicit substance use disorder has only been successful if abstinence is maintained	2.25 (1.39)	2.00 (1.27)	–0.25	.43

Self-Efficacy

In total, 2 self-efficacy statements resulted in statistically significant increases: choosing a medication management strategy for nicotine use disorder (*P*=.002) and alcohol use disorder (*P*=.02). Additionally, all self-efficacy domain statements resulted in an increased change in mean score from preseries to postseries. The results of the self-efficacy domain preseries and postseries scores are presented in Table 6.

Table 6. The results of the paired-samples *t* test for the self-efficacy domain.

Statement	Preseries mean (SD; 1-5)	Postseries mean (SD; 1-5)	Change in mean	*P* value
Providing trauma-informed care	3.20 (1.01)	3.67 (0.90)	+0.47	.15
Using motivational interviewing techniques	3.31 (1.08)	3.81 (0.83)	+0.50	.12
Creating SMART[a] goals with patients	3.43 (1.09)	3.79 (0.80)	+0.36	.29
Managing co-occurring conditions	3.40 (1.30)	3.93 (1.16)	+0.53	.16
Choosing an appropriate medication management strategy for—nicotine use disorder	2.75 (1.49)	3.88 (1.13)	+0.13	.002
Choosing an appropriate medication management strategy for—alcohol use disorder	2.63 (1.19)	3.75 (1.28)	+1.12	.02
Choosing an appropriate medication management strategy for—stimulant use disorder	1.86 (1.46)	2.00 (1.00)	+0.14	.79
Choosing an appropriate medication management strategy for—opioid use disorder	3.67 (1.41)	3.78 (1.48)	+0.11	.76

[a]SMART: specific, measurable, achievable, relevant, timely.

Skill

In total, 1 skill domain statement resulted in a statistically significant increase: using the stages of change theory to provide stage-based interventions to patients experiencing SUDs (*P*=.03). Additionally, all skill domain statements resulted in an increased change in mean score from preseries to postseries. The results of the skill domain preseries and postseries scores are presented in Table 7.

XSL•FO
RenderX

Table 7. The results of the paired-samples *t* test for the skill domain.

Statement	Preseries mean (SD; 1-5)	Postseries mean (SD; 1-5)	Change in mean	*P* value
Screening patients experiencing substance use disorders for trauma	3.53 (1.13)	4.07 (0.96)	+0.54	.06
Using the stages of change theory to provide stage-based interventions to patients experiencing substance use disorders	3.06 (1.29)	3.69 (0.87)	+0.63	.03
Collaborating with peer support specialists when working with patients experiencing substance use disorders	3.33 (1.29)	3.93 (1.03)	+0.60	.06
Referring patients to a higher level of care, such as IOP[a], if needed	3.79 (0.98)	4.36 (0.75)	+0.57	.06
Preventing drug overdose of my patients experiencing a substance use disorder	2.83 (1.27)	3.25 (0.97)	+0.42	.10

[a]IOP: intensive outpatient.

Discussion

Principal Findings

Ohio's annual average prevalence of tobacco use, heroin use, use of prescription pain relievers, OUDs, illicit drug use disorder, and SUD have been higher compared to both regional and national averages [2]. Considering the need to address this public health concern, CSUDC ECHO was implemented to train Ohio providers and care team members in substance use care. CSUDC ECHO enhanced the Project ECHO work in this field by focusing content and learning objectives on a comprehensive, team-based lens and a harm reduction philosophy of care to address multiple illicit and nonillicit substances including opioids, alcohol, nicotine, cannabis, and stimulants. To assess the ability of the CSUDC ECHO program to meet its 7 program learning objectives (Textbox 2) and address multiple substances, this study analyzed (1) the frequency of exposure to learning objective themes and substance types during case presentations and discussions and (2) participating providers' change in knowledge, attitudes, self-efficacy, and skills related to the treatment of SUDs.

Study results demonstrate that all 7 learning objectives were frequently addressed in the content of case presentations and discussions throughout the program, with team-based care being the most frequently mentioned, 3 objectives appearing in 100% (n=22) of case discussions (eg, team-based care, harm reduction, and co-occurring conditions), and all 7 objectives appearing in >81% (n=18) of all cases discussed. This may have resulted in the learner outcome improvement pre- to postprogram for multiple learner domains (eg, knowledge, self-efficacy, and skill) for the following themes: team-based care, MAT, polysubstance use, and behavioral techniques. No pattern emerged among the participants included in the paired-samples *t* test analysis exposure to didactic topics and changes in learner outcomes.

Alcohol, stimulants, opioids, cannabis, and nicotine were addressed in the content of case presentations and discussions throughout CSUDC ECHO with alcohol being the most frequently mentioned and most common substance appearing in cases, 4 substances appearing in >59% (n=13) of case discussions (eg, alcohol, stimulants, opioid, and cannabis), and all coded substances appearing in at least a quarter of cases. The dialogue about these substances during case discussions likely resulted in improvements to the following learner outcomes related to medication management: alcohol use disorder, OUD, and nicotine use disorder. Medication management of cannabis use disorder was not assessed in the pre- to postsurveys. Additionally, the didactic presentation topics that centered on alcohol, opioid, and nicotine use disorder resulted in a higher attendance rate with about 40% (n=6) to 50% (n=8) of the participants included in the paired-samples *t* test analysis attending the sessions, as compared to only 20% (n=3) of the aforementioned participant sample having attended the session centered on stimulant use disorder.

These findings indicate that the ECHO program's content aligned with its stated learning objectives; met its learning objectives for the 3 themes where significant improvements were measured; and met its intent to address multiple substances in case presentations and discussions. While case presentations and discussions comprise from half to the majority of content in the sessions (30-35 minutes of a 60-minute session), content during sessions also includes faculty didactic presentations (20-25 minutes), which also addresses these 7 learning objectives and various substances but was not a part of the content analysis. Therefore, learner outcome improvements may also be a result of content addressed in didactic presentations.

While the Project ECHO model has been shown to be effective in training the primary care workforce [9], specifically on OUD [10,11] and addiction medicine [12,13], there has been no documentation, to our knowledge, of the ability of a team-based, comprehensive SUD and polysubstance-focused Project ECHO designed to improve learner outcomes (eg, knowledge, self-efficacy, and skills). Although Komaromy and colleagues [14] investigated the frequency of cases presented based on substance type in a comprehensive SUD-focused ECHO, a content analysis of the case presentation and discussion transcripts was not analyzed to either assess the frequency of substances or learning objectives. Furthermore, to our knowledge, this process has not been combined in a mixed method approach to compare learner outcomes with a content analysis to gauge the ability of an SUD-focused Project ECHO program to meet its stated learning objectives. Our results reported here align with this literature and expand to demonstrate that Project ECHO is a potential tool to effectively educate

XSL•FO
RenderX

multidisciplinary providers in a comprehensive approach to SUD care.

Strengths

This study has several strengths which promote the ability of the Project ECHO model in enhancing health care providers' knowledge, self-efficacy, and skill associated with comprehensive SUD care. The focus of this study is unique as there is minimal research exploring the benefits and training ability of Project ECHO with a comprehensive SUD care focus. This study's noteworthy strength is the use of a mixed methods design that presents a comprehensive evaluation correlating the content addressed in the case presentations and discussions to statistically significant learner outcomes to demonstrate how this telementoring continuing education series improved provider's knowledge, skills, and self-efficacy to benefit participating providers and their practices.

Limitations

This study faced several limitations during data collection and analysis. The first limitation of this study was the limited sample size and low response rate. There was a decline between the number of participants who completed the preseries survey and postseries survey, resulting in a low comparative sample, which restricted the options for statistical analysis. Another limitation was generalizability; the results of this Project ECHO are limited to the target audience of medical providers, behavioral health providers, and care team members from the state of Ohio, which is not a representative sample of broader populations nationally. Additionally, participants self-selected to take part in the Project ECHO series, which presents the potential for self-selection bias. Another limitation this study faced was the lack of available or reliable data on Project ECHO and its ability to meet learning objectives and address multiple substances through providers' knowledge, self-efficacy, skill, and attitudes. Furthermore, self-reported data to assess knowledge and skills, and self-reported data in general, could present participant biases and is difficult to corroborate with outcomes. The use of internally designed survey instruments instead of using validated instruments presents as a limitation. In light of these limitations, future studies in this subject matter should include a larger data set. Additionally, future studies using a nested analysis approach might provide more insight into how the learning objective themes coincide with the various illicit and nonillicit substance types and would be a useful analysis to contribute to the knowledge base. Another recommendation for future studies in this subject matter should include a deeper analysis of attendance dose and exposure to didactic topics to better understand the impact on changes in learner outcomes. Future research with greater validity will contribute to the significant gaps in literature regarding this subject.

Conclusions

The purpose of this research study was to assess the ability of CSUDC ECHO to both address and meet 7 learning objectives (Textbox 2) and address multiple substances by analyzing (1) the frequency of exposure to the learning objective topics and substance types during case presentations and discussions and (2) participants' knowledge, self-efficacy, skills, and attitudes related to the treatment of SUDs from pre- to postprogram. The results of this study indicate that CSUDC ECHO was able to both address and meet its learning objectives while addressing multiple substances, as demonstrated by improvements in learner knowledge, self-efficacy, and skills. All learning objective themes resulted in high frequencies and doses, appearing in a majority of case presentations throughout the series. These promising results suggest that Project ECHO is a potential tool to educate primary care providers, behavioral health providers, and care team members in a comprehensive approach to SUD assessment and treatment through complex case discussions combined with didactic learning for certain settings. As Project ECHO programs continue to be established globally and existing programs strengthen, further research examining the model's ability to achieve positive learning outcomes and factors that may contribute to these outcomes (eg, frequency of topic dose) is needed to confirm the outcomes in larger population samples, additional topics of focus, and other geographical settings.

Acknowledgments

The authors would like to acknowledge our partners at Buckeye Health Plan and Ohio University Heritage College of Osteopathic Medicine. We would like to thank our funders, Centene Corporation through its subsidiary, Buckeye Health Plan; without their financial support, this work would not have been possible. We would like to thank the faculty that led the ECHO sessions, delivered didactic presentations, and provided high-quality case recommendations, including core faculty members Dana Vallangeon, doctor of medicine, Tracy Plouck, master of public administration, Amy Black, master of science in nursing, advanced practice registered nurse, nurse practitioner-certified, Ericka Ludwig, licensed professional clinical counselors applying for training supervision designation, licensed independent chemical dependency counselor, as well as guest faculty members. We would also like to thank our Weitzman Institute colleagues who helped with the content analysis: Zeba Kokan, Claire Newby, and Reilly Orner. To learn more about Weitzman Extension for Community Healthcare Outcomes programs, visit their website [15]. This project was supported by Buckeye Health Plan, a subsidiary of Centene Corporation. The views, opinions, and content expressed in this paper do not necessarily reflect the views, opinions, or policies of Buckeye Health Plan or Centene Corporation. The authors did not use generative artificial intelligence in any portion of this paper.

XSL•FO
RenderX

Data Availability

The data sets generated and analyzed during this study are not publicly available due to a portion of the data being deemed as exempt by the institutional review board and the institutional review board approving a waiver of informed consent for the exempt data, as well as the sensitive nature of the data, but are available from the corresponding author on reasonable request.

Authors' Contributions

MK wrote this paper, reviewed this paper, designed the evaluation plan, and performed the qualitative and statistical analyses. AP wrote this paper, reviewed this paper, and assisted with the evaluation design and approval. RM wrote this paper, reviewed this paper, performed the literature review, and assisted with the evaluation design and approval. NRN wrote this paper, reviewed this paper, and performed the literature review. KA critically reviewed this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Weitzman Extension for Community Healthcare Outcomes: Comprehensive Substance Use Disorder Care preseries survey instrument.

[DOCX File , 22 KB - mededu_v10i1e48135_app1.docx]

Multimedia Appendix 2

Weitzman Extension for Community Healthcare Outcomes: Comprehensive Substance Use Disorder Care postseries survey instrument.

[DOCX File , 19 KB - mededu_v10i1e48135_app2.docx]

References

1. Drug overdose mortality by state. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/nchs/pressroom/sosmap/drug_poisoning_mortality/drug_poisoning.htm [accessed 2023-01-30]
2. Behavioral health barometer: Ohio, Volume 6: indicators as measured through the 2019 national survey on drug use and health and the national survey of substance abuse treatment services. Substance Abuse and Mental Health Services Administration. 2020. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt32852/Ohio-BH-Barometer_Volume6.pdf [accessed 2023-01-30]
3. Arora S, Thornton K, Murata G, Deming P, Kalishman S, Dion D, et al. Outcomes of treatment for hepatitis C virus infection by primary care providers. N Engl J Med 2011;364(23):2199-2207 [FREE Full text] [doi: 10.1056/NEJMoa1009370] [Medline: 21631316]
4. Fact sheet: Biden-Harris administration announces new actions and funding to address the overdose epidemic and support recovery. The White House. 2022. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2022/09/23/fact-sheet-biden-harris-administration-announces-new-actions-and-funding-to-address-the-overdose-epidemic-and-support-recovery/ [accessed 2023-01-27]
5. Key substance use and mental health indicators in the United States: results from the 2019 national survey on drug use and health. Substance Abuse and Mental Health Services Administration. 2020. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt29393/2019NSDUHFFRPDFWHTML/2019NSDUHFFR1PDFW090120.pdf [accessed 2023-01-25]
6. Abroms LC, Padmanabhan N, Thaweethai L, Phillips T. iPhone apps for smoking cessation: a content analysis. Am J Prev Med 2011;40(3):279-285 [FREE Full text] [doi: 10.1016/j.amepre.2010.10.032] [Medline: 21335258]
7. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
8. Moore DE, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]
9. Zhou C, Crawford A, Serhal E, Kurdyak P, Sockalingam S. The impact of project ECHO on participant and patient outcomes: a systematic review. Acad Med 2016;91(10):1439-1461 [FREE Full text] [doi: 10.1097/ACM.0000000000001328] [Medline: 27489018]
10. Tofighi B, Isaacs N, Byrnes-Enoch H, Lakew R, Lee JD, Berry C, et al. Expanding treatment for opioid use disorder in publicly funded primary care clinics: exploratory evaluation of the NYC health + hospitals buprenorphine ECHO program. J Subst Abuse Treat 2019;106:1-3 [FREE Full text] [doi: 10.1016/j.jsat.2019.08.003] [Medline: 31540604]
11. Alford DP, Zisblatt L, Ng P, Hayes SM, Peloquin S, Hardesty I, et al. SCOPE of pain: an evaluation of an opioid risk evaluation and mitigation strategy continuing education program. Pain Med 2016;17(1):52-63 [FREE Full text] [doi: 10.1111/pme.12878] [Medline: 26304703]

XSL•FO
RenderX

12. Sagi MR, Aurobind G, Chand P, Ashfak A, Karthick C, Kubenthiran N, et al. Innovative telementoring for addiction management for remote primary care physicians: a feasibility study. Indian J Psychiatry 2018;60(4):461-466 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_211_18] [Medline: 30581211]
13. Englander H, Patten A, Lockard R, Muller M, Gregg J. Spreading addictions care across Oregon's rural and community hospitals: mixed-methods evaluation of an interprofessional telementoring ECHO program. J Gen Intern Med 2021;36(1):100-107 [FREE Full text] [doi: 10.1007/s11606-020-06175-5] [Medline: 32885371]
14. Komaromy M, Duhigg D, Metcalf A, Carlson C, Kalishman S, Hayes L, et al. Project ECHO (Extension for Community Healthcare Outcomes): A new model for educating primary care providers about treatment of substance use disorders. Subst Abus 2016;37(1):20-24 [FREE Full text] [doi: 10.1080/08897077.2015.1129388] [Medline: 26848803]
15. Weitzman Institute. URL: https://www.weitzmaninstitute.org/education/weitzman-echo/ [accessed 2024-03-12]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
CSUDC: Comprehensive Substance Use Disorder Care
ECHO: Extension for Community Healthcare Outcomes
IRB: institutional review board
MAT: medication-assisted treatment
OUD: opioid use disorder
SUD: substance use disorder
WI: Weitzman Institute

Edited by T de Azevedo Cardoso; submitted 12.04.23; peer-reviewed by A Arbabisarjou, J Ford Ii; comments to author 12.09.23; revised version received 06.11.23; accepted 29.02.24; published 01.04.24.

Please cite as:
Koester M, Motz R, Porto A, Reyes Nieves N, Ashley K
Using Project Extension for Community Healthcare Outcomes to Enhance Substance Use Disorder Care in Primary Care: Mixed Methods Study
JMIR Med Educ 2024;10:e48135
URL: https://mededu.jmir.org/2024/1/e48135
doi:10.2196/48135
PMID:38557477

©MacKenzie Koester, Rosemary Motz, Ariel Porto, Nikita Reyes Nieves, Karen Ashley. Originally published in JMIR Medical Education (https://mededu.jmir.org), 01.04.2024. 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.

XSL•FO
RenderX

Original Paper

Gender and Sexuality Awareness in Medical Education and Practice: Mixed Methods Study

Rola Khamisy-Farah[1], MD; Eden Biras[1], MD; Rabie Shehadeh[1], MD; Ruba Tuma[1,2], MD; Hisham Atwan[3], MD; Anna Siri[4,5], PhD; Manlio Converti[6], MD; Francesco Chirico[7], MD; Łukasz Szarpak[8,9], MD; Carlo Biz[10], MD, PhD; Raymond Farah[1], MD; Nicola Bragazzi[11,12], MPH, MD, PhD

[1]Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel
[2]Department of Obstetrics and Gynecology, Galilee Medical Center, Nahariya, Israel
[3]Department of Internal Medicine, Kaplan Medical Centre, Hebrew University, Rehovot, Israel
[4]United Nations Educational, Scientific and Cultural Organization, Chair Anthropology of Health, Biosphere and healing systems, University of Genoa, Genoa, Italy
[5]Department of Wellbeing, Nutrition and Sport, Pegaso University, Naples, Italy
[6]ASL Napoli 2 Nord, Naples, Italy
[7]Post-Graduate School of Occupational Health, Università Cattolica del Sacro Cuore, Rome, Italy
[8]Henry JN Taub Department of Emergency Medicine, Baylor College of Medicine, Houston, TX, United States
[9]Institute of Outcomes Research, Maria Sklodowska-Curie Medical Academy, Warsaw, Poland
[10]Orthopedics and Orthopedic Oncology, Department of Surgery, Oncology and Gastroenterology (DiSCOG), University of Padua, Padua, Italy
[11]Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, ON, Canada
[12]Department of Food & Drug, University of Parma, Parma, Italy

Corresponding Author:
Nicola Bragazzi, MPH, MD, PhD
Laboratory for Industrial and Applied Mathematics
Department of Mathematics and Statistics
York University
4700 Keele Street
Toronto, ON, M3J 1P3
Canada
Phone: 1 416 736 2100
Email: robertobragazzi@gmail.com

Abstract

Background: The integration of gender and sexuality awareness in health care is increasingly recognized as vital for patient outcomes. Despite this, there is a notable lack of comprehensive data on the current state of physicians' training and perceptions in these areas, leading to a gap in targeted educational interventions and optimal health care delivery.

Objective: The study's aim was to explore the experiences and perceptions of attending and resident physicians regarding the inclusion of gender and sexuality content in medical school curricula and professional practice in Israel.

Methods: This cross-sectional survey targeted a diverse group of physicians across various specializations and experience levels. Distributed through Israeli Medical Associations and professional networks, it included sections on experiences with gender and sexuality content, perceptions of knowledge, the impact of medical school curricula on professional capabilities, and views on integrating gender medicine in medical education. Descriptive and correlational analyses, along with gender-based and medical status-based comparisons, were used, complemented, and enhanced by qualitative analysis of participants' replies.

Results: The survey, encompassing 189 respondents, revealed low-to-moderate exposure to gender and sexuality content in medical school curricula, with a similar perception of preparedness. A need for more comprehensive training was widely recognized. The majority valued training in these areas for enhancing professional capabilities, identifying 10 essential gender-related knowledge areas. The preference for integrating gender medicine throughout medical education was significant. Gender-based analysis indicated variations in exposure and perceptions.

Conclusions: The study highlights a crucial need for the inclusion of gender and sexuality awareness in medical education and practice. It suggests the necessity for curriculum development, targeted training programs, policy advocacy, mentorship initiatives,

XSL•FO
RenderX

and research to evaluate the effectiveness of these interventions. The findings serve as a foundation for future directions in medical education, aiming for a more inclusive, aware, and prepared medical workforce.

(JMIR Med Educ 2024;10:e59009) doi:10.2196/59009

KEYWORDS

gender medicine; medical education; clinical practice; gender-sensitive care; gender awareness; sexuality awareness; awareness; medical education and practice; healthcare; patient outcomes; patient; patients; medical professionals; training; educational interventions; status-based; survey; effectiveness; medical workforce

Introduction

The contemporary health care landscape is undergoing a significant transformation, with a growing recognition of the importance of integrating gender and sexuality awareness into both medical education and clinical practice [1-5]. This shift in perspective acknowledges that gender and sexuality are not just marginal issues but, on the contrary, are central determinants of health outcomes [6], influencing patient care in complex and diverse ways, affecting various aspects, from the prevalence and presentation of diseases to treatment responses and patient interactions [1,6].

Despite this growing awareness, there remains a significant gap in our understanding of how well attending and resident physicians are trained in these areas [7-10]. This includes a lack of comprehensive data on the depth and breadth of their knowledge, the extent of their exposure to gender and sexuality issues during their training, and their perceptions and attitudes toward these crucial aspects of patient care [1,9]. This paucity of information is problematic because it hinders the ability of medical education institutions and health care organizations to develop targeted educational interventions [9].

Without a clear understanding and an updated picture of the current state of medical education, training, and knowledge, it becomes challenging to craft effective strategies to enhance the competencies and skills of health care providers in dealing with gender- and sexuality-related health issues [10].

Our study aims to fill this critical gap by exploring the experiences and perceptions of attending and resident physicians regarding the inclusion of gender and sexuality content in their education and subsequent professional practice. We intend to paint a clearer snapshot of the current state of awareness and understanding in the medical community. Our objective is to identify not only the strengths but also the potential areas for improvement in medical education regarding gender and sexuality. This will enable us to contribute valuable insights to the ongoing discourse on personalized and gender-sensitive health care. In doing so, we seek to influence the future direction of medical education and practice, steering it toward a more inclusive, aware, and responsive model that takes into account the diverse needs of patients. This, in turn, is expected to lead to more effective, personalized patient care, better health outcomes, and a health care system that is more attuned to the complexities of human diversity.

Methods

Survey Design and Participants

This study used a mixed methods, cross-sectional survey design. We targeted a diverse group of Israeli attending and resident physicians, encompassing various specializations, professional statuses, and levels of experience. The survey was distributed through multiple Israeli Medical Associations and professional networks to ensure a wide reach.

Demographics

The demographic section of the survey covered age, sex and gender, medical specialization, professional status (attending or resident physician), and years of experience in the medical field. This information provided a sociodemographic context for the subsequent analysis.

Survey Content

The survey was devised based on a previous systematic review of the literature [1] and consisted of several sections, each focusing on different aspects of gender and sexuality in medical education and practice.

Experiences With Gender and Sexuality Content

This section assessed the respondents’ exposure to and preparedness in gender and sexuality topics during their attendance in medical schools and residency training. More specifically, participants were asked whether medical school curricula and residency programs included content related to gender and sexuality. They were then asked to rate their levels of exposure to gender and sexuality content during their academic studies and residency on a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

Perceptions of Knowledge and Tools

Respondents rated their current levels of knowledge and the adequacy of tools available to address gender and sexuality issues in their professional practice on a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

Impact of Medical School Curricula and Residency Programs on Professional Capabilities

This part evaluated the perceived impact of gender and sexuality training on enhancing professional capabilities and identified essential areas of knowledge in this domain. Participants were asked to what extent they felt they lacked training in the field of gender and sexuality, and to what extent they believed training in these areas would contribute to their professional

XSL•FO
RenderX

capabilities, using a Likert scale from 1 (no contribution) to 5 (very great extent).

Integration of Gender Medicine in Medical Education

Respondents shared their views on when and how gender medicine should be incorporated into medical education, with options including preclinical years, clinical years, both preclinical and clinical years, or not at all.

Integration of Gender Medicine in Clinical Practice

In this section, participants were asked whether they considered the patient's sex and gender when choosing drug treatments and whether they considered the effects of treatment on the patient's life course in relation to sex and gender, both rated on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). They were also asked if they had observed differences in the presentation and nature of symptoms based on the patient's sex and gender and whether they believed there is a distinction in treating the LGBTQI (lesbian, gay, bisexual, transgender, queer/questioning, and intersex) population in terms of common conditions and emotional impacts.

Finally, the survey inquired if participants had mentored students in the past year, with a simple yes or no response.

This comprehensive survey aimed to gather quantitative detailed information on physicians' experiences and perceptions regarding gender and sexuality content in their education and professional practice, highlighting areas for potential improvement in medical training.

Statistical Analysis

Descriptive statistics were used to summarize the sociodemographic data. We used correlational analysis to explore relationships between different aspects of gender and sexuality awareness and training. A gender-based analysis was conducted to discern any differences in responses based on sex and gender. In addition, we compared responses between attending physicians and residents to identify variations based on professional status. Multivariate analyses were performed to uncover associations between survey responses and demographic variables such as age, sex and gender, years of experience, and medical specialization. Effect sizes were computed as Cohen *d* and odds ratios. For all analyses, a significance level of .05 was used as the statistical threshold. To control for the increased risk of type I error due to multiple comparisons, the Bonferroni correction was applied where necessary. All statistical analyses were conducted in the open-source R environment (R Foundation for Statistical Computing).

Qualitative Analysis

In addition to the quantitative survey, qualitative data were collected through the inclusion of 2 open-ended items within the survey itself. These items aimed to gain deeper insights into the experiences and perceptions of attending and resident physicians regarding essential gender knowledge for medical education. More specifically, the 2 open-ended items included in the survey asked the participants to select terms or concepts related to gender knowledge that they believed are essential for medical students to study.

Thematic analysis was used to analyze the qualitative data. The process involved steps, such as familiarization, coding, theme development, defining themes, and reporting. First, responses were read multiple times to become familiar with the data. Initial codes were generated by systematically identifying key terms and concepts mentioned by respondents. Codes were then grouped into potential themes based on commonalities and patterns in the responses. Each theme was defined and named, providing a detailed analysis of its significance in the context of gender and sexuality education. The final themes were integrated into a coherent narrative, illustrating the respondents' views on essential gender knowledge for medical students.

Ethical Considerations

The study was conducted in compliance with the ethical standards for research involving human participants. They were informed about the purpose of the study, and participation was voluntary. Informed, written consent was collected before the commencement of the study. Anonymity and confidentiality of responses were maintained throughout the study.

Results

Sociodemographic Data

The survey data encompassed 189 respondents with an average age of 39.8 (SD 12.1, range 22-70, median 36; Table 1). Gender distribution showed a majority of women, totaling 57.1% (108/189) of participants. In terms of medical specialization, the respondents represented 27 different fields, with internal medicine being the most common, reported by 37% (70/189) of participants. Regarding professional status, the sample was almost evenly split between attending physicians and residents, with the former category being slightly more common at 104 out of 189 (55%) instances. Concerning their tenure in the medical field across 5 distinct experience categories (0-5 years, 6-10 years, 11-15 years, 16-20 years, and more than 20 years), the survey reflected a less experienced demographic, with the 0-5 years category reported most frequently by 86 out of 189 (45.5%) respondents, thus highlighting a considerable representation of early-career physicians within the surveyed population.

XSL•FO
RenderX

Table 1. Demographics of the survey's sample.

Demographic variable	Details
Sample size	189 respondents
Age	39.8 (range 22-70, median 36) years
Gender distribution	Women: 108 (57.1%) respondents
Medical specialization	27 different medical fields; most common: Internal medicine (37%)
Professional status	Attending physicians: 104 (55%) respondents; residents: 85 (45%) respondents
Years of experience	86 (45.5%) respondents with 0-5 years of experience
Mentorship	Active involvement in student mentorship reported by 65.1% of respondents

Respondents' Experiences With Gender and Sexuality Content in Their Education and Professional Training

The average level of exposure to gender and sexuality content during academic studies was rated at 2.03 (SD 0.98), suggesting a low-to-moderate exposure among participants. Respondents rated their academic program's preparedness in imparting gender and sexuality awareness at an average of 1.99 (SD 1.06), indicating a similar low-to-moderate perception of preparedness. The readiness provided by specialization or residency programs had a slightly higher average rating of 2.18 (SD 1.10). Regarding current knowledge and tools to address gender and sexuality issues, the average rating was 2.74 (SD 0.96). Finally, the extent of perceived lack of training in gender and sexuality fields averaged at 3.26 (SD 1.16), suggesting that respondents generally felt a need for more training in these areas.

On average, respondents rated 2.84 (SD 1.20) on the importance of considering the patient's sex and gender when choosing drug treatment, indicating moderate agreement and some variability in responses. When it comes to accounting for the effects of treatment on a patient's life course in relation to sex and gender, the mean rating was higher at 3.39 (SD 1.17), suggesting a generally higher agreement on this consideration. Observations of differences in symptom presentation based on the patient's sex and gender had a mean rating of 2.97 (SD 1.12), reflecting that the respondents somewhat agreed that they noticed such differences.

Respondents' Assessment of the Impact of Training in Gender and Sexuality on Professional Capabilities

The majority of respondents valued such training highly. Approximately 47.1% (89/189) believed it can contribute to a great extent, while 20.1% (38/189) felt it can contribute to a very great extent. A further 16.4% (31/189) saw it as moderately impactful, whereas only 9% (17/189) considered its potential contribution small. Notably, 6.3% (12/189) perceived no contribution from this training. A small fraction of respondents (2/189, 1.1%) had mixed views. Overall, these findings indicate a strong consensus on the positive impact of gender and sexuality training in enhancing professional capabilities (Table 2).

Table 2. Major findings of the survey.

Major finding	Details
Exposure to gender and sexuality content	Low-to-moderate exposure during academic studies (Average rating: 2.03 out of a Likert scale from 1 to 5)
Preparedness to gender-sensitive care	Low-to-moderate perception of preparedness (Academic: 1.99; Specialization/Residency: 2.18, out of a Likert scale from 1 to 5)
Perceived need for training	General consensus on the need for more training (Average lack of training rating: 3.26, out of a Likert scale from 1 to 5)
Impact on professional capabilities	Majority see training as beneficial (47.1%=great extent; 20.1%=very great extent; 16.4%=moderate extent)
Essential gender-related knowledge areas	Ten areas identified, including patriarchy, LGBTQI[a] awareness, gender awareness, sexual and domestic violence, gender-specific diseases and symptoms, pharmacology and gender differences, treatment compliance and gender, psychological and social effects of gender, and sex and gender-aware research; Table 3 for further details
Preference for integration of gender medicine	Majority prefer integration throughout medical education (Preclinical and clinical years: 55.6%)

[a]LGBTQI: lesbian, gay, bisexual, transgender, queer/questioning, and intersex.

XSL•FO
RenderX

According to the respondents, this training should cover 10 essential gender-related knowledge areas, as reported in Table 3.

Table 3. Ten essential gender-related knowledge areas that should be covered by the training, according to survey's participants.

Gender-related knowledge area	Brief description
Patriarchy	Understanding the social organization where power is primarily held by men and its impact on health care access and treatment outcomes, which is crucial to recognize how patriarchal structures can affect both patient care and the work environment in health care settings
LGBTQI[a] awareness	Knowledge about the health needs and challenges faced by the LGBTQI community, which includes understanding diverse sexual orientations and gender identities, and how these factors influence health risks, disease prevalence, and access to health care
Gender awareness	Recognizing and addressing gender biases and stereotypes in health care, which involves understanding how societal gender roles and expectations can impact health and health care delivery
Sexual violence	Awareness of the medical, psychological, and social implications of sexual violence, which includes understanding how to provide sensitive and appropriate care to survivors
Domestic violence	Recognizing signs of domestic violence and understanding its health implications, which also involves knowing how to provide support and resources to survivors
Gender-specific diseases and symptoms	Understanding differences in disease presentation, symptom onset, and diagnosis between sexes and genders, which is essential for accurate diagnosis and effective treatment
Pharmacology and gender differences	Acknowledging how drugs may affect sexes and genders differently in terms of efficacy, side effects, and treatment response, which is vital for personalized medicine
Treatment compliance and gender	Recognizing that gender can influence treatment adherence and response, with factors such as societal roles, communication styles, and access to health care varying between genders and impacting treatment outcomes
Psychological and social effects of gender	Understanding the broader psychological and social implications of gender on health, which includes the impact of gender roles, expectations, and discrimination on mental health and social well-being
Sex and gender-aware research	Promoting and using research that takes into account sex and gender differences, ensuring that medical knowledge and practice are based on inclusive and comprehensive data

[a]LGBTQI: lesbian, gay, bisexual, transgender, queer/questioning, and intersex.

According to the respondents, these topics provide a broad and nuanced understanding of how sex and gender affect health and health care, equipping medical students to deliver more compassionate, informed, and effective care to all patients, regardless of their sex and gender.

Finally, respondents had various opinions on when gender medicine should be incorporated into medical education. The majority believed it should be taught during both preclinical and clinical study years, with 55.6% (105/189) respondents endorsing this approach. Furthermore, 50 respondents, out of 189, felt it should be specifically included in the clinical study years (26.5%), and 23 participants argued for its introduction in the preclinical years (12.2%). A minority of 3.7% (7/189) subjects believed there was little need to teach gender medicine. There are also a few isolated responses that combine these categories or indicate no need at all for such education, each with 1 respondent. Overall, this distribution indicates a strong preference for integrating gender medicine throughout the entire span of medical education, with a significant emphasis on its presence in both foundational and advanced stages of the medical curriculum.

Among the 189 survey participants, 123 respondents indicated that they have been mentoring students in the past year (65.1%), while 66 respondents have not engaged in student mentorship during that time (34.9%). This suggests a significant portion of the respondents are actively involved in the mentorship and educational development of students.

Correlations and Trends: Insights From Correlational Analysis

There was a strong correlation (r=0.70) between respondents' perceptions of their academic program's preparation in terms of gender and sexuality awareness and their views on the preparation provided by their specialization and residency program. Furthermore, respondents' levels of exposure to gender and sexuality content in their academic program strongly correlated (r=0.68) with their perception of how well the program prepared them in these areas. There was a moderate

XSL•FO
RenderX

correlation (r=0.48) between how well respondents feel their specialization and residency program prepared them and their current perception of having sufficient knowledge and tools to deal with gender and sexuality issues in their field. Furthermore, respondents who felt their academic program prepared them well in gender and sexuality awareness also tend to feel they currently have sufficient knowledge and tools in this area, with a moderate correlation (r=0.40).

Gender-Based Analysis

The analysis based on gender reveals null-to-small effect sizes, with only 1 medium effect size concerning the sex- and gender-specific choice of a treatment. Men reported a higher level of exposure to gender and sexuality content during their academic studies (mean 2.23, SD 1.06) compared with women (mean 1.87, SD 0.89; d=0.38). Similarly, they also rated their academic study program's preparation in gender and sexuality awareness higher (mean 2.22, SD 1.15 for men vs mean 1.82, SD 0.97 for women; d=0.38). In terms of how well respondents felt their specialization and residency program prepared them in gender and sexuality awareness, men's responses were only slightly higher (mean 2.28, SD 1.15) than women's (mean 2.10, SD 1.05; d=0.17; not statistically significant). When asked if they currently have sufficient knowledge and tools to deal with issues of gender and sexuality in their field, men's responses were on average comparable (mean 2.79, SD 1.02) with those of women (mean 2.70, SD 0.92; d=0.09; not statistically significant). Regarding the extent to which they feel they lack training in the field of gender and sexuality, men had a lower average (mean 3.10, SD 1.21) than women (mean 3.38, SD 1.12; d=–0.24), suggesting that women perceive a greater need for training in these areas. Regarding the consideration of patient's sex and gender in drug treatment, men reported an average score of 3.10 (SD 1.22), higher than those reported by women (mean 2.64, SD 1.16), with d=0.39. When choosing the treatment, men were more likely to take into account its effects on the patient's life course in relation to sex and gender (mean 3.85, SD 1.02 vs mean 3.07, SD 1.16; d=0.71). When questioned about the observation of differences in the presentation and nature of symptoms based on patient's sex and gender, the responses from both men and women participants were similar (mean 2.99, SD 1.18 vs mean 2.96, SD 1.08; d=0.02). Similarly, recognition of the unique health care needs and challenges faced by LGBTQI individuals did not differ between men and women (mean 3.11, SD 1.31 vs mean 3.21, SD 1.22; d=–0.08).

In terms of subjects mentioned by respondents, men were more likely to mention certain topics like "domestic violence," "homophobia," "LGBTQI awareness," and "gender awareness." Other subjects such as differences in symptom onset and diagnosis of gender-specific diseases, topics related to pharmacology, treatment compliance, gender aspects of heart health, disease prevention, psychological and social effects, and feminism were more common among men respondents. Among women, respondents' topics covered more medically gender-related subjects, such as sexually transmitted infections, sexual education, and safer sex, among others.

To get more insights about gender-specific differences in the responses, further gender-based analyses were conducted. No gender-specific differences could be found in terms of age (mean 41.0, SD 14.3 vs mean 38.9, SD 10.1; P=.23) and the status of the respondent, that is, attending versus resident (χ^2_1=0.02, P=.90). On the contrary, there were some gender imbalances concerning the different medical specializations of the respondents (χ^2_{26}=36.76, P=.08). In particular, in the field of pediatrics all subjects were practically women (10 vs 1; P=.03 at the post hoc test), and in the field of internal medicine, men were overrepresented compared with women (42 vs 28; P<.001 at the post hoc test). In terms of years of practice and experience, some slight gender imbalances could be noted (χ^2_4=6.94, P=.14), with men being overrepresented in the category "more than 20 years" (24 vs 16; P=.02 at the post hoc test). Finally, women were more likely to report not having mentored students in the last year (χ^2_1=3.76, P=.05, with an odds ratio of 1.84 [95% CI 0.99-3.43]).

Medical Status-Based Analysis

The comparison between attending physicians and residents yields the following insights. On average, attending physicians reported a lower level of exposure to gender and sexuality content during academic studies (mean 1.88) compared with residents (mean 2.21). Similarly, they rated their academic study program's preparation in gender and sexuality awareness lower (mean 1.86 for attending doctors vs 2.16 for residents). Attending physicians also rated the preparation provided by their specialization and residency program slightly higher (mean 2.21) than residents did (mean 2.14). In assessing whether they have sufficient knowledge and tools to deal with issues of gender and sexuality, attending physicians' average response was higher (mean 2.87) compared with residents (mean 2.59). When it comes to the extent of lacking training in gender and sexuality, attending physicians feel slightly less deficient (mean 3.20) than residents (mean 3.33), indicating that residents may perceive a greater need for training in these areas.

The comparison among the different medical specializations revealed that respondents in the field of internal medicine perceived themselves as relatively well-prepared or exposed to gender and sexuality topics.

Multivariate Analysis

At the multivariate analysis, the reported level of exposure to gender and sexuality content during academic studies was associated with gender ($F_{2,186}$=8.89, P=.003), with women reporting lower exposure than men (β=–.46, 95% CI –0.77 to –0.16). Similarly, perceived academic preparedness in terms of gender and sexuality awareness was found to be associated with gender ($F_{2,186}$=7.33, P=.007), with women scoring lower than men (β=–.43, 95% CI –0.74 to –0.12), while thinking of currently having sufficient knowledge and tools to deal with issues of gender and sexuality in one's field was associated with years of experience in a statistically significant way ($F_{5,183}$=2.48, P=.045 at the ANOVA omnibus test). In particular, the category "over 20 years" versus "0-5 years" was more likely to report a higher score (β=1.30, 95% CI 0.27-2.33; P=.014). The perceived lack of training in the field of gender and sexuality was found to be associated with medical status ($F_{2,186}$=4.06, P=.045 at the

XSL•FO
RenderX

ANOVA omnibus test), with residents scoring higher than attending doctors (β=.54, 95% CI 0.01-1.08). The perceived impact of training in gender and sexuality on professional skills was once again associated with gender ($F_{2,186}$=4.89, *P*=.028), with women reporting greater perceived impact (β=.34, 95% CI 0.04-0.65) than men. Accounting for the person's sex and gender in the choice of the treatment was associated with the gender ($F_{2,186}$=5.26, *P*=.023), with women reporting this practice less (β=–.37, 95% CI –0.68 to –0.05) than men. When choosing the treatment, taking into account its effects on the patient's life course in relation to sex and gender was associated with gender ($F_{2,186}$=17.12, *P*<.001), medical specialization ($F_{27,161}$=19.62, *P*<.001), and increasing years of practice and experience ($F_{5,183}$=2.21, *P*=.07). This practice was less reported by women (β=–.54, 95% CI –0.79 to –0.28), and doctors non specialist in internal medicine (β=–.74, 95% CI –1.07 to –0.41). No significant predictors could be found for the other items of the questionnaire.

Qualitative Analysis

Participants highlighted 10 essential gender-related knowledge areas that should be covered by training, as identified by survey participants. First, understanding patriarchy is crucial for recognizing how power dynamics, predominantly controlled by men, can impact health care access and treatment outcomes. This knowledge helps in identifying the influence of patriarchal structures on both patient care and the work environment in health care settings. Awareness of LGBTQI health needs is also essential, encompassing knowledge about diverse sexual orientations and gender identities, and their influence on health risks, disease prevalence, and access to health care. Recognizing and addressing gender biases and stereotypes in health care, known as gender awareness, involves understanding how societal gender roles and expectations affect health and health care delivery. Awareness of sexual violence includes understanding its medical, psychological, and social implications to provide sensitive and appropriate care to survivors. Similarly, recognizing signs of domestic violence and understanding its health implications is vital, along with knowing how to provide support and resources to survivors. In addition, understanding gender-specific diseases and symptoms is essential for accurate diagnosis and effective treatment, as is acknowledging how drugs may affect sexes and genders differently in terms of efficacy, side effects, and treatment response. Recognizing that gender can influence treatment adherence and response is important, with factors such as societal roles, communication styles, and access to health care varying between genders. Understanding the broader psychological and social effects of gender on health includes considering the impact of gender roles, expectations, and discrimination on mental health and social well-being. Finally, promoting and using sex- and gender-aware research ensures that medical knowledge and practice are based on inclusive and comprehensive data, leading to improved health care outcomes for all sexes and genders.

Discussion

Principal Findings

This survey offered a rich and nuanced view of physicians' experiences and perceptions related to gender and sexuality in their education and practice. The demographic data revealed an average respondent age of nearly 40 years, with a notable majority of women. Diversity was evident in their medical specializations, with internal medicine emerging as the most common field, whereas the professional status of the respondents was well-balanced between attending physicians and residents, although slightly skewed toward the former. A significant portion of the survey population was relatively new to the medical field, emphasizing the presence of early-career physicians.

In terms of their experiences with gender and sexuality content, the data suggested that the exposure during academic studies was generally low to moderate. This was mirrored in their perception of preparedness in these areas, indicating a gap in the curriculum.

The sociological landscape of LGBTQI rights in Israel has been marked by both significant progress and notable contradictions. Since the early 2000s, there have been considerable advancements for LGBTQI individuals. However, progress has been uneven, especially for the transgender community, which continues to face significant discrimination, violence, and material disadvantages. In key institutions like health care and education, LGBTQI individuals encounter barriers that reflect broader societal tensions. Privatization and economic disparities exacerbate these challenges, particularly for those without the resources to navigate these systems effectively [11].

Our survey findings can be interpreted and discussed against this framework. However, they are challenging to directly compare due to the scarcity of research in Israeli contexts, which is predominantly limited to specific populations, such as physiotherapy students [12]. On the other hand, the findings well align with the broader trends identified in the literature, emphasizing a widespread issue in medical education regarding the adequacy of training on gender and sexuality. According to a survey by Obedin-Maliver et al [13], medical schools in the United States and Canada devote a small amount of time in their curricula to LGBTQI health and other topics related to sexuality, indicating a need for more comprehensive education in these areas. This survey was conducted more than a decade ago (between May 2009 and March 2010) and replicated recently, finding that, while the median time allocated to LGBTQI health-related topics increased in US and Canadian undergraduate medical education institutions, the scope, effectiveness, and quality of this instruction varied significantly. Despite the rise in hours, the total remains below the number recommended by the Association of American Medical Colleges (AAMC) for LGBTQI health competencies [14].

Of note, when it comes to the application of this knowledge in professional practice, there was a moderate level of self-assessed competence, coupled with a general consensus on the need for more comprehensive training. The importance of considering

XSL•FO
RenderX

the patient's sex and gender in treatment decisions showed moderate agreement among the respondents, with a slightly higher emphasis on the impact of treatment in the context of the patient's sex and gender. The impact of training in gender and sexuality on professional capabilities was largely acknowledged by the majority. This aligns with studies suggesting that insufficient training on gender and sexuality issues can lead to lower confidence among physicians when addressing the health needs of LGBTQI patients and other gender-diverse populations. For instance, a study by Marr et al [15] highlighted that many medical residents feel unprepared to provide high-quality care to LGBTQI patients, mirroring our survey's findings on the perceived preparedness gap.

Of note, our survey's participants expressed confidence in their ability to handle gender and sexuality issues in clinical practice, despite a low self-reported exposure to gender and sexuality content during their medical training. This apparent paradox, where a curricular gap was identified but respondents still felt prepared to deliver care, has been observed in other studies. What is particularly intriguing in our study is the differentiation between residents and fully licensed physicians, with responses further stratified by years of practice. These data seem to suggest a pattern; the longer a physician has been in practice, the less gender and sexuality content they recall from their training, yet the more confident they feel in their knowledge and ability to address these issues. This pattern could point to several possibilities. It may suggest that while formal education on gender and sexuality issues is lacking, the day-to-day experiences and challenges of medical practice provide ample opportunities for physicians to develop the necessary competencies. Alternatively, it could imply that confidence increases with experience, even if knowledge gaps persist, potentially leading to overconfidence in areas where additional training would be beneficial. This distinction between learning through experience *versus* feeling prepared due to increased confidence is a critical area for further exploration, as it has significant implications for medical education and ongoing professional development.

Furthermore, the respondents pinpointed 10 critical areas of gender-related knowledge, encompassing a broad spectrum from LGBTQI awareness to the specifics of gendered pharmacology, pointing to the multifaceted nature of sex and gender in medical practice. In discussing the incorporation of gender medicine into medical education, there was a clear preference for its integration across all stages of learning, reflecting a progressive approach toward medical training. The data also highlighted active involvement in student mentorship by a substantial number of respondents. The literature increasingly supports the integration of gender medicine and education on sexuality and gender diversity throughout the entire medical education continuum, from undergraduate education to continuing medical education for practicing physicians. This approach is advocated to ensure that medical doctors are well-equipped to meet the diverse needs of all patients, recognizing the significant role of sex and gender in health outcomes. For example, a consensus statement by the AAMC on the inclusion of gender awareness and LGBTQI health in medical education curriculum frameworks emphasizes the need for longitudinal integration rather than isolated modules or electives [16,17].

Furthermore, the survey revealed intriguing correlations, particularly between perceptions of academic program preparation and specialization and residency program views on gender and sexuality awareness. The gender-based analysis presented a complex picture, with variations in exposure and perceptions between men and women. Certain topics showed gender imbalances, while others exhibited more parity. Comparing attending physicians and residents, differences emerged in their perceptions of exposure to and preparedness in gender and sexuality content, suggesting variations in training across different stages of medical careers. The multivariate analysis further unraveled associations between various factors such as gender, years of experience, and medical status in relation to the survey responses. Similarly, a survey [18] conducted in Taiwan identified several shortcomings in present medical education and the lack of readiness among medical students and trainees to offer improved care for LGBTQI individuals.

In summary, this survey underscored the growing recognition of gender and sexuality as pivotal components in medical education and practice. It highlighted existing gaps in training and varying perceptions based on demographic and professional factors, pointing toward a need for a more inclusive and comprehensive approach in medical training and practice.

This survey offers valuable guidance for medical teachers and institutional stakeholders on developing and applying effective curricula and training programs, as well as faculty development initiatives. These strategies should aim to furnish medical students and trainees with the self-awareness and skills necessary to deliver gender-sensitive care, including comprehensive care to sexual and gender minorities, align with societal advancements, and advance health equity for a broader range of patients.

Effective communication is crucial for medical doctors, involving active listening, clear explanations, recognition of nonverbal cues, and patient education. By actively listening, health care providers can fully understand patients' symptoms and concerns, leading to more accurate diagnoses and tailored treatments. Clear explanations about diagnoses and treatment options ensure patients can make informed decisions. Recognizing and responding to nonverbal cues enhance understanding and trust, while effective patient education ensures patients comprehend their health conditions and necessary treatments. Cultural competence is essential, requiring awareness of diverse cultural backgrounds, including values and beliefs. Sensitivity to cultural differences and avoiding stereotypes help build trust and provide respectful care. Adapting health care practices to meet cultural needs improves health outcomes and patient satisfaction. Ongoing cultural competence training enhances inclusive care. Empathy involves understanding and valuing patients' feelings and experiences and building therapeutic relationships. Providing compassionate care alleviates anxiety and improves the health care experience. Offering emotional support and reassurance is crucial, and reflective practice helps physicians improve empathetic

XSL•FO
RenderX

interactions. Navigating complex social and ethical considerations related to gender and sexuality is vital in gender-based medicine. Recognizing and respecting diverse gender identities ensures all patients receive appropriate care. Implementing inclusive practices, such as using correct pronouns and offering gender-neutral facilities, supports patient well-being. Addressing ethical dilemmas requires careful consideration of patient autonomy and confidentiality. Staying informed about gender issues, advocating for patients' rights, and working to eliminate health care disparities are integral to ethical medical practice. These competencies—effective communication, cultural competence, empathy, and navigating gender and sexuality issues—are fundamental for medical doctors to provide comprehensive, sensitive, and effective care. Ensuring all patients feel understood, respected, and valued is the cornerstone of excellent gender-based medicine.

A further point that should be stressed is that our findings revealed significant variations in exposure and preparedness between men and women concerning gender and sexuality content in medical education. Men reported a higher level of perceived preparedness in dealing with gender and sexuality issues than women. This discrepancy highlights a crucial point often overlooked in discussions about educational interventions, that is, those in positions of privilege (in this case, men) may report more comfort and may not perceive the existing gaps as those in less privileged positions (women). Men's higher self-reported comfort could stem from their generally more prominent status within the medical community, which may afford them more confidence in professional settings. Conversely, women, who historically and structurally face more barriers in the medical field, may experience and recognize these gaps more acutely. This perception gap is critical as it underscores the need for more targeted and inclusive educational programs that not only address the specific needs of female physicians but also raise awareness among male physicians about these disparities. In addition, it is essential to acknowledge the role of implicit biases and structural inequalities that contribute to these differing perceptions. Training programs must be designed to bridge this gap by fostering an environment where both male and female physicians can gain a more balanced and comprehensive understanding of gender and sexuality issues. This approach can lead to a more equitable and effective health care delivery system, where all practitioners are equally prepared to address the diverse needs of their patients. By incorporating these considerations into the development of medical curricula and professional training, we can work toward reducing the perception and comfort gap between male and female physicians, ultimately leading to improved patient outcomes and a more inclusive medical community.

Future Directions

The World Federation for Medical Education (WFME) sets global standards for quality improvement in medical education. These standards include explicit requirements for integrating gender and sexuality education into medical school curricula. The WFME's standards ensure that medical schools worldwide provide education that prepares physicians to address diverse patient needs, including those related to gender and sexuality. Our study's findings indicate a significant gap in the integration of gender and sexuality content within medical education, highlighting a discrepancy between current practices and the WFME's curricular requirements. As such, the findings of this survey highlight the need for a comprehensive overhaul of medical education curricula. Future efforts should focus on integrating gender and sexuality content more thoroughly and consistently across all stages of medical training. This includes both preclinical and clinical years, ensuring that medical doctors are equipped with the necessary knowledge and skills from the onset of their careers.

Given the reported gap in preparedness and exposure, there is a clear need for targeted training programs that address specific areas of gender and sexuality in health care. These programs should cover the 10 critical areas identified by respondents, ranging from LGBTQI awareness to gender-specific diseases and symptoms.

Further research is necessary to continuously monitor and evaluate the effectiveness of implemented educational strategies. Longitudinal studies could be beneficial in assessing the impact of improved gender and sexuality training on health care outcomes. In addition, research should explore the evolving needs and perceptions of medical residents and practicing physicians in these areas.

The study's results can be used to advocate for policy changes at institutional and national levels. This involves lobbying for mandatory inclusion of gender and sexuality topics in medical education accreditation standards and continuous professional development requirements.

Furthermore, mentorship programs that emphasize gender and sexuality awareness should be encouraged. This point is crucial and experienced medical doctors who are well-versed in these topics should mentor younger colleagues, fostering a culture of continuous learning and sensitivity toward these issues. Efforts should be made to promote diversity and inclusion within the medical community, addressing gender imbalances in various medical specializations and ensuring that medical education and practice are inclusive of all sexes, genders, sexual orientations, and gender identities.

The latest technological advancements can be leveraged. Using, for instance, virtual reality and e-learning platforms, can provide innovative ways to teach and engage medical students and practicing doctors in gender and sexuality topics [19,20]. This approach can supplement traditional learning methods and offer flexible training opportunities. Future studies investigating the effectiveness of these technological methods in gender medicine education are crucial as they would help in understanding how well these technologies enhance learning outcomes, their impact on the practical skills of medical doctors, and how they compare with traditional teaching methods. Implementing technology in medical education, especially for topics like gender and sexuality, represents a significant step forward in creating a more informed and sensitive health care environment.

Conclusions

This study underscores the critical need for integrating gender and sexuality awareness into medical education and practice,

XSL•FO
RenderX

finding that, despite the recognized importance, there is a notable gap in the current training and preparedness of medical residents and practicing physicians in these areas. The survey results reveal a consensus on the necessity for more comprehensive training, reflecting the evolving landscape of health care where gender and sexuality play a significant role in patient care and outcomes. The variations in exposure and perceptions based on gender, professional status, and years of experience highlight the diversity of learning and training needs within the medical community. This calls for a tailored approach in educational interventions, ensuring that they are relevant and effective for various groups within the medical profession.

Overall, the study contributes significantly to the ongoing discourse on personalized, gender-sensitive health care, by providing valuable insights for educators, policy makers, and health care providers, emphasizing the need for a more inclusive, aware, and well-prepared medical workforce to cater to the diverse health care needs of the population.

Conflicts of Interest

None declared.

References

1. Khamisy-Farah R, Bragazzi NL. How to integrate sex and gender medicine into medical and allied health profession undergraduate, graduate, and post-graduate education: insights from a rapid systematic literature review and a thematic meta-synthesis. J Pers Med 2022;12(4):612 [FREE Full text] [doi: 10.3390/jpm12040612] [Medline: 35455728]
2. Yang HC. What should be taught and what is taught: integrating gender into medical and health professions education for medical and nursing students. Int J Environ Res Public Health 2020;17(18):6555 [FREE Full text] [doi: 10.3390/ijerph17186555] [Medline: 32916861]
3. Yang HC. Education first: promoting LGBT+ friendly healthcare with a competency-based course and game-based teaching. Int J Environ Res Public Health 2019;17(1):107 [FREE Full text] [doi: 10.3390/ijerph17010107] [Medline: 31877850]
4. Morris M, Cooper RL, Ramesh A, Tabatabai M, Arcury TA, Shinn M, et al. Training to reduce LGBTQ-related bias among medical, nursing, and dental students and providers: a systematic review. BMC Med Educ 2019;19(1):325 [FREE Full text] [doi: 10.1186/s12909-019-1727-3] [Medline: 31470837]
5. Danckers M, Nusynowitz J, Jamneshan L, Shalmiyev R, Diaz R, Radix AE. The sexual and gender minority (LGBTQ+) medical trainee: the journey through medical education. BMC Med Educ 2024;24(1):67 [FREE Full text] [doi: 10.1186/s12909-024-05047-4] [Medline: 38233849]
6. Lego VD. Uncovering the gender health data gap. Cad Saude Publica 2023;39(7):e00065423 [FREE Full text] [doi: 10.1590/0102-311XEN065423] [Medline: 37585901]
7. Kling JM, Rose SH, Kransdorf LN, Viggiano TR, Miller VM. Evaluation of sex- and gender-based medicine training in post-graduate medical education: a cross-sectional survey study. Biol Sex Differ 2016;7(Suppl 1):38 [FREE Full text] [doi: 10.1186/s13293-016-0097-3] [Medline: 27790362]
8. Rrustemi I, Locatelli I, Schwarz J, Lagro-Janssen T, Fauvel A, Clair C. Gender awareness among medical students in a Swiss university. BMC Med Educ 2020;20(1):156 [FREE Full text] [doi: 10.1186/s12909-020-02037-0] [Medline: 32487129]
9. Jenkins MR, Herrmann A, Tashjian A, Ramineni T, Ramakrishnan R, Raef D, et al. Sex and gender in medical education: a national student survey. Biol Sex Differ 2016;7(Suppl 1):45 [FREE Full text] [doi: 10.1186/s13293-016-0094-6] [Medline: 27785347]
10. Lindsay S, Kolne K. The training needs for gender-sensitive care in a pediatric rehabilitation hospital: a qualitative study. BMC Med Educ 2020;20(1):468 [FREE Full text] [doi: 10.1186/s12909-020-02384-y] [Medline: 33238977]
11. Blus-Kadosh I, Rogel A, Blatt R, Hartal G. Progress and challenges of the LGBT+ community in Israel. In: Kumaraswamy PR, editor. The Palgrave International Handbook of Israel. Singapore: Palgrave Macmillan; 2023.
12. Elboim-Gabyzon M, Klein R. Lesbian, gay, bisexual, and transgender clinical competence of physiotherapy students in Israel. BMC Med Educ 2024;24(1):729 [FREE Full text] [doi: 10.1186/s12909-024-05679-6] [Medline: 38970017]
13. Obedin-Maliver J, Goldsmith ES, Stewart L, White W, Tran E, Brenman S, et al. Lesbian, gay, bisexual, and transgender-related content in undergraduate medical education. JAMA 2011;306(9):971-977. [doi: 10.1001/jama.2011.1255] [Medline: 21900137]
14. Streed Jr CG, Michals A, Quinn E, Davis JA, Blume K, Dalke KB, et al. Sexual and gender minority content in undergraduate medical education in the United States and Canada: current state and changes since 2011. BMC Med Educ 2024;24(1):482 [FREE Full text] [doi: 10.1186/s12909-024-05469-0] [Medline: 38693525]
15. Marr MC, Bunting SR, Blansky BA, Dickson L, Gabrani A, Sanchez NF. Graduate medical education curriculum regarding the health and healthcare of older lesbian, gay, bisexual, transgender, and queer (LGBTQ+) adults. J Gay Lesbian Soc Serv 2023;35(4):420-433 [FREE Full text] [doi: 10.1080/10538720.2023.2172122] [Medline: 38107508]
16. Implementing curricular and institutional climate changes to improve health care for individuals who are LGBT, gender nonconforming, or born with DSD: a resource for medical educators. Assoc Am Med Coll. 2014. URL: https://store.aamc.org/downloadable/download/sample/sample_id/129/ [accessed 2024-08-23]

XSL•FO
RenderX

17. Jewell TI, Petty EM. LGBTQ+ health education for medical students in the United States: a narrative literature review. Med Educ Online 2024;29(1):2312716 [FREE Full text] [doi: 10.1080/10872981.2024.2312716] [Medline: 38359164]
18. Lu PY, Hsu ASC, Green A, Tsai JC. Medical students' perceptions of their preparedness to care for LGBT patients in Taiwan: is medical education keeping up with social progress? PLoS One 2022;17(7):e0270862 [FREE Full text] [doi: 10.1371/journal.pone.0270862] [Medline: 35797357]
19. Khamisy-Farah R, Gilbey P, Furstenau LB, Sott MK, Farah R, Viviani M, et al. Big data for biomedical education with a focus on the COVID-19 era: an integrative review of the literature. Int J Environ Res Public Health 2021;18(17):8989 [FREE Full text] [doi: 10.3390/ijerph18178989] [Medline: 34501581]
20. Lewis KO, Popov V, Fatima SS. From static web to metaverse: reinventing medical education in the post-pandemic era. Ann Med 2024;56(1):2305694 [FREE Full text] [doi: 10.1080/07853890.2024.2305694] [Medline: 38261592]

Abbreviations

AAMC: Association of American Medical Colleges
LGBTQI: lesbian, gay, bisexual, transgender, queer/questioning, and intersex
WFME: World Federation for Medical Education

Edited by K Prairie, D Chartash; submitted 30.03.24; peer-reviewed by SD Stryker, B Schuster, R Primavesi; comments to author 01.06.24; revised version received 28.07.24; accepted 16.08.24; published 08.10.24.

Please cite as:
Khamisy-Farah R, Biras E, Shehadeh R, Tuma R, Atwan H, Siri A, Converti M, Chirico F, Szarpak Ł, Biz C, Farah R, Bragazzi N
Gender and Sexuality Awareness in Medical Education and Practice: Mixed Methods Study
JMIR Med Educ 2024;10:e59009
URL: https://mededu.jmir.org/2024/1/e59009
doi:10.2196/59009
PMID:39152652

©Rola Khamisy-Farah, Eden Biras, Rabie Shehadeh, Ruba Tuma, Hisham Atwan, Anna Siri, Manlio Converti, Francesco Chirico, Łukasz Szarpak, Carlo Biz, Raymond Farah, Nicola Bragazzi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 08.10.2024. 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.

XSL•FO
RenderX

Utilization of, Perceptions on, and Intention to Use AI Chatbots Among Medical Students in China: National Cross-Sectional Study

Wenjuan Tao[1,*], PhD; Jinming Yang[2,3,*], MS; Xing Qu[1], PhD

1

2

3

*these authors contributed equally

Corresponding Author:
Xing Qu, PhD

Abstract

Background: Artificial intelligence (AI) chatbots are poised to have a profound impact on medical education. Medical students, as early adopters of technology and future health care providers, play a crucial role in shaping the future of health care. However, little is known about the utilization of, perceptions on, and intention to use AI chatbots among medical students in China.

Objective: This study aims to explore the utilization of, perceptions on, and intention to use generative AI chatbots among medical students in China, using the Unified Theory of Acceptance and Use of Technology (UTAUT) framework. By conducting a national cross-sectional survey, we sought to identify the key determinants that influence medical students' acceptance of AI chatbots, thereby providing a basis for enhancing their integration into medical education. Understanding these factors is crucial for educators, policy makers, and technology developers to design and implement effective AI-driven educational tools that align with the needs and expectations of future health care professionals.

Methods: A web-based electronic survey questionnaire was developed and distributed via social media to medical students across the country. The UTAUT was used as a theoretical framework to design the questionnaire and analyze the data. The relationship between behavioral intention to use AI chatbots and UTAUT predictors was examined using multivariable regression.

Results: A total of 693 participants were from 57 universities covering 21 provinces or municipalities in China. Only a minority (199/693, 28.72%) reported using AI chatbots for studying, with ChatGPT (129/693, 18.61%) being the most commonly used. Most of the participants used AI chatbots for quickly obtaining medical information and knowledge (631/693**,** 91.05%) and increasing learning efficiency (594/693, 85.71%). Utilization behavior, social influence, facilitating conditions, perceived risk, and personal innovativeness showed significant positive associations with the behavioral intention to use AI chatbots (all *P* values were <.05).

Conclusions: Chinese medical students hold positive perceptions toward and high intentions to use AI chatbots, but there are gaps between intention and actual adoption. This highlights the need for strategies to improve access, training, and support and provide peer usage examples to fully harness the potential benefits of chatbot technology.

(JMIR Med Educ 2024;10:e57132) doi:10.2196/57132

KEYWORDS

medical education; artificial intelligence; UTAUT model; utilization; medical students; cross-sectional study; AI chatbots; China; acceptance; electronic survey; social media; medical information; risk; training; support

Introduction

The rapid advancements in artificial intelligence (AI) have significantly transformed various sectors, including health care. Among these advancements, AI chatbots have emerged as a promising tool with potential applications in medical education [1]. These intelligent systems use natural language processing and machine learning algorithms to engage in human-like dialogues, providing information in an understandable, efficient, interactive, and scenario-specific format, such as ChatGPT, Claude, Google Bard, and Bing's AI [2]. The chatbots can assist medical students in medical research support, personalized learning, comprehending complex medical topics, developing clinical decision-making skills, and so forth [1,3]. A recent study demonstrated the efficacy of AI chatbots in answering complex medical questions and providing valuable medical educational support [4].

In China, integrating AI technology into medical education is particularly important, given the country's substantial investment in AI development and its growing emphasis on innovative educational methodologies [5,6]. AI chatbots would facilitate personalized learning experiences when facing the situation of

XSL•FO
RenderX

rigorous curricula and high student to teacher ratios in China. Medical students are a crucial target group for AI chatbot technology, as they are early adopters of technology and future health care providers who will play a vital role in shaping the future of health care. While research on AI chatbot applications in medical students has emerged [7-10], the utilization of, perceptions on, and intention to use AI chatbots among Chinese medical students are still unknown.

The adoption and effective utilization of AI chatbots among medical students depend on various factors, including their perceptions, attitudes, and behavioral intentions. The Unified Theory of Acceptance and Use of Technology (UTAUT), developed by Venkatesh et al [11], provides a comprehensive framework to understand the determinants of technology acceptance and usage, which is widely used in health care. Applying the UTAUT model in the context of AI chatbots can yield valuable insights into the factors that drive or hinder their adoption among medical students.

This study aims to explore the utilization of, perceptions on, and intention to use generative AI chatbots among medical students in China, using the UTAUT framework. By conducting a national cross-sectional survey, we seek to identify the key determinants that influence medical students' acceptance of AI chatbots, thereby providing a basis for enhancing their integration into medical education. Understanding these factors is crucial for educators, policy makers, and technology developers to design and implement effective AI-driven educational tools that align with the needs and expectations of future health care professionals.

Methods

Participants and Procedure

The target population was medical students enrolled in Chinese medical colleges or universities. An electronic survey was developed through a web-based survey platform named Wenjuanxing Questionnaire Star (Ranxing Technology Corp.), and the survey link was distributed via WeChat (Tencent Holdings Ltd) to medical college students across the country. Using a convenience sampling method, the questionnaire was posted on WeChat Moments and sent to WeChat groups from the research team's WeChat accounts. We identified relevant WeChat groups that consisted of medical students across various regions in China. These groups were selected based on their active participation in medical education discussions and their membership of medical students from diverse backgrounds and institutions. The research team directly contacted a total of 15 WeChat groups. To further enhance the reach, we used a snowball sampling method by requesting initial respondents to forward the survey link to other medical students in their network. Questionnaires that were considered valid included only the following: (1) each account responded only once, and (2) the total response time for completing the questionnaire was more than 300 seconds. Participants were recruited between June 2023 and July 2023.

To ensure adequate statistical power and precision for the intended analyses, we conducted a sample size calculation using G*Power software (version 3.1.9.7) [12]. The calculation was based on the following parameters: a small effect size (f_2=0.05) was chosen for the multivariable regression analysis; the number of predictors was set at 15; the desired statistical power was set at 0.95; and the significance level was set at .05 (2-tailed). Based on these parameters, the minimum required sample size for the multivariable regression analysis was calculated to be 566 participants. However, to account for 20% missing data and increase the generalizability of the findings, we aimed to recruit 680 participants. Ultimately, we were able to collect 715 questionnaires across China, with 693 determined valid, representing a 96.9% final response rate.

The Theoretical Framework

The study used the UTAUT as a theoretical framework for the research. The UTAUT describes 4 key independent variables: performance expectancy (PE), effort expectancy (EE), social influence (SI), and facilitating conditions (FC). In this study, PE measured the participants' expectation that an AI chatbot will be useful for the study; EE measured the expectation that an AI chatbot is user-friendly and easy to use; SI measured the degree to which a user perceives that important others believe that he or she should use the new technology; and FC measured the degree to which a user believes that an organizational and technical infrastructure exists to support AI chatbot use [13]. The dependent variable BI was determined by PE, EE, SI, and FC. BI measured participants' intention to use the AI chatbot in their future study.

The intention was used as an outcome instead of the actual use of an AI chatbot, because the application of AI chatbot services has not been widely commercialized in China. Most medical students may not have experience with an AI chatbot in their study. Also, BI is a good representation of actual behavior [14]. Previous studies confirmed that these 4 variables (PE, EE, SI, and FC) have a positive influence on the intention to use the AI technology [15-17]. The original UTAUT validation study found that the UTAUT model is robust in explaining a high degree of variance (70%) in BI [13]. Moderating effects of age, gender, and experience were not tested in this study.

In addition, we added perceived risk (PR), resistance bias (RB), and personal innovativeness (PI) as 3 variables to the original UTAUT model. PR is defined as the potential for loss in the pursuit of the desired outcome of using a technology and identified for 7 facets of PR [18]. Here, PR was measured for performance risk, time risk, and privacy risk. RB is resistance to change, referring to people's attempts to maintain previous behaviors or habits that are connected to their past experiences when facing changes [16]. PI was designed to measure an individual's willingness to try out any new information technology [19]. Since AI chatbots are an emerging technology in health care, a user's inherent innovativeness may impact his or her intention to adopt this innovation, and some users may be accompanied by concerns and resistance to change when embracing the new technology. Previous studies found that PR and RB have been regarded as major barriers to health care information technology adoption [20,21], and PI has been statistically significant in predicting the BI of the user [22].

XSL•FO
RenderX

Questionnaire and Instrument

The developed questionnaire consisted of 3 parts (see questionnaire in Multimedia Appendix 1): (1) participants' sociodemographic information, such as age, gender, and grade level; (2) participants' cognition of, attitude toward, and experience with AI chatbots (these items were designed as categorical variables and were derived through a comprehensive process, including literature review and expert consultation, to ensure their relevance and clarity); and (3) the scale of the research model. The model covered 7 constructs with 29 questionnaire items (Table 1). The items of the survey were ordered such that items measuring each construct were grouped. The responses were recorded using a 5-point Likert scale (ranging from 1=totally disagree to 5=totally agree) in which the higher score values indicated a higher level of a construct and a higher score of the outcome (BI) indicated greater intention to use the AI chatbot.

In the third part of the questionnaire, each item in the scale was sourced from relevant literature related to new technology acceptance research. The main modifications to the original instrument were made to fit the context of an AI chatbot used for medical students, such as changing the word "system" to "AI Chatbot." The items that assessed PE, EE, SI, FC, and BI were adopted from the original instrument developed by Venkatesh et al [11]. The original survey was validated and applied to previous studies based on the UTAUT model [16,23,24]. The items that assessed PR and RB were adopted from the validated questionnaire developed by Zhai et al [16]. The reliability of the items' scales was tested by Cronbach α coefficient analysis. The results of Cronbach α are considered to have acceptable reliability (Table 1), as the generally accepted rule is that α values of 0.6 - 0.7 indicate an acceptable level of reliability, and 0.8 or greater is a very good level [25].

After we developed the questionnaire and before implementing the survey, we conducted a consensus panel of 5 experts to review the questionnaire and ensure clarity of the survey and content validity. We then conducted a pilot study of 20 students to clarify phrasing and eliminate items that were not identifiable in the questionnaire.

XSL•FO
RenderX

Table . The model constructs and its measuring scale items.

Constructs and items		Cronbach α
PE [a]		0.920
	PE 1: I would find AI[b] Chatbot useful in my study.	
	PE 2: Using AI Chatbot will enable me to accomplish tasks more quickly.	
	PE 3: Using AI Chatbot will increase my productivity.	
	PE 4: If I use AI Chatbot, I will increase my chances of getting better grades.	
EE [c]		0.904
	EE 1: My interaction with AI Chatbot will be clear and understandable.	
	EE 2: It would be easy for me to become skillful at using AI Chatbot.	
	EE 3: I would find AI Chatbot easy to use.	
	EE 4: Learning to operate AI Chatbot is easy for me.	
SI [d]		0.871
	SI 1: People who influence my behavior (eg, classmates, colleagues, and friends) think that I should use AI Chatbot.	
	SI 2: People who are important to me (eg, department heads, supervisors, and hospital leaders) think that I should use AI Chatbot.	
	SI 3: The senior health administration has been helpful in the use of AI Chatbot.	
	SI 4: In general, my university and hospital have supported the use of AI Chatbot.	
FC [e]		0.756
	FC 1: I have the resources necessary to use AI Chatbot.	
	FC 2: I have the knowledge necessary to use AI Chatbot.	
	FC 3: AI Chatbot is not compatible with other systems I use.	
	FC 4: A specific person (or group) is available for assistance with the AI Chatbot difficulties.	
PR [f]		0.643
	PR 1: There is a possibility of malfunction and performance failure, so the AI Chatbot fails to deliver accurate information and could mislead my study.	
	PR 2: There is a probability that I need more time to fix the errors and nuances of the AI Chatbot.	
	PR 3: I am worried that AI chatbots will reveal my private information.	
RB [g]		0.879

XSL•FO
RenderX

Constructs and items		Cronbach α
	RB 1: I do not want AI chatbots to change the way I study or work because the new AI tools are unfamiliar to me.	
	RB 2: I do not want to use the AI chatbots because of past experiences; these new high-tech products always fall flat during practical application.	
	RB 3: I do not want to use the AI chatbots because there is a possibility of losing my job, as artificial intelligence–assisted technology may do my work better than me.	
PI [h]		0.634
	PI 1: If I heard about a new information technology, I would look for ways to experiment with it.	
	PI 2: Among my peers, I am usually the first to try out new information technologies.	
	PI 3: In general, I am hesitant to try out new information technologies.	
	PI 4: I like to experiment with new information technologies.	
BI [i]		0.946
	BI 1: I intend to use the AI chatbots in the next 2 months.	
	BI 2: I predict I would use the AI chatbots in the next 2 months.	
	BI 3: I plan to use the AI chatbots in the next 2 months.	

[a]PE: performance expectancy.
[b]AI: artificial intelligence.
[c]EE: effort expectancy.
[d]SI: social influence.
[e]FC: facilitating conditions.
[f]PR: perceived risk.
[g]RB: resistance bias.
[h]PI: personal innovativeness.
[i]BI: behavioral intention.

Data Analysis

The statistical software SPSS 25.0 (IBM Corp) was used to calculate the Cronbach α coefficient. Data analysis was carried out using descriptive statistics, such as means, frequencies, and percentages, as well as inferential statistics, such as multiple linear regression, to explore the relationships between the dependent variable (BI) and the set of predictors (PE, EE, SI, FC, PR, RB, and PI). The α level was set at .05 for all analyses. Data analysis was performed using Stata (version 17.0; StataCorp LLC).

Ethical Considerations

Ethical approval was obtained from the Ethics Committee on Biomedical Research, West China Hospital of Sichuan University (approval number: 2023 - 834). The research purpose; methods; and participants' rights, including that they could cease participation at any point without penalty, were explained. All the participants read and signed the electronic informed consent before completing the questionnaire. The detailed information on the informed consent form is given in the questionnaire in Multimedia Appendix 1. This survey was anonymous and voluntary. To promote survey completion and ensure an adequate response rate, postsurvey gifts were randomly raffled as an incentive.

Results

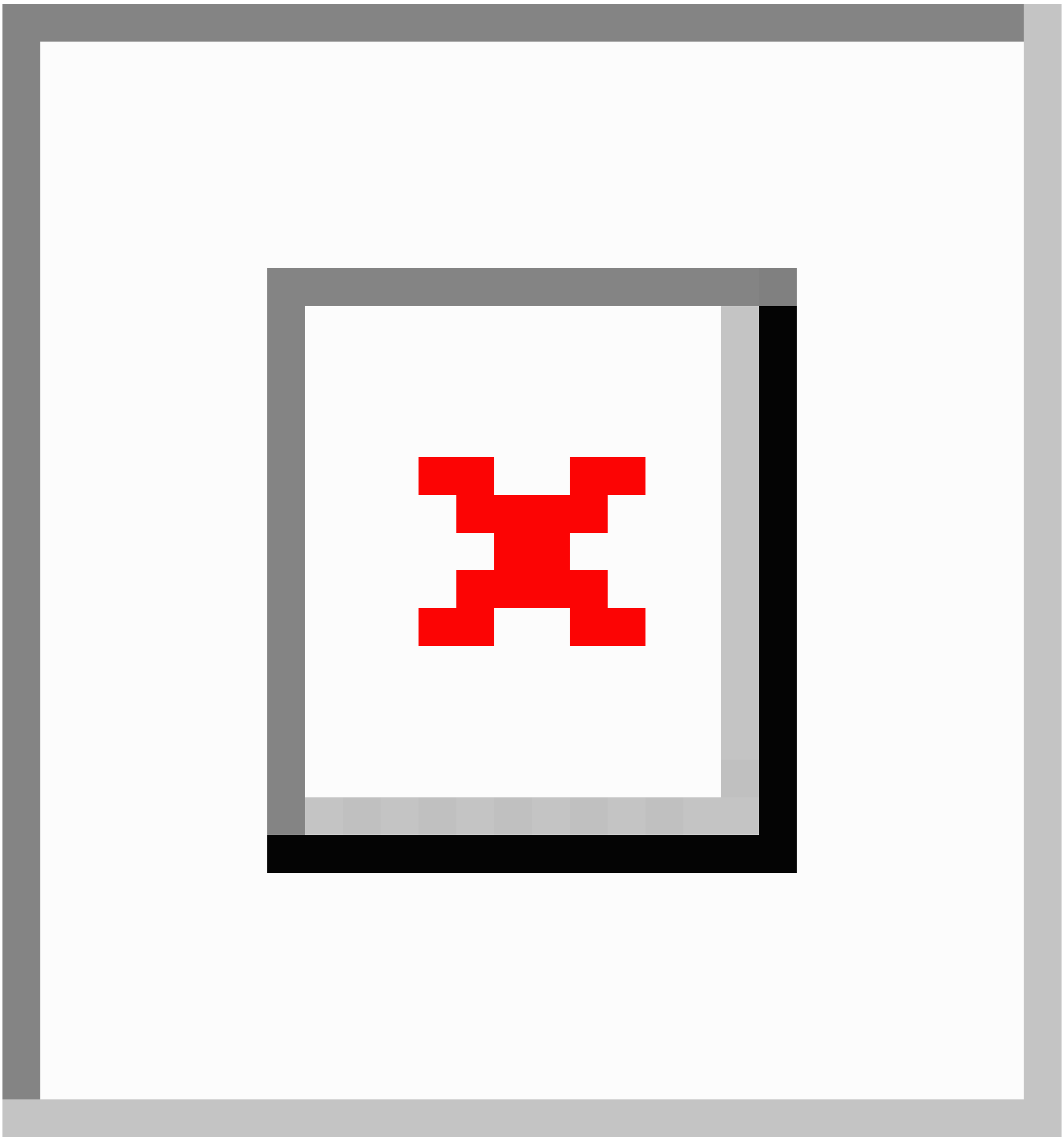
Participants' Information

A total of 693 participants were from 57 universities covering 21 provinces or municipalities. The sample distribution is shown in Figure 1. The demographic characteristics of the participants are shown in Table 2. The majority of participants (251/693,

XSL•FO
RenderX

63.78%) were female, while 36.22% (442/693) were male. The average age was 22.6 (SD 5.2) years, and more than half of the participants were in the 20 - to 24-year age range (413/693, 59.60%). The majority (543/693, 78.35%) were undergraduate students. The mean self-reported academic score was 73.7 (SD 14.8), and the most common self-reported score range was 80 - 89 (247/693, 35.64%).

Figure 1. Sample distribution.



XSL•FO
RenderX

Table . Demographic characteristics of participants (N=693).

Characteristics	Participants, n (%)
Sex	
Male	251 (36.22)
Female	442 (63.78)
Age (years)	
<20	141 (20.35)
20 - 24	413 (59.60)
25 - 29	87 (12.55)
≥30	52 (7.50)
Hukou type [a]	
Urban	364 (52.53)
Rural	329 (47.47)
Education level	
Undergraduate	543 (78.35)
Master student	101 (14.57)
Doctor student	49 (7.07)
Self-reported academic scores[b]	
90 - 100	70 (10.10)
80 - 89	247 (35.64)
70 - 79	163 (23.52)
60 - 69	144 (20.78)
<60	69 (9.96)

[a]"Hukou type" refers to the classification within the Chinese household registration system. This system classifies individuals based on their place of household registration and typically includes 2 main categories: Urban Hukou and Rural Hukou.

[b]"Academic scores" for medical students refer to the grades or marks they receive in various courses throughout their medical education in college or university. Academic scores are important indicators of a student's academic performance, reflecting his or her learning effectiveness.

Participants' Perception of AI Chatbots

Table 3 presents the participants' cognition, attitudes, usage behavior, and willingness to pay for AI chatbots. While only 24.68% (171/693) of participants reported being fairly familiar with AI chatbots and 4.47% (31/693) were very familiar with AI chatbots, 59.88% (415/693) agreed or strongly agreed with using them for study or work purposes. Of the 28.72% (199/693) who have used AI chatbots for studying, mainly ChatGPT (129/693, 18.61%), 50.25% (100/199) reported occasional usage as needed. Among nonusers, 55.06% (272/494) expressed willingness to learn AI chatbot usage, with the main reasons for unwillingness being no need (15/29, 51.72%) and no interest (15/29, 51.72%). In addition, 36.45% (242/664) preferred to use AI chatbots without charge.

Table 4 summarizes the participants' purposes for using AI chatbots and their perceived advantages and disadvantages. The primary purposes were quickly obtaining medical information and knowledge (631/693, 91.05%) and increasing learning efficiency (594/693, 85.71%). Perceived advantages included effectively helping medical students learn (631/693, 91.05%) and providing fast and accurate medical information (624/693, 90.04%). However, data privacy breaches (635/693, 91.63%) and risks of misdiagnosis or underdiagnosis (619/693, 89.32%) were predominant concerns.

XSL•FO
RenderX

Table . Participants' cognition, attitude, usage behavior, and willingness to pay for AI chatbots (exchange rate: US $1=¥7.22, July 9, 2023).

Items	Participants, n (%)
Do you know what an AI[a] chatbot is? (N=693)	
Completely unfamiliar	17 (2.45)
Unfamiliar	103 (14.86)
Average	371 (53.54)
Fairly familiar	171 (24.68)
Very familiar	31 (4.47)
Do you agree with the use of AI chatbot applications for study or work? (N=693)	
Strongly disagree	16 (2.31)
Disagree	31 (4.47)
Neutral	231 (33.33)
Agree	328 (47.33)
Strongly agree	87 (12.55)
Have you used AI chatbots in your study? (N=693)	
No	494 (71.28)
Yes	199 (28.72)
ChatGPT	129 (18.61)
New Bing	20 (2.89)
Others	31 (4.47)
Missing	19 (2.74)
How often do you use this AI chatbot? (N=199)	
Every day	18 (9.05)
Several times a week	36 (18.09)
About once a week	6 (3.02)
Occasionally, as needed	100 (50.25)
Rarely, only in specific situations	39 (19.60)
If you have not used it, would you be willing to learn how to use AI chatbots? (N=494)	
Strongly unwilling	6 (1.21)
Unwilling	23 (4.66)
Neutral	193 (39.07)
Somewhat willing	217 (43.93)
Very willing	55 (11.13)
If you are unwilling to use AI chatbots, what is the main reason? (N=29)	
No need	15 (51.72)
No interest	15 (51.72)
Inconvenient operation	8 (27.59)
Worries about privacy issues	14 (48.28)
Worries about inaccurate information provided	13 (44.83)
If a high-quality and convenient AI chatbot were available to assist you in your learning, how much would you be willing to pay per month to use it? (N=664)	
Free	242 (36.45)
<¥20	188 (28.31)
¥20 to ¥50	150 (22.59)

XSL•FO
RenderX

Items	Participants, n (%)
¥50 to ¥100	59 (8.89)
>¥100	25 (3.77)

[a]AI: artificial intelligence.

XSL•FO
RenderX

Table . Participants' purpose of using artificial intelligence (AI) chatbots and perceived advantages or disadvantages.[a]

Items		Total choices, n (%)	First choice, n (%)
What is your main purpose in using an AI chatbot?			
	Quickly obtaining basic medical information and knowledge	631 (91.05)	434 (62.63)
	Increasing learning efficiency.	594 (85.71)	69 (9.96)
	Seeking answers and guidance for complex medical questions.	583 (84.13)	68 (9.81)
	Exploring new research and academic resources.	564 (81.39)	43 (6.20)
	Self-health management and self-diagnosis.	520 (75.04)	22 (3.17)
	Improving the experience of medical learning and training.	509 (73.45)	12 (1.73)
	Retrieving various information, such as regular search engines.	443 (63.92)	30 (4.33)
	Chatting and entertainment.	377 (54.40)	12 (1.73)
	Others[b]	—[c]	—
What advantages do you think AI chatbots have?			
	They can effectively help medical students learn and master medical knowledge.	631 (91.05)	157 (22.66)
	They can provide fast and accurate medical information and diagnosis results.	624 (90.04)	405 (58.44)
	They can improve the efficiency and quality of health care services.	575 (82.97)	53 (7.65)
	They can reduce the workload and burden of doctors.	574 (82.83)	72 (10.39)
	Others[d]	—	—
What disadvantages or risks do you think AI chatbots have?			
	There may be risks of data privacy breaches.	635 (91.63)	406 (58.59)
	There may be risks of misdiagnosis or underdiagnosis.	619 (89.32)	137 (19.77)
	They may potentially lead to the degradation or unemployment of medical professionals.	570 (82.25)	87 (12.55)
	They may potentially reduce the personal touch and humanization of health care services.	559 (80.66)	58 (8.37)
	Others[e]	—	—

[a]The questions in this part of the survey were ranking questions. "Total choices" provide an overall measure of how often an option was selected. "First choice" represents the preference for an option as the most preferred or prioritized choice among respondents.

[b]Response examples: "Make code modifications," "Polishing the content of the documents," "Complete some unimportant homework," and "Online operation training."

[c]Not applicable.

[d]Response examples: "Reduce feelings of loneliness," "Regulate emotions of healthcare workers," "Provide arguments for the group work," and "Provide timely and patient answers."

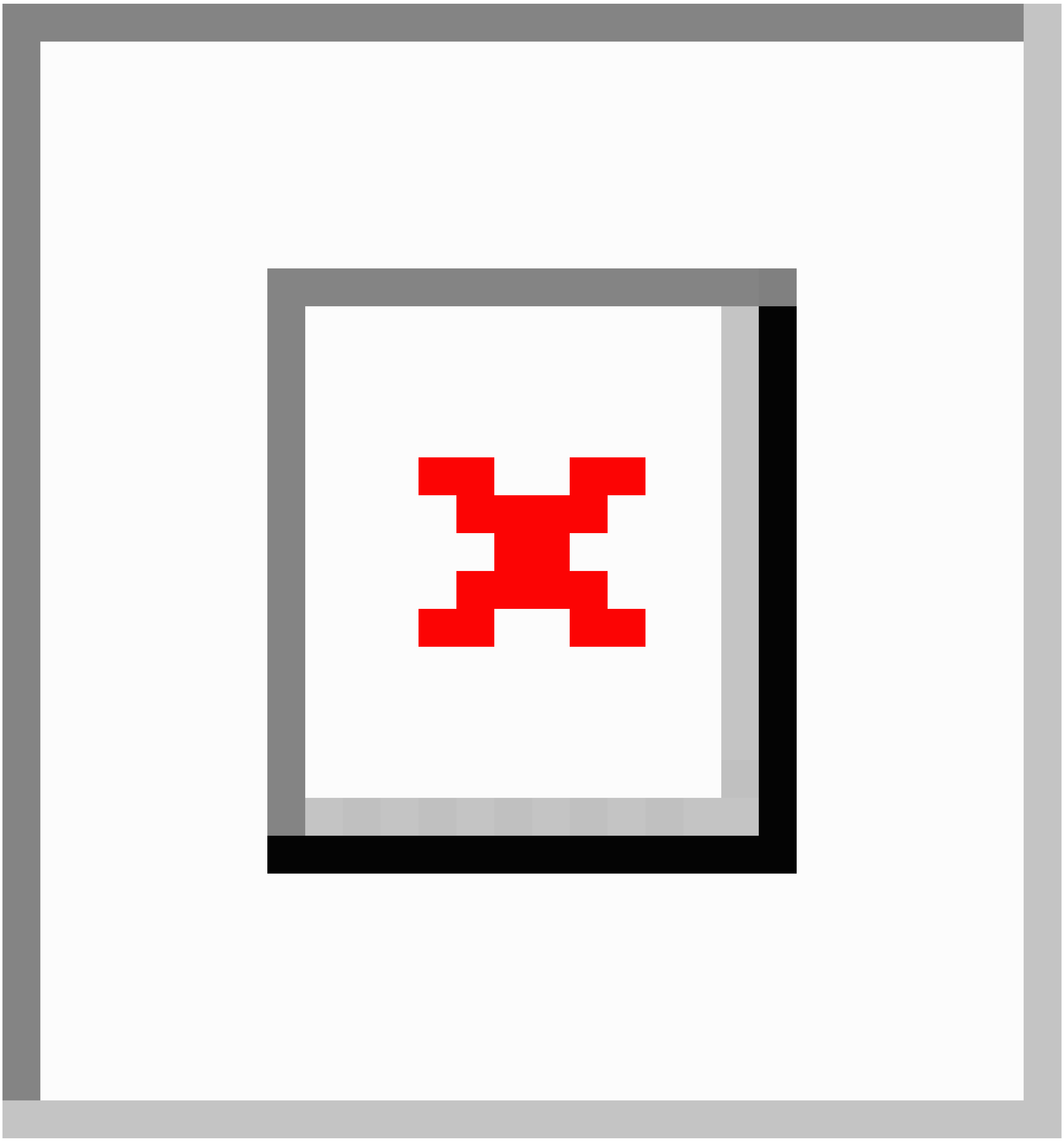
[e]Response examples: "Provide misleading information, such as fabricating references," "Patients may have doubts and mistrust towards these technological products," "AI currently cannot reflect the artistic elements required in medicine," and "It may not be able to provide the desired, high-quality answers."

XSL•FO
RenderX

Descriptive Statistics of the UTAUT Constructs

Descriptive statistics (mean [SD]) were reported to explain and describe the UTAUT constructs (Table S1 in Multimedia Appendix 1). The value of each construct ranges from 1 to 5 (1=strongly disagree, 5=strongly agree). As shown in Figure 2, the mean for PE, EE, SI, FC, PR, and PI were higher than 3 and the mean for RB was <3. The highest score was PE at 3.66, followed by EE at 3.56. The mean of BI was 3.26, which shows a higher level of intention to use AI chatbots among Chinese medical students.

Figure 2. Descriptive statistics of the Unified Theory of Acceptance and Use of Technology constructs. BI: behavioral intention; EE: effort expectancy; FC: facilitating conditions; PE: performance expectancy; PI: personal innovativeness; PR: perceived risk; RB: resistance bias; SI: social influence.



Determinant Factors of Intention to Use AI Chatbots

A multiple linear regression analysis was conducted to identify factors influencing medical students' intentions to use AI chatbots (Table 5). Utilization behavior (β=.27; *P*<.001), SI (β=.32; *P*<.001), FC (β=.29; *P*<.001), PR (β=.27; *P*<.001), and PI (β=.35; *P*<.001) were significantly positively associated with BI. PI had the largest positive regression coefficient (β=.35) compared with the other significant variables. PE (β=.09; *P*=.12), EE (β=.03; *P*=.60), and RB (β=−0.04; *P*=.32) did not significantly affect BI.

XSL•FO
RenderX

Table . Analysis of influence factors of medical students' behavioral intention to use artificial intelligence chatbots.[a]

Variables	Coefficient	SE	*t*	*P* value	95% CI
Age (years)	0.01	0.01	1.89	.06	0.00 to 0.02
Gender	−0.04	0.05	−0.83	.41	−0.15 to 0.06
Hukou type	0.10	0.05	1.93	.06	0.00 to 0.20
Education level	−0.02	0.05	−0.46	.65	−0.12 to 0.08
Academic scores	0.00	0.00	0.51	.61	0.00 to 0.00
Cognition[b]	0.01	0.04	0.27	.79	−0.06 to 0.08
Attitude[c]	0.03	0.04	0.89	.37	−0.04 to 0.11
Utilization behavior[d]	0.27	0.07	4.19	<.001	0.15 to 0.40
Performance expectancy	0.09	0.06	1.56	.12	−0.02 to 0.21
Effort expectancy	0.03	0.06	0.53	.60	−0.08 to 0.14
Social influence	0.32	0.05	6.32	<.001	0.22 to 0.42
Facilitating conditions	0.29	0.07	4.13	<.001	0.15 to 0.42
Perceived risk	0.27	0.05	5.33	<.001	0.17 to 0.37
Resistance bias	−0.04	0.04	−0.99	.32	−0.12 to 0.04
Personal innovativeness	0.35	0.07	5.33	<.001	0.22 to 0.48
Constant term	−1.98	0.31	−6.31	<.001	−2.60 to −1.36

[a]Model parameters: Probability>F=0, R^2=0.518, adjusted R^2=0.507, and Root Mean Square Error=0.651. df_Total=692, df_Model=15, df_Residual=677. The results of multicollinearity diagnostics showed that there is no multicollinearity among all independent variables in the multiple linear regression (Table S2 in Multimedia Appendix 1).

[b]The variable "Cognition" is measured by "Do you know what an AI chatbot is?"

[c]The variable "Attitude" is measured by "Do you agree with the use of AI chatbot applications for study or work?"

[d]The variable "Utilization behavior" is measured by "Have you used AI chatbots in your study?"

Discussion

Principal Findings

In this study, we examined the perceptions of Chinese medical students toward Natural Language Processing–based AI chatbots and investigated the factors that may influence their intention to use such technology based on the UTAUT model. This research yielded several key findings. First, the medical students demonstrated positive perceptions and expressed a high BI to use AI chatbots. Second, among the factors considered, SI and FC emerged as more influential in the adoption of AI chatbots among medical students than PE and EE. However, PE and EE were not found to have a significant relationship with BI. Third, PR and PI positively influenced BI, while RB did not show a significant association with BI.

This study revealed that although most medical students have limited knowledge about AI chatbots at an early time, they hold positive perceptions and demonstrate a strong intention to use this innovative technology. The overall sample displayed high BIs, with a mean score of 3.26 out of 5.00. Furthermore, 81.63% (1697/2079) of participants rated their intention as 3 or higher, indicating their plans to use the technology within the next 2 months. These findings align with previous research indicating that while medical students may lack knowledge about AI and its applications, they maintain a favorable view of AI in the medical field and are willing to adopt it [26,27]. The majority of participants believe that AI chatbots have the potential to enhance their study or work performance, improve efficiency, and provide fast and accurate medical information, among other benefits. However, limited availability and coverage of AI chatbots in China have resulted in less than one-third of participants actually using these tools and only a few using them on a daily basis. This indicates a gap between intention to use and actual adoption. Practical barriers, such as inadequate technical infrastructure and lack of support, may hinder the actual implementation and use of AI chatbots. In addition, our regression analysis revealed that utilization behavior significantly influences medical students' intentions to use AI chatbots. User experience may impact their perceptions of the technology from multiple aspects, thereby affecting their usage intentions.

This study found that SI and FC have a stronger impact on BI than PE and EE. This finding aligns with research examining the perceptions of Chinese radiation oncologists toward adopting AI-assisted contouring technology [16]. However, it contradicts some prior studies that have established a positive and significant relationship between PE and EE with students' BI to use AI-assisted learning environments [28] or chatbots [29]. This suggests that factors such as PE and EE may be less critical

XSL•FO
RenderX

for the population in this study, although they rated PE and EE higher than other dimensions. It is possible that as medical students are still in training, they rely more on the experiences of their peers and the infrastructure provided by their educational institution to guide their technology use. Thus, demonstrating adoption and endorsement from fellow students, professors, and the academic medical system may be more influential in persuading them to use AI chatbots than emphasizing use and usability. Ensuring accessibility within the educational context appears to shape students' willingness to use AI chatbots more than their individual perceptions of performance and efficiency.

Interestingly, this study found that PR and PI positively influenced the intentions of medical students to use AI chatbots, while RB was not found to be a significant factor. This suggests that concerns regarding the risks associated with adopting AI chatbots were outweighed by the students' openness to embracing new technologies. Those with a greater inclination toward innovation recognized the potential benefits despite the potential risks involved. This finding aligns with previous research indicating that perceived usefulness can override PR when it comes to determining acceptance of technology [18]. It also reflects a growing understanding that AI systems present both opportunities and risks, necessitating ethical analysis and oversight [30], including privacy breaches and the possibility of misinformation. Notably, we found that PI emerged as a key determinant of user behavior intentions, which is consistent with a similar study [31]. However, RB did not negatively impact intentions, suggesting that medical students may have fewer biases against AI than health care professionals in hospitals who may fear a loss of professional autonomy and challenges in integrating AI into clinical workflows [32]. Encouraging PI while addressing risk concerns through testing and regulation may further bolster the adoption of AI chatbots.

Implications for Practice

Based on our findings, we recommend the following specific strategies for educational institutions and AI chatbot developers to enhance the adoption rate among medical students. First, medical schools and health care organizations should prioritize efforts to improve FC and leverage SI to drive the adoption of AI chatbots, rather than solely focusing on performance benefits. Providing integrated access, training, and IT support and sharing peer usage examples can help translate positive intentions into actual usage behaviors. In addition, demonstrating value through pilot studies and addressing valid risk concerns will promote responsible and open adoption of the technology. Targeted training in AI competencies can further equip students to become champions of safe and effective adoption. The key lies in creating optimal environments and processes to enable the proficient use of AI systems such as chatbots as students transition into practice.

Strengths and Limitations

This study was the first to use the UTAUT theoretical framework to analyze medical students' intention to use AI chatbots. It possesses several strengths, including the robust technology adoption model used, the focus on an important user population, and the identification of key variables influencing intentions. However, there are some limitations that need to be addressed in future studies. First, the unbalanced research sample primarily from Sichuan province may limit the generalizability of the findings, potentially overrepresenting specific regional experiences. Although we distributed the survey widely, future studies should use stratified sampling for better regional representation. Second, the cross-sectional design offers only a snapshot of adoption, which may change over time as participants accumulate knowledge and experience. Future research should consider longitudinal designs to track these changes. Third, it is crucial to acknowledge that the field of AI chatbots is rapidly evolving, and our findings capture perceptions and attitudes at a specific point in time. As AI chatbot capabilities continue to advance, the external validity of our findings may need to be reevaluated.

Future studies with larger samples using longitudinal methods would enhance our understanding of actual perceptions and usage patterns over time. For example, a longitudinal study could follow a cohort of medical students from their entry into medical school until graduation, periodically assessing their perceptions, intentions, and actual usage of AI chatbots. This longitudinal approach would capture how their adoption and experiences with AI chatbots evolve as they progress through their medical education and gain more exposure to clinical settings. Furthermore, mixed methods designs, combining quantitative surveys with qualitative interviews or focus groups, could provide more in-depth insights into specific barriers, challenges, and facilitators influencing AI chatbot adoption among medical students. Overall, this study lays the foundation for a wide range of future research, which can deepen knowledge and generate evidence to guide the implementation of AI in education and health care.

Conclusions

This study offers valuable insights into medical students' utilization of, perceptions on, and intention to use AI chatbots in health care. The results indicate that these medical students have positive perceptions and strong intentions to use chatbots, primarily influenced by SI and FC rather than PE and EE. However, despite these intentions, there remains a gap between intention and actual adoption, signaling the need for strategies that improve access, training, and support and provide peer usage examples to enhance the realization of the potential benefits of chatbots. While concerns about risks exist, the students' general openness to innovation suggests that the integration of AI with proper oversight is well received. As future health care professionals, students serve as early adopters who can shape wider acceptance if barriers to adoption are actively addressed. This research provides a foundation for understanding the technology needs and motivations of this important user population in order to guide the successful implementation of AI.

XSL•FO
RenderX

Acknowledgments

We thank all the respondents who participated in the interviews and survey, as well as the ones who assisted with participant recruitment on the web. This work was supported by China Postdoctoral Science Foundation (grant 2023M732419), National Natural Science Foundation of China (grant 72374146), and Sichuan University Healthy City Development Research Center (grant 2022ZC003).

Data Availability

The datasets analyzed during this study are not publicly available due to privacy concerns in informed consent. The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Authors' Contributions

WJT contributed to conceptualization, data curation, formal analysis, investigation, methodology, and writing—original draft. JMY participated in the conceptualization, data curation, investigation, and writing—review and editing. XQ contributed to conceptualization, data curation, funding acquisition, investigation, supervision, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Survey questionnaire on medical students' use of artificial intelligence chatbots (translated version), descriptive statistics of the Unified Theory of Acceptance and Use of Technology constructs, and multicollinearity diagnostics.

[DOCX File, 33 KB - mededu_v10i1e57132_app1.docx]

References

1. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus 2023 Aug;15(8):e43271. [doi: 10.7759/cureus.43271] [Medline: 37692629]
2. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. ANZ J Surg 2024 Feb;94(1-2):68-77. [doi: 10.1111/ans.18666] [Medline: 37602755]
3. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ 2024 Apr 10;10:e52483. [doi: 10.2196/52483] [Medline: 38598263]
4. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. JMIR Med Educ 2023 Nov 1;9:e51421. [doi: 10.2196/51421] [Medline: 37910155]
5. The State Council's notice on issuing the next generation artificial intelligence development plan. The State Council of the People's Republic of China. 2017. URL: https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm [accessed 2024-06-09]
6. The Ministry of Education releases four initiatives to promote artificial intelligence empowerment in education. Xinhua News Agency. 2024. URL: http://edu.people.com.cn/n1/2024/0329/c1006-40205772.html [accessed 2024-06-09]
7. Han JW, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ 2022 Dec 1;22(1):830. [doi: 10.1186/s12909-022-03898-3] [Medline: 36457086]
8. Suárez A, Adanero A, Díaz-Flores García V, Freire Y, Algar J. Using a virtual patient via an artificial intelligence chatbot to develop dental students' diagnostic skills. Int J Environ Res Public Health 2022 Jul 18;19(14):8735. [doi: 10.3390/ijerph19148735] [Medline: 35886584]
9. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659. [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
10. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
11. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425-478. [doi: 10.2307/30036540]
12. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
13. Ammenwerth E. Technology acceptance models in health informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul 30;263:64-71. [doi: 10.3233/SHTI190111] [Medline: 31411153]

XSL•FO
RenderX

14. Eccles MP, Hrisos S, Francis J, et al. Do self- reported intentions predict clinicians' behaviour: a systematic review. Implement Sci 2006 Nov 21;1(28):28. [doi: 10.1186/1748-5908-1-28] [Medline: 17118180]
15. Jain R, Garg N, Khera SN. Adoption of AI-enabled tools in social development organizations in India: an extension of UTAUT model. Front Psychol 2022;13:893691. [doi: 10.3389/fpsyg.2022.893691] [Medline: 35795409]
16. Zhai H, Yang X, Xue J, et al. Radiation oncologists' perceptions of adopting an artificial intelligence–assisted contouring technology: model development and questionnaire study. J Med Internet Res 2021 Sep 30;23(9):e27122. [doi: 10.2196/27122] [Medline: 34591029]
17. García de Blanes Sebastián M, Sarmiento Guede JR, Antonovica A. Application and extension of the UTAUT2 model for determining behavioral intention factors in use of the artificial intelligence virtual assistants. Front Psychol 2022;13:993935. [doi: 10.3389/fpsyg.2022.993935] [Medline: 36329748]
18. Featherman MS, Pavlou PA. Predicting e-services adoption: a perceived risk facets perspective. Int J Hum Comput Stud 2003 Oct;59(4):451-474. [doi: 10.1016/S1071-5819(03)00111-3]
19. Agarwal R, Prasad J. A conceptual and operational definition of personal innovativeness in the domain of information technology. Inf Syst Res 1998 Jun;9(2):204-215. [doi: 10.1287/isre.9.2.204]
20. Bhattacherjee A, Hikmet N. Physicians' resistance toward healthcare information technology: a theoretical model and empirical test. Eur J Inf Syst 2007 Dec;16(6):725-737. [doi: 10.1057/palgrave.ejis.3000717]
21. Hsieh PJ. Physicians' acceptance of electronic medical records exchange: an extension of the decomposed TPB model with institutional trust and perceived risk. Int J Med Inform 2015 Jan;84(1):1-14. [doi: 10.1016/j.ijmedinf.2014.08.008] [Medline: 25242228]
22. Simarmata MTA, Hia IJ. The role of personal innovativeness on behavioral intention of information technology. J Econ Bus 2020;1(2):18-29. [doi: 10.36655/jeb.v1i2.169]
23. Kijsanayotin B, Pannarunothai S, Speedie SM. Factors influencing health information technology adoption in Thailand's community health centers: applying the UTAUT model. Int J Med Inform 2009 Jun;78(6):404-416. [doi: 10.1016/j.ijmedinf.2008.12.005] [Medline: 19196548]
24. Alabdullah JH, Van Lunen BL, Claiborne DM, Daniel SJ, Yen CJ, Gustin TS. Application of the unified theory of acceptance and use of technology model to predict dental students' behavioral intention to use teledentistry. J Dent Educ 2020 Nov;84(11):1262-1269. [doi: 10.1002/jdd.12304] [Medline: 32705688]
25. Ursachi G, Horodnic IA, Zait A. How reliable are measurement scales? External factors with indirect influence on reliability estimators. Proc Econ Finance 2015;20:679-686. [doi: 10.1016/S2212-5671(15)00123-9]
26. Ahmed Z, Bhinder KK, Tariq A, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Pakistan: a cross-sectional online survey. Ann Med Surg 2022;76. [doi: 10.1016/j.amsu.2022.103493]
27. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep 2023 Mar;6(3):e1138. [doi: 10.1002/hsr2.1138] [Medline: 36923372]
28. Wu W, Zhang B, Li S, Liu H. Exploring factors of the willingness to accept AI-assisted learning environments: an empirical investigation based on the UTAUT model and perceived risk theory. Front Psychol 2022;13:870777. [doi: 10.3389/fpsyg.2022.870777] [Medline: 35814061]
29. Almahri FAJ, Bell D, Merhi M. Understanding student acceptance and use of chatbots in the United Kingdom universities: a structural equation modelling approach. Presented at: 2020 6th International Conference on Information Management (ICIM); Mar 27-29, 2020; London, United Kingdom. [doi: 10.1109/ICIM49319.2020.244712]
30. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 2020 Aug;26(4):2141-2168. [doi: 10.1007/s11948-019-00165-5] [Medline: 31828533]
31. Tian W, Ge J, Zhao Y, Zheng X. AI chatbots in Chinese higher education: adoption, perception, and influence among graduate students-an integrated analysis utilizing UTAUT and ECM models. Front Psychol 2024;15:1268549. [doi: 10.3389/fpsyg.2024.1268549] [Medline: 38384353]
32. Lambert SI, Madi M, Sopka S, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. NPJ Digit Med 2023 Jun 10;6(1):111. [doi: 10.1038/s41746-023-00852-5] [Medline: 37301946]

Abbreviations

AI: artificial intelligence
BI: behavioral intention
EE: effort expectancy
FC: facilitating conditions
PE: performance expectancy
PI: personal innovativeness
PR: perceived risk
RB: resistance bias

XSL•FO
RenderX

SI: social influence
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by B Lesselroth; submitted 26.02.24; peer-reviewed by L Lu, Y Yan; revised version received 23.06.24; accepted 15.08.24; published 28.10.24.

Please cite as:
Tao W, Yang J, Qu X
Utilization of, Perceptions on, and Intention to Use AI Chatbots Among Medical Students in China: National Cross-Sectional Study
JMIR Med Educ 2024;10:e57132
URL: https://mededu.jmir.org/2024/1/e57132
doi:10.2196/57132

© Wenjuan Tao, Jinming Yang, Xing Qu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.10.2024. 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.

XSL•FO
RenderX

Leveraging the Electronic Health Record to Measure Resident Clinical Experiences and Identify Training Gaps: Development and Usability Study

Vasudha L Bhavaraju[1], MEd, MD; Sarada Panchanathan[2], MD; Brigham C Willis[3], MEd, MD; Pamela Garcia-Filion[2], MPH, PhD

1

2

3

Corresponding Author:
Vasudha L Bhavaraju, MEd, MD

Abstract

Background: Competence-based medical education requires robust data to link competence with clinical experiences. The SARS-CoV-2 (COVID-19) pandemic abruptly altered the standard trajectory of clinical exposure in medical training programs. Residency program directors were tasked with identifying and addressing the resultant gaps in each trainee's experiences using existing tools.

Objective: This study aims to demonstrate a feasible and efficient method to capture electronic health record (EHR) data that measure the volume and variety of pediatric resident clinical experiences from a continuity clinic; generate individual-, class-, and graduate-level benchmark data; and create a visualization for learners to quickly identify gaps in clinical experiences.

Methods: This pilot was conducted in a large, urban pediatric residency program from 2016 to 2022. Through consensus, 5 pediatric faculty identified diagnostic groups that pediatric residents should see to be competent in outpatient pediatrics. Information technology consultants used *International Classification of Diseases*, *Tenth Revision* (*ICD-10*) codes corresponding with each diagnostic group to extract EHR patient encounter data as an indicator of exposure to the specific diagnosis. The frequency (volume) and diagnosis types (variety) seen by active residents (classes of 2020‐2022) were compared with class and graduated resident (classes of 2016‐2019) averages. These data were converted to percentages and translated to a radar chart visualization for residents to quickly compare their current clinical experiences with peers and graduates. Residents were surveyed on the use of these data and the visualization to identify training gaps.

Results: Patient encounter data about clinical experiences for 102 residents (N=52 graduates) were extracted. Active residents (n=50) received data reports with radar graphs biannually: 3 for the classes of 2020 and 2021 and 2 for the class of 2022. Radar charts distinctly demonstrated gaps in diagnoses exposure compared with classmates and graduates. Residents found the visualization useful in setting clinical and learning goals.

Conclusions: This pilot describes an innovative method of capturing and presenting data about resident clinical experiences, compared with peer and graduate benchmarks, to identify learning gaps that may result from disruptions or modifications in medical training. This methodology can be aggregated across specialties and institutions and potentially inform competence-based medical education.

(JMIR Med Educ 2024;10:e53337) doi:10.2196/53337

KEYWORDS

clinical informatics; electronic health record; pediatric resident; COVID-19; competence-based medical education; pediatric; children; SARS-CoV-2; clinic; urban; diagnosis; health informatics; EHR; individualized learning plan

Introduction

Medical education is traditionally time-based, which presumes that learners will meet professional standards in a predetermined period of time, whereas competence-based medical education proposes an outcomes-based approach framed by competencies [1]. This latter approach requires robust data to measure outcomes and link them to competence. One such data set is the number and variety of clinical diagnoses that learners see, grounded in Kolb's framework that emphasizes hands-on experiences and reflection as a basis for experiential learning [2].

Unlike surgical specialties requiring minimum case numbers for procedural competence, nonprocedural specialties do not

XSL•FO
RenderX

endorse minimum numbers of diagnoses trainees should see to be considered competent. Literature exists on this topic across medical specialties [3-9]; however, the methods used to collect volume and variety of clinical cases are frequently incomplete, limiting their use. The variability of literature on this topic may stem from the challenge of capturing these patient experiences in a straightforward, accurate, and abstractable form.

The electronic health record (EHR) is useful for collecting patient encounters for quality improvement and business analytics. It has been incorporated in continuing professional development for practicing physicians to drive practice change [10]. Graduate medical education has also used EHR data to measure training outcomes. In 2023, Lees et al [3] performed a systematic review of the published uses of EHR data to measure competencies in medical trainees. The most common study theme identified was “trainee condition experience,” or the trainees’ involvement in patients with specific medical conditions. While study authors commonly mapped raw EHR data to diagnostic groupings and compared them with national standards or in-training examinations to identify gaps in training, there were limitations in utilization of these data. For example, studies documenting residents’ exposure to patient experiences, such as reporting the volume of diagnoses seen, often excluded important variables such as variety of diagnoses. Others examined data in aggregate rather than individualized data as is needed to link exposure and resident competence [11-16]. There are additional studies that compared individual resident clinical exposures with peer averages using the EHR and most commonly displayed these data using dashboards; however, they did not include benchmark data which provide a necessary framework to analyze the information [3,17-19].

Since 2020, the SARS-CoV-2 (COVID-19) pandemic has provided an opportunity to examine variability in diagnoses exposure for residents and to extrapolate its impact on their education. From June 2020 to February 2021, Yarahuan et al [20] noted a significant decrease in notes authored by pediatric interns on common inpatient diagnoses, on both respiratory and nonrespiratory conditions, compared with the prepandemic group. This variability resulted from the shifting prevalence of seasonal diagnoses and altered patient exposure due to practices such as “platooning” trainees for workforce preservation, shifting trainees from ambulatory to inpatient settings, and implementing telehealth [20-24]. In response, medical education leaders and learners were tasked with identifying gaps in clinical exposure compared with prepandemic standards and creating individualized learning plans; this needs assessment, however, was largely based on recall of clinical experiences in training rather than objective data [22,25,26].

In 2019, Sebok-Syer et al [27] analyzed resident and faculty feedback about the potential use of EHR data to assess gaps and inform trainees’ learning plans. The authors found that while these data may be valuable to support formative assessment practices, the data, in isolation, would portray an incomplete picture of the trainee and require context for interpretation. Meaningful analysis and presentation of EHR data are necessary in order to explore how volume and variety of clinical experiences may objectively identify gaps and inform competence.

The purpose of this pilot was to establish a process in our residency to extract meaningful EHR data for measuring clinical exposure and address associated gaps in the literature specifically to (1) develop a feasible and efficient method to capture EHR data that measure patient experiences of individual residents; (2) offer context to these data by comparing individual resident metrics to classmates and aggregated graduate residents’ data; and (3) create a visualization that provides residents and program directors with a snapshot of the volume and variety of trainees’ clinical experiences to allow quick identification of training gaps to inform focused learning plans.

Methods

This pilot study was conducted from 2019 to 2021 in a large, urban, pediatric residency program with multiple institutional sites. To assess feasibility, we focused on ambulatory diagnoses at 1 pediatric continuity clinic site. We chose this site since it had a larger volume of general pediatric patients with fewer complex medical needs than the other continuity clinics. Subjects were limited to pediatric residents and excluded rotating residents and students as we were seeking longitudinal clinical experiences and these latter 2 groups completed only 1 block rotation in the clinic. The resident patient panels at this clinic site are a combination of patients assigned by schedulers and those recruited by residents from other settings within the health care system, such as the nursery or inpatient unit. The residents generally stay with the same faculty preceptor for 3 years and have increasing levels of autonomy during the patient visit including billing and coding; however, billing and coding are always verified by the preceptor prior to closing the encounter.

This was a retrospective analysis of EPIC (Epic Systems) EHR metadata of ambulatory clinic notes authored by pediatric residents at this clinic site from 2013 to 2020. This was true EHR metadata attached to the note, not extracted from administrative claims data. Resident data from the graduating classes of 2016, 2017, 2018, and 2019 were used as the graduation benchmarks. Data were extracted up to April 2020, representing nearly 3 years of data from the class of 2020, 2 years from the class of 2021, and 1 year from the class of 2022 (Table 1).

XSL•FO
RenderX

Table . Resident electronic health record data.

Class of	Count, n	Resident category at the time of study	Dates of data extraction	Data used for	Report distribution dates
2016	8	Graduated	7/2013-6/2016	Graduate benchmark	N/A[a]
2017	15	Graduated	7/2014-6/2017	Graduate benchmark	N/A
2018	18	Graduated	7/2015-6/2018	Graduate benchmark	N/A
2019	11	Graduated	7/2016-6/2019	Graduate benchmark for 9/2019 and 4/2020 reports	N/A
2020	18	Active	7/2017–6/2020	Individual reports and 2020 class benchmarks	4/2019, 9/2019, 4/2020
2021	19	Active	7/2018-4/2020	Individual reports and 2021 class benchmarks	4/2019, 9/2019, 4/2020
2022	13	Active	7/2019-4/2020	Individual reports and 2022 class benchmarks	9/2019, 4/2020

[a]N/A: not applicable.

Ethical Considerations

The Phoenix Children's Hospital institutional review board has determined that this project involves quality improvement and does not meet the definition of research; therefore, the approval of the institutional review board was not required and this study was deemed exempt.

Key Stakeholders

Project stakeholders included the residency program director, residency program coordinator, and ambulatory clinic faculty preceptors. For information technology (IT) support, we engaged data analysts who recognized that graduate medical education was connected to the hospital business model and therefore supported this opportunity for improved billing and coding through EHR data analysis. Our pediatric residents were also vital participants in this pilot and were aware of its planning and rollout.

Diagnoses Set

Five general pediatricians from 3 clinic sites determined the key diagnostic groups that pediatric residents should see to be competent for independent outpatient practice. The group created a shared mental model with inclusion and exclusion criteria. For example, high-volume diagnoses (eg, pediatric well-checks) and low-volume, yet important diagnoses (eg, gait abnormality) were included. Common, self-limited conditions (eg, upper respiratory infections) were intentionally excluded presuming that residents in our busy clinics receive adequate exposure of these during residency, and the addition of these common diagnoses in a data report may distract from the more actionable data. The final list was generated through several rounds of review and consensus. Each diagnostic group was converted to *International Classification of Diseases, Tenth Revision* (*ICD-10*) codes (Table 2).

XSL•FO
RenderX

Table . Diagnostic groups for clinical experiences of pediatric residents and associated *ICD-10*[a] codes.

Group name	*ICD-10* codes	Notes (inclusions)
Well check	Z00.129, Z00.121	N/A[b]
Anemia	D50-D64	N/A
Constipation	K59.xx	N/A
Vomiting/diarrhea	R11.1, R11.2, R19.7, A09	N/A
Underweight/failure to thrive	R62.51	N/A
Gait problem/limp	R26.xx	In toeing, limp, genu varus, genu valgus
Genitourinary concerns	N43.xx, N47.xx, N48.xx, N90.89, K40.xx, K41.xx, Q53.xx, Q54.xx, Q55.xx	Hydroceles, phimosis, labial adhesions, hernias
Overweight/obesity, increased BMI	Z68.51-Z68.54, E66.3, E66.9, E66.09	N/A
Sexually transmitted infections	A50.0-A64, Z11.3-Z11.9	Screening and management
Asthma	J45.xx	N/A
Eczema	L30.8, L30.9, L20.9, L20.82, L20.83, L20.84, L21.1	N/A
Heart murmurs	R01.xx, I35.8, Q21.xx-Q24.xx	Functional and pathologic
Ear infections	H65.xx-H66.xx, H60.xx	Otitis media and variants, otitis externa
Urinary tract infection	N10, N30.xx, N39	N/A
Developmental delay	F80.xx, F82, F88, F89, R62.50	N/A
Behavioral/ADHD[c]	F90.0 - 90.2, F90.8 - 90.9, F91.0 - 91.3, F91.8 - 91.9, F93.0, F93.8 - 94.2, F94.8 - 94.9, F95, F98, F30-39.9999, F40-48.9999	Depression, anxiety, ADHD with all variants
Young women’s health	Z30.xx, N92.6, N93.9, N94.3 - 97	Contraception, menstrual concerns
Headache	G43.xx-G44.xx, R51	N/A
Autism spectrum disorder	F84.xx	N/A
Genetic and chromosomal disorders	Q90-Q99.xx	N/A
Specific congenital nongenetic disorders	Q35.xx-37.xx, Q05.xx	Spina bifida and variants, cleft lip and palate
Vaccine hesitancy	Z28.xx	N/A

[a]*ICD-10*: *International Classification of Diseases, Tenth Revision*.

[b]N/A: not applicable.

[c]ADHD: attention-deficit/hyperactivity disorders.

EHR Data Extraction

The project data set, including resident data, patient data, and encounter data, was abstracted from the EHR. To ensure accuracy and completeness of the data, we performed an iterative process with our IT consultants, starting with smaller subsets of data with 1 resident, to ensure that each piece of data pulled was relevant to the pilot before expanding to larger subsets and more residents. Variety was determined through *ICD-10* codes for all visit diagnoses per encounter and volume was measured as the number of unique visits. When residents authored an encounter note, they were attributed to that patient and his or her associated diagnoses.

A flat file of the EHR data was imported to Microsoft Excel and analyzed using the pivot table function. Pivot tables were created to enumerate the volume of patients for each diagnosis by individual resident. Individual resident data were aggregated to the respective class level of postgraduate year (PGY1, PGY2, or graduated) to calculate average volumes. An Excel worksheet was created for each individual resident to summarize the resident’s volume (column) and variety (rows) of clinical experiences. For comparison, the class and graduated average volumes were appended as columns. These worksheet data were used to generate the visualizations.

Visualization and Data Report

To facilitate the assessment of training gaps, we used Microsoft Excel to translate tabular data into a radar chart (or spider graph) visualization, which is a 2D graphical method of illustrating multiple quantitative variables on axes (eg, diagnostic categories) with the same starting point. Since these data have vastly different scales on the same chart, the data were converted from raw numbers into percentages using the number of patient encounters seen for each diagnostic category (numerator) divided by the average number seen by graduated residents (denominator). The radar chart was rendered to illustrate the percentage of clinical experiences per diagnostic category for

XSL•FO
RenderX

the individual resident and the class average compared with graduated residents (benchmark) as the maximal total area. Not only was this an ideal prototype for this pilot, our residents and faculty were already familiar with using radar graphs for milestone data, in which individual resident progress across a range of competencies is compared with classmates and graduates. We did trial other visualizations, including a dot plot with error bars, but found that these did not give an accurate picture because not every diagnosis was normalized at the same value.

Two comparative radar chart visualizations were created: (1) the percent volume of clinical experiences by the individual resident versus the class average and graduated residents (for individual resident review), and (2) the percent volume of clinical experiences between the aggregated classes (PGY1 and PGY2) and graduated residents (for program leadership review).

Reports with visualizations were distributed approximately every 6 months to align with semiannual reviews with clinic preceptors. Prior to distribution, residents and preceptors were educated on using the reports to stimulate discussion on learning goals.

To enable ongoing data extraction and reports, we identified 3 vital team members: the IT champion to initiate the data extraction into an Excel spreadsheet, the residency program coordinator to provide an updated list of residents at the start of each academic year and transform the raw data into individual reports with radar graph visualizations, and the clinic champion to distribute the reports and educate faculty and residents. The process was semiautomated as once the query was set up, it could be run the same way semiannually to produce a spreadsheet of every patient visit by every resident with benchmark averages calculated.

Resident Postimplementation Surveys

Residents from the classes of 2021 and 2022 were surveyed on their individual data report readability and specific utilization for setting clinical and coding goals. The survey was homegrown by authors without validity evidence. It contained 6 multiple-choice and 2 open-ended questions and was delivered via email with instructions for completion. Results were collected anonymously. Multiple-choice questions were analyzed using frequency of responses and open-ended questions were grossly interpreted for themes and representative quotes.

Results

We extracted information about clinical experiences for 102 residents including 52 graduated residents for the graduate benchmark and 50 active residents for individual reports and class benchmarks (Table 1). Residents from the classes of 2020 and 2021 received 3 data reports and those from the class of 2022 received 2.

Table 3 displays data for an individual resident alongside the class and graduated residents' averages to enable residents to follow their progress against internal benchmarks. Figure 1 uses data from Table 3 to visualize these data in a radar chart. This visualization method makes deficient areas immediately apparent to the resident and identifies which experiences must be intentionally pursued. The second semiannual report was created in Spring of the academic year when equity in clinical rotation experiences is assumed within a class and peer averages are more accurate.

The change in the class-level volumes versus graduate class volumes (Table 4) demonstrated that average volume for each diagnosis group increased progressively with each year. The visualization of this aggregated information (Figure 2) can be used by program leaders for tracking general trends in diagnoses exposure year-to-year.

XSL•FO
RenderX

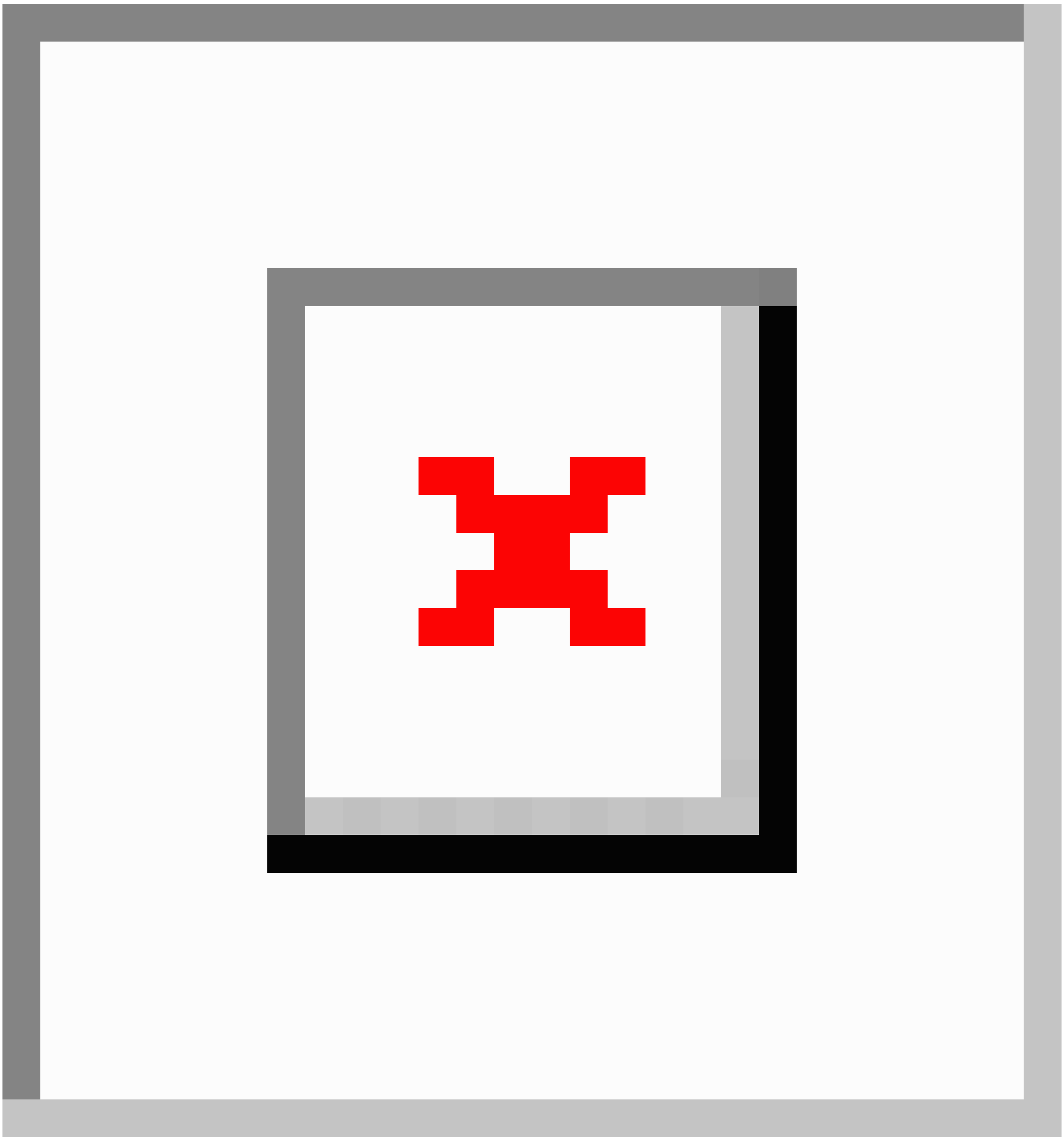
Table . Sample table with average numbers of patients seen with each diagnosis in continuity clinic, by a single postgraduate year 2 (PGY2) resident, compared with the PGY2 class average and with the average numbers seen by recently graduated residents in this program.

Diagnosis description	PGY2 resident	Class average (PGY2)	Graduated residents
Well check	463	454	863
Anemia	5	7	11
Constipation	32	33	54
Vomiting/diarrhea	9	9	13
Underweight/failure to thrive	51	17	28
Gait problem/limp	7	4	4
Genitourinary concerns	23	14	20
Overweight/obesity/increased BMI	69	79	104
Sexually transmitted infections	0	3	6
Asthma	18	25	58
Eczema	26	28	47
Heart murmurs	16	11	24
Ear infections	15	11	25
Urinary tract infection	0	2	3
Developmental delay	33	24	54
Behavioral/ADHD[a]	5	9	15
Young women’s health	0	8	8
Headache	3	8	16
Autism spectrum disorder	1	3	5
Genetic and chromosomal disorders	1	7	10
Specific congenital nongenetic disorders	0	10	4
Vaccine hesitancy	10	19	38

[a]ADHD: attention-deficit/hyperactivity disorders.

XSL•FO
RenderX

Figure 1. Individual resident profile with class average benchmarked against graduated residents. The diagnostic categories were deliberately placed randomly rather than ordered from high to low percentage so that residents would focus on individual categories rather than the extremes. ADHD: attention-deficit/hyperactivity disorders; FTT: failure to thrive; GU: genitourinary; UTI: urinary tract infection.



XSL•FO
RenderX

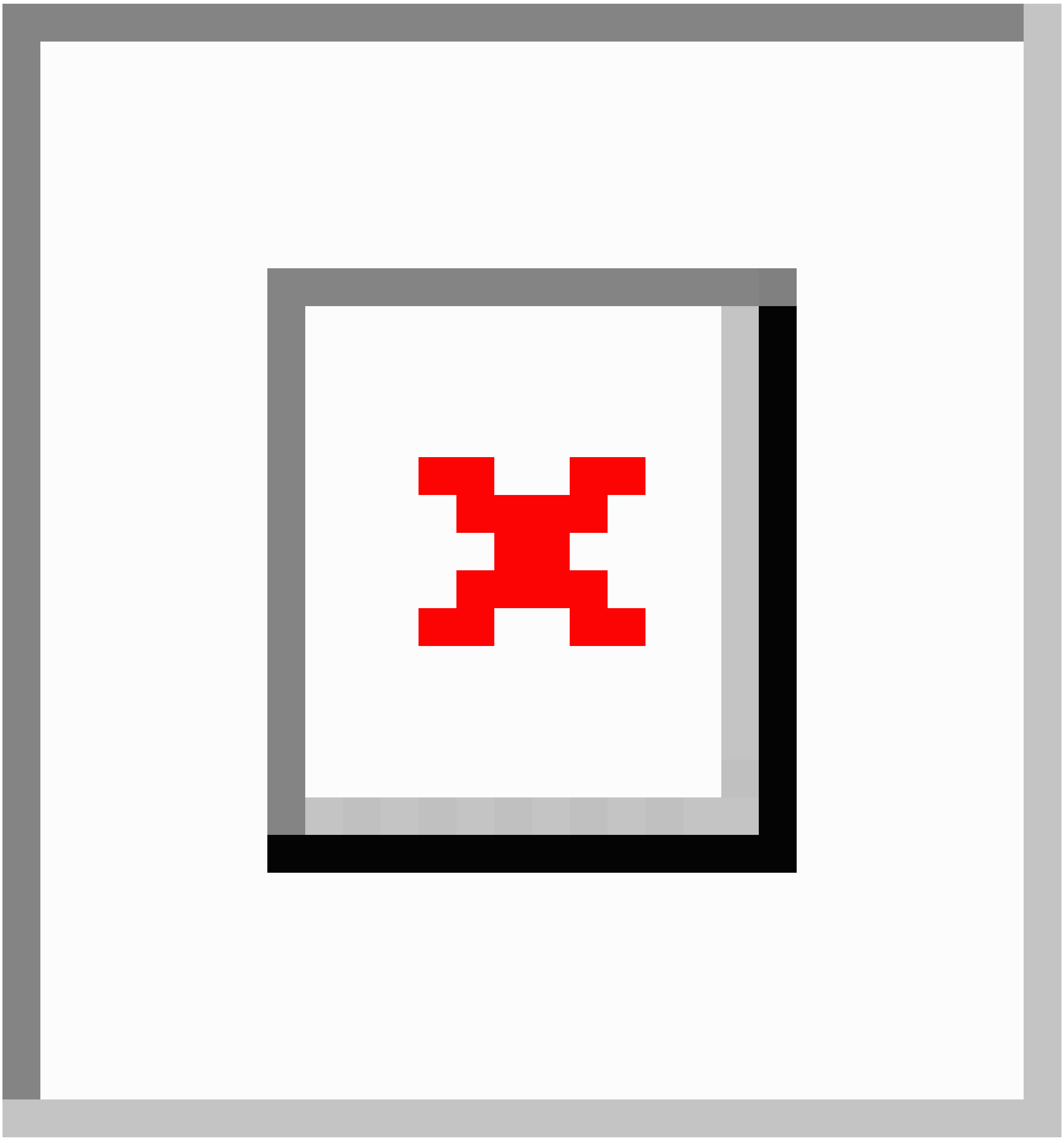
Table . Average numbers of patients seen with each diagnosis in continuity clinic by postgraduate year class and graduated residents.

Diagnosis description	Class: PGY1[a]	Class: PGY2[b]	Graduated residents
Well check	150	454	863
Anemia	2	7	11
Constipation	11	33	54
Vomiting/diarrhea	3	9	13
Underweight/failure to thrive	5	17	28
Gait problem/limp	3	4	4
Genitourinary concerns	4	14	20
Overweight/obesity/Increased BMI	27	79	104
Sexually transmitted infections	2	3	6
Asthma	10	25	58
Eczema	9	28	47
Heart murmurs	7	11	24
Ear infections	4	11	25
Urinary tract infection	2	2	3
Developmental delay	8	24	54
Behavioral/ADHD[c]	4	9	15
Young women's health	3	8	8
Headache	3	8	16
Autism spectrum disorder	2	3	5
Genetic and chromosomal disorders	3	7	10
Specific congenital nongenetic disorders	3	10	4
Vaccine delay/refusal	7	19	38

[a]PGY1: postgraduate year 1.
[b]PGY2: postgraduate year 2.
[c]ADHD: attention-deficit/hyperactivity disorders.

XSL•FO
RenderX

Figure 2. Class averages benchmarked against graduated residents. ADHD: attention-deficit/hyperactivity disorders; FTT: failure to thrive; GU: genitourinary; PGY1: postgraduate year 1; PGY2: postgraduate year 2; UTI: urinary tract infection.



We surveyed graduating classes of 2021 and 2022 following implementation of this project and report distribution; 53% (17/32) of the residents responded. Of respondents, 69% (11/16) reviewed their reports with their clinic preceptors, 44% (7/16) used their reports to make clinical goals, such as "see more adolescent patients" and "increase comfort dealing with vaccine hesitancy," and 38% (6/16) used the reports to make coding goals, such as "include Z-codes regarding counseling" and "bill more on top of well-checks." Gross interpretation of open-ended comments showed that residents found the radar chart easy to interpret and to identify in which areas they have had less exposure than their classmates.

Discussion

Principal Findings

Our pilot study demonstrated an innovative method to collaborate with IT and leverage EHR data to measure and display the volume and variety of clinical experiences, relative to peers and previous program graduates, in a pediatric residency program continuity clinic. We presented the data in a functional manner to pinpoint gaps that may result from disruptive events such as the SARS-CoV-2 pandemic. For example, in Figure 1, this resident may recognize that he or she has not seen nor coded patients with young women's health concerns, specific genetic or congenital disorders, or vaccine delay or refusal. He or she

XSL•FO
RenderX

could, therefore, intentionally choose an adolescent or genetics elective as well as ensure that when seeing patients for well-checks, any additional diagnoses, such as vaccine refusal, are coded. While radar charts are not novel for representing competence in medical education, our visualization has not been previously used for this purpose [28,29]. Utilization of reports for this type of goal setting does require preceptor education and comfort with the tool, which can be achieved through faculty development and consistent use.

We believe that this methodology can be used for programs of any specialty, size, or setting. This list of diagnoses can be easily created in internal medicine or dermatology or in subspecialties such as pediatric cardiology. While the amount of resident data used in class or graduated benchmarks may be decreased in smaller-sized programs, an individual resident can still use the data to evaluate progress and set goals. Community settings may have less breadth of diagnoses than academic settings, but this substantiates the value of this innovation; if these training programs were able to review similar data from larger programs, they may be able to examine trends and program gaps that require supplementary clinical or nonclinical experiences (eg, focused didactics, external specialty rotations, or simulation). Notably, while these reports are not intended to inform summative evaluation of resident performance or any high-stakes training decisions, the data provide objectivity and specificity in resident experiences that may enrich the feedback between preceptor and resident.

The ability to access objective data on the clinical experiences of current residents compared with prior years is indispensable for program-wide or individual events that disrupt patient exposure during training, such as rotation-site closures or extended leaves of absence. Obtaining these data is feasible and can be automated with each new class. Moreover, this process may be modified to accommodate the changing landscape of medicine. New diagnoses, such as "exposure to COVID-19," can be added to the EHR reports. Additional metadata, such as number of telehealth visits or time-to-note-completion, can also be extracted to create a comprehensive individualized "report card" of metrics, as described by Sebok-Syer et al [30], to enhance resident feedback and assessment. For these metrics, data from periodic reports (eg, semiannually), rather than from real-time dashboards, appear to be more beneficial for the recipient to set learning goals, as these data represent trends in experiences or practices over time. Furthermore, peer benchmarks are more reliable in the periodic reports, as residents in the same class will generally complete similar clinical rotations as the academic year progresses.

Limitations

One limitation in our methodology is its dependence on accurate and complete coding of all diagnoses addressed at a patient encounter, which is often performed by residents in the clinic setting. Some diagnoses for which only discussion was required (eg, vaccine refusal) may be underrepresented and lead to gaps as noted in Figure 1. As trainees become familiar with the data, they can differentiate a lack of coding from lack of clinical exposure. In addition, a true lack of clinical exposure may be seen with important but uncommon diagnoses, and it may be harder to estimate a consistent goal number of patients to seek with these diagnoses. This may lead residents to presume that they will be less successful in managing these diagnoses should they encounter them in the future. In these cases, we rely on our faculty preceptors, when reviewing the reports with residents, to offer perspective and strategies to gain knowledge in advance or "in the moment" when encountering rare diagnoses.

Since this was a pilot study to determine feasibility, we opted to use small-group consensus to determine the diagnostic categories rather than established resources, such as certifying board examination content specifications. We also acknowledge that many diagnostic categories identified, such as urinary tract infection and asthma, are seen in other settings where the residents rotate (eg, emergency department and urgent care) thus offering an incomplete number of total exposures. In addition, there are common diagnoses, such as pain management and mental health disorders, which are not present on the list. We made these decisions as this was a pilot study limited to a single setting with a finite list of diagnoses to demonstrate proof of concept. We anticipate that expanding the list of diagnoses, designating specific categories by age groups, and implementing the process across other clinical settings would offer more representative data.

Another limitation is that the resident survey measuring acceptability and utilization of the reports was not a validated tool and was sent 1 year after the last report distribution, likely leading to recall bias and a lower rate of return. A standardized usability survey distributed in a timelier manner would have strengthened these results. The authors also recognize that while we found our business analysts, rather than clinical informaticians, to be our IT champions, this is institution specific. We encourage readers to explore all potential partnerships between IT and graduate medical education if embarking on a similar project. Finally, despite efforts to automate the process to semiannually extract data for individual resident reports, the project stalled after our 3 main team members, the residency coordinator and IT and faculty champions, left the institution within a short period of time. We were, therefore, unable to study the outcomes of learning goals set by residents and the distribution and utilization of reports for future classes. We learned that expanding teams to allow for cross-training of tasks, proper timing and transitions of responsibilities, and creating standardized operating procedures are essential for sustainability.

Next Steps

Within our residency program, we have identified new champions to reinvigorate this process for our clinics and expand to the inpatient and emergency department settings using a set of diagnoses unique for each location. With additional data sets across varied clinical settings, we anticipate that the trends in volume and variety will be more reflective of the complete resident experience. The authors understand that comparing data internally within a program is not the ideal "gold standard" to measure competence when compared with more standardized benchmarks. Moving forward, this method can be shared across specialties and institutions to develop national benchmarks on

XSL•FO
RenderX

the average volume and variety of patient encounters trainees see and provide a measure for programs to compare their experiences with others and identify gaps in training. Once these benchmarks are compared with other measures of competence, such as milestone assessment ratings, certifying examination scores, and postgraduate performance, we can better inform competence-based medical education and fill the gap in the literature on this topic.

Conclusions

Medical education requires robust data to measure outcomes but gathering data about clinical encounters and making them meaningful can be challenging. This pilot describes a feasible method of capturing resident clinical experiences from the EHR, setting internal benchmarks using class and graduated residents' averages, and creating a radar chart visualization that allows learners to quickly identify gaps in their training.

Acknowledgments

The authors wish to acknowledge the Valleywise Health business analytics team for providing the information technology expertise and resources for this project.

Disclaimer

This article reflects the views of the authors and should not be construed to represent views or policies of the Food and Drug Administration.

Authors' Contributions

VLB contributed to data curation and writing (original draft, review, and editing). SP contributed to conceptualization, data curation, methodology, and writing (original draft, review, and editing). BCW contributed to writing (review and editing). PG-F contributed to formal analysis, visualization, and writing (original draft, review, and editing).

Conflicts of Interest

None declared.

References

1. Frank JR, Mungroo R, Ahmad Y, Wang M, De Rossi S, Horsley T. Toward a definition of competency-based education in medicine: a systematic review of published definitions. Med Teach 2010 Aug;32(8):631-637. [doi: 10.3109/0142159X.2010.500898]
2. Kolb D. Experiential Learning: Experiences as the Source of Learning and Development: Prentice-Hall; 1984.
3. Lees AF, Beni C, Lee A, et al. Uses of electronic health record data to measure the clinical learning environment of graduate medical education trainees: a systematic review. Acad Med 2023 Nov 1;98(11):1326-1336. [doi: 10.1097/ACM.0000000000005288] [Medline: 37267042]
4. Williams RG, Swanson DB, Fryer JP, et al. How many observations are needed to assess a surgical trainee's state of operative competency? Ann Surg 2019 Feb;269(2):377-382. [doi: 10.1097/SLA.0000000000002554] [Medline: 29064891]
5. Emans SJ, Bravender T, Knight J, et al. Adolescent medicine training in pediatric residency programs: are we doing a good job? Pediatrics 1998 Sep;102(3 Pt 1):588-595. [doi: 10.1542/peds.102.3.588] [Medline: 9738181]
6. Trainor JL, Krug SE. The training of pediatric residents in the care of acutely ill and injured children. Arch Pediatr Adolesc Med 2000 Nov 1;154(11):1154. [doi: 10.1001/archpedi.154.11.1154]
7. Kinoshita K, Tsugawa Y, Shimizu T, et al. Impact of inpatient caseload, emergency department duties, and online learning resource on General Medicine In-Training Examination scores in Japan. Int J Gen Med 2015;8:355-360. [doi: 10.2147/IJGM.S81920] [Medline: 26586961]
8. Mizuno A, Tsugawa Y, Shimizu T, et al. The impact of the hospital volume on the performance of residents on the general medicine in-training examination: a multicenter study in Japan. Intern Med 2016;55(12):1553-1558. [doi: 10.2169/internalmedicine.55.6293]
9. Sclafani A, Currier P, Chang Y, Eromo E, Raemer D, Miloslavsky EM. Internal medicine residents' exposure to and confidence in managing hospital acute clinical events. J Hosp Med 2019 Apr;14(4):218-223. [doi: 10.12788/jhm.3168] [Medline: 30933672]
10. Bucalon B, Whitelock-Wainwright E, Williams C, et al. Thought leader perspectives on the benefits, barriers, and enablers for routinely collected electronic health data to support professional development: qualitative study. J Med Internet Res 2023 Feb 16;25:e40685. [doi: 10.2196/40685] [Medline: 36795463]
11. Agarwal V, Bump GM, Heller MT, et al. Resident case volume correlates with clinical performance: finding the sweet spot. Acad Radiol 2019 Jan;26(1):136-140. [doi: 10.1016/j.acra.2018.06.023] [Medline: 30087064]
12. Li J, Roosevelt G, McCabe K, et al. Pediatric case exposure during emergency medicine residency. AEM Educ Train 2018 Oct;2(4):317-327. [doi: 10.1002/aet2.10130] [Medline: 30386842]

XSL•FO
RenderX

13. Li J, Roosevelt G, McCabe K, et al. Critically ill pediatric case exposure during emergency medicine residency. J Emerg Med 2020 Aug;59(2):278-285. [doi: 10.1016/j.jemermed.2020.04.047] [Medline: 32536497]
14. Douglass A, Yip K, Lumanauw D, Fleischman RJ, Jordan J, Tanen DA. Resident clinical experience in the emergency department: patient encounters by postgraduate year. AEM Educ Train 2019 Jul;3(3):243-250. [doi: 10.1002/aet2.10326] [Medline: 31360817]
15. McCoy CP, Stenerson MB, Halvorsen AJ, Homme JH, McDonald FS. Association of volume of patient encounters with residents' in-training examination performance. J Gen Intern Med 2013 Aug;28(8):1035-1041. [doi: 10.1007/s11606-013-2398-0] [Medline: 23595933]
16. Bischof JJ, Emerson G, Mitzman J, Khandelwal S, Way DP, Southerland LT. Does the emergency medicine in-training examination accurately reflect residents' clinical experiences? AEM Educ Train 2019 Oct;3(4):317-322. [doi: 10.1002/aet2.10381] [Medline: 31637348]
17. Sequist TD, Singh S, Pereira AG, Rusinak D, Pearson SD. Use of an electronic medical record to profile the continuity clinic experiences of primary care residents. Acad Med 2005 Apr;80(4):390-394. [doi: 10.1097/00001888-200504000-00017] [Medline: 15793025]
18. Levin JC, Hron J. Automated reporting of trainee metrics using electronic clinical systems. J Grad Med Educ 2017 Jun;9(3):361-365. [doi: 10.4300/JGME-D-16-00469.1] [Medline: 28638518]
19. Rajkomar A, Ranji SR, Sharpe B. Using the electronic health record to identify educational gaps for internal medicine interns. J Grad Med Educ 2017 Feb 1;9(1):109-112. [doi: 10.4300/JGME-D-16-00272.1]
20. Yarahuan JW, Bass L, Hess LM, Singhal G, Lo HY. COVID-19 impact on intern exposure to common inpatient diagnoses. Hosp Pediatr 2021 Nov 4. [doi: 10.1542/hpeds.2021-006077] [Medline: 34737218]
21. Geanacopoulos AT, Sundheim KM, Greco KF, et al. Pediatric intern clinical exposure during the COVID-19 pandemic. Hosp Pediatr 2021 Jul;11(7):e106-e110. [doi: 10.1542/hpeds.2021-005899] [Medline: 33863816]
22. Antoon JW, Williams DJ, Thurm C, et al. The COVID ‐ 19 pandemic and changes in healthcare utilization for pediatric respiratory and nonrespiratory illnesses in the United States. J Hosp Med 2021 May;16(5):294-297. [doi: 10.12788/jhm.3608]
23. Blankenburg R, Gonzalez Del Rey J, Aylor M, et al. The impact of the COVID-19 pandemic on pediatric graduate medical education: lessons learned and pathways forward. Acad Med 2022 Mar 1;97(3S):S35-S39. [doi: 10.1097/ACM.0000000000004532] [Medline: 34817400]
24. Mallon D, Pohl JF, Phatak UP, et al. Impact of COVID-19 on pediatric gastroenterology fellow training in North America. J Pediatr Gastroenterol Nutr 2020 Jul;71(1):6-11. [doi: 10.1097/MPG.0000000000002768] [Medline: 32369320]
25. American Board of Medical Subspecialties, American Osteopathic Association. Transitions in medical education practical guidance to support important transitions: residency to fellowship. 2022. URL: https://www.acgme.org/globalassets/documents/covid-19/medicaleducationtransitions.residencyfellowship.pdf [accessed 2024-10-11]
26. American Board of Medical Subspecialties, American Osteopathic Association. Transitions to clinical practice. practical guidance to support important transitions: residency and fellowship to practice. 2022. URL: https://www.acgme.org/globalassets/documents/covid-19/medicaleducationtransitions.gmetopractice.pdf [accessed 2024-10-11]
27. Sebok-Syer SS, Goldszmidt M, Watling CJ, Chahine S, Venance SL, Lingard L. Using electronic health record data to assess residents' clinical performance in the workplace: the good, the bad, and the unthinkable. Acad Med 2019 Jun;94(6):853-860. [doi: 10.1097/ACM.0000000000002672] [Medline: 30844936]
28. Harrington DT, Miner TJ, Ng T, Charpentier KP, Richardson P, Cioffi WG. What shape is your resident in? Using a radar plot to guide a milestone clinical competency discussion. J Surg Educ 2015;72(6):e294-e298. [doi: 10.1016/j.jsurg.2015.04.005] [Medline: 26143521]
29. Keister DM, Larson D, Dostal J, Baglia J. The radar graph: the development of an educational tool to demonstrate resident competency. J Grad Med Educ 2012 Jun;4(2):220-226. [doi: 10.4300/JGME-D-11-00163.1] [Medline: 23730445]
30. Sebok-Syer SS, Shaw JM, Sedran R, et al. Facilitating residents' understanding of electronic health record report card data using faculty feedback and coaching. Acad Med 2022 Nov 1;97(11S):S22-S28. [doi: 10.1097/ACM.0000000000004900] [Medline: 35947480]

Abbreviations

EHR: electronic health record
ICD-10: *International Classification of Diseases, Tenth Revision*
IT: information technology
PGY: postgraduate year

XSL•FO
RenderX

Edited by B Lesselroth; submitted 04.10.23; peer-reviewed by A Azizi, J Chaparro, KF Hollis, M Leu; revised version received 01.05.24; accepted 19.08.24; published 06.11.24.

Please cite as:
Bhavaraju VL, Panchanathan S, Willis BC, Garcia-Filion P
Leveraging the Electronic Health Record to Measure Resident Clinical Experiences and Identify Training Gaps: Development and Usability Study
JMIR Med Educ 2024;10:e53337
URL: https://mededu.jmir.org/2024/1/e53337
doi:10.2196/53337

© Vasudha L Bhavaraju, Sarada Panchanathan, Brigham C Willis, Pamela Garcia-Filion. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.11.2024. 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.

XSL•FO
RenderX

Original Paper

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

Sergio Guinez-Molinos[1*], MSc, PhD; Sonia Espinoza[2*], BEng; Jose Andrade[2*], BEng; Alejandro Medina[2*], BEng

[1]School of Medicine, Universidad de Talca, Talca, Chile

[2]Interoperability Area, National Center for Health Information System, Santiago, Chile

[*]all authors contributed equally

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

Abstract

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

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

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

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

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

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e45413) doi:10.2196/45413

KEYWORDS

interoperability; health information system; Health Level Seven International; HL7; Fast Healthcare Interoperability Resource; FHIR; certification; training; interoperable; e-learning; application programming interface; API

Introduction

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

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

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

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

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

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

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

In this sense, our goal was to develop an API that allows us to integrate and communicate a set of loosely coupled modules that enable teachers to manage FHIR training programs, designing courses, questions, and scenarios. In addition, learners can interact through a web client for self-learning sessions, where the API, in conjunction with an assessment engine, provides feedback for each attempt the learner makes. This undoubtedly streamlines the self-learning process and automates the correction of hundreds of CRUD operations and the submission of learner responses within a context that the platform delivers.

XSL•FO
RenderX

The design and development of a platform called Huemul support the creation of courses associated with multiple questions (which expect a CRUD operation as an answer), automate the evaluation of the responses, and provide automatic feedback to the students in each exercise. We have also created an administrator that allows us to create and manage courses, questions, and users.

Methods

Study Design

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

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

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

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

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

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

Requirements

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

XSL•FO
RenderX

Textbox 1. Functional requirements to design the system for teaching FHIR (Fast Healthcare Interoperability Resource).

1. Test and question manager:

- Users' management
- FHIR create, read, update, and delete (CRUD)-oriented test management
- FHIR CRUD operations
- CRUD courses
- Create and manage a database with questions, tests, and courses

For an FHIR test (where the context and the problem are explained), examples of questions could be:

- Create the patient with the information given in the description
- Create a medical encounter
- Modify the phone number and address of the doctor
- Delete the patient

2. Application Programming Interface (API) for orchestrating components:

- Users' authentication management
- Call up tests and questions
- Validate user answers
- Save user answers
- Execute FHIR CRUD operation on the server

3. Test evaluator:

- Evaluate answers
- Compare questions and answers
- Build resources with the HAPI FHIR library
- Validate resources with standard

The expected answer should be a CRUD operation for a FHIR test (where the context and the problem are explained). For example, for the creation of a patient, the student must complete the following:

- Method for creating a FHIR resource (post)
- [FHIR Endpoint]/patient (URL server and resource name)
- Patient data (JSON format; patient information)

4. Client application:

- Create responsive front end
- Communicate using the Huemul API
- Decoupled other components

XSL•FO
RenderX

Table 1. Tools, libraries, and relation with each software component.

Development area and tools or libraries	Related component				
	Engine	Admin	API[a]	Client	FHIR[b] server
Environment					
NetBeans	✓	✓	✓	✓	✓
IntelliJ CE	✓	✓	✓	✓	✓
Backend					
Python					
Python 3.6		✓	✓		
Celery		✓			
Django 3.1		✓	✓		
Django DRF 3.1			✓		
Java					
OpenJDK 11	✓				✓
Apache Maven	✓				✓
Apache Tomcat 9	✓				✓
HapiFhir 5.3	✓				✓
Front end					
Bootstrap 4.3				✓	
jQuery 3.1.1				✓	
Deployment					
Docker					
Docker Compose					
Database					
MySQL 5.7	✓	✓	✓	✓	✓

[a]API: application programming interface.

[b]FHIR: Fast Healthcare Interoperability Resource.

Software Development Methodology

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

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

1. Modeling and management: in the first iteration, a functional database model was generated with the objective that it can support the definition of models related to tests, users, questions, and courses and the creation of FHIR learning tests. In addition, an administration application (Huemul Admin) was created to maintain the generated models. Once the model was built, a REST API (Huemul API) was developed to consult the information.
2. Improvements to the data model and API: in the second iteration, improvements to the model were included with the analysis of the previous iterations, authentication and security features of the REST API, and the creation of a web client (Huemul Client) for the consumption and interaction of the REST API.
3. Response processing and evaluation: in the third iteration, models for response processing are included, an interface for sending responses to the web client is added, and an engine (Huemul Engine) for response evaluation is created. The administrator creates a test planning mechanism, setting start and end times.
4. Functional improvements and feedback: in the fourth iteration, modifications are introduced in the processing of answers, feedback in case of incorrect answers, and the enabling of a natural resource query interface.

Each developed component has a set of tools described in Table 1, the languages used are Python (Python Software Foundation) and Java (Oracle Corporation) in the backend, and all interaction between components involves using a REST API. In addition, the front end group has some traditional libraries for client development, as it uses another API to consume resources independently and does not restrict alternative clients.

XSL•FO
RenderX

Three full-time computer engineers and the leader of the CENS interoperability area worked on the platform to create the software. It took 6 months to develop the prototype and 1 month to make modifications during the pilot implementation.

Ethical Considerations

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

Results

Overview

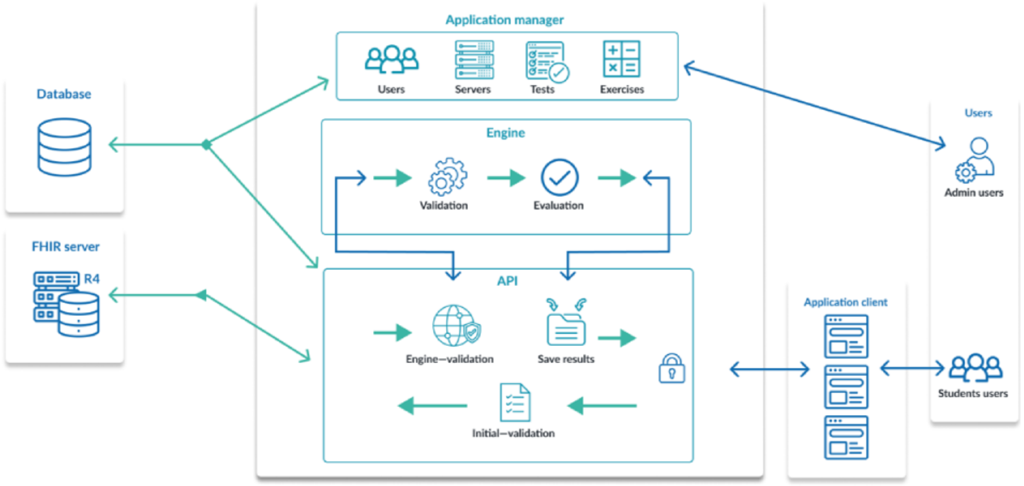
Huemul has 4 components that were designed and named considering the functional and nonfunctional requirements. Therefore, the following modules are necessary to develop a scalable and robust system:

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

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

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

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



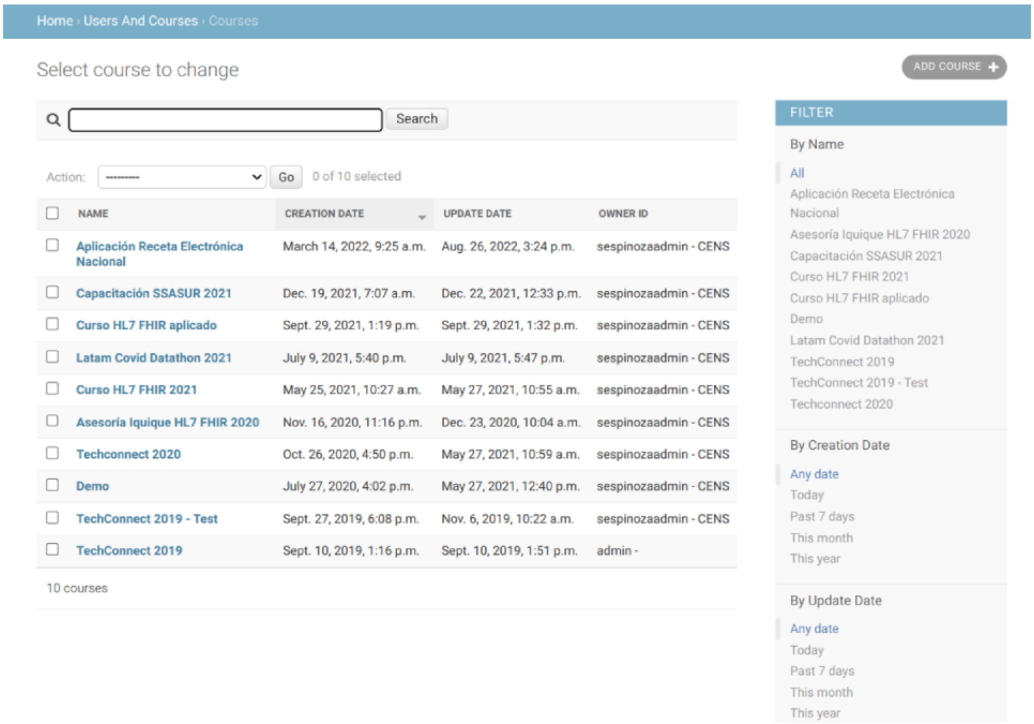
Huemul Admin

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

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

XSL•FO
RenderX

Figure 2. Huemul Admin component with active FHIR courses in the platform. CENS: National Center for Health Information System; FHIR: Fast Healthcare Interoperability Resource; HL7: Health Level Seven International.



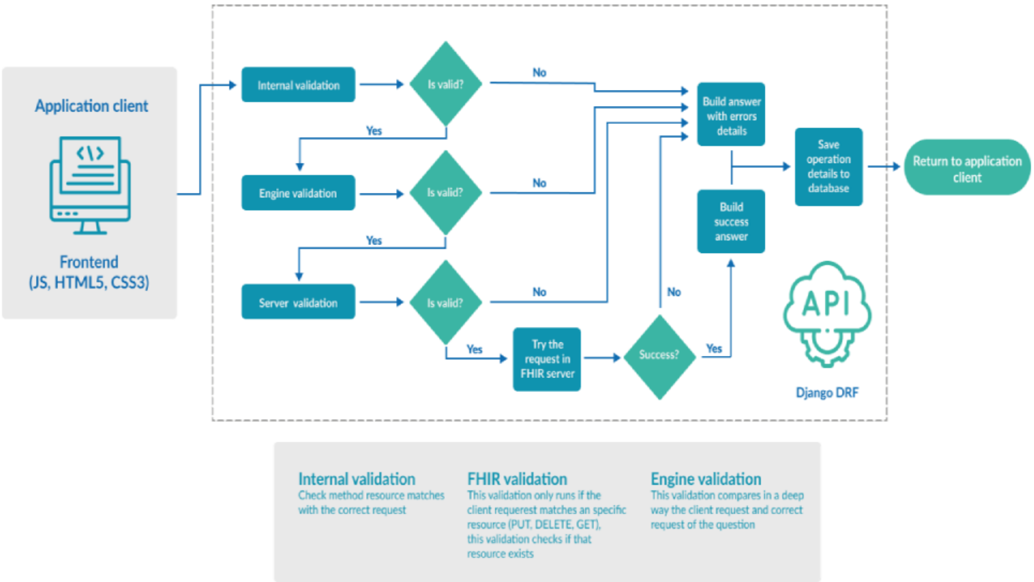
Huemul API

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

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

XSL•FO
RenderX

Figure 3. Huemul API component that communicates with all the components of the system. API: application programming interface; FHIR: Fast Healthcare Interoperability Resource.



Huemul Engine

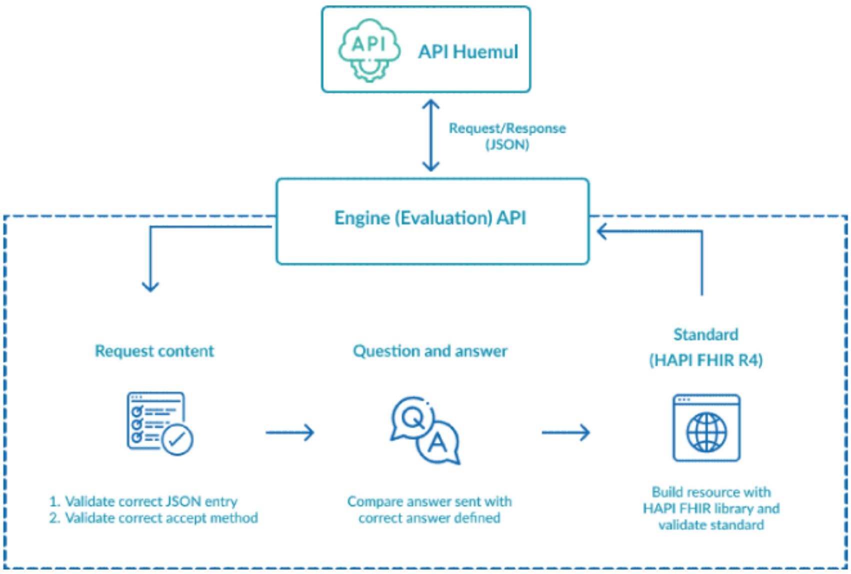
This component has the function of response evaluation, for which it evaluates 2 responses, the expected response and the user's response. The processing comprises 3 subprocesses to finally have an evaluation result that allows us to assess if the answer is correct or to assess the percentage of completeness (Figure 4).

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

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

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

Figure 4. Huemul Engine component that validates, evaluates, and builds the response. API: application programming interface; FHIR: Fast Healthcare Interoperability Resource.



Huemul Client

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

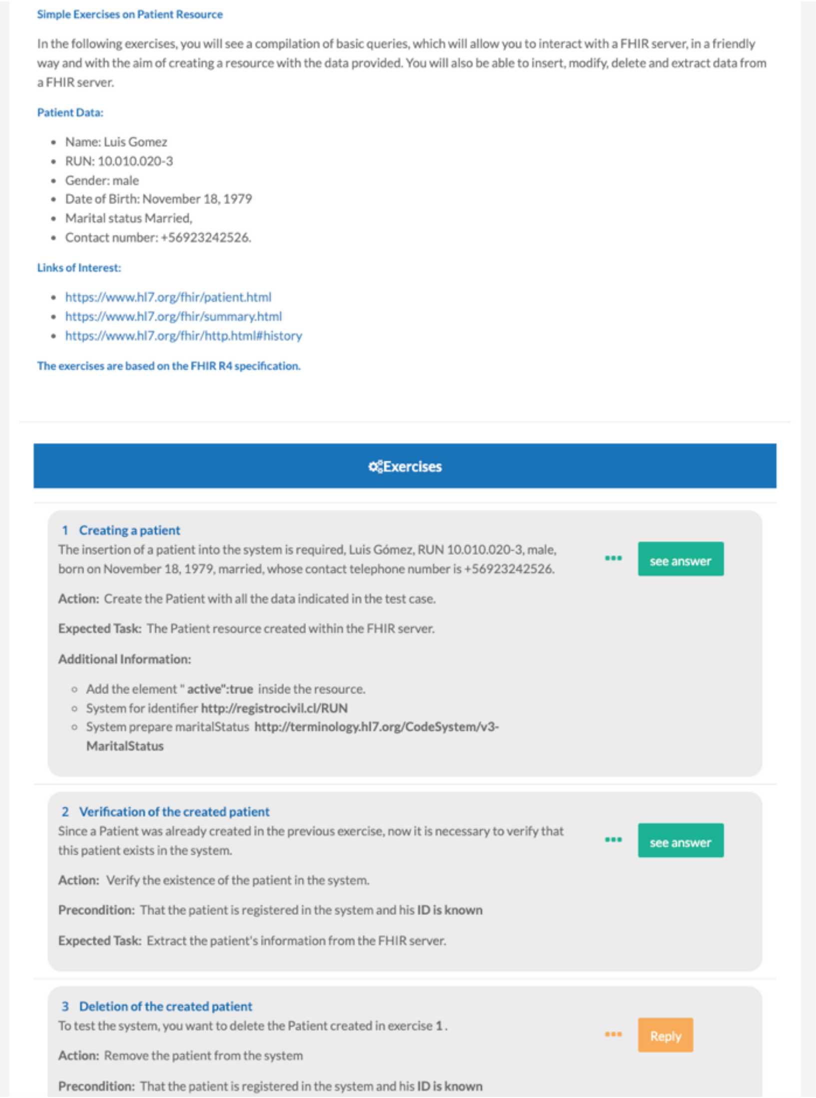
When the user enters a course, the client presents the complete scenario, including information relevant to the test. Below is a

XSL•FO
RenderX

list of the exercises to be answered; each activity has an associated answer button with different colors.

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

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



Initial Evaluation of Huemul Use

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

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

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

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

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

XSL•FO
RenderX

concerning the navigation between the questions and the test, the font and size of the text, and the lack of information to support formatting.

Currently, Huemul has the following usage statistics:

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

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

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

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

Discussion

Principal Findings

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

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

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

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

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

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

Most trained professionals are leading interoperability projects with FHIR from the government, universities, and public or

XSL•FO
RenderX

private health institutions. CENS continues to support capacity building for both professionals and institutions. In this sense, Huemul is an effective tool to support practical activities, enabling the teaching of FHIR. We expect to continue advancing and complementing Huemul with new functionalities and modules in future work.

Future Work

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

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

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

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

Conclusions

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

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

Acknowledgments

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

Conflicts of Interest

None declared.

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

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

References

1. About Information Systems for Health (IS4H). Pan American Health Organization. 2022. URL: https://www3.paho.org/ish/index.php/en/about-is4h [accessed 2022-12-15]
2. Otero P, Leikam M, Gonzalez Z, de Fatima Marin H, Aravena IP, Zawadzki S. Informatics education in Latin America. In: Berner ES, editor. Informatics Education in Healthcare: Lessons Learned. Cham: Springer International Publishing; 2020:167-182.
3. HL7 FHIR release 4B. Standard Developing Organization Health Level Seven International (HL7). 2022. URL: http://hl7.org/implement/standards/fhir/index.html [accessed 2022-10-20]

XSL•FO
RenderX

4. Fielding RT. Architectural styles and the design of network-based software architectures. University of California, Irvine. 2000. URL: https://ics.uci.edu/~fielding/pubs/dissertation/top.htm [accessed 2023-11-21]
5. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
6. Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. 2013 Presented at: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; June 20-22, 2013; Porto, Portugal p. 326-331. [doi: 10.1109/cbms.2013.6627810]
7. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
8. Shull JG. Digital health and the state of interoperable electronic health records. JMIR Med Inform 2019;7(4):e12712 [FREE Full text] [doi: 10.2196/12712] [Medline: 31682583]
9. Guinez-Molinos S, Andrade JM, Negrete AM, Vidal SE, Rios E. Interoperable platform to report polymerase chain reaction SARS-CoV-2 tests from laboratories to the Chilean government: development and implementation study. JMIR Med Inform 2021;9(1):e25149 [FREE Full text] [doi: 10.2196/25149] [Medline: 33417587]
10. Benson T, Grieve G. Implementing FHIR. In: Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. Cham: Springer International Publishing; 2016:397-416.
11. Khan L, Dieter MG, Berner ES, Valenta AL. Managing unspoken assumptions in online education. In: Berner ES, editor. Informatics Education in Healthcare: Lessons Learned. Cham: Springer International Publishing; 2020:263-275.
12. Kraleva R, Sabani M, Kralev V. An analysis of some learning management systems. Int J Adv Sci Eng Inf Technol 2019;9(4):1190-1198 [FREE Full text] [doi: 10.18517/ijaseit.9.4.9437]
13. Hussain MA, Langer SG, Kohli M. Learning HL7 FHIR using the HAPI FHIR server and its use in medical imaging with the SIIM dataset. J Digit Imaging 2018;31(3):334-340 [FREE Full text] [doi: 10.1007/s10278-018-0090-y] [Medline: 29725959]
14. Gonzalez-Cid Y, Guerrero C, Lobo J, Tarbal A, Picking R, Castell N, et al. eHealth Eurocampus: an innovative educational framework to train qualified professionals in the emerging ehealth sector. 2019 Presented at: EDULEARN19: 11th International Conference on Education and New Learning Technologies Proceedings; July 1-3, 2019; Palma, Spain p. 9938-9947. [doi: 10.21125/edulearn.2019.2476]
15. National Center for Health Information Systems. CENS. 2020. URL: https://cens.cl [accessed 2020-02-07]
16. Sandars J, Lafferty N. Twelve tips on usability testing to develop effective e-learning in medical education. Med Teach 2010;32(12):956-960. [doi: 10.3109/0142159X.2010.507709] [Medline: 21090948]
17. Zahabi M, Kaber DB, Swangnetr M. Usability and safety in electronic medical records interface design: a review of recent literature and guideline formulation. Hum Factors 2015;57(5):805-834. [doi: 10.1177/0018720815576827] [Medline: 25850118]
18. Boehm BW. A spiral model of software development and enhancement. Computer 1988;21(5):61-72. [doi: 10.1109/2.59]
19. Leff A, Rayfield JT. Web-application development using the model/view/controller design pattern. 2001 Presented at: Proceedings Fifth IEEE International Enterprise Distributed Object Computing Conference; September 04-07, 2001; Seattle, WA, USA p. 118-127. [doi: 10.1109/edoc.2001.950428]
20. Gutierrez S, Erazo C, Guinez-Molinos S, Taramasco C, Galindo C, Figueroa R, et al. No CENS: towards a Latin American proposal for core competencies in health information systems. 2018 Presented at: AMIA 2018 Informatics Educators Forum; June 19-21, 2018; New Orleans, LA.
21. Techconnect interoperability challenge. CENS. 2022. URL: https://www.techconnect.cl [accessed 2022-12-15]
22. Smile CDR. HAPI FHIR. 2022. URL: https://hapifhir.io [accessed 2022-12-15]
23. Dixon BE, Rahurkar S, Apathy NC. Interoperability and health information exchange for public health. In: Magnuson JA, Dixon BE, editors. Public Health Informatics and Information Systems. Cham: Springer International Publishing; 2020:307-324.

Abbreviations

API: application programming interface
CENS: National Center for Health Information System
CRUD: create, read, update, and delete
FHIR: Fast Healthcare Interoperability Resource
HIS: Health Information Systems
HL7: Health Level Seven International
LMS: learning management system
REST: Representational State Transfer

XSL•FO
RenderX

Edited by G Eysenbach; submitted 29.12.22; peer-reviewed by R Saripalle, F Besoain, F Buendia, JL Sierra, D Chrimes; comments to author 30.01.23; revised version received 27.03.23; accepted 16.11.23; published 29.01.24.

Please cite as:
Guinez-Molinos S, Espinoza S, Andrade J, Medina A
Design and Development of Learning Management System Huemul for Teaching Fast Healthcare Interoperability Resource: Algorithm Development and Validation Study
JMIR Med Educ 2024;10:e45413
URL: https://mededu.jmir.org/2024/1/e45413
doi:10.2196/45413
PMID:38285492

©Sergio Guinez-Molinos, Sonia Espinoza, Jose Andrade, Alejandro Medina. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.01.2024. 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.

XSL•FO
RenderX

Original Paper

The Effects of Immersive Virtual Reality–Assisted Experiential Learning on Enhancing Empathy in Undergraduate Health Care Students Toward Older Adults With Cognitive Impairment: Multiple-Methods Study

Justina Yat Wa Liu[1,2], PhD; Pui Ying Mak[1], BSc; Kitty Chan[1], PhD; Daphne Sze Ki Cheung[1,2], PhD; Kin Cheung[1], PhD; Kenneth N K Fong[3], PhD; Patrick Pui Kin Kor[1], PhD; Timothy Kam Hung Lai[1], MSc; Tulio Maximo[4], PhD

[1]School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[2]Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[3]Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[4]School of Design, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

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

Abstract

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

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

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

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

XSL•FO
RenderX

outcomes of understanding the challenges faced by older adults with cognitive impairment, the importance of providing person-centered care, and the need for an age-friendly society.

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

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

KEYWORDS

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

Introduction

Background

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

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

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

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

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

Previous studies have recognized the effectiveness of IVR-assisted experiential learning in improving empathy among undergraduate health care students [29]. The Cognitive Affective Model of Immersive Learning by Makransky and Petersen [30] suggests that the mental state of perceiving a virtual self as one's actual self with a heightened sense of embodiment refers to the sensation of possessing a virtual body. Hence, using a first-person viewpoint with a virtual environment through IVR

XSL•FO
RenderX

as a “perspective taking machine” could lead to a feeling of immersion and improve a participant’s level of embodiment, leading to an increase in empathy [31-33]. Scholars have also recommended that medical students participate in IVR experiential learning to improve their empathy before starting their clinical placement [34].

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

Objectives

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

Methods

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

Participants

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

Design of the IVR-Assisted Experiential Tutorial

Overview

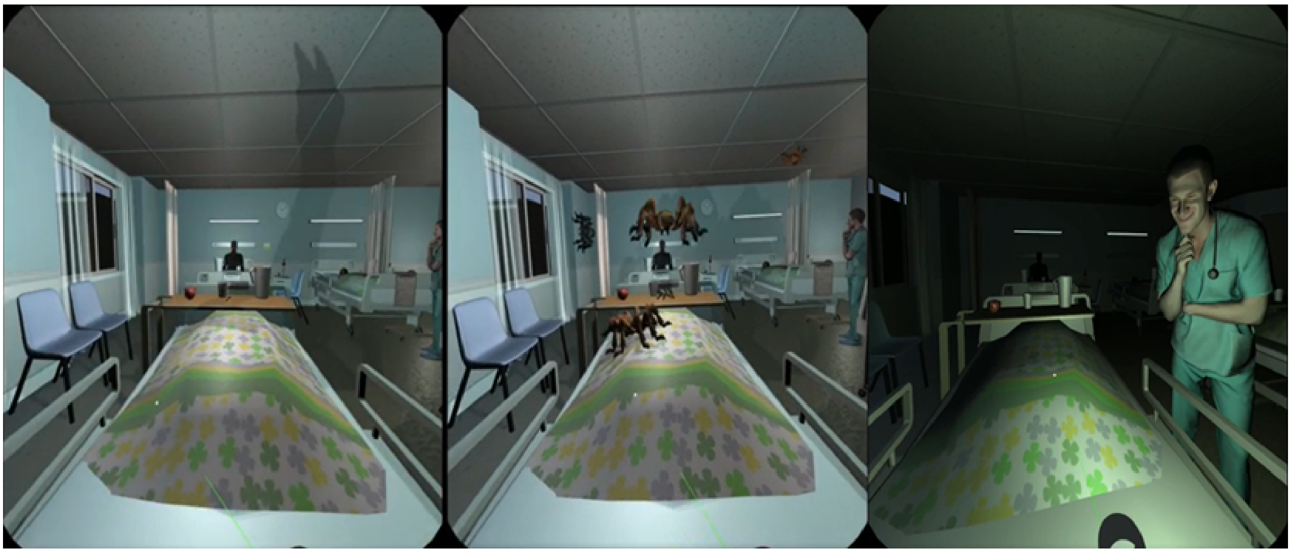
To ensure that students had a solid grasp of the foundational knowledge in the subjects of gerontological nursing (for nursing students) and human occupations (for OT students), a 2-hour IVR-assisted experiential tutorial was arranged in week 7, halfway through the 13-week courses. Only the nursing students were mandated to complete and submit a group assignment within 2 weeks following the IVR tutorial.

The research team developed 2 IVR games that simulated experiences commonly encountered by older adults with cognitive impairment. The first IVR game simulated a scenario in which an individual with cognitive impairment gets lost in a community setting (Figure 1). The second IVR game simulated the distorted auditory and visual perceptions commonly experienced by older adults with delirium (Figure 2). These are common challenges faced on a daily basis by older adults with cognitive impairments. These 2 IVR games were used in the 2-hour IVR-assisted experiential tutorial. Each tutorial comprised students aged between 25 and 30 years who were divided into 7 to 8 subgroups. Each subgroup underwent concurrent IVR–assisted experiential learning.

Figure 1. Scenarios simulating getting lost when looking for a supermarket as experienced by individuals with cognitive impairment.

XSL•FO
RenderX

Figure 2. Scenarios simulating the hallucinations experienced by older adults with delirium.



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

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

Stage 1: Concrete Experience Through Experiential Learning

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

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

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

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

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

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

XSL•FO
RenderX

Outcome Measures

Empathy Toward Older Adults (Objective 1)

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

Learning Experience (Objective 2)

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

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

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

Learning Outcomes (Objective 3)

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

Data Analysis

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

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

Ethical Considerations

This study was approved by the Human Subjects Ethics Application Review System of the Hong Kong Polytechnic University (HSEARS20200423001) and conducted between June 2021 and May 2022. It was carried out in accordance with

XSL•FO
RenderX

the Declaration of Helsinki. This included but was not limited to guaranteeing the anonymity of participants and obtaining the informed consent of the participating students. The participation of the students was voluntary, and their academic results were not affected by their decision to participate in the study.

Results

Overview

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

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

Empathy Toward Older Adults (Objective 1)

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

Table 1. Changes in the Kiersma-Chen Empathy Scale (KCES) score before and after immersive virtual reality experiential learning (n=344)

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

[a]Items with negative wordings are scored in reverse.

XSL•FO
RenderX

Learning Experience (Objective 2)

Students' Satisfaction and Self-Confidence in Learning

Students were satisfied with the current learning innovation, as reflected by a high satisfaction mean score of 20.68 (SD 2.55) out of 25. For example, 92.7% (319/344) of the students agreed or strongly agreed that IVR-assisted experiential learning was suitable for the way they learned (item 5). The same percentage of students agreed or strongly agreed that the IVR learning experience provided an alternative learning experience to promote their learning interests (item 2). A total of 91.6% (315/344) of the students agreed or strongly agreed that the IVR simulation was motivating and helped them learn better (item 4).

They also showed a high level of self-confidence in their IVR experiential learning, with a mean score of 32.04 (SD 3.52) out of 40. Approximately 85.5% (294/344) of the students agreed or strongly agreed that they were confident that they would obtain the necessary skills and knowledge through learning with the IVR simulation (items 6-8). A total of 95.1% (327/344) of the participants agreed or strongly agreed that students should take responsibility for their learning (items 10-11; Table 2).

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

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

[a]IVR: immersive virtual reality.

IVR Fidelity

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

XSL•FO
RenderX

84), and sensory fidelity (mean 31.47, SD 5.23 out of 42) (Table 3).

On the basis of the focus group discussions with the students about their experiences of experiential learning with IVR, 4 themes were identified: *enhanced sympathetic feeling toward older adults with cognitive impairment*, *improved engagement in IVR learning*, *confidence in understanding key concepts in the IVR experiential learning process*, and *limitations of IVR technology*.

XSL•FO
RenderX

Table 3. The findings of the Presence Questionnaire version 2.0 (n=344).

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

XSL•FO
RenderX

Item	Participants, n (%)						
	1	2	3	4	5	6	7
25. How completely were your senses engaged in this experience?	1 (0.3)	2 (0.6)	5 (1.5)	54 (15.7)	111 (32.3)	119 (34.6)	52 (15.1)
27. Were there moments during the virtual environment experience when you felt completely focused on the task or environment?	2 (0.6)	0 (0)	5 (1.5)	66 (19.2)	122 (35.5)	98 (28.5)	51 (14.8)
28. How easily did you adjust to the control devices used to interact with the virtual environment?	0 (0)	6 (1.7)	8 (2.3)	91 (26.5)	139 (40.4)	58 (16.9)	42 (12.2)
29. Was the information provided through different senses in the virtual environment (eg, vision, hearing, touch) consistent?	1 (0.3)	4 (1.2)	5 (1.5)	74 (21.5)	111 (32.3)	93 (27)	56 (16.3)
Interface quality							
19. How much delay did you experience between your actions and the expected outcomes?[b]	7 (2)	32 (9.3)	64 (18.6)	102 (29.7)	57 (16.6)	58 (16.9)	24 (7)
22. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?[b]	17 (4.9)	79 (23)	66 (19.2)	96 (27.9)	44 (12.8)	24 (7)	18 (5.2)
23. How much did the control devices interfere with the performance of assigned tasks or with other activities?[b]	21 (6.1)	74 (21.5)	106 (30.8)	73 (21.2)	37 (10.8)	15 (4.4)	18 (5.2)

[a]IVR: immersive virtual reality.

[b]Reverse items.

Enhanced Sympathetic Feelings Toward Older Adults With Cognitive Impairment

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

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

Another student added the following:

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

Improved Engagement in IVR Learning

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

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

Confidence in Understanding Key Concepts in the IVR Experiential Learning Process

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

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

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

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

Limitations of IVR Technology

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

> *I encountered some technical issues during the IVR experience, which interrupted the flow of the scenario*

XSL•FO
RenderX

> *and disrupted my immersion in the experience. It was frustrating, and I felt like I missed out on some important learning opportunities as a result.*

Another student added the following:

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

Learning Outcomes (Objective 3)

Overview

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

Understanding the Challenges Faced by Older People With Cognitive Impairment

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

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

Another statement also said the following:

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

Person-Centered Care

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

> *Person-centered care is essential to ensure that older people with cognitive impairment receive care that is tailored to their unique needs and preferences.*

"Effective communication," "family involvement," and "supportive care with patience" were 3 critical aspects of person-centered care that were frequently discussed in the assignments:

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

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

> *Family members play a critical role in providing support and care to older people with cognitive impairment. This is particularly the case during delirium.*
>
> *Their involvement can promote continuity of care and provide emotional support to their families with cognitive impairment, especially when they are in a distressing situation, such as delirium.*

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

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

Creation of an Age-Friendly Society

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

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

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

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

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

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

Another group echoed this with the following suggestion:

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

Discussion

Principal Findings

The results suggest that IVR-assisted experiential learning is effective in enhancing empathy toward older people among undergraduate nursing and OT students, as reflected in their higher scores on the KCES after the IVR simulation. The

XSL•FO
RenderX

students reported a high level of satisfaction with the IVR learning experience, citing its suitability, ability to motivate, and innovativeness in the self-administered survey. In addition, the findings from the survey suggest that the students experienced a strong sense of presence in the IVR learning environment, enabling them to gain a deeper understanding of the challenges involved in caring for older adults with cognitive impairment. In total, 3 major themes were identified from the focus groups with 23 nursing students: *enhanced sympathetic feelings toward older adults with cognitive impairment*, *improved engagement in IVR learning*, and *confidence in understanding the key concepts through the learning process*. The thematic findings supplement and are in line with the results from the survey. The analysis of the written assignments showed that the students attained the learning outcomes of understanding the challenges faced by older people with cognitive impairment, the importance of providing person-centered care, and the need to create an age-friendly society.

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

Empathy Toward Older Adults and Learning Experience

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

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

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

Learning Outcomes

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

Implications

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

Limitations and Challenges of IVR Learning

Although IVR-assisted experiential learning has shown positive results in enhancing health care education, it is important to acknowledge the limitations and challenges associated with adopting this technology in teaching. Technical issues such as equipment malfunctions and slow processing times could result in missed learning opportunities, as noted by some students during the focus group discussions. Similar technical issues mentioned in previous studies disrupted the flow of scenarios

XSL•FO
RenderX

and limited the effectiveness of the IVR learning experience [57,58]. These technical limitations must be addressed to ensure that IVR can be used effectively for teaching. Other main challenges that we experienced include the cost of implementing and maintaining the IVR technology, including hardware and software [50]. Another challenge is the need for technical support to develop and maintain IVR simulations, which requires collaboration between educators and technologists [59]. This may be prohibitive for some educational institutions to undertake.

Study Limitations

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

Future Directions

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

Conclusions

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

Acknowledgments

The authors sincerely thank the students and teachers who participated in this study. They also thank Ms Jay Wong, the research assistant, for her excellent work in ensuring that this project ran smoothly. This study was funded by a Large-Scale Collaborative Teaching Development Grant (2019-2022) from the Learning and Teaching Committee and matching funds from the School of Nursing (.53.XX.49LP), Hong Kong Polytechnic University.

Data Availability

The data sets are not publicly available owing to pending further analysis of the data. The virtual reality scenarios used in this study were just a part of the scenarios present in the complete virtual reality training system. As we may need to conduct further comparisons, we regret not being able to disclose the data sets.

Authors' Contributions

All the authors were involved in the design of the virtual reality (VR) games and the study. JYWL contributed to the conceptualization of the study. JYWL, DSKC, PPKK, KNKF, and TM implemented the VR games in their subjects. JYWL was responsible for data collection, data analysis, and quality control of the study. JYWL and PYM wrote the original draft of the manuscript. All coauthors commented on and rewrote the manuscript. All the authors have read and approved the final version of the manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1

The findings on the between-group changes in Kiersma-Chen Empathy Scale scores across the time points between nursing and occupational therapy students.

[PDF File (Adobe PDF File), 13 KB - mededu_v10i1e48566_app1.pdf]

References

1. Hojat M, Spandorfer J, Louis DZ, Gonnella JS. Empathic and sympathetic orientations toward patient care: conceptualization, measurement, and psychometrics. Acad Med 2011 Aug;86(8):989-995. [doi: 10.1097/ACM.0b013e31822203d8] [Medline: 21694570]
2. Glaser KM, Markham FW, Adler HM, McManus PR, Hojat M. Relationships between scores on the Jefferson scale of physician empathy, patient perceptions of physician empathy, and humanistic approaches to patient care: a validity study. Med Sci Monit 2007 Jul;13(7):CR291-CR294. [Medline: 17599021]
3. Hojat M, Louis DZ, Markham FW, Wender R, Rabinowitz C, Gonnella JS. Physicians' empathy and clinical outcomes for diabetic patients. Acad Med 2011 Mar;86(3):359-364. [doi: 10.1097/ACM.0b013e3182086fe1] [Medline: 21248604]
4. Menendez ME, Chen NC, Mudgal CS, Jupiter JB, Ring D. Physician empathy as a driver of hand surgery patient satisfaction. J Hand Surg Am 2015 Sep;40(9):1860-5.e2. [doi: 10.1016/j.jhsa.2015.06.105] [Medline: 26231482]
5. Kim SS, Kaplowitz S, Johnston MV. The effects of physician empathy on patient satisfaction and compliance. Eval Health Prof 2004 Sep;27(3):237-251. [doi: 10.1177/0163278704267037] [Medline: 15312283]
6. Kaplan-Liss E, Lantz-Gefroh V, Bass E, Killebrew D, Ponzio NM, Savi C, et al. Teaching medical students to communicate with empathy and clarity using improvisation. Acad Med 2018 Mar;93(3):440-443. [doi: 10.1097/ACM.0000000000002031] [Medline: 29059072]
7. Abreu M, Caldevilla N. Attitudes toward aging in Portuguese nursing students. Procedia Soc Behav Sci 2015 Jan 16;171:961-967. [doi: 10.1016/j.sbspro.2015.01.215]
8. Celik SS, Kapucu S, Tuna Z, Akkus Y. Views and attitudes of nursing students towards ageing and older patients. Aust J Adv Nurs 2010 Jun;27(4):24-30 [FREE Full text]
9. Lambrinou E, Sourtzi P, Kalokerinou A, Lemonidou C. Attitudes and knowledge of the Greek nursing students towards older people. Nurse Educ Today 2009 Aug;29(6):617-622. [doi: 10.1016/j.nedt.2009.01.011] [Medline: 19243864]
10. Usta YY, Demir Y, Yönder M, Yildiz A. Nursing students' attitudes toward ageism in Turkey. Arch Gerontol Geriatr 2012;54(1):90-93. [doi: 10.1016/j.archger.2011.02.002] [Medline: 21353316]
11. Petry H, Ernst J, Steinbrüchel-Boesch C, Altherr J, Naef R. The acute care experience of older persons with cognitive impairment and their families: a qualitative study. Int J Nurs Stud 2019 Aug;96:44-52. [doi: 10.1016/j.ijnurstu.2018.11.008] [Medline: 30660445]
12. Jang I, Oh D, Kim YS. Factors associated with nursing students' willingness to care for older adults in Korea and the United States. Int J Nurs Sci 2019 Sep 06;6(4):426-431 [FREE Full text] [doi: 10.1016/j.ijnss.2019.09.004] [Medline: 31728396]
13. Peng X, Wu L, Xie X, Dai M, Wang D. Impact of virtual dementia tour on empathy level of nursing students: a quasi-experimental study. Int J Nurs Sci 2020 Jun 24;7(3):258-261 [FREE Full text] [doi: 10.1016/j.ijnss.2020.06.010] [Medline: 32817846]
14. Lam TC, Kolomitro K, Alamparambil FC. Empathy training: methods, evaluation practices, and validity. J Multidiscip Eval 2011;7(16):162-200. [doi: 10.56645/jmde.v7i16.314]
15. Hoover JD. Experiential learning: conceptualization and definition. In: Proceedings of the Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference. 1974 Mar Presented at: Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference; March 1974; Japan.
16. Hoover JD, Whitehead CJ. An experiential-cognitive methodology in the first course in management: some preliminary results. In: Proceedings of the Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference. 1975 Mar Presented at: Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference; March 1975; Bloomington, IN.
17. Johnson CE, Jilla AM, Danhauer JL. Didactic content and experiential aging simulation for developing patient-centered strategies and empathy for older adults. Semin Hear 2018 Feb;39(1):74-82 [FREE Full text] [doi: 10.1055/s-0037-1613707] [Medline: 29422715]
18. Buffum MD, Hutt E, Chang VT, Craine MH, Snow AL. Cognitive impairment and pain management: review of issues and challenges. J Rehabil Res Dev 2007;44(2):315-330 [FREE Full text] [doi: 10.1682/jrrd.2006.06.0064] [Medline: 17551882]
19. Underberg KE. Experiential learning and simulation in health care education. SSM 2003 Aug;9(4):31-4,36.
20. Lowell VL, Alshammari A. Experiential learning experiences in an online 3D virtual environment for mental health interviewing and diagnosis role-playing: a comparison of perceived learning across learning activities. Educ Tech Res Dev 2018 Nov 1;67:825-854. [doi: 10.1007/s11423-018-9632-8]

XSL•FO
RenderX

21. Burnard P. Teaching Interpersonal Skills: A Handbook of Experiential Learning for Health Professionals. New York City, NY: Springer; Dec 14, 2013.
22. Hodgson P, Lee VW, Chan JC, Fong A, Tang CS, Chan L, et al. Immersive virtual reality (IVR) in higher education: development and implementation. In: tom Dieck M, Jung T, editors. Augmented Reality and Virtual Reality. Cham, Switzerland: Springer; Feb 20, 2019:161-173.
23. Kavanagh S, Luxton-Reilly A, Wuensche B, Plimmer B. A systematic review of virtual reality in education. Themes Sci Technol Educ 2017;10(2):85-119.
24. Hannans JA, Nevins CM, Jordan K. See it, hear it, feel it: embodying a patient experience through immersive virtual reality. Inf Learn Sci 2021 Jun;122(7/8):565-583. [doi: 10.1108/ILS-10-2020-0233]
25. Buchman S, Henderson D. Qualitative study of interprofessional communication through immersive virtual reality 360 video among healthcare students. Int J Nurs Health Care Res 2019 Apr 23;3:76. [doi: 10.29011/IJNHR-076.1000076]
26. Aebersold M. Simulation-based learning: no longer a novelty in undergraduate education. Online J Issues Nurs 2018 Apr 03;23(2):1. [doi: 10.3912/OJIN.Vol23No02PPT39]
27. Ai-Lim Lee E, Wong KW, Fung CC. How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. Comput Educ 2010 Dec;55(4):1424-1442. [doi: 10.1016/j.compedu.2010.06.006]
28. Appel L, Peisachovich E, Sinclair D. CVRriculum program: outcomes from an exploratory pilot program incorporating virtual reality into existing curricula and evaluating its impact on empathy-building and experiential education. Can J Nurs Inform 2021;16(1).
29. Wan WH, Lam AH. The effectiveness of virtual reality-based simulation in health professions education relating to mental illness: a literature review. Health 2019 Jun;11(06):646-660. [doi: 10.4236/health.2019.116054]
30. Makransky G, Petersen GB. The cognitive affective model of immersive learning (CAMIL): a theoretical research-based model of learning in immersive virtual reality. Educ Psychol Rev 2021 Jan 06;33:937-958. [doi: 10.1007/s10648-020-09586-2]
31. Johnson-Glenberg MC. Immersive VR and education: embodied design principles that include gesture and hand controls. Front Robot AI 2018 Jul 24;5:81 [FREE Full text] [doi: 10.3389/frobt.2018.00081] [Medline: 33500960]
32. Barbot B, Kaufman JC. What makes immersive virtual reality the ultimate empathy machine? Discerning the underlying mechanisms of change. Comput Hum Behav 2020 Oct;111:106431. [doi: 10.1016/j.chb.2020.106431]
33. Han I, Shin HS, Ko Y, Shin WS. Immersive virtual reality for increasing presence and empathy. J Comput Assist Learn 2022 Apr 07;38(4):1115-1126. [doi: 10.1111/jcal.12669]
34. Donnelly F, McLiesh P, Bessell SA. Using 360° video to enable affective learning in nursing education. J Nurs Educ 2020 Jul 01;59(7):409-412. [doi: 10.3928/01484834-20200617-11] [Medline: 32598013]
35. Buchman S, Henderson D. Interprofessional empathy and communication competency development in healthcare professions' curriculum through immersive virtual reality experiences. J Interprof Educ Pract 2019 Jun;15:127-130. [doi: 10.1016/j.xjep.2019.03.010]
36. Alexander V, Thomas H, Cronin A, Fielding J, Moran-Ellis J. Mixed methods. In: Gilbert N, Stoneman P, editors. Researching Social Life. Thousand Oaks, CA: SAGE Publications, Inc; 2016.
37. Davis DF, Golicic SL, Boerstler CN. Benefits and challenges of conducting multiple methods research in marketing. J Acad Mark Sci 2010 Jun 16;39:467-479. [doi: 10.1007/s11747-010-0204-7]
38. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. Hoboken, NJ: Prentice-Hall; 1984.
39. Rudolph JW, Raemer DB, Simon R. Establishing a safe container for learning in simulation: the role of the presimulation briefing. Simul Healthc 2014 Dec;9(6):339-349. [doi: 10.1097/SIH.0000000000000047] [Medline: 25188485]
40. Donaghy M, Morss K. An evaluation of a framework for facilitating and assessing physiotherapy students' reflection on practice. Physiother Theory Pract 2007;23(2):83-94. [doi: 10.1080/09593980701211952] [Medline: 17530538]
41. Arseven I. The use of qualitative case studies as an experiential teaching method in the training of pre-service teachers. Int J High Educ 2018 Jan;7(1):111. [doi: 10.5430/ijhe.v7n1p111]
42. Kiersma ME, Chen AM, Yehle KS, Plake KS. Validation of an empathy scale in pharmacy and nursing students. Am J Pharm Educ 2013 Jun 12;77(5):94 [FREE Full text] [doi: 10.5688/ajpe77594] [Medline: 23788805]
43. Hojat M, Mangione S, Nasca TJ, Cohen MJ, Gonnella JS, Erdmann JB, et al. The Jefferson scale of physician empathy: development and preliminary psychometric data. Educ Psychol Meas 2001;61(2):349-365. [doi: 10.1177/00131640121971158]
44. Kitaoka‐Higashiguchi K, Nakagawa H, Morikawa Y, Ishizaki M, Miura K, Naruse Y, et al. Construct validity of the Maslach burnout inventory-general survey. Stress Health 2004 Dec 06;20(5):255-260. [doi: 10.1002/smi.1030]
45. Jeffries PR, Rizzolo MA. Designing and implementing models for the innovative use of simulation to teach nursing care of ill adults and children: a national, multi-site, multi-method study. National League for Nursing and Laerdal Medical. 2006. URL: https://www.nln.org/docs/default-source/uploadedfiles/professional-development-programs/read-the-nln-laerdal-project-summary-report-pdf.pdf [accessed 2024-01-16]
46. Witmer BG, Singer MJ. Measuring presence in virtual environments: a presence questionnaire. Presence Teleoperators Virtual Environ 1998 Jun 01;7(3):225-240. [doi: 10.1162/105474698565686]
47. Witmer BG, Jerome CJ, Singer MJ. The factor structure of the presence questionnaire. Presence Teleoperators Virtual Environ 2005 Jun 01;14(3):298-312. [doi: 10.1162/105474605323384654]

XSL•FO
RenderX

48. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
49. Alieldin R, Borasi R, Nofziger A, DeAngelis K, Peyre S. Can we walk in our patients' shoes? Immersive virtual reality as an empathy training tool for medical students. Frameless 2021 Oct;4(1).
50. Dyer E, Swartzlander BJ, Gugliucci MR. Using virtual reality in medical education to teach empathy. J Med Libr Assoc 2018 Oct;106(4):498-500 [FREE Full text] [doi: 10.5195/jmla.2018.518] [Medline: 30271295]
51. Liu JY, Yin YH, Kor PP, Cheung DS, Zhao IY, Wang S, et al. The effects of immersive virtual reality applications on enhancing the learning outcomes of undergraduate health care students: systematic review with meta-synthesis. J Med Internet Res 2023 Mar 06;25:e39989 [FREE Full text] [doi: 10.2196/39989] [Medline: 36877550]
52. Logeswaran A, Munsch C, Chong YJ, Ralph N, McCrossnan J. The role of extended reality technology in healthcare education: towards a learner-centred approach. Future Healthc J 2021 Mar;8(1):e79-e84 [FREE Full text] [doi: 10.7861/fhj.2020-0112] [Medline: 33791482]
53. Hu-Au E, Lee J. Virtual reality in education: a tool for learning in the experience age. Int J Innov Educ 2018 Apr;4(4):215-226. [doi: 10.1504/IJIIE.2017.10012691]
54. Meyer K, James D, Amezaga B, White C. Simulation learning to train healthcare students in person-centered dementia care. Gerontol Geriatr Educ 2022;43(2):209-224 [FREE Full text] [doi: 10.1080/02701960.2020.1838503] [Medline: 33081626]
55. Breen H, Jones M. Experiential learning: using virtual simulation in an online RN-to-BSN program. J Contin Educ Nurs 2015 Jan;46(1):27-33. [doi: 10.3928/00220124-20141120-02] [Medline: 25401340]
56. Adefila A, Graham S, Clouder DL, Bluteau P, Ball S. myShoes – the future of experiential dementia training? J Ment Health Train 2016 May;11(2):91-101. [doi: 10.1108/JMHTEP-10-2015-0048]
57. Adhikari R, Kydonaki C, Lawrie J, O'Reilly M, Ballantyne B, Whitehorn J, et al. A mixed-methods feasibility study to assess the acceptability and applicability of immersive virtual reality sepsis game as an adjunct to nursing education. Nurse Educ Today 2021 Aug;103:104944. [doi: 10.1016/j.nedt.2021.104944] [Medline: 34015677]
58. Fertleman C, Aubugeau-Williams P, Sher C, Lim AN, Lumley S, Delacroix S, et al. A discussion of virtual reality as a new tool for training healthcare professionals. Front Public Health 2018 Feb 26;6:44 [FREE Full text] [doi: 10.3389/fpubh.2018.00044] [Medline: 29535997]
59. Wang R, DeMaria SJ, Goldberg A, Katz D. A systematic review of serious games in training health care professionals. Simul Healthc 2016 Feb;11(1):41-51. [doi: 10.1097/SIH.0000000000000118] [Medline: 26536340]

Abbreviations

IVR: immersive virtual reality
KCES: Kiersma-Chen Empathy Scale
OT: occupational therapy
PQ2: Presence Questionnaire version 2.0
VR: virtual reality

Edited by G Eysenbach, T Leung; submitted 28.04.23; peer-reviewed by CJR Siah, YH Yin; comments to author 25.07.23; revised version received 08.09.23; accepted 28.12.23; published 15.02.24.

Please cite as:
Liu JYW, Mak PY, Chan K, Cheung DSK, Cheung K, Fong KNK, Kor PPK, Lai TKH, Maximo T
The Effects of Immersive Virtual Reality–Assisted Experiential Learning on Enhancing Empathy in Undergraduate Health Care Students Toward Older Adults With Cognitive Impairment: Multiple-Methods Study
JMIR Med Educ 2024;10:e48566
URL: https://mededu.jmir.org/2024/1/e48566
doi:10.2196/48566
PMID:38358800

©Justina Yat Wa Liu, Pui Ying Mak, Kitty Chan, Daphne Sze Ki Cheung, Kin Cheung, Kenneth N K Fong, Patrick Pui Kin Kor, Timothy Kam Hung Lai, Tulio Maximo. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

A Language Model–Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study

Friederike Holderried[1], MD, MME; Christian Stegemann-Philipps[1], Dr rer nat; Anne Herrmann-Werner[1], Prof Dr Med, MME; Teresa Festl-Wietek[1], Dr rer nat; Martin Holderried[2], Prof Dr, Dr med; Carsten Eickhoff[3], Prof Dr; Moritz Mahling[1,2], MD, MHBA

[1]Tübingen Institute for Medical Education (TIME), Medical Faculty, University of Tübingen, Tübingen, Germany

[2]Department of Medical Development, Process and Quality Management, University Hospital Tübingen, Tübingen, Germany

[3]Institute for Applied Medical Informatics, University of Tübingen, Tübingen, Germany

Corresponding Author:
Friederike Holderried, MD, MME
Tübingen Institute for Medical Education (TIME)
Medical Faculty, University of Tübingen
Elfriede-Aulhorn-Strasse 10
Tübingen, 72076
Germany
Phone: 49 707129 ext 73688
Email: friederike.holderried@med.uni-tuebingen.de

Abstract

Background: Although history taking is fundamental for diagnosing medical conditions, teaching and providing feedback on the skill can be challenging due to resource constraints. Virtual simulated patients and web-based chatbots have thus emerged as educational tools, with recent advancements in artificial intelligence (AI) such as large language models (LLMs) enhancing their realism and potential to provide feedback.

Objective: In our study, we aimed to evaluate the effectiveness of a Generative Pretrained Transformer (GPT) 4 model to provide structured feedback on medical students' performance in history taking with a simulated patient.

Methods: We conducted a prospective study involving medical students performing history taking with a GPT-powered chatbot. To that end, we designed a chatbot to simulate patients' responses and provide immediate feedback on the comprehensiveness of the students' history taking. Students' interactions with the chatbot were analyzed, and feedback from the chatbot was compared with feedback from a human rater. We measured interrater reliability and performed a descriptive analysis to assess the quality of feedback.

Results: Most of the study's participants were in their third year of medical school. A total of 1894 question-answer pairs from 106 conversations were included in our analysis. GPT-4's role-play and responses were medically plausible in more than 99% of cases. Interrater reliability between GPT-4 and the human rater showed "almost perfect" agreement (Cohen κ=0.832). Less agreement (κ<0.6) detected for 8 out of 45 feedback categories highlighted topics about which the model's assessments were overly specific or diverged from human judgement.

Conclusions: The GPT model was effective in providing structured feedback on history-taking dialogs provided by medical students. Although we unraveled some limitations regarding the specificity of feedback for certain feedback categories, the overall high agreement with human raters suggests that LLMs can be a valuable tool for medical education. Our findings, thus, advocate the careful integration of AI-driven feedback mechanisms in medical training and highlight important aspects when LLMs are used in that context.

(JMIR Med Educ 2024;10:e59213) doi:10.2196/59213

KEYWORDS

virtual patients communication; communication skills; technology enhanced education; TEL; medical education; ChatGPT; GPT: LLM; LLMs; NLP; natural language processing; machine learning; artificial intelligence; language model; language models; communication; relationship; relationships; chatbot; chatbots; conversational agent; conversational agents; history; histories; simulated; student; students; interaction; interactions

XSL•FO
RenderX

Introduction

For most medical problems, history taking is the cornerstone of the diagnostic journey. Despite the increase in diagnostic tools such as advanced imaging and molecular and laboratory assays, a comprehensive history is necessary to guide further steps and may sometimes even be sufficient for diagnosing a disease without further testing [1,2]. Conversely, insufficient history taking can risk patients' safety [3,4]. Due to its importance, history taking is taught to health care students worldwide, usually as part of a communication-focused curriculum or clinical clerkship [5-8] and mostly relying on real patients [9].

To enable more student-patient interactions without increasing costs, staff's workload, or the burden on patients, virtual simulated patients have emerged as an adjunctive approach [10,11]. For communication skills in particular, web-based chatbots have been developed to offer an additional learning format [12], and recent advances in artificial intelligence (AI) such as large language models (LLMs) have helped those tools to achieve a new level of realism [13-15]. Indeed, recent work has demonstrated that OpenAI's Generative Pretrained Transformer (GPT) model is capable of providing realistic, positively perceived patient experiences as well as scenarios requiring the breaking of bad news, all of which are simulated [13,16].

However, patient experiences alone are hardly sufficient to develop competence. Indeed, no matter the amount of their exposure to patients, medical students have to have feedback in order to progress in their performance [17,18]. Traditional teaching methods require teachers' significant involvement in providing feedback, either while history taking is performed or in assessing the results afterward. LLM-based education, by contrast, offers the opportunity for repeated, unsupervised exposure to simulated patients. Whereas traditional virtual patients often yield low levels of feedback [10], the linguistic capabilities of LLMs can provide students with higher-quality feedback [19]. LLMs have also demonstrated the capability of providing feedback in other circumstances, including argumentation [20], writing [21], and scientific papers [22]. However, their capability to provide feedback on the quality of history taking has not been elucidated on a large scale, and concerns about the accuracy of AI-based feedback persist [23].

Building on our previous work showing that GPT-3.5 can provide simulated patient experiences [13], we evaluated the extension of our chatbot with an integrated feedback system while using the latest LLM model, GPT-4. In particular, we aimed to investigate whether GPT-4 can provide structured feedback on medical students' performance during history-taking dialogs with a simulated patient, with special focus on such feedback's realism and educational use. We hypothesized that GPT-4, given its capabilities in medical knowledge [24-26] and reasoning [13], can accurately assess students' performance in history taking despite potential limitations such as logical errors [27] and AI's propensity to generate nonsensical content, known as "hallucinations" [28]. Our objective was to evaluate feedback on medical students' history taking provided by GPT and compare it with human feedback, all to contribute to the broader discourse on integrating AI into medical education.

Considering all of the above, we formulated the following research questions for our study:

1. What are the characteristics of medical students' history-taking conversations (ie, question length and chain questions) with a GPT-4–powered simulated patient chatbot?
2. What is the quality of the GPT-4–powered chatbot's role-play during such conversations (ie, are the questions answered and are the answers medically plausible)?
3. How is the history-taking dialog rated by GPT-4 and a human rater in terms of feedback topics covered?
4. How does GPT-4's feedback compare with the feedback of a human rater (ie, interrater reliability)?
5. How can significantly different feedback between GPT-4 and the human rater regarding certain topics be explained?

Methods

Study Outline

We conducted a prospective study in which students performed a written history-taking exercise with a GPT-powered simulated patient (for more information, see [13]). Afterward, GPT-4 was prompted to provide the students feedback on the topics covered in the history taking. The chat history was analyzed in detail, and the GPT model's feedback was compared with feedback from a human rater.

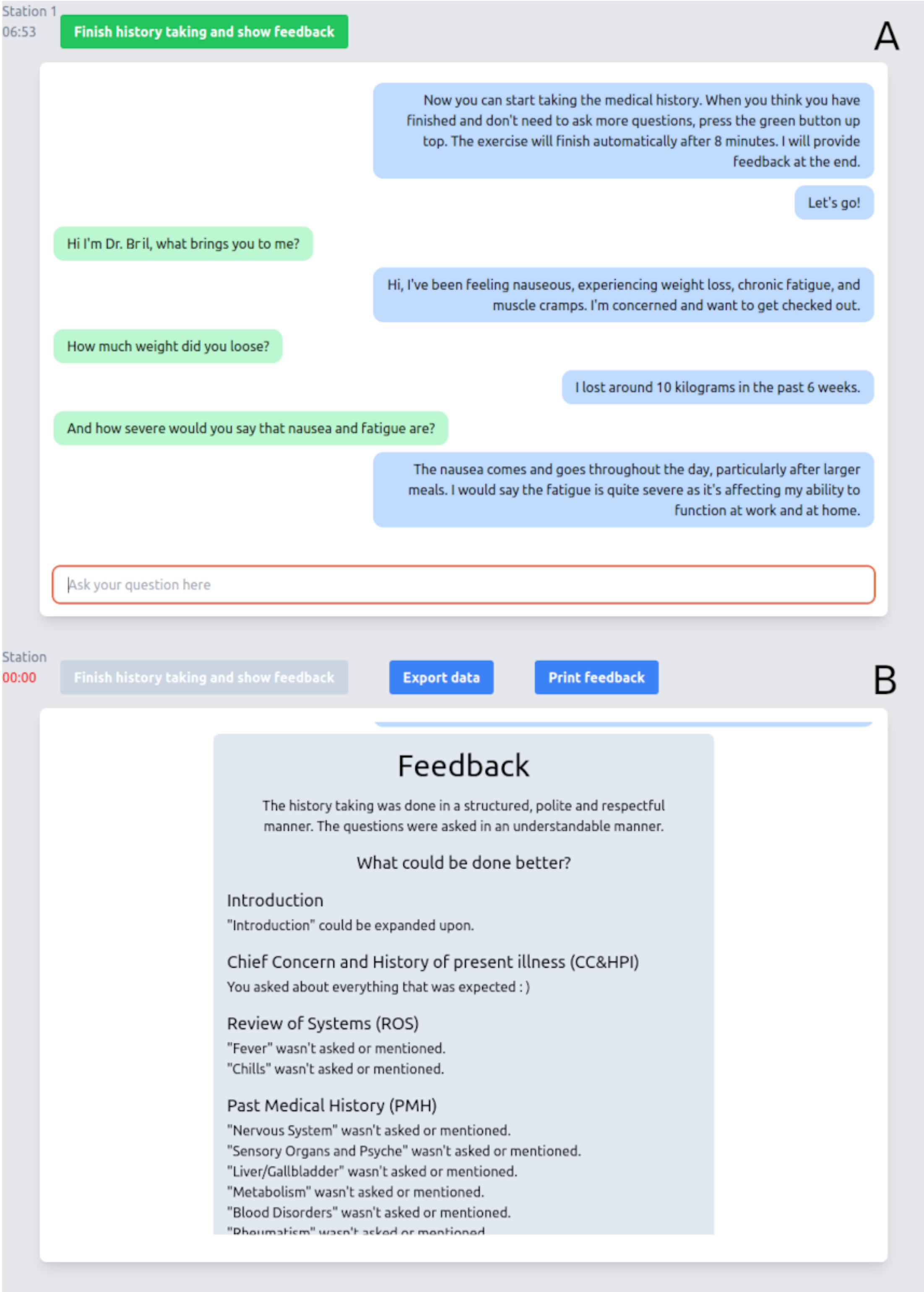
Setting and Participants

During a scheduled break in a skills training course involving multiple opportunities for practice, medical students were asked to participate at an additional training station affording the opportunity to participate in history taking with our GPT-powered chatbot. Participation was voluntary. Given our study's exploratory nature and aim to broadly assess the use of GPT-powered feedback in medical education, we did not impose any specific inclusion or exclusion criteria on participation beyond the willingness and ability to engage with the chatbot. Neither of those components was associated with any examination outcomes.

The training station consisted of a laptop with the chat interface already prepared (Figure 1A). Given the course in which our station was embedded, the time limit for history taking was set to reflect the time limit of other stations (ie, 8 minutes). After finishing history taking, students were presented with AI-generated feedback (Figure 1B) and proceeded to the next practice station.

XSL•FO
RenderX

Figure 1. Screenshot of the chatbot interface as presented to participants (translated from German): (A) the interface during the interactive dialog and (B) the interface presenting the feedback.



Our chat platform was a major update to the platform for history taking previously detailed by our group [13]. In short, we embedded GPT-4, accessed via an application programming interface (API), in a web page in order to enable participants to ask questions to a virtual simulated patient. Model parameters were left at their default settings, and the full chat history was anonymized and saved for further analysis.

Prompt Development

Two prompts were developed: one for providing the interactive history-taking dialog, and the other for giving feedback.

Behavioral Prompt

For the interactive history taking, we used an updated version of the prompt previously developed by our group [13]. In brief, we provided the model with a script describing an illness

XSL•FO
RenderX

("illness script") and used an additional behavioral prompt to make the model behave as a virtual simulated patient. For the updated version of the behavioral prompt used in our study, the prompts for history taking were mostly upgraded by adding sentences describing intended or unintended behavior. We made those upgrades because the earlier prompts made the model too verbose or willing to provide assistance only in certain cases. We added more specific instructions, including that the model should generally answer in 1 sentence or 2, never ask a question unless specifically asked to do so, and never offer assistance.

Moreover, we provided tailored examples of how the simulated patient should respond to certain inputs—for instance, to respond with "OK" if no question was asked. Such modifications aimed to correct for intrusive model behavior in which the model sometimes provided its own question in response to a participant simply writing an affirmation or "OK."

Feedback Prompt

To make the GPT model generate feedback, we used an entirely different prompt. By calling the API, it is possible to gain full control of any message history that the model can access, as opposed to the common web interfaces of chatbots. In our case, that meant that the prompt for history taking and the prompt for feedback could not influence one another unless we intentionally reused parts of one in the other.

We used the illness script, as described in [13] and already used in the prompt for history taking, to define the categories by which to provide feedback, called "feedback categories." Next, for each category of the illness script, the model's task was to judge whether the information had appeared in the chat between the user and the simulated patient or whether it had been asked about. The main dilemma was, thus, the existence of 2 primary sources of information—the illness script and the chat—which complicated what the model paid attention to.

Our strategy was to begin with a description of the task, namely that the model needs to check whether the dialog that follows contains certain information and needs to answer a few questions at the end. We then provided categories from the illness script as fully phrased requirements in the form of "There should have been mention of X in the dialogue, with possible mention of Y", in which case "X" was a category and "Y" the information given in the illness script. We used that strategy to guide the attention of the model before providing the chat. An example of such a construct was "In the dialogue, 'Previous illnesses related to the main symptom' should have been discussed, including information such as 'I've never been like this before. I was usually healthy before.'"

We next pasted the complete chat, scaffolded with "=== START DIALOG ===" and "=== END DIALOG ===" to indicate that the content was a single long block quotation. As previously described [13], we inserted additional formatting into the chat to be presented in the prompt for history taking. However, those modifications were unnecessary and thus absent in the chat reproduced in the feedback prompt—that is, the chat was reproduced in the same way it would be shown to participants.

Following the dialog part, we again described the task of checking the dialog for certain information. We subsequently told the model that we would repeat the feedback categories and information from the illness script in a highly compact format, which we also added to the prompt. Last, we formulated the main question—"Did these categories appear in the dialog?"—and asked the model to give its answer in the form of a JSON dictionary, a computer-readable, structured way of representing key-value pairs and special feature available in recent GPT models [29]. Using the JSON dictionary allowed us to parse the answer of the model in our interface in order to compute scores for participants.

Another problem was that the amount of information in the prompt was liable to led to exceptionally long prompts. We also observed that inquiring about all categories simultaneously led to a high probability of scrambled answers, in which categories were not fully reproduced in the answers or were simply wrong. Despite the plausibility of asking about 1 category of the illness script at a time and issuing different API calls for each, sometimes called the "divide-and-conquer" strategy [30], doing so in our case may have easily overloaded the limits set by OpenAI for model usage or led to very high computing cost. We, therefore, decided to ask about a certain number of categories at a time and issue prompts for each of those small lists. In small initial experiments, limiting the number of categories to 8 tended to provide a good balance between accuracy and cost.

The full prompt is available as Multimedia Appendix 1.

Analysis of the Characteristics of Conversations

Descriptive methods were used to characterize the conversations and question-answer pairs (QAPs), in which each question was inputted by participants and the answer was outputted by GPT. First, we calculated basic metrics to describe those QAPs, including the number of QAPs per conversation and the number of words per question and answer. To analyze the students' questions in detail, they were scored for the presence of the following undesirable behavior when asking for a medical history:

1. Foreign terms (eg, use of medical terminology);
2. Chain questions (ie, asking more than one question);
3. Single-word questions (eg, typing simply "smoking" as a question); and
4. Incomprehensive wording or barely understandable input (eg, nonconclusive series of words or spelling mistakes).

Analysis of the Quality of GPT-4's Role-Play

To assess whether the quality of the GPT's role-play was suitable for a simulated patient experience, as we have shown and analyzed in detail using GPT-3.5 [13], we had the answers outputted by GPT in the first 20% of the conversations assessed by rater FH and discussed with rater MM in equivocal cases. We assessed whether the participant's questions were covered by our illness script, whether each answer matched the question, and whether the answer was medically plausible. That process was guided by categories and definitions developed in previous research [13]. Table 1 shows the categories, possible scores, and examples from the conversations.

XSL•FO
RenderX

Table 1. Areas, categories, and examples for the dialogs for the rating process, adapted from [13].

Areas, categories, and examples	Dialogues
Student's question covered by the illness script	
Question not covered by the script	• Student: "Or was someone in your surroundings ill?" • GPT: "No, as far as I know, no one in my immediate surroundings was ill. Both my sons are currently fit, and my wife hasn't been feeling sick either."
Question partly or indirectly covered by the script	• Student: "And you mentioned feeling listless as well. Are you also experiencing a loss of interest and a low mood?" • GPT: "I primarily feel listless, and it's difficult for me to attend to my family duties after a long day at work. But I wouldn't say that I'm experiencing a loss of interest or a low mood."
Question explicitly covered by the script	• Student: "And do you have any other symptoms like fever or night sweats?" • GPT: "No, no fever. But I do occasionally experience episodes of increased sweating, though not only at night and not very intensely."
Statement (no question)	• Student: "Thank you and see you soon!" • GPT: "See you soon!"
GPT's response matches the question	
GPT's response matches the question	• Student: "Do you have a loss of appetite?" • GPT: "Yes, my appetite is indeed much worse than before. Maybe also because I feel frequently more nauseous."
GPT's response does not match the question	• No example occurred
Not applicable	• For QAPs with answers provided for a previous question
Answer medically plausible (regarding case and illness script)	
+1: Answer plausible, relevant to case	• Student: "What was your starting weight?" • GPT: "I previously weighed 115 kg. Now I weigh only 105 kg."
0: Answer plausible, not case-specific	• Student: "All right!" • GPT: "Okay."
–1: Answer implausible	• No example occurred
Not applicable	• For QAPs with missing answers due to technical reasons

Analysis of the Completeness of History Taking

The completeness of the medical history for the prespecified topics was assessed by GPT-4 (see "Feedback Prompt") and by a human rater (FH). To extract the feedback from GPT-4, we used the JSON file. For the human feedback, the rater assessed each QAP for the categories covered in a Microsoft Excel (version 16.0.10394.20022) spreadsheet. Both data sets were imported into R (version 4.3.1; The R Foundation) [31] for statistical analysis and figure generation. We calculated Cohen κ to compare the feedback from GPT-4 and the human rater on the chat using the R function "CohenKappa" from the "DescTools" package. Categories with $\kappa < 0.6$ were further examined by raters FH and MM in order to identify possible explanations.

All numerical data were assessed for normal distribution and, in this article, are presented as means and standard deviations. If the data deviated from a Gaussian distribution, then we provided the median and interquartile range (Q25-Q75).

Ethical Considerations

This study was approved by the Ethics Committees of the Faculty of Medicine at Tübingen University Hospital (605/2023BO2). Participation in the study was voluntary, without any compensation, and data was collected anonymized. All methods were implemented in accordance with the Declaration of Helsinki.

Results

Participants' Demographic Data

Of the 111 students asked to participate, 5 could not due to experiencing technical problems with the interview platform. All remaining 106 students agreed to participate; 78 (73.6%) identified as female, 25 (23.6%) as male, and 3 (2.8%) as nonbinary, and participants were 22.8 (SD 3.7) years old on average. As for progress in medical school, 93% of participants (N=99) were in their third year of medical school, whereas the remaining participants were in their first (2/106, 2%), second (1/106, 1%), or fourth (3/106, 3%) years, and one student

XSL•FO
RenderX

provided an implausible answer (1/106, 1%). No student had to be excluded from the analysis.

Characteristics of Conversations

In a total of 106 conversations, 1920 QAPs were recorded. Of them, 26 QAPs (1.4%) had to be excluded due to a missing server response, which left 1894 QAPs for analysis. Each conversation yielded a median number of 18 QAPs (IQR 15-23). Whereas questions consisted of a median of 6 words (IQR 4-9), the answers consisted of a median of 22 words (IQR 15-29).

In our analysis of the participants' wordings of questions, most questions did not show any abnormality (1673/1894, 88.3%). Foreign terms were found in 6.3% of the questions (119/1894), chain questions in 3.3% (n=62/1894), single-word questions in 1.2% (23/1894), and incomprehensible wording in 0.7% (13/1894). Four questions (0.2%) contained both a chain question and foreign terms.

Quality of GPT-4's Role-Play

To further assess GPT-4's accuracy in providing a simulated patient chatbot, we assessed the quality of the role-play in the first 20% of conversations, which resulted in the analysis of 410 QAPs, as previously described [13].

Our script covered the majority of questions asked by participants (354/410, 86.3%), with 28 questions (6.8%) partly covered and 13 questions (3.2%) not covered at all by the script (not applicable: 15/410, 3.7%—that is, when no question was asked).

As for the answers provided by GPT-4, 99.3% of them matched the question (n=407), and no answer failed to match the question altogether (not applicable: n=3, 0.7%—that is, provided an answer to a previous question).

Regarding the plausibility of the answers provided by GPT-4, 99.3% (n=407) were rated as plausible, none as implausible, and 0.7% (3/410) as neither implausible nor plausible.

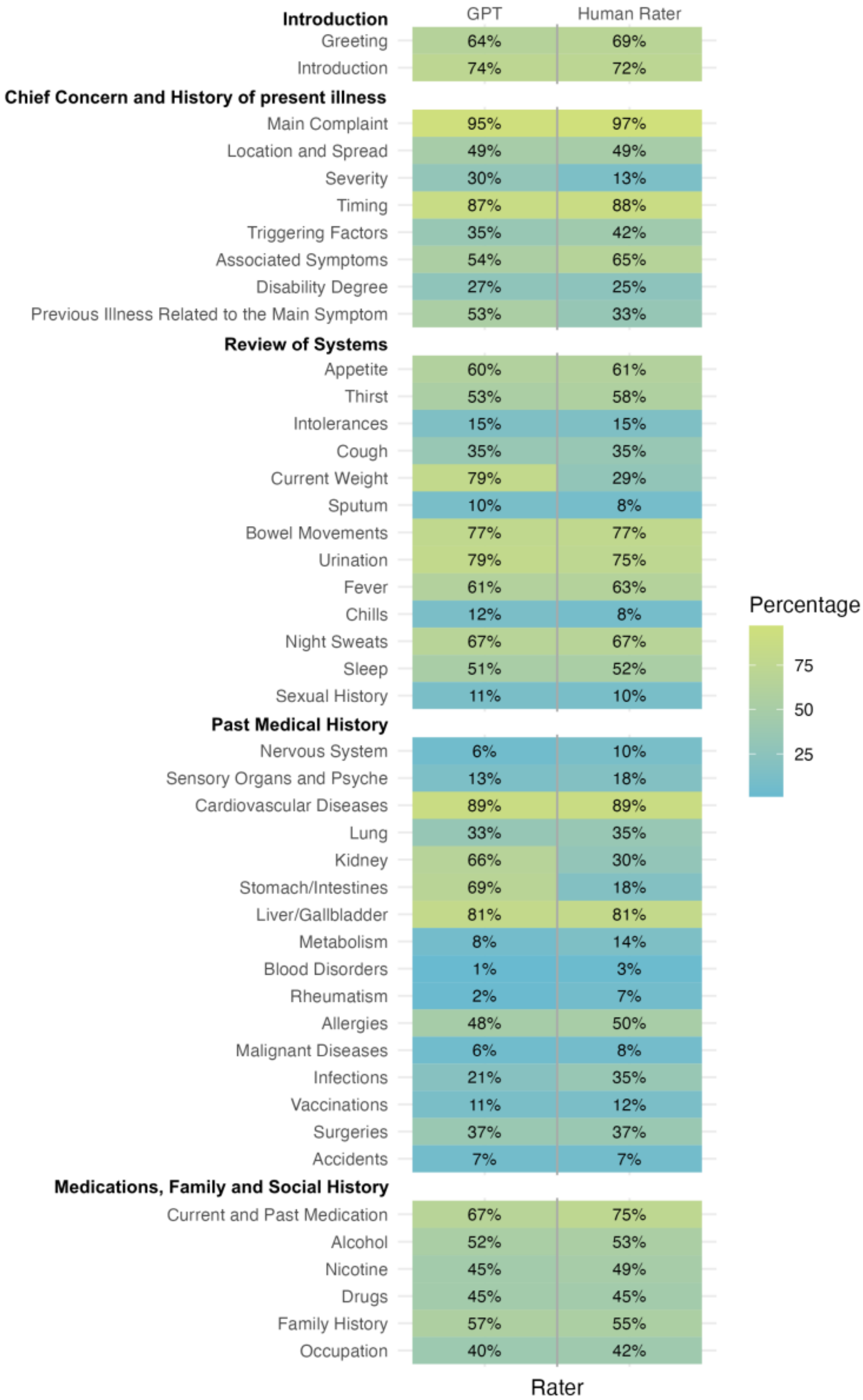
Assessment of History Taking

Coverage of Feedback Categories and Items

Participants' history taking was assessed by both GPT-4 and the human rater (Figure 2). Combining both raters, the first feedback category (ie, introduction) was mentioned by 69.6% of participants, whereas the second category (ie, main complaint) was addressed by 52.7%. A total of 45.1% of participants asked about the vegetative system, and a system assessment was performed by 29.7% of participants. The fifth feedback category (ie, medication, family, social environment, and drugs) was addressed by 52% of participants.

XSL•FO
RenderX

Figure 2. Heat map showing the percentage of conversations mentioning the feedback categories for both raters: Generative Pretrained Transformer (GPT) in the first column, and human rater in the second.

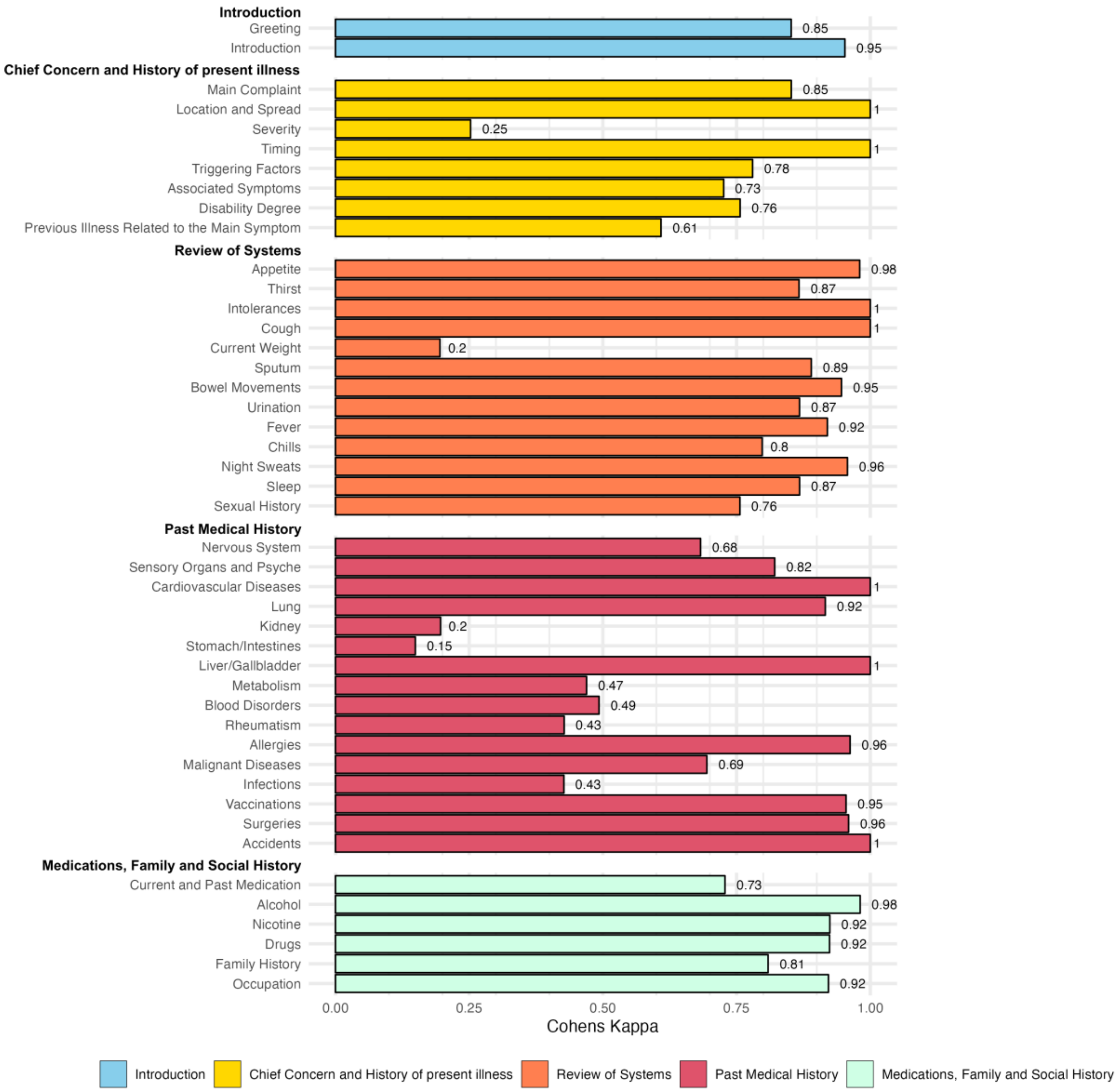


Interrater Reliability

For total feedback, we found an interrater reliability, measured by Cohen κ, of 0.832 (95% CI 0.816-0.848), indicating an "almost perfect" agreement [32]. We further analyzed Cohen κ for each individual category of feedback, displayed in Figure 3.

XSL•FO
RenderX

Figure 3. Cohen κ for every category of feedback for the human rater and Generative Pretrained Transformer (GPT) as a rater, with the different feedback topics displayed in different colors.



Analysis of Divergent Ratings

As displayed in Figure 3, we found at least substantial interrater agreement for most categories of feedback. If conversations had divergent ratings, then we first inspected them in detail to evaluate whether agreement between the human rater and the GPT rating could be achieved. After corrections, 8 out of 42 categories still demonstrated lower-than-expected agreement (κ<0.6) and were, thus, further inspected (Table 2). For those categories, we performed a throughout analysis of the ratings and discussed possible reasons for the divergent ratings.

XSL•FO
RenderX

Table 2. List of feedback categories with Cohen κ<0.6.

Feedback category	Cohen κ	Mentioned (GPT)	Mentioned (Human rater)	Probable explanations for low Cohen's κ with suggested solution and specific example (if appropriate)
Severity	0.25	30%	13%	• The category "Severity" derived from a pain history. In the context of the illness script, there was overlap with the category "disability degree." • Suggested solution: Clarify category "Severity" and possibly rename it "Pain, Numeric Analogue Scale." • Specific example (from the illness script): • Severity: "Recently, I've been significantly restricted. In the evenings after a long workday, I can't do anything, and I've also noticed that I keep forgetting things at work." • Disability degree: "By now, I feel severely limited. This can't continue. I can't manage either my work or the tasks at home with my family like this!"
Current weight	0.20	79%	29%	• Probably different interpretation: • GPT was more liberal than the human rater. For example, when students asked any question related to weight, GPT rated it as "yes," whereas the human rater rated it as yes only when actual weight was mentioned. • Suggested solution: Define category more precisely or split category in "Current Weight" and "Weight Dynamics." • Specific example (from the illness script): • "Overweight, previously 115 kg at a height of 178 cm, but now I only weigh 105 kg."
Kidney	0.20	66%	30%	• Polyuria has been repeated in the category "Kidney" because it was deemed highly important information. However, it resulted in an overlap with the category "Urination." • Suggested solution: Give information only once and precisely. • Specific example (from the illness script): • Urination: "Lately, I've been experiencing frequent urination during the day and at night. There's no pain during urination, and the urine looks normal, as usual." • Kidney: "No pre-existing conditions, but now I constantly have to go to the toilet at night. However, I also haven't been to a urologist in a long time."
Stomach or intestines	0.15	69%	18%	• Overlap exists with the category "Bowel Movements," however, medically challenging to separate clearly. • Suggested solution: Amend prompt to instruct GPT to rate both categories as "Yes" when a question or its answer covers both categories clearly and completely. • Specific example (from the illness script): • Stomach or intestines: "Mild tendency towards constipation" • Bowel Movements: "Tending more towards constipation, but recently having a regular bowel movement once a day. Stool is otherwise normal: brown, without blood, without mucus, and without diarrhoea."
Metabolism	0.47	8%	14%	• Probably a different interpretation: GPT did not rate conversations positively when students asked for "metabolism disorders" and "diabetes." Because we could not explain those ratings, we prompted GPT to explain its reasoning. The answer included that metabolism "encompasses all the chemical reactions that occur in the body" and includes aspects on "how [the] body converts food into energy," thereby confirming our suspicion of different interpretations. • Example of a question rated "Yes" by human rater and "No" by GPT: "Are you aware of having diabetes or hypercholesterolemia?"
Blood disorders	0.49	1%	3%	• Low prevalence of "Yes" in the feedback category [33]
Rheumatism	0.43	2%	7%	• Low prevalence of "Yes" in the feedback category [33]

XSL•FO
RenderX

Feedback category	Cohen κ	Mentioned (GPT)	Mentioned (Human rater)	Probable explanations for low Cohen’s κ with suggested solution and specific example (if appropriate)
Infections	0.43	21%	35%	• Category not defined clearly enough with overlaps between “recent infections” and “infectious diseases.” • Suggested solution: Amend illness script to include both categories and define both categories clearly. • Example of statement rated “Yes” by GPT and “No” by human rater: “Additionally, I suffer from many simple infections, an increased sense of thirst, and dizziness.”

Discussion

In our study, we assessed GPT-4’s performance in providing automatic feedback on learners’ history taking in a large cohort of medical students. Our findings suggest that GPT-4, accessed via an API, is capable of not only simulating patient experiences through a chatbot-like interface but also of providing accurate feedback on medical history-taking dialogs.

Principal Results

Extending the line of our group’s previous research, the study presented here confirmed GPT-4’s capability of offering medically plausible responses in more than 99% of interactions, with a negligible rate of missing server responses (1.4%) that showcases its high reliability and availability in medical training [13]. That technical capability is particularly relevant when considering the asynchronous nature of such feedback systems in educational settings [34]. Building on our past work [13], we have demonstrated that GPT-4 can not only act as a simulated patient chat bot but can also assist the learner in providing structured feedback on the topics covered or not covered by the student.

The high level of agreement (Cohen κ=0.832) between GPT ratings and human ratings of students’ input that we observed indicates GPT-4’s capabilities in evaluating history-taking dialogs. It also supports GPT-4’s potential to enhance medical education by providing immediate, accurate feedback to students, thereby potentially fostering the learning process by enabling more practice opportunities and instant feedback. Given the importance of feedback for the learning process, the result offers an encouraging perspective on how LLMs such as GPT-4 can be used to cultivate the skills acquisition of medical students [17,18].

At the same time, we also found 8 feedback categories that yielded a Cohen κ of less than 0.6. For those items, in some cases we found GPT-4 to be “overly specific” in its rating. For example, in the category “Current Weight,” GPT-4 rated the occurrence of the topic “weight” positively (ie, disregarding whether the actual weight was mentioned), whereas the human rater focused on whether the actual weight was present in the chat. Those cases can probably be attributed to different interpretations of the items rated, and they indicate that the prompting should be as specific as possible in order to achieve higher interrater reliability.

We further hypothesize that those ratings can be improved by providing more detailed specifications for every category—for instance, by including examples and using more advanced prompting techniques such as chain-of-thought prompting [35]. However, longer prompts might be problematic when using models such as GPT-4, for the context window is limited to 8192 tokens [36]. Although our prompts (ie, system prompt of 2303 tokens and feedback prompt of 1336 tokens) fit well within those limits, longer prompts could require more advanced LLMs with longer context windows.

Furthermore, some lower κ values could have been caused by certain categories overlapping with other categories (eg, “Kidney” and “Urination”). Because medical cases often affect multiple topics, future studies should focus on the clear separation of feedback items. In our study, we did not prompt GPT-4 to provide any reasoning for the ratings (eg, in “chain-of-thought” prompting [37]), which researchers could improve upon in the future in order to better understand the models’ output.

Regarding the performance of the participating students, completeness scores for the feedback topics ranged from 31.0% to 68.9%. Although such rates might seem to indicate only modest performance, students also had a time restriction of 8 minutes maximum (ie, owing to the practising circuit that our chatbot was embedded in), which made a complete history-taking dialog exceptionally difficult.

Comparison With Prior Work

Since the development of digital learning systems, automatic feedback has emerged as a topic of interest. Covering the pre-LLM era, a systematic review from 2021 analyzed 63 studies, most of them examining programming and mathematical skills [23]. While the review’s authors concluded that automatic feedback can foster students’ performance, the main method of generating automatic feedback was a comparison with a desired answer [23]. Further developments then included sophisticated dialog management systems [38], although those systems still performed below the level of feedback generated by LLMs. Because those pre-LLM technologies have been shown to help students [23], it can be expected that properly employed LLMs might provide even more benefits to learners (although the comparison was not investigated in our study).

Consequently, the recent emergence of LLMs such as GPT has been heralded as having the potential to revolutionize how students learn [39]. For example, Dai et al [40] found that ChatGPT was capable of generating more detailed feedback than human instructors while also achieving high agreement with the instructor. Beyond that, and in line with our results, in a study with students learning English as a new language, feedback from GPT-4 was found to be of similar quality to

XSL•FO
RenderX

human feedback regarding learning outcomes and students' perception [41]. Furthermore, LLM-based feedback has been shown to elucidate secondary effects, including increasing positive emotions and task motivation [42]. Indeed, the high motivation of students to participate in our study and in past investigations supports that motivational aspect [13]. Another essential aspect is the curricular implementation of the feedback, which is important for learners to develop a widespread understanding and develop mastery [10]. However, when implemented correctly, LLMs offer new tools for education and can be further improved when combined with speech-to-text tools and personalized databases [43].

However, some studies have also revealed problems with AI-generated feedback. For example, one showed that some participants might have negative attitudes toward the feedback due to being AI-generated feedback [44]. Such attitudes could affect learning outcomes considering that students' perception of feedback is associated with self-regulated learning [45]. Furthermore, LLMs might elicit unexpected behaviors and escape prompts, thereby resulting in problematic interactions [46]. Although we did not observe that unexpected behavior in our study, the feedback provided by the AI might ultimately be understood as "official" feedback and should thus be rigorously assessed for its quality. Last, incorporating AI in teaching might lead students to rely on AI instead of learning from it [47], which indicates the importance of keeping the complete learning task in mind when designing AI-based learning opportunities.

Limitations

Our findings have some limitations that deserve discussion. First, we relied on 1 LLM (ie, GPT-4) and a single prompt in our study. Although our study has demonstrated GPT-4's potential in medical education, our reliance on a single LLM and type of prompt means that our findings might not apply to all educational contexts. Future research should, therefore, explore a variety of prompts and LLMs. Second, we chose a specific case for the history-taking dialog. Although we believe that GPT-4's observed performance is transferable, our data cannot corroborate that assumption. Exploring a variety of cases and conditions would provide a more robust understanding of GPT-4's applicability and limitations. Third, we used binary criteria (ie, "yes" or "no") for the completeness of history taking in order to provide students with a simple checklist on what was asked or not asked. However, real-world clinical dialogs and history taking are complex and might benefit from more nuanced evaluation in order to accurately reflect which skills and topics students need to improve upon. Beyond that, it is important for students to receive feedback from the AI-generated tool on their social skills (eg, nonverbal communication and comprehensible language) during patient-physician encounters, which should be further investigated in future research. Last, we did not measure any educational outcomes (ie, skill acquisition), and thus, cannot state whether the AI-generated feedback in fact improved students' performance.

Conclusions

In sum, the LLM GPT-4 can provide a simulated patient experience and generate tailored, unsupervised feedback for medical students. The feedback given by GPT-4 was mostly accurate and had few minor flaws, most of which likely stemmed from our prompts. Our findings support the implementation of the system and the evaluation of its effectiveness in subsequent assessments.

Acknowledgments

We wish to thank the Open Access Publishing Fund of the University of Tübingen for supporting our study and Eric Nazarenus for his assistance with our analysis.

Data Availability

The data sets used and analyzed in our study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full prompt.
[PDF File (Adobe PDF File), 83 KB - mededu_v10i1e59213_app1.pdf]

References

1. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J 1975;2(5969):486-489 [FREE Full text] [doi: 10.1136/bmj.2.5969.486] [Medline: 1148666]
2. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 1992;156(2):163-165 [FREE Full text] [Medline: 1536065]
3. Dorr Goold S, Lipkin M. The doctor-patient relationship: challenges, opportunities, and strategies. J Gen Intern Med 1999;14(Suppl 1):S26-S33 [FREE Full text] [doi: 10.1046/j.1525-1497.1999.00267.x] [Medline: 9933492]

XSL•FO
RenderX

4. Hausberg MC, Hergert A, Kröger C, Bullinger M, Rose M, Andreas S. Enhancing medical students' communication skills: development and evaluation of an undergraduate training program. BMC Med Educ 2012;12:16 [FREE Full text] [doi: 10.1186/1472-6920-12-16] [Medline: 22443807]
5. Deveugele M, Derese A, De Maesschalck S, Willems S, Van Driel M, De Maeseneer J. Teaching communication skills to medical students, a challenge in the curriculum? Patient Educ Couns 2005;58(3):265-270. [doi: 10.1016/j.pec.2005.06.004] [Medline: 16023822]
6. Noble LM, Scott-Smith W, O'Neill B, Salisbury H, UK Council of Clinical Communication in Undergraduate Medical Education. Consensus statement on an updated core communication curriculum for UK undergraduate medical education. Patient Educ Couns 2018;101(9):1712-1719. [doi: 10.1016/j.pec.2018.04.013] [Medline: 29706382]
7. Borowczyk M, Stalmach-Przygoda A, Doroszewska A, Libura M, Chojnacka-Kuraś M, Małecki Ł, et al. Developing an effective and comprehensive communication curriculum for undergraduate medical education in Poland—the review and recommendations. BMC Med Educ 2023;23(1):645 [FREE Full text] [doi: 10.1186/s12909-023-04533-5] [Medline: 37679670]
8. Laidlaw A, Hart J. Communication skills: an essential component of medical curricula. Part I: Assessment of clinical communication: AMEE guide No. 51. Med Teach 2011;33(1):6-8. [doi: 10.3109/0142159X.2011.531170] [Medline: 21182378]
9. Kaplonyi J, Bowles KA, Nestel D, Kiegaldie D, Maloney S, Haines T, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ 2017;51(12):1209-1219. [doi: 10.1111/medu.13387] [Medline: 28833360]
10. Kelly S, Smyth E, Murphy P, Pawlikowska T. A scoping review: virtual patients for communication skills in medical undergraduates. BMC Med Educ 2022;22(1):429 [FREE Full text] [doi: 10.1186/s12909-022-03474-9] [Medline: 35659213]
11. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Med Educ 2022;22(1):365 [FREE Full text] [doi: 10.1186/s12909-022-03410-x] [Medline: 35550085]
12. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023;25:e43311 [FREE Full text] [doi: 10.2196/43311] [Medline: 37335593]
13. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt J, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024;10:e53961 [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
14. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ 2020;54(9):786-795. [doi: 10.1111/medu.14152] [Medline: 32162355]
15. Chung K, Park RC. Chatbot-based heathcare service with a knowledge base for cloud computing. Cluster Comput 2018;22(S1):1925-1937. [doi: 10.1007/s10586-018-2334-5]
16. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023;15(5):e38755 [FREE Full text] [doi: 10.7759/cureus.38755] [Medline: 37303324]
17. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME guide No. 7. Med Teach 2006;28(2):117-128. [doi: 10.1080/01421590600622665] [Medline: 16707292]
18. Bing-You R, Hayes V, Varaklis K, Trowbridge R, Kemp H, McKelvy D. Feedback for learners in medical education: what is known? A scoping review. Acad Med 2017;92(9):1346-1354. [doi: 10.1097/ACM.0000000000001578] [Medline: 28177958]
19. Yan L, Sha L, Zhao L, Li Y, Martinez‐Maldonado R, Chen G, et al. Practical and ethical challenges of large language models in education: a systematic scoping review. Brit J Educational Tech 2023;55(1):90-112. [doi: 10.1111/bjet.13370]
20. Wang L, Chen X, Wang C, Xu L, Shadiev R, Li Y. ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: a case study. Think Ski Creat 2024;51:101440. [doi: 10.1016/j.tsc.2023.101440]
21. Carlson M, Pack A, Escalante J. Utilizing OpenAI's GPT-4 for written feedback. TESOL J 2023;15(2):e759. [doi: 10.1002/tesj.759]
22. Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI 2024. [doi: 10.1056/aioa2400196]
23. Cavalcanti AP, Barbosa A, Carvalho R, Freitas F, Tsai YS, Gašević D, et al. Automatic feedback in online learning environments: a systematic literature review. Comput Educ Artif Intell 2021;2:100027. [doi: 10.1016/j.caeai.2021.100027]
24. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
25. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]

XSL•FO
RenderX

26. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
27. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's mastery of bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024;26:e52113 [FREE Full text] [doi: 10.2196/52113] [Medline: 38261378]
28. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1-38. [doi: 10.1145/3571730]
29. OpenAI Platform. URL: https://platform.openai.com [accessed 2024-02-03]
30. Zhang Y, Du L, Cao D, Fu Q, Liu Y. Prompting large language models with divide-and-conquer program for discerning problem solving. arXiv 2024. [doi: 10.48550/arXiv.2402.05359]
31. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023.
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]
33. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46(5):423-429. [doi: 10.1016/0895-4356(93)90018-v] [Medline: 8501467]
34. Memarian B, Doleck T. ChatGPT in education: methods, potentials, and limitations. Comput Hum Behav Artif Hum 2023;1(2):100022. [doi: 10.1016/j.chbah.2023.100022]
35. Fagbohun O, Harrison RM, Dereventsov A. An empirical categorization of prompting techniques for large language models: a practitioner's guide. arXiv. 2024. URL: http://arxiv.org/abs/2402.14837 [accessed 2024-03-25]
36. GPT-4. URL: https://openai.com/research/gpt-4 [accessed 2024-03-25]
37. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv 2023. [doi: 10.48550/arXiv.2201.11903]
38. Haut K, Wohn C, Kane B, Carroll T, Guigno C, Kumar V, et al. Validating a virtual human and automated feedback system for training doctor-patient communication skills. : IEEE; 2023 Presented at: 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII); 2023 June 27; MA, USA p. 1-8. [doi: 10.1109/acii59096.2023.10388213]
39. 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;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
40. Dai W, Lin J, Jin H, Li T, Tsai Y, Gašević D, et al. Can large language models provide feedback to students? A case study on ChatGPT. 2023 Presented at: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT); 2023 July 10; Orem, UT, USA p. 323-325. [doi: 10.1109/icalt58122.2023.00100]
41. Escalante J, Pack A, Barrett A. AI-generated feedback on writing: insights into efficacy and ENL student preference. Int J Educ Technol High Educ 2023;20(1):57. [doi: 10.1186/s41239-023-00425-2]
42. Meyer J, Jansen T, Schiller R, Liebenow LW, Steinbach M, Horbach A, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Comput Educ Artif Intell 2024;6:100199. [doi: 10.1016/j.caeai.2023.100199]
43. Barker LA, Moore JD, Cook HA. Generative artificial intelligence as a tool for teaching communication in nutrition and dietetics education—a novel education innovation. Nutrients 2024;16(7):914 [FREE Full text] [doi: 10.3390/nu16070914] [Medline: 38612948]
44. Häkkinen J, Ramadan Z. A Study on the Perception of Feedback with Varying Sentiment Generated Using a Large Language Model. Stockholm, Swedem; 2023. URL: https://www.diva-portal.org/smash/get/diva2:1779789/FULLTEXT01.pdf [accessed 2024-03-25]
45. He J, Liu Y, Ran T, Zhang D. How students' perception of feedback influences self-regulated learning: the mediating role of self-efficacy and goal orientation. Eur J Psychol Educ 2022;38(4):1551-1569. [doi: 10.1007/s10212-022-00654-5]
46. Bowman SR. Eight things to know about large language models. arXiv 2023.
47. Darvishi A, Khosravi H, Sadiq S, Gašević D, Siemens G. Impact of AI assistance on student agency. Comput Educ 2024;210:104967. [doi: 10.1016/j.compedu.2023.104967]

Abbreviations

AI: artificial intelligence
API: application programming interface
GPT: Generative Pretrained Transformer
LLM: large language model
QAP: question-answer pair

XSL•FO
RenderX

Edited by B Lesselroth; submitted 05.04.24; peer-reviewed by SC Tan; comments to author 02.05.24; revised version received 21.05.24; accepted 27.06.24; published 16.08.24.

Please cite as:
Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, Mahling M
A Language Model–Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study
JMIR Med Educ 2024;10:e59213
URL: https://mededu.jmir.org/2024/1/e59213
doi:10.2196/59213
PMID:

©Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, Moritz Mahling. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.08.2024. 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.

XSL•FO
RenderX

Original Paper

Knowledge Mapping and Global Trends in the Field of the Objective Structured Clinical Examination: Bibliometric and Visual Analysis (2004-2023)

Hongjun Ba[1], MD; Lili Zhang[1], MM; Xiufang He[1], MM; Shujuan Li[1], MD

Department of Pediatric Cardiology, First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

Corresponding Author:
Shujuan Li, MD
Department of Pediatric Cardiology
First Affiliated Hospital of Sun Yat-sen University
58# Zhongshan Road 2
Guangzhou, 510080
China
Phone: 86 13430329103
Email: lishuj2@mail.sysu.edu.cn

Abstract

Background: The Objective Structured Clinical Examination (OSCE) is a pivotal tool for assessing health care professionals and plays an integral role in medical education.

Objective: This study aims to map the bibliometric landscape of OSCE research, highlighting trends and key influencers.

Methods: A comprehensive literature search was conducted for materials related to OSCE from January 2004 to December 2023, using the Web of Science Core Collection database. Bibliometric analysis and visualization were performed with VOSviewer and CiteSpace software tools.

Results: Our analysis indicates a consistent increase in OSCE-related publications over the study period, with a notable surge after 2019, culminating in a peak of activity in 2021. The United States emerged as a significant contributor, responsible for 30.86% (1626/5268) of total publications and amassing 44,051 citations. Coauthorship network analysis highlighted robust collaborations, particularly between the United States and the United Kingdom. Leading journals in this domain—*BMC Medical Education*, *Medical Education*, *Academic Medicine*, and *Medical Teacher*—featured the highest volume of papers, while *The Lancet* garnered substantial citations, reflecting its high impact factor (to be verified for accuracy). Prominent authors in the field include Sondra Zabar, Debra Pugh, Timothy J Wood, and Susan Humphrey-Murto, with Ronaldo M Harden, Brian D Hodges, and George E Miller being the most cited. The analysis of key research terms revealed a focus on “education,” “performance,” “competence,” and “skills,” indicating these are central themes in OSCE research.

Conclusions: The study underscores a dynamic expansion in OSCE research and international collaboration, spotlighting influential countries, institutions, authors, and journals. These elements are instrumental in steering the evolution of medical education assessment practices and suggest a trajectory for future research endeavors. Future work should consider the implications of these findings for medical education and the potential areas for further investigation, particularly in underrepresented regions or emerging competencies in health care training.

(JMIR Med Educ 2024;10:e57772) doi:10.2196/57772

KEYWORDS

Objective Structured Clinical Examination; OSCE; medical education assessment; bibliometric analysis; academic collaboration; health care professional training; medical education; medical knowledge; medical training; medical student

Introduction

Objective Structured Clinical Examinations (OSCEs) have emerged as indispensable tools for assessing health care professionals, providing structured evaluations of clinical competencies, communication skills, and decision-making abilities [1,2]. Despite their widespread adoption since the 1970s, the landscape of OSCE research remains multifaceted and dynamic, reflecting ongoing innovations in medical, nursing, and allied health education [3].

XSL•FO
RenderX

While numerous studies have explored various aspects of OSCEs, gaps persist in our understanding of the overarching trends and global dynamics shaping this field. A comprehensive review of the existing literature highlights the need for a systematic approach to mapping the knowledge landscape and identifying emerging trends through bibliometric analysis [4-6]. By applying quantitative methods to scholarly publications, bibliometric analysis offers a unique opportunity to uncover hidden patterns, elucidate research trajectories, and forecast future directions in OSCE research.

Building on this rationale, our study aims to bridge these gaps by conducting a bibliometric analysis of OSCE literature from 2004 to 2023. We hypothesize that this analysis will reveal distinct patterns of publication output, collaboration networks, and thematic clusters within the OSCE research domain. Specifically, we seek to (1) identify key research themes, including but not limited to assessment methodologies, educational interventions, and technological innovations in OSCEs; (2) map the global distribution of OSCE research, highlighting geographic hotspots and areas of collaboration; and (3) explore the interconnections between different disciplines within medical education, shedding light on interdisciplinary collaborations and knowledge diffusion.

By elucidating these aspects, our study aims to provide stakeholders in medical education with valuable insights into the current state and future directions of OSCE research. Ultimately, this knowledge mapping exercise seeks to inform evidence-based decision-making, guide educational practices, and stimulate further research in the field of clinical skills assessment.

Methods

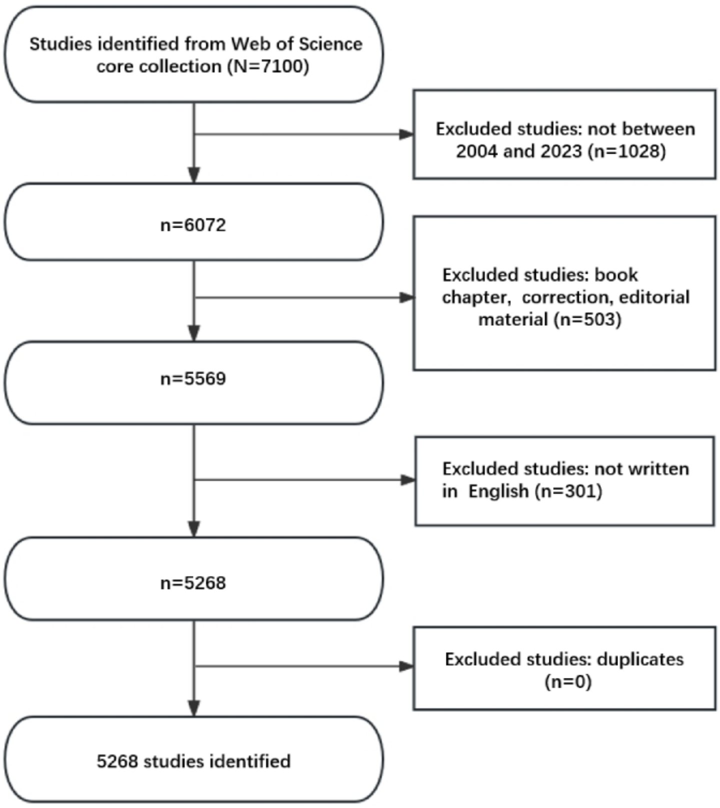
Data Acquisition and Search Strategy

The bibliographic accuracy of literature types in the Web of Science Core Collection (WoSCC) database is superior to any other database, making it the optimal choice for conducting literature analysis [7,8]. Therefore, we opted to perform our search within this database. We conducted a search in the Web of Science (WoS) for all relevant papers published between January 1, 2004, and December 31, 2023. The search formula "(TS=(The Objective Structured Clinical Examination)) or TS=(OSCE)" was used. The literature screening for this study was based on the inclusion criteria: (1) full-text publications related to the OSCEs; (2) papers and review manuscripts written in English; and (3) papers published between January 1, 2004, and December 31, 2023. The exclusion criteria included (1) topics not related to the OSCEs and (2) papers in the form of conference abstracts, news briefs, and so on. A plain text version of the papers was exported.

General Data

Figure 1 shows the process of literature searching and bibliometric analysis. The results indicate that from January 1, 2004, to December 31, 2023, there were a total of 5268 publications related to the OSCE in the WoSCC database, including 1800 papers (84.96%) and 384 reviews (15.04%). The literature involved 133 countries and regions, 5291 institutions, and 24,478 authors.

Figure 1. The workflow of data collection and bibliometric analysis.



XSL•FO
RenderX

Data Analysis

To depict annual publication trends and the distribution of national contributions, we used GraphPad Prism (version 8.0.2; Dotmatics). For the bibliometric analysis and the visualization of scientific knowledge maps, the study used both CiteSpace (6.2.4R, 64 bit advanced edition; Chaomei Chen, Drexel University) [9] and VOSviewer (version 1.6.18; Leiden University) [10]. These tools were selected for their robustness in handling extensive bibliometric data and their ability to graphically represent complex networks.

VOSviewer, a Java-based software pioneered by van Eck and Waltman [9] in 2009, facilitates the construction of various types of network maps, such as bibliographic coupling, cocitation, and coauthorship networks. CiteSpace, developed by Professor Chaomei Chen, provides a dynamic and computer-based platform for identifying and visualizing patterns and trends in scientific literature, thereby enabling the exploration of knowledge domains and predictive analysis of research trajectories [10].

Our methodological approach within these applications involved setting specific parameters for network density, threshold values for the inclusion of nodes, and time-slicing techniques to analyze temporal changes. The references corresponding to the software applications were verified against our citation list to ensure accuracy [9,10].

In our study using VOSviewer and CiteSpace software tools for bibliometric analysis, the criteria for defining country-based collaborations were established based on specific considerations. Collaborations were determined by considering the first authors and corresponding authors listed in the paper bylines. This approach was chosen to ensure inclusivity and to capture the entirety of collaborative efforts between researchers from different countries.

The burst detection in CiteSpace is based on the Kleinberg algorithm, which is based on modeling the stream using an infinite-state automaton to extract a meaningful structure from document streams that arrive continuously over time [11]. These analyses can show the fast-growing topics that last for multiple years as well as a single year.

Rationale for Analysis Selection

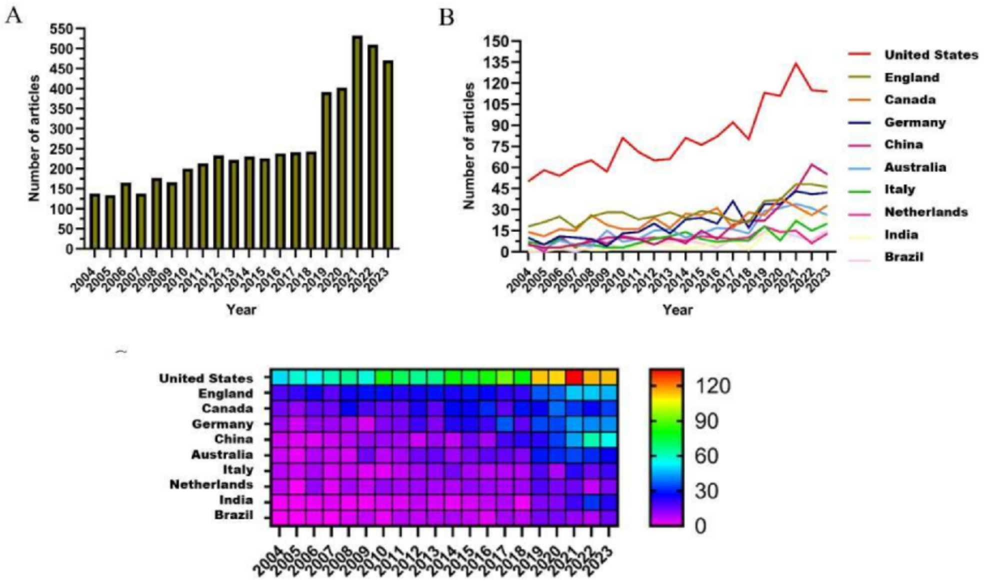
The aforementioned techniques were chosen a priori due to their widespread use and effectiveness in bibliometric studies. They provide robust and complementary insights into productivity, impact, and collaborative patterns within the research field.

Results

Publication Trend

Since 2004, there has been a gradual increase in the number of papers published annually (Figure 2A). We have divided this into 3 periods: from 2004 to 2010, there was a slow growth, with fewer than 150 papers published per year, indicating that the field had not yet captured researchers' attention. From 2011 to 2018, the volume of publications gradually increased, indicating growing interest in the field. After 2019, there was a rapid rise in the number of publications, peaking in 2021, which suggests that the field has received widespread attention since then.

Figure 2. Trend chart of publications in the past 20 years. (A) Annual publication count chart. (B) Line chart of national publication count. (C) Heatmap of national publication count.



Country or Region and Institution Contributions

Figure 2B and C show the annual number of publications from the top 10 countries over the past decade. The top 5 countries in the field are the United States, the United Kingdom, Canada, Germany, and China, respectively. The United States accounts for 30.86% (1626/5268) of the total volume of publications, significantly surpassing other countries.

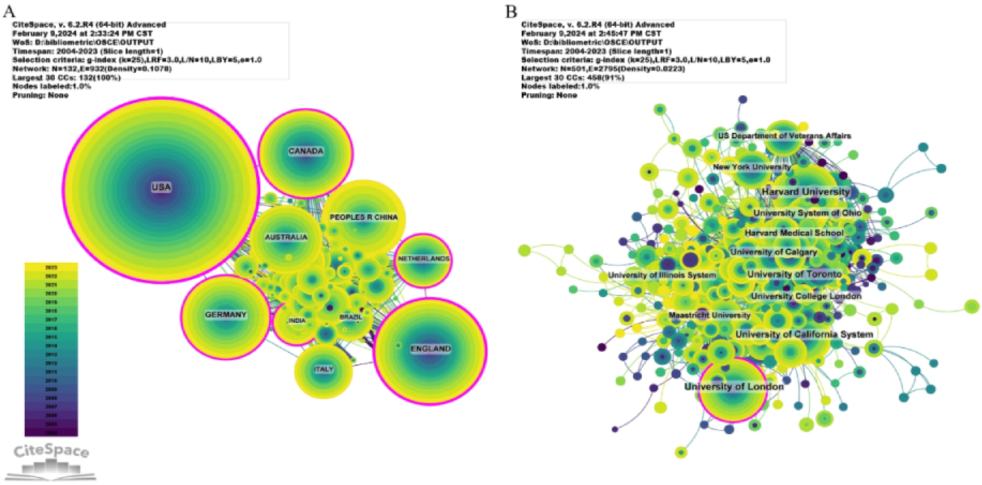
Among the top 10 countries or regions in terms of the number of published papers, the United States had a citation count of 44,051, far exceeding all other countries or regions. Its citation-per-publication ratio (27.13) ranks third among all

XSL•FO
RenderX

countries or regions, which suggests a generally high quality of the published papers. The United Kingdom had the second-highest number of published papers (576 papers) and ranked second in terms of citation count (15,929 citations). The cooperation network, as shown in Figure 3A, indicates close collaboration between the United States and the United Kingdom, which are the highest producers.

A total of 5291 institutions have systematically published papers related to the OSCE. Among the top 10 institutions in terms of publication volume, 6 are from the United States, 2 are from the United Kingdom, and 2 are from Canada (Figure 3B).

Figure 3. Network graph of national and institutional collaborations. (A) Network graph of national collaborations. (B) Network graph of institutional collaborations. The bubble size represents the number of publications. WoS: Web of Science.



Journals' Contributions

Tables 1 and 2 list the top 10 journals with the highest outputs and the most citations, respectively. *BMC Medical Education*, with 227 out of 5268 papers, accounting for 4.31% of publications in the field, is the journal with the most published papers, followed by *Medical Teacher* (179/5268, 3.40%), *Medical Education* (132/5268, 2.51%), and *Journal of Surgical Education* (66/5268, 1.25%). Among the top 10 most productive journals, *Annals of the Rheumatic Diseases* has the highest impact factor at 27.6. All journals are categorized within either Q1 or Q2 quartiles.

The influence of a journal is determined by the frequency with which it is cocited, which indicates whether the journal has made a significant impact on the scientific community. According to Table 2, the most commonly cocited journal is *Medical Education* with 1868 citations, followed by *Academic Medicine* with 1775 citations, and *Medical Teacher* with 1597 citations. Among the top 10 journals by cocitation count, *The Lancet* was cited 697 times and has the highest impact factor of 168.9 within these top journals. All journals within the most cocited list are in the Q1 or Q2 zone.

Table 1. Top 10 most productive journals.

Rank	Journals	Papers (N=5268), n (%)	IF[a]	Quartile in category
1	*BMC Medical Education*	227 (4.31)	3.6	Q1
2	*Medical Teacher*	179 (3.40)	4.7	Q1
3	*Medical Education*	132 (2.51)	7.1	Q1
4	*Journal of Surgical Education*	66 (1.25)	2.9	Q2
5	*Academic Medicine*	64 (1.21)	7.4	Q1
6	*Patient Education and Counseling*	64 (1.21)	3.5	Q2
7	*Advances in Health Sciences Education*	60 (1.14)	4.0	Q1
8	*American Journal of Pharmaceutical Education*	59 (1.12)	3.3	Q2
9	*PLoS One*	59 (1.12)	3.7	Q2
10	*Nurse Education Today*	56 (1.06)	3.9	Q1

[a]IF: impact factor.

XSL•FO
RenderX

Table 2. Top 10 journals with the highest number of cocitations. Cocited journals refer to 2 or more journals that are simultaneously cited in the reference lists of other research papers.

Rank	Cited journals	Cocitations, n	IF[a] (2020)	Quartile in category
1	*Medical Education*	1868	4.7	Q1
2	*Academic Medicine*	1775	7.4	Q1
3	*Medical Teacher*	1597	4.7	Q1
4	*BMC Medical Education*	941	3.6	Q1
5	*JAMA—Journal of American Medical Association*	931	120.7	Q1
6	*British Medical Journal*	827	107.7	Q1
7	*Advances in Health Sciences Education*	802	4.0	Q1
8	*The Lancet*	697	168.9	Q1
9	*New England Journal of Medicine*	694	158.5	Q1
10	*Teaching and Learning Medicine*	599	2.5	Q3

[a]IF: impact factor.

Authors and Cocited Authors' Contributions

Among all authors who have published literature related to OSCE, Tables 3 and 4 list the top 10 authors with the most published papers. Together, these top 10 authors have published 185 papers, accounting for 3.51% of all papers (N=5268) in the field. Sondra Zabar has 26 publications, which is the highest number of published research papers, followed by Debra Pugh with 22, Timothy J Wood with 20, and Susan Humphrey-Murto with 19. Further analysis indicates that among the top 10 ranked authors, 4 are from the United States, 3 are from Canada, 2 are from Australia, and 1 is from China. CiteSpace visualizes the network of relationships between authors (Figure 4).

Table 4 displays the top 10 authors who have been cocited and cited the most, respectively. A total of 148 authors have been cited more than 50 times, indicating that their research has a high reputation and influence. The largest nodes are associated with the authors who have been cocited the most, including Ronald M Harden with 751 citations, Brian D Hodges with 330 citations, and George E Miller with 222 citations.

Table 3. Top 10 most productive authors.

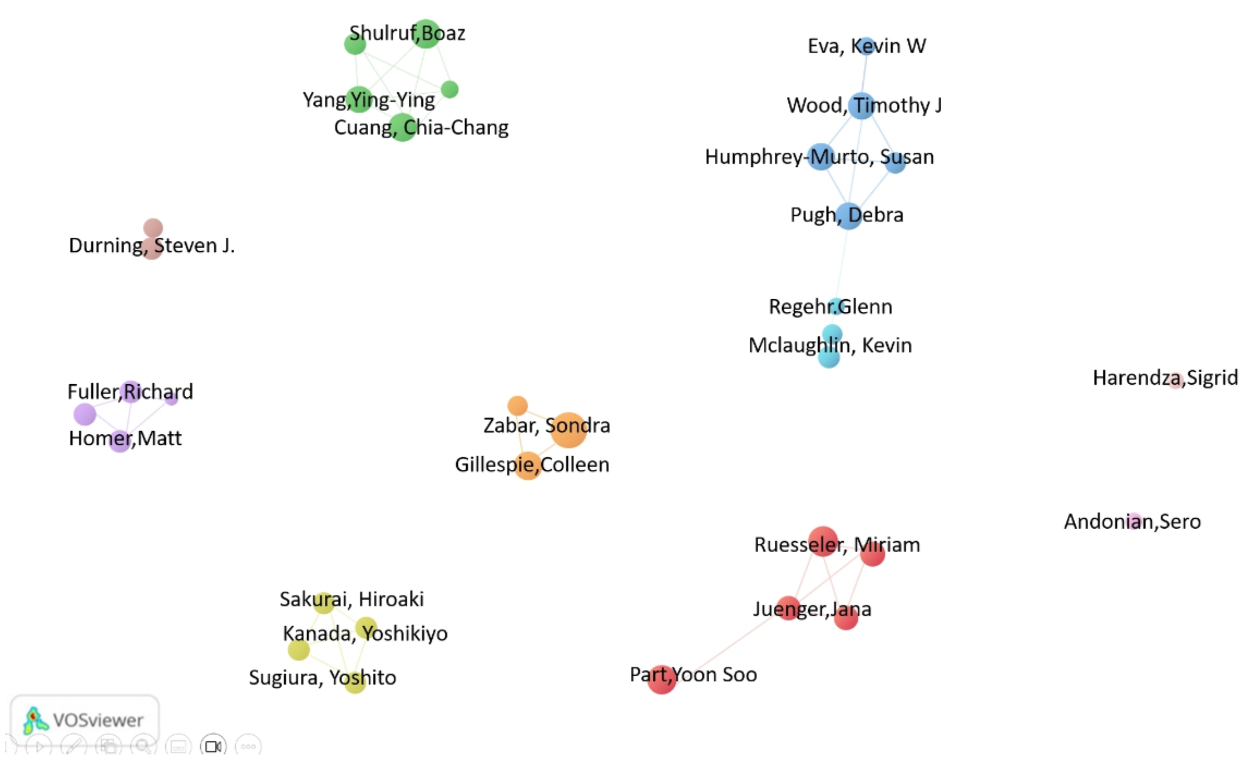
Rank	Authors	Papers, n	Locations
1	Zabar, Sondra	26	United States
2	Pugh, Debra	22	Canada
3	Wood, Timothy J	20	Canada
4	Humphrey-Murto, Susan	19	Canada
5	Gillespie, Colleen	17	United States
6	Shulruf, Boaz	17	Australia
7	Yang, Ying-Ying	17	China
8	Durning, Steven J	16	United States
9	Fuller, Richard	16	Australia
10	Park, Yoon Soo	15	United States

XSL•FO
RenderX

Table 4. Top 10 most cocited authors.

Rank	Cocited authors	Citations, n
1	Harden, Ronald M	751
2	Hodges, Brian D	330
3	Miller, George E	222
4	Epstein, Ronald M	194
5	van der Vleuten, Cees PM	173
6	Wass, Valerie	172
7	Khan, Kamran Z	164
8	Regehr, Glenn	162
9	Cook, David A	160
10	Downing, Steven M	156

Figure 4. Network diagram of author collaborations. The bubble size represents the number of publications.



Analysis of Highly Cited References

Over the time span from 2004 to 2023, the cocitation network comprised 1053 nodes and 3508 links (Figure 5). According to the top 10 papers by cocitation frequency (Table 5), the most cocited reference is from the journal *Advances in Medical Education and Practice* (impact factor=2.0), titled "An evaluative study of Objective Structured Clinical Examination (OSCE): students and examiners perspectives" [12]. The first author of this paper is Md Anwarul Azim Majumder. The paper posits that OSCE is the gold standard and universal form for assessing medical students' clinical competence in a comprehensive, reliable, and effective manner.

XSL•FO
RenderX

Figure 5. Network diagram of cocited references.

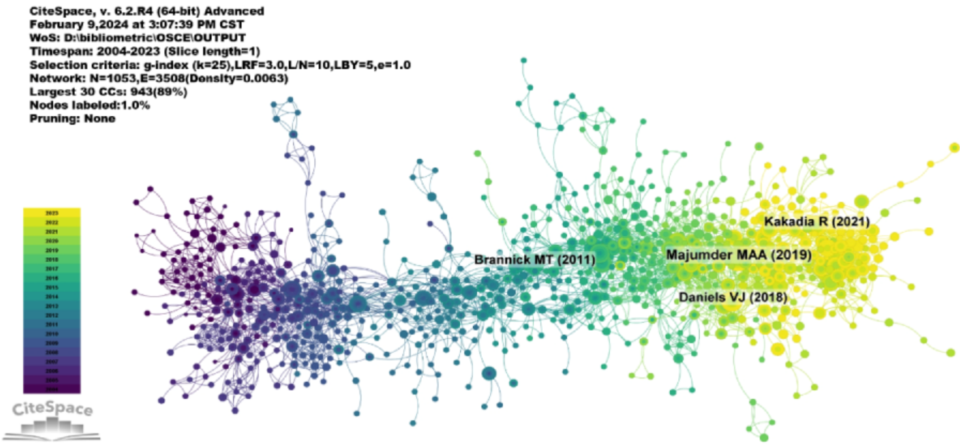


Table 5. Top 10 highest cited references.

Rank	Titles	Journals	IF[a] (2021)	First authors	Total citations, n
1	An evaluative study of Objective Structured Clinical Examination (OSCE): students and examiners perspectives [12]	*Advances in Medical Education and Practice*	2.0	Majumder, Md Anwarul Azim	38
2	Implementing an online OSCE during the COVID-19 pandemic [13]	*Journal of Dental Education*	2.3	Kakadia, Rahen	31
3	Diagnostic and statistical manual of mental disorders [14]	*Psychiatry Research*	11.3	Mittal, Vijay A	31
4	A systematic review of the reliability of Objective Structured Clinical Examination scores [15]	*Medical Education*	7.1	Brannick, Michael T	30
5	Twelve tips for developing an OSCE that measures what you want [16]	*Medical Teacher*	4.7	Daniels, Vijay John	30
6	Is the OSCE a feasible tool to assess competencies in undergraduate medical education? [17]	*Medical Teacher*	4.7	Patricio, Madalena F	29
7	Techniques for measuring clinical competence: Objective Structured Clinical Examinations [18]	*Medical Education*	7.1	Newble, David	26
8	Assessment in medical education [19]	*New England Journal of Medicine*	158.5	Epstein, Ronald M	26
9	Assessing communication skills of medical students in Objective Structured Clinical Examinations (OSCE)-a systematic review of rating scales [20]	*PLoS One*	3.7	Cömert, Musa	26
10	Twelve tips for conducting a virtual OSCE [21]	*Medical Teacher*	4.7	Hopwood, Jenny	26

[a]IF: impact factor.

Keyword Analysis

Through the analysis of keywords, we can quickly understand the situation and development direction of a field. Based on the co-occurrence of keywords in VOSviewer, the hottest keyword is “education” (n=677 occurrences), followed by “performance” (n=536), “competence” (n=458), and “skills” (n=449; Table 6).

XSL•FO
RenderX

Table 6. Top 20 keywords co-occurrence frequencies.

Rank	Keywords	Co-occurrences, n
1	Education	677
2	Performance	536
3	Competence	458
4	Skills	449
5	Reliability	371
6	Assessment	342
7	Students	337
8	Validity	329
9	Simulation	284
10	Medical education	264
11	Diagnosis	228
12	Care	217
13	Prevalence	207
14	Medical students	197
15	Management	196
16	Medical education	171
17	Curriculum	168
18	Communication	161
19	Impact	156
20	Clinical skills	147

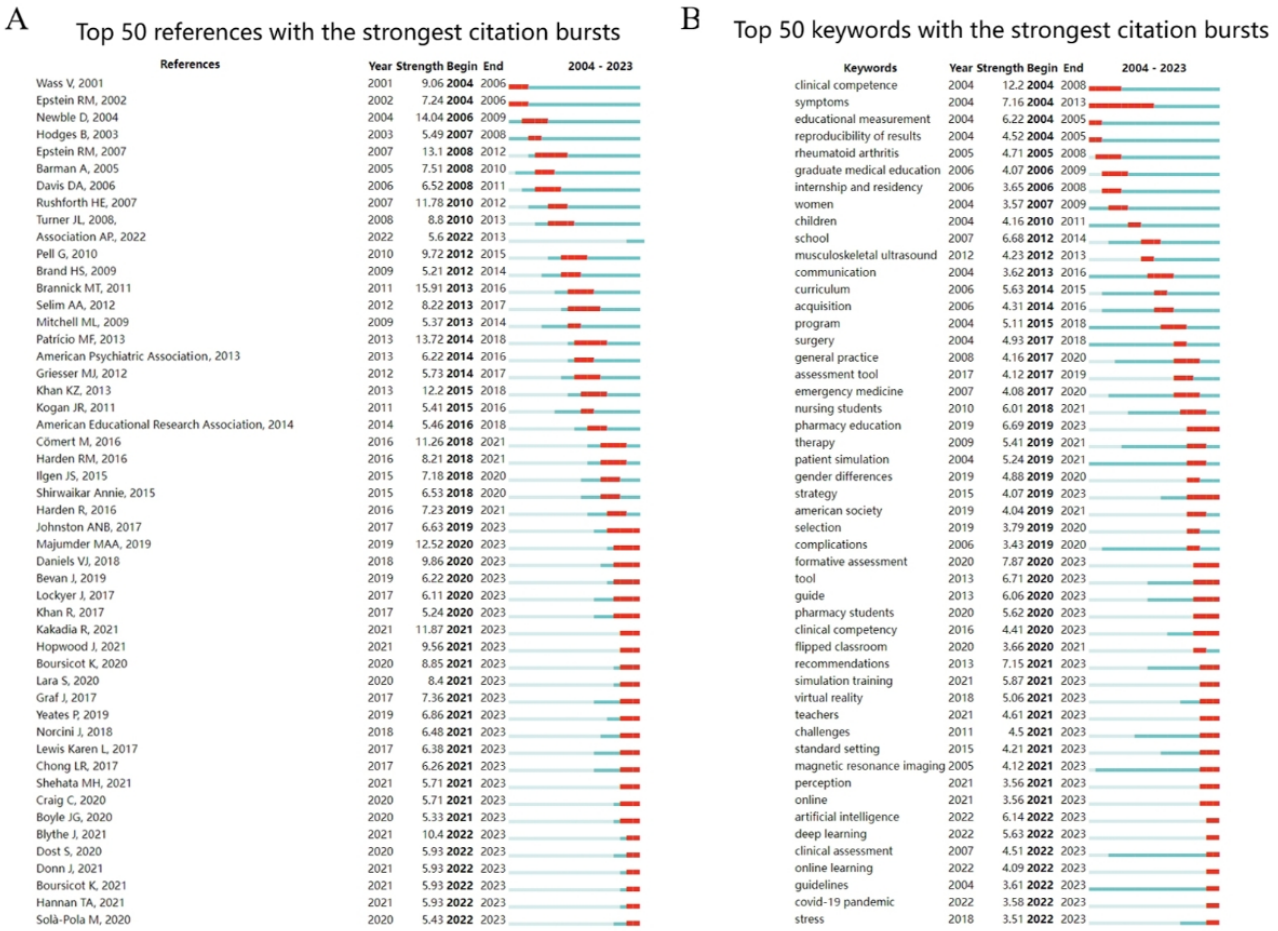
The Burst of Cocited References and Keywords

With CiteSpace, we identified 50 of the most reliable citation bursts in the field related to OSCE [12,13,15-62]. The most frequently cited reference, with a burst strength of 15.91, is a paper published in *Medical Education* titled "A systematic review of the reliability of Objective Structured Clinical Examination scores" [15], whose first author is Michael T Brannick. The paper suggests that OSCEs consist of a series of simulated tasks to assess medical practitioners' skills in diagnosing and treating patients. Of the 50 references, 47 (94%) were published between 2004 and 2023, indicating that these papers have been frequently cited over nearly 20 years. Notably, 24 of these papers are currently at a citation peak (Figure 6A [12,13,15-62]), meaning that research related to OSCE is expected to continue receiving significant attention in the future.

Among the 768 strongest emerging keywords in the field, we focused on the 50 with the most significant surges (Figure 6B), representing the current hotspots in the field and likely future research directions.

XSL•FO
RenderX

Figure 6. Citation burst graph (A), and keyword burst graph (B; sorted by the beginning year of the burst). The blue bars mean the reference has been published; the red bars mean citation burstness.



Discussion

Principal Findings

This study is pioneering in its bibliometric approach to OSCE, encapsulating a comprehensive view of the dynamic research trends in this field. By analyzing the bibliometric data internationally, we have mapped out collaboration networks, identified prevailing research directions, and forecasted potential future developments in OSCE scholarship. The surge in OSCE-related publications since 2019 underscores the recognition of OSCEs as essential for evaluating health care practitioners, meeting the demands of modern medicine for more robust and comprehensive assessment methods to gauge clinical competency [22,63].

Despite this growth, the concentration of research output in countries like the United States, the United Kingdom, and Canada may reflect deeper issues of resource allocation and priority setting in medical education globally [64,65]. This suggests a need for a more nuanced discussion on the uneven geographical spread of OSCE research and its implications. The disparity in research contribution could hinder the global exchange of innovative practices and perspectives in medical education [66,67].

Furthermore, the bibliometric data point to the importance of technology in OSCEs, particularly the integration of virtual and augmented reality. However, to fully understand the implications of technological advances, a more detailed analysis is warranted. This should include how technology shapes the development of OSCEs, its impact on the validity and reliability of assessments, and the potential barriers to its widespread adoption [68-70].

The high concentration of publications in Q1 and Q2 quartile journals, especially those with a significant impact factor, attests to the intersection of OSCE research with impactful clinical education and outcomes. The association with prestigious journals underlines the extensive influence and critical importance of OSCEs across multiple medical specialties [71-73].

The prominence of a core group of scholars leading OSCE research suggests a centralization of expertise that could be diversified through broader international collaboration. Such collaboration could introduce various cultural and pedagogical perspectives into the OSCE discourse, thereby enriching both the practice and the research of OSCEs worldwide [74,75].

The keyword analysis reflects a continual focus on the foundational elements of clinical education, such as "education," "performance," "competence," and "skills," which are at the heart of the OSCE methodology. Emerging research trends

XSL•FO
RenderX

suggest a shift toward the integration of innovative educational technologies and methodologies, enhancing both the OSCE process and its outcomes [76,77].

Comparison to the Literature

Our findings align with those of Lim et al [78], who identified issues with construct, content, and predictive validity in OSCEs in pharmacy education, as well as significant resource challenges. These concerns are echoed in our analysis, where similar validity issues and logistical constraints were observed. Other studies, such as those by Hodges et al [79], have highlighted persistent challenges in psychiatric OSCEs, emphasizing the need for continuous refinement and adaptation. Our study extends these discussions by mapping global trends and collaboration networks, underscoring the necessity for continuous re-evaluation and innovation in OSCE methodologies.

Implications of Findings

The challenges associated with OSCEs suggest a need for evolving assessment methods that incorporate simulations, peer assessments, and reflective practices. The resource-intensive nature of OSCEs underscores the necessity for scalable and sustainable alternatives, such as virtual simulations. Policymakers and educators should leverage global collaboration networks to share best practices and develop adaptable, technology-enhanced assessment frameworks. This approach will help address validity concerns and logistical constraints, ensuring that educational assessments remain robust and relevant in the ever-evolving landscape of health care education.

Limitations

Our bibliometric analysis has limitations that may affect our findings. We only used data from the WoSCC database, potentially excluding studies not indexed there and leading to bias toward English-language literature. This limits the scope of our analysis and overlooks valuable contributions from non-English sources.

Suggestions

To address this, future research should involve a wider range of databases and languages [80,81]. Moreover, the data quality in our study may vary, affecting the credibility of our knowledge mapping. Therefore, caution is needed when interpreting results, and complementary research methods should be considered for a more comprehensive understanding of the field. Longitudinal studies are crucial to assess the impact of OSCEs on medical performance, connecting educational assessments with clinical practice and patient care [82,83].

Moreover, understanding how OSCEs adapt to different health care systems, cultural contexts, and specializations will provide insights into their scalability and adaptability. This is particularly relevant as the health care sector grapples with rapid changes and as medical education seeks to prepare health care professionals for diverse practice environments [19,84].

Conclusions

In conclusion, this bibliometric study not only reaffirms the enduring importance and evolutionary path of OSCEs within medical education but also emphasizes the need for OSCEs to evolve in step with broader health care transformations. The data-driven insights from this analysis should inform future research directions, influence policymaking, and refine educational strategies. By doing so, OSCEs can continue to serve as a dynamic, relevant, and innovative tool in the arsenal of clinical education and evaluation methods.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

HB conceived and designed the ideas for the paper. HB, LZ, XH, and SL participated in all data collection and processing. HB was the major contributor in organizing records and drafting the manuscript. All authors proofread and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Criscione-Schreiber L. Turning Objective Structured Clinical Examinations into Reality. Rheum Dis Clin North Am 2020 Feb;46(1):21-35. [doi: 10.1016/j.rdc.2019.09.010] [Medline: 31757285]
2. Alkhateeb N, Salih AM, Shabila N, Al-Dabbagh A. Objective structured clinical examination: Challenges and opportunities from students' perspective. PLoS One 2022;17(9):e0274055 [FREE Full text] [doi: 10.1371/journal.pone.0274055] [Medline: 36054202]
3. Jünger J, Schäfer S, Roth C, Schellberg D, Friedman Ben-David M, Nikendei C. Effects of basic clinical skills training on objective structured clinical examination performance. Med Educ 2005 Oct;39(10):1015-1020. [doi: 10.1111/j.1365-2929.2005.02266.x] [Medline: 16178828]
4. Gauthier É. Bibliometric analysis of scientific and technological research: a user's guide to the methodology. Science and Technology Redesign Project, CiteSeer. 1998. URL: https://www150.statcan.gc.ca/n1/en/catalogue/88F0006X1998008 [accessed 2024-08-26]

XSL•FO
RenderX

5. Birch S, Lee MS, Alraek T, Kim T. Overview of Treatment Guidelines and Clinical Practical Guidelines That Recommend the Use of Acupuncture: A Bibliometric Analysis. J Altern Complement Med 2018 Aug;24(8):752-769. [doi: 10.1089/acm.2018.0092] [Medline: 29912569]
6. Wilson M, Sampson M, Barrowman N, Doja A. Bibliometric Analysis of Neurology Articles Published in General Medicine Journals. JAMA Netw Open 2021 Apr 01;4(4):e215840 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.5840] [Medline: 33856477]
7. Wu H, Li Y, Tong L, Wang Y, Sun Z. Worldwide research tendency and hotspots on hip fracture: a 20-year bibliometric analysis. Arch Osteoporos 2021 Apr 17;16(1):73. [doi: 10.1007/s11657-021-00929-2] [Medline: 33866438]
8. Vargas JS, Livinski AA, Karagu A, Cira MK, Maina M, Lu Y, et al. A bibliometric analysis of cancer research funders and collaborators in Kenya: 2007-2017. J Cancer Policy 2022 Sep;33:100331 [FREE Full text] [doi: 10.1016/j.jcpo.2022.100331] [Medline: 35792397]
9. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
10. Chen C. CiteSpace: A Practical Guide for Mapping Scientific Literature. New York, NY: Nova Science Publishers; 2016.
11. Kleinberg J. Bursty and hierarchical structure in streams. Data Min Knowl Discov 2003;7:373-397. [doi: 10.1023/A:1024940629314]
12. Majumder MAA, Kumar A, Krishnamurthy K, Ojeh N, Adams OP, Sa B. An evaluative study of Objective Structured Clinical Examination (OSCE): students and examiners perspectives. Adv Med Educ Pract 2019 Jun 5;10:387-397 [FREE Full text] [doi: 10.2147/AMEP.S197275] [Medline: 31239801]
13. Kakadia R, Chen E, Ohyama H. Implementing an online OSCE during the COVID-19 pandemic. J Dent Educ 2020 Jul 15;85(Suppl 1):1006-1008 [FREE Full text] [doi: 10.1002/jdd.12323] [Medline: 32666512]
14. Mittal VA, Walker EF. Diagnostic and statistical manual of mental disorders. Psychiatry Res 2011 Aug 30;189(1):158-159 [FREE Full text] [doi: 10.1016/j.psychres.2011.06.006] [Medline: 21741095]
15. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of Objective Structured Clinical Examination scores. Med Educ 2011 Dec;45(12):1181-1189. [doi: 10.1111/j.1365-2923.2011.04075.x] [Medline: 21988659]
16. Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. Med Teach 2018 Dec;40(12):1208-1213. [doi: 10.1080/0142159X.2017.1390214] [Medline: 29069965]
17. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach 2013 Jun;35(6):503-514. [doi: 10.3109/0142159X.2013.774330] [Medline: 23521582]
18. Newble D. Techniques for measuring clinical competence: Objective Structured Clinical Examinations. Med Educ 2004 Feb;38(2):199-203. [doi: 10.1111/j.1365-2923.2004.01755.x] [Medline: 14871390]
19. Epstein RM. Assessment in medical education. N Engl J Med 2007;356(4):387-396. [doi: 10.1056/NEJMra054784] [Medline: 17251535]
20. Cömert M, Zill JM, Christalle E, Dirmaier J, Härter M, Scholl I. Assessing communication skills of medical students in Objective Structured Clinical Examinations (OSCE)-a systematic review of rating scales. PLoS One 2016 Mar 31;11(3):e0152717 [FREE Full text] [doi: 10.1371/journal.pone.0152717] [Medline: 27031506]
21. Hopwood J, Myers G, Sturrock A. Twelve tips for conducting a virtual OSCE. Med Teach 2021 Jun;43(6):633-636. [doi: 10.1080/0142159X.2020.1830961] [Medline: 33078984]
22. Harden RM. Revisiting 'assessment of clinical competence using an objective structured clinical examination (OSCE)'. Med Educ 2016;50(4):376-379. [doi: 10.1111/medu.12801] [Medline: 26995470]
23. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001;357(9260):945-949. [doi: 10.1016/S0140-6736(00)04221-5] [Medline: 11289364]
24. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287(2):226-235. [doi: 10.1001/jama.287.2.226] [Medline: 11779266]
25. Hodges B. OSCE! Variations on a theme by Harden. Med Educ 2003;37(12):1134-1140. [doi: 10.1111/j.1365-2923.2003.01717.x] [Medline: 14984124]
26. Barman A. Critiques on the objective structured clinical examination. Ann Acad Med Singap 2005;34(8):478-482 [FREE Full text] [Medline: 16205824]
27. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 2006;296(9):1094-1102. [doi: 10.1001/jama.296.9.1094] [Medline: 16954489]
28. Rushforth HE. Objective Structured Clinical Examination (OSCE): review of literature and implications for nursing education. Nurse Educ Today 2007;27(5):481-490. [doi: 10.1016/j.nedt.2006.08.009] [Medline: 17070622]
29. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. Fam Med 2008;40(8):574-578. [Medline: 18988044]
30. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). Washington, DC: American Psychiatric Publishing; 2022.

XSL•FO
RenderX

31. Pell G, Fuller R, Homer M, Roberts T, International Association for Medical Education. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach 2010;32(10):802-811 [FREE Full text] [doi: 10.3109/0142159X.2010.507716] [Medline: 20854155]
32. Brand HS, Schoonheim-Klein M. Is the OSCE more stressful? Examination anxiety and its consequences in different assessment methods in dental education. Eur J Dent Educ 2009;13(3):147-153. [doi: 10.1111/j.1600-0579.2008.00554.x] [Medline: 19630933]
33. Selim AA, Ramadan FH, El-Gueneidy MM, Gaafer MM. Using Clinical Examination (OSCE) in undergraduate psychiatric nursing education: is it reliable and valid? Nurse Educ Today 2012;32(3):283-288. [doi: 10.1016/j.nedt.2011.04.006] [Medline: 21555167]
34. Mitchell ML, Henderson A, Groves M, Dalton M, Nulty D. The Objective Structured Clinical Examination (OSCE): optimising its value in the undergraduate nursing curriculum. Nurse Educ Today 2009;29(4):398-404. [doi: 10.1016/j.nedt.2008.10.007] [Medline: 19056152]
35. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychiatric Publishing; 2013.
36. Griesser MJ, Beran MC, Flanigan DC, Quackenbush M, Van Hoff C, Bishop JY. Implementation of an Objective Structured Clinical Exam (OSCE) into orthopedic surgery residency training. J Surg Educ 2012;69(2):180-189. [doi: 10.1016/j.jsurg.2011.07.015] [Medline: 22365863]
37. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach 2013;35(9):e1437-e1446. [doi: 10.3109/0142159X.2013.818634] [Medline: 23968323]
38. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ 2011;45(10):1048-1060. [doi: 10.1111/j.1365-2923.2011.04025.x] [Medline: 21916943]
39. American Educational Research Association. Standards for Educational & Psychological Testing (2014 Edition). 2024. URL: https://www.aera.net/publications/books/standards-for-educational-psychological-testing-2014-edition [accessed 2024-09-26]
40. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ 2015;49(2):161-173. [doi: 10.1111/medu.12621] [Medline: 25626747]
41. Shirwaikar A. Objective Structured Clinical Examination (OSCE) in pharmacy education - a trend. Pharm Pract (Granada) 2015;13(4):627-630 [FREE Full text] [doi: 10.18549/PharmPract.2015.04.627] [Medline: 26759616]
42. Harden HR. OSC Guide. 2016. URL: https://www.osc.ca/en/news-events/subscribe/osc-guide [accessed 2024-09-26]
43. Johnston ANB, Weeks B, Shuker M, Coyne E, Niall H, Mitchell M, et al. Nursing students' perceptions of the Objective Structured Clinical Examination: an integrative review. Clin Simul Nurs 2017;13(3):127-142 [FREE Full text] [doi: 10.1016/j.ecns.2016.11.002]
44. Bevan J, Russell B, Marshall B. A new approach to OSCE preparation - PrOSCEs. BMC Med Educ 2019;19(1):126 [FREE Full text] [doi: 10.1186/s12909-019-1571-5] [Medline: 31046773]
45. Lockyer J, Carraccio C, Chan MK, Hart D, Smee S, Touchie C, ICBME Collaborators. Core principles of assessment in competency-based medical education. Med Teach 2017;39(6):609-616. [doi: 10.1080/0142159X.2017.1315082] [Medline: 28598746]
46. Khan R, Payne MWC, Chahine S. Peer assessment in the Objective Structured Clinical Examination: a scoping review. Med Teach 2017;39(7):745-756. [doi: 10.1080/0142159X.2017.1309375] [Medline: 28399690]
47. Boursicot K, Kemp S, Ong TH, Wijaya L, Goh SH, Freeman K, et al. Conducting a high-stakes OSCE in a COVID-19 environment. MedEdPublish (2016) 2020;9:54 [FREE Full text] [doi: 10.15694/mep.2020.000054.1] [Medline: 38058921]
48. Lara S, Foster CW, Hawks M, Montgomery M. Remote assessment of clinical skills during COVID-19: a virtual, high-stakes, summative pediatric Objective Structured Clinical Examination. Acad Pediatr 2020;20(6):760-761 [FREE Full text] [doi: 10.1016/j.acap.2020.05.029] [Medline: 32505690]
49. Graf J, Smolka R, Simoes E, Zipfel S, Junne F, Holderried F, et al. Communication skills of medical students during the OSCE: gender-specific differences in a longitudinal trend study. BMC Med Educ 2017;17(1):75 [FREE Full text] [doi: 10.1186/s12909-017-0913-4] [Medline: 28464857]
50. Yeates P, Cope N, Hawarden A, Bradshaw H, McCray G, Homer M. Developing a video-based method to compare and adjust examiner effects in fully nested OSCEs. Med Educ 2019;53(3):250-263 [FREE Full text] [doi: 10.1111/medu.13783] [Medline: 30575092]
51. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, et al. 2018 Consensus framework for good assessment. Med Teach 2018;40(11):1102-1109. [doi: 10.1080/0142159X.2018.1500016] [Medline: 30299187]
52. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The association of standardized patient educators (ASPE) standards of best practice (SOBP). Adv Simul (Lond) 2017;2:10 [FREE Full text] [doi: 10.1186/s41077-017-0043-4] [Medline: 29450011]
53. Chong L, Taylor S, Haywood M, Adelstein BA, Shulruf B. The sights and insights of examiners in Objective Structured Clinical Examinations. J Educ Eval Health Prof 2017;14(3):34-242 [FREE Full text] [doi: 10.3352/jeehp.2017.14.34] [Medline: 29278906]

XSL•FO
RenderX

54. Shehata MH, Kumar AP, Arekat MR, Alsenbesy M, Mohammed Al Ansari A, Atwa H, et al. A toolbox for conducting an online OSCE. Clin Teach 2021;18(3):236-242. [doi: 10.1111/tct.13285] [Medline: 33063427]
55. Craig C, Kasana N, Modi A. Virtual OSCE delivery: the way of the future? Med Educ 2020;54(12):1185-1186 [FREE Full text] [doi: 10.1111/medu.14286] [Medline: 32627218]
56. Boyle JG, Colquhoun I, Noonan Z, McDowall S, Walters MR, Leach JP. Viva la VOSCE? BMC Med Educ 2020;20(1):514 [FREE Full text] [doi: 10.1186/s12909-020-02444-3] [Medline: 33334327]
57. Blythe J, Patel NSA, Spiring W, Easton G, Evans D, Meskevicius-Sadler E, et al. Undertaking a high stakes virtual OSCE ("VOSCE") during Covid-19. BMC Med Educ 2021;21(1):221 [FREE Full text] [doi: 10.1186/s12909-021-02660-5] [Medline: 33879139]
58. Dost S, Hossain A, Shehab M, Abdelwahed A, Al-Nusair L. Perceptions of medical students towards online teaching during the COVID-19 pandemic: a national cross-sectional survey of 2721 UK medical students. BMJ Open 2020;10(11):e042378 [FREE Full text] [doi: 10.1136/bmjopen-2020-042378] [Medline: 33154063]
59. Donn J, Scott JA, Binnie V, Bell A. A pilot of a virtual Objective Structured Clinical Examination in dental education. A response to COVID-19. Eur J Dent Educ 2021;25(3):488-494 [FREE Full text] [doi: 10.1111/eje.12624] [Medline: 33185919]
60. Boursicot K, Kemp S, Wilkinson T, Findyartini A, Canning C, Cilliers F, et al. Performance assessment: consensus statement and recommendations from the 2020 Ottawa conference. Med Teach 2021;43(1):58-67. [doi: 10.1080/0142159X.2020.1830052] [Medline: 33054524]
61. Hannan TA, Umar SY, Rob Z, Choudhury RR. Designing and running an online Objective Structured Clinical Examination (OSCE) on zoom: a peer-led example. Med Teach 2021;43(6):651-655. [doi: 10.1080/0142159X.2021.1887836] [Medline: 33626286]
62. Solà-Pola M, Morin-Fraile V, Fabrellas-Padrés N, Raurell-Torreda M, Guanter-Peris L, Guix-Comellas E, et al. The usefulness and acceptance of the OSCE in nursing schools. Nurse Educ Pract 2020;43:102736. [doi: 10.1016/j.nepr.2020.102736] [Medline: 32058920]
63. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13(1):41-54. [doi: 10.1111/j.1365-2923.1979.tb00918.x] [Medline: 763183]
64. Lee GB, Chiu AM. Assessment and feedback methods in competency-based medical education. Ann Allergy Asthma Immunol 2022;128(3):256-262. [doi: 10.1016/j.anai.2021.12.010] [Medline: 34929390]
65. Mathew MM, Thomas KA. Medical aptitude and its assessment. Natl Med J India 2018;31(6):356-363 [FREE Full text] [doi: 10.4103/0970-258X.262905] [Medline: 31397372]
66. Zayyan M. Objective structured clinical examination: the assessment of choice. Oman Med J 2011;26(4):219-222 [FREE Full text] [doi: 10.5001/omj.2011.55] [Medline: 22043423]
67. Jiang Z, Ouyang J, Li L, Han Y, Xu L, Liu R, et al. Cost-effectiveness analysis in performance assessments: a case study of the objective structured clinical examination. Med Educ Online 2022;27(1):2136559 [FREE Full text] [doi: 10.1080/10872981.2022.2136559] [Medline: 36250891]
68. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
69. Bajpai S, Semwal M, Bajpai R, Car J, Ho AHY. Health professions' digital education: review of learning theories in randomized controlled trials by the digital health education collaboration. J Med Internet Res 2019;21(3):e12912 [FREE Full text] [doi: 10.2196/12912] [Medline: 30860483]
70. Cheng A, Lang T, Starr S, Pusic M, Cook D. Technology-enhanced simulation and pediatric education: a meta-analysis. Pediatrics 2014;133(5):e1313-e1323. [doi: 10.1542/peds.2013-2139] [Medline: 24733867]
71. Wilkinson TJ, Wade WB, Knock LD. A blueprint to assess professionalism: results of a systematic review. Acad Med 2009;84(5):551-558. [doi: 10.1097/ACM.0b013e31819fbaa2] [Medline: 19704185]
72. Preez RRD, Pickworth GE, van Rooyen M. Teaching professionalism: a South African perspective. Med Teach 2007;29(9):e284-e291. [doi: 10.1080/01421590701754128] [Medline: 18158653]
73. Mueller PS. Teaching and assessing professionalism in medical learners and practicing physicians. Rambam Maimonides Med J 2015;6(2):e0011 [FREE Full text] [doi: 10.5041/RMMJ.10195] [Medline: 25973263]
74. Alinier G. A typology of educationally focused medical simulation tools. Med Teach 2007;29(8):e243-e250. [doi: 10.1080/01421590701551185] [Medline: 18236268]
75. Fox-Robichaud AE, Nimmo GR. Education and simulation techniques for improving reliability of care. Curr Opin Crit Care 2007;13(6):737-741. [doi: 10.1097/MCC.0b013e3282f1bb32] [Medline: 17975400]
76. Eva KW, Regehr G. "I'll never play professional football" and other fallacies of self-assessment. J Contin Educ Health Prof 2008;28(1):14-19. [doi: 10.1002/chp.150] [Medline: 18366120]
77. Colthart I, Bagnall G, Evans A, Allbutt H, Haig A, Illing J, et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME guide no. 10. Med Teach 2008;30(2):124-145. [doi: 10.1080/01421590701881699] [Medline: 18464136]

XSL•FO
RenderX

78. Lim AS, Ling YL, Wilby KJ, Mak V. What's been trending with OSCEs in pharmacy education over the last 20 years? A bibliometric review and content analysis. Curr Pharm Teach Learn 2024;16(3):212-220 [FREE Full text] [doi: 10.1016/j.cptl.2023.12.028] [Medline: 38171979]
79. Hodges BD, Hollenberg E, McNaughton N, Hanson MD, Regehr G. The psychiatry OSCE: a 20-year retrospective. Acad Psychiatry 2014;38(1):26-34. [doi: 10.1007/s40596-013-0012-8] [Medline: 24449223]
80. Boulet J, Durning S. What we measure … and what we should measure in medical education. Med Educ 2019;53(1):86-94. [doi: 10.1111/medu.13652] [Medline: 30216508]
81. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical education's wicked problem: achieving equity in assessment for medical learners. Acad Med 2020;95(12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments):S98-S108. [doi: 10.1097/ACM.0000000000003717] [Medline: 32889943]
82. Tormey W. Education, learning and assessment: current trends and best practice for medical educators. Ir J Med Sci 2015;184(1):1-12. [doi: 10.1007/s11845-014-1069-4] [Medline: 24549647]
83. Gröne O, Mielke I, Knorr M, Ehrhardt M, Bergelt C. Associations between communication OSCE performance and admission interviews in medical education. Patient Educ Couns 2022;105(7):2270-2275. [doi: 10.1016/j.pec.2021.11.005] [Medline: 34801337]
84. Min Simpkins AA, Koch B, Spear-Ellinwood K, St John P. A developmental assessment of clinical reasoning in preclinical medical education. Med Educ Online 2019;24(1):1591257 [FREE Full text] [doi: 10.1080/10872981.2019.1591257] [Medline: 30935299]

Abbreviations

OSCE: Objective Structured Clinical Examination
WoS: Web of Science
WoSCC: Web of Science Core Collection

Edited by B Lesselroth; submitted 26.02.24; peer-reviewed by S Alkan, W Chou; comments to author 15.05.24; revised version received 17.05.24; accepted 19.08.24; published 30.09.24.

Please cite as:
Ba H, Zhang L, He X, Li S
Knowledge Mapping and Global Trends in the Field of the Objective Structured Clinical Examination: Bibliometric and Visual Analysis (2004-2023)
JMIR Med Educ 2024;10:e57772
URL: https://mededu.jmir.org/2024/1/e57772
doi:10.2196/57772
PMID:

©Hongjun Ba, Lili Zhang, Xiufang He, Shujuan Li. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.09.2024. 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.

XSL•FO
RenderX

Original Paper

Integrating Digital Assistive Technologies Into Care Processes: Mixed Methods Study

Sebastian Hofstetter[1,2*], BSc, MA, PhD; Max Zilezinski[1,2*], MSc; Dominik Behr[1,2], BSc; Bernhard Kraft[1,3], MA; Christian Buhtz[2], MSc; Denny Paulicke[2], Prof Dr; Anja Wolf[1], PhD; Christina Klus[2], MA; Dietrich Stoevesandt[2], PhD; Karsten Schwarz[2], PhD; Patrick Jahn[1], Prof Dr

[1]AG Versorgungsforschung Pflege im Krankenhaus, Departement of Internal Medicine, University Medicine Halle (Saale), Halle (Saale), Germany

[2]Dorothea-Erxleben-Lernzentrum, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

[3]Institute for History and Ethics of Medicine, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

[*]these authors contributed equally

Corresponding Author:
Sebastian Hofstetter, BSc, MA, PhD
AG Versorgungsforschung Pflege im Krankenhaus
Departement of Internal Medicine
University Medicine Halle (Saale)
Ernst-Grube-Str. 40, 06120 Halle (Saale)
Halle (Saale)
Germany
Phone: 49 345 557 4064
Email: sebastian.hofstetter@medizin.uni-halle.de

Abstract

Background: Current challenges in patient care have increased research on technology use in nursing and health care. Digital assistive technologies (DATs) are one option that can be incorporated into care processes. However, how the application of DATs should be introduced to nurses and care professionals must be clarified. No structured and effective education concepts for the patient-oriented integration of DATs in the nursing sector are currently available.

Objective: This study aims to examine how a structured and guided integration and education concept, herein termed the sensitization, evaluative introduction, qualification, and implementation (SEQI) education concept, can support the integration of DATs into nursing practices.

Methods: This study used an explanatory, sequential study design with a mixed methods approach. The SEQI intervention was run in 26 long-term care facilities oriented toward older adults in Germany after a 5-day training course in each. The participating care professionals were asked to test 1 of 6 DATs in real-world practice over 3 days. Surveys (n=112) were then administered that recorded the intention to use DATs at 3 measurement points, and guided qualitative interviews with care professionals (n=12) were conducted to evaluate the learning concepts and effects of the intervention.

Results: As this was a pilot study, no sample size calculation was carried out, and *P* values were not reported. The participating care professionals were generally willing to integrate DATs—as an additional resource—into nursing processes even before the 4-stage SEQI intervention was presented. However, the intervention provided additional background knowledge and sensitized care professionals to the digital transformation, enabling them to evaluate how DATs fit in the health care sector, what qualifies these technologies for correct application, and what promotes their use. The care professionals expressed specific ideas and requirements for both technology-related education concepts and nursing DATs.

Conclusions: Actively matching technical support, physical limitations, and patients' needs is crucial when selecting DATs and integrating them into nursing processes. To this end, using a structured process such as SEQI that strengthens care professionals' ability to integrate DATs can help improve the benefits of such technology in the health care setting. Practical, application-oriented learning can promote the long-term implementation of DATs.

(JMIR Med Educ 2024;10:e54083) doi:10.2196/54083

KEYWORDS

digital assistive technologies; education concept; intention to use; learning effects; digital transformation

XSL•FO
RenderX

Introduction

Background

Digital assistive technologies (DATs) offer novel possibilities for nursing and health care. Therefore, health care institutions must determine how to successfully implement digitization and adapt to the digital transformation in the health sector. Requirements include strategic planning, governing, organizing, controlling, orchestrating, and training technology-intense processes and services. Therefore, developing competencies among care professionals is necessary. Providing extra education on this topic will help care professionals appropriately implement and integrate DATs—as an additional resource—into their professional practices.

As no structured education concept for the implementation of DATs in nursing care currently exists, this study developed the sensitization, evaluative introduction, qualification, and implementation (SEQI) education concept. It then evaluated how the implementation of the 4-stage SEQI benefited nursing care by measuring the changes in care professionals' intention to use DATs in long-term care facilities after its implementation. Long-term care facilities often fail to implement DATs because users have only a brief opportunity to familiarize themselves with and apply such technology. The population of interest for this study was defined as individuals trained and registered in a health or social profession who had worked in long-term inpatient facilities (care professionals). While the initial intention was to study only registered nurses as a target group, it quickly became apparent that this specification was an unrealistic inclusion criterion in current care practice. The care situation demonstrates that, in addition to nurses, other health care professionals, such as care assistants, nursing assistants, social workers, and physiotherapists, also provide a significant amount of care. It is evident that these individuals benefit equally from the implemented educational approach and, as a result, should not be excluded. In this study, a sample of care professionals from 26 long-term care facilities in Germany were examined using a mixed methods approach to understand changes in their intention to use DATs after the SEQI intervention. This study fills a gap in the literature by closely monitoring and evaluating the implementation of DATs using a structured approach.

Worldwide, health care systems are responding to the pressure created by increasing demand for care and the digital transformation of the health care sector [1,2]. Notably, DATs offer novel opportunities for promoting the independence and participation of older adults in long-term care [3] and improving their quality of life [4]. DATs can facilitate a range of activities associated with daily living, including smart medication management, digitally assisted fall prevention, and communication. DATs represent an evolution of assistive technologies that have been enhanced with digital capabilities [5]. Heinemann and Matusiewicz [6] pointed out that the digital transformation of health care and use of DATs can also be an opportunity to address the health care crisis. The term *health care crisis* is used to describe a phenomenon that has now affected many countries regardless of the structure of their social security systems, which all vary considerably yet are facing similar crises. These include an aging population, falling birth rates, increasing care needs, social isolation, declining social support networks, a shortage of nursing staff, and so on. The use of DATs is expected to benefit patients, care professionals, physicians, and health care organizations. Nursing care professionals are the leading user group among health care professionals; however, DATs can only be beneficial if care professionals accept and use them.

Some studies have discussed the use of DATs to support care professionals [7] in acquiring transformative competencies [8] and the relevant technological knowledge [9,10]. Care professionals' expertise can reduce the risk of developing impractical and ill-suited technologies [2,11,12]. Thus, to realize DATs' full potential, tailored solutions that address functional limitations are necessary, and DATs' use should be preplanned and problem oriented. Research has established that systems intended for patient interactions, such as those that primarily support individuals in their activities of daily living (ADLs), should be the focus of DAT developments [5] as ADL assistance and support are frequent targets of nursing care. Therefore, it is important to take a differentiated look at the topic of robotics in nursing and, in addition to the technical shortcomings, examine other reasons why colleagues are not yet using robots in nursing care to the extent required. The question arises as to how DATs can be classified in the existing system of support for carers. This research focuses on systems to be used in direct patient interactions to support those affected, for example, in such ADLs as communication, self-feeding, and mobility. The targeted use for each application then has an assistive, supporting character and, therefore, is not new for care professionals.

According to the World Health Organization, *assistive technology* is an umbrella term that includes all adaptive and rehabilitative devices for supporting people with health impairments as well as their selection and use [13,14]. Assistive technologies have expanded to include a digital component that makes it possible to improve quality of life and opportunities for participation by promoting greater independence. Support always occurs when people can complete tasks and perform movements that would otherwise be difficult without technical support. At the same time, DATs reduce the need for formal health, support, and long-term care services by providing dependent care when self-care is no longer fully possible due to limitations.

An educated and trained use of DATs has the potential to reduce the high workload of nursing staff and other health care professionals.

Usability refers to the context-sensitive, application-oriented, and effective relationship between people and technology [15]. The lack of usability and implementation strategies for new technologies frequently results in unsuccessful implementation [16]. The most frequent causes of these failures are the lack of fundamental knowledge of the availability of DATs and poor understanding of their possible uses [9]. To date, innovative nursing technologies, such as those designed to prevent pressure ulcers and falls, provide supportive care for diabetes mellitus, address disorientation, and have predominantly been used as

XSL•FO
RenderX

unconnected individual solutions on a person-by-person basis. To implement these technologies more broadly, the willingness of health care professionals to use them, as well as their actual use of such technology, must be analyzed. Providing opportunities to learn about the digitization of care approaches in the digital transformation era is also necessary.

Education on DATs that offers practical and actionable interpretations that help users better understand and use DATs in their daily practice should be provided to caregivers, care professionals, and health care professionals, which will ultimately lead to better care outcomes [17]. These interpretations are not directly applicable to caring practice. For practicality, arguments need to be transformed and excluded from their social scientific identity; they can then be reconstructed based on practice conditions in a way that is relevant and applicable to the practical situation. Doing so involves translating theoretical concepts and knowledge into practical actions that can be implemented in real-world situations. The transformation of theoretical knowledge into practical applications is an important component of any education approach aimed at integrating DATs into caregiving practices [17]. No theory-guided education concepts introducing DATs have been developed yet. Despite the identification of numerous challenges, strategies for the sound implementation of technology in nursing care can be formulated. These strategies should include training on digital skills [18], and the creation of a positive attitude among health professionals toward technology is essential. Nadav et al [19] argue that this can be achieved through the extensive introduction of professionals to technology operation. Albrecht et al [20] propose that this should be accompanied by a positive attitude toward technology. Therefore, understanding the factors influencing users' intention to use DATs is key to ensuring the optimal integration of DATs within the health care system and achieving measurable benefits. A review of the German health care system (in which this study is also situated) reveals that surveys indicate a positive, open-minded, and inquisitive attitude toward new technologies among care professionals. For instance, one study demonstrated that care professionals tend to embrace DATs and perceive them as beneficial and user-friendly [21]. Conversely, respondents exhibited a more reserved attitude toward the use of robotics, with negative expectations of its use being more frequently expressed [21].

To address this issue, we propose a model based on the extended Technology Usage Inventory (TUI) to explore how users' technology readiness and perceptions of DATs influence their intention to use them. This study's results improved our understanding of caregivers and care professionals' intention to use DATs and contributed to innovative research on the adoption of such technology. The proposed 4-stage SEQI is a novel education concept introduced as a form of learning. Its "evaluative introduction" and "implementation" stages are influenced by previous work [9]. *Evaluative introduction* refers to acquiring the competencies to assess whether DATs are suitable for addressing a functional nursing problem. *Implementation* is defined as on-site testing over a longer period in real use conditions. *Sensitization* means becoming aware of the digital transformation. *Qualification* is defined as the proficiency and expertise that individuals acquire to operate and use DATs in the given care context effectively and confidently. In summary, this mixed methods study examined whether a structured and guided education concept (ie, SEQI) can change care professionals' intention to use DATs. The changes in intention to use and learning effects were captured through a quantitative survey of 112 care professionals. This study also evaluated the learning effects of the SEQI intervention based on guided interviews with 12 care professionals.

Intention to Use DATs

In the health care context, previous studies have applied the technology acceptance model (TAM) and TAM 2 to explore acceptance and examine physicians' [22-24] and nurses' [25-29] intention to use DATs. Those studies have revealed nurses [30] and nursing care students' intention to accept health care technologies [9,11,31,32]. Theories have conceptualized those factors (ie, barriers and facilitators) that influence the outcomes of implementation efforts spanning both generalized theory building and the development of practical approaches [33]. In particular, previous studies have found that user training and technology acceptance are key factors to successful implementation [33-35]. However, technical specifications and standards cannot simply be transferred to professional nursing from other fields [5]. The practical application of DATs or even the intention to use DATs is also significantly influenced by factors such as usability, usefulness, accessibility, and immersion; these factors are assessed using the TAM. Furthermore, psychological factors that affect the actual use of technology include acceptance, which is determined by personal attitudes such as curiosity, fear of technology, skepticism, and social norms [36], meaning that the individual characteristics of care professionals determine their intention to use DATs.

The acquisition of digital skills is essential for addressing the psychological factors that affect intention to use. The US health care system has slowly evolved from a system built on episodic and ambulatory in-person encounters to one that is digitally based, technology rich, and data informed [37]. The practical use of DATs requires an understanding of not only how to use them but also what opportunities and possibilities they offer for a customized and problem-oriented use in nursing care [37,38]. To achieve this, digital skills must be integrated more strongly into nursing curricula [2,12,39,40] and then expanded through advanced training and further education in later professional life. The nursing profession, for example, is facing a crisis in its education pipeline and professional development. Traditionally, formal health care education and the postgraduate novice-to-expert continuum have not emphasized technology and informatics as integral components of nursing [37]. Hence, solutions that enhance educational pathways should be explored, allowing care professionals to effectively practice in today's health care environment under the digital transformation. It is helpful to distinguish the term *digitization* from the term *digital transformation*. *Digitization* describes a technical concept that includes software programs, corresponding hardware, and the translation of analog values into bits and bytes, whereas the term *digital transformation* generally describes changes that also relate to the values, attitudes, and mindsets of the professional groups concerned [41,42]. Targeted education

XSL•FO
RenderX

programs should provide care professionals with the tools necessary for their profession, including the available DATs and their possible applications, as well as reflective competencies to evaluate and adapt these DATs for practical use in nursing. When properly used, DATs can provide opportunities for relief from providing care [29,33].

While implementing the SEQI education concept, researching and testing a possible education approach to introduce care professionals in long-term care to DATs was challenging. To develop a transformative implementation concept for DATs in the long-term care field, it was necessary to test care professionals' understanding of DATs. The overarching project goal was to increase both care professionals' intention to use DATs and, simultaneously, obtain their assessment of the practicality of the SEQI education concept. Currently available DATs are already an additional resource shaping the digital transformation of nursing processes. However, it was important for care professionals to understand the need to fit a specific DAT to a patient's physical (functional) limitation or relevant care problem. Furthermore, initiating reflection and discussion among care professionals was crucial. As such, the following project goals were highlighted:

1. Focus on further education on DATs in real-world working conditions
2. Link theoretical and practical knowledge (theory-practice transfer)
3. Test currently available DATs
4. Focus on care professionals' acquisition of knowledge and competencies
5. Build on care professionals' expertise, knowledge, and learning habits
6. Improve practical testing experience to reduce uncertainty
7. Prepare for the digital transformation in the health care sector

Theoretical Framework

Changes in professional, technical, and organizational conditions often lead to shifts in employees' competencies or even require new competencies. The potential to sustainably change professional requirements, tasks, activities, and job profiles is a factor in the digital transformation of workplaces. In this situation, requirements for health care professionals' competencies should also be evaluated. For example, the German Ethics Council calls for curricula to be supplemented to include new nursing techniques, including their ethical implications [12], to ensure the continuing education of care professionals. While transformative learning on how to manage the digital transformation of the health care sector is scarce, there is a great deal of interest in DATs as part of this transformation. Hence, the demand for professionals to acquire expertise in the field is increasing [9,43]. Thus, creating structured guidance to introduce care professionals to the use of DATs in nursing processes and planning is sensible. The theoretical considerations of transformative learning [44,45] provide a suitable framework for developing this type of guidance.

Beginning with the concept of lifelong learning [46], transformative learning allows earlier experiences to be reinterpreted and re-evaluated through the lens of experience-based assumptions and attitudes. Developing new perspectives through the dynamics of learning (intentional, intuitive learning) embedded in a problem-solving process is a focus of the concept [8,44,45,47]. This learning initiates a process of transformation, that is, the development of new perspectives on previously unquestionable attributions of meaning [8,47]. Care professionals can then embed the abstract construct of the digital transformation into their own experiences and reflect on it both as a starting point and as an end point for integrating technology into nursing processes [9]. By drawing on existing and experience-based knowledge, the possible applications for currently available DATs can be better assessed. The precise and solution-oriented use of currently available DATs as well as the development of new and innovative DATs offer great potential for defining nursing care problems both now and in the future.

Problem Statement and Study Objectives

Overview

Despite the potential benefits of DATs in improving health care, their integration into nursing practices remains underdeveloped. The principal challenge is the absence of structured and efficacious educational methodologies that facilitate the patient-oriented integration of DATs in the nursing sector. The implementation of DATs is frequently impeded by the limited opportunities available to care professionals to become acquainted with and use these technologies effectively in actual clinical settings. This gap underscores the necessity for a comprehensive educational framework to guide the integration process.

Research Gap and Objectives

This study addressed the critical need for a structured approach to the education on and integration of DATs into nursing practices. The SEQI education concept was developed with the objective of providing a structured approach to support care professionals in understanding, adopting, and effectively using DATs. The objective of this study was 2-fold: first, to evaluate the impact of the SEQI education concept on care professionals' intention to use DATs and, second, to assess the practical application and learning effects of this intervention in long-term care facilities.

Research Question

The primary research question guiding this study was as follows: how does the structured SEQI education concept affect care professionals' intention to use DATs in long-term care settings?

A Priori Hypotheses

The first hypothesis was as follows: the implementation of the SEQI education concept is expected to significantly increase care professionals' intention to use DATs.

The second hypothesis was as follows: care professionals who undergo the SEQI intervention will demonstrate enhanced understanding and practical application of DATs, leading to an improved theory-to-practice transfer of DATs.

XSL•FO
RenderX

This study aimed to provide a validated educational framework that can be widely adopted to facilitate the digital transformation in nursing and health care, ensuring that care professionals are well equipped to leverage DATs in their daily practices.

Methods

Overview

This mixed methods study used a combination of qualitative and quantitative data collection methods. The sequential explanatory design facilitates the determination of which quantitative results require further elucidation [48]. The sequential explanatory design with mixed methods comprises 2 distinct phases: a quantitative phase and a qualitative phase. The researcher initiates the study with a quantitative phase, which is followed by a second qualitative phase. The purpose of this second phase is to provide a more in-depth explanation of the initial results [48]. This research and development project ran from September 2019 to September 2022. It began with the question of what multimodal, transformative learning concepts presented in a structured manner could aid the digital transformation in the health care sector and encourage the use of DATs by health care professionals. It is recommended to understand multimodal educational offerings on DAT should present caregivers with a variety of approaches to addressing and thematically dealing with the topic. To ensure methodological quality, the Good Reporting of a Mixed Methods Study checklist [49] and the Mixed Methods Appraisal Tool [50] were used. The analysis of the in-depth interviews served to ascertain the needs and requirements of caregivers regarding an educational concept, thereby enabling the development of an intention to use DATs in the first place. This process of “connecting” the qualitative and quantitative phases was achieved through the sampling design. At the same time, the subject areas and subjects identified in the analysis of the interviews were used to estimate the dimensions and indicators of the results derived from the TUI questionnaire used in the quantitative part of the study. As noted previously, this study’s primary objective was to determine changes in the intention to use DATs. The secondary objective was to assess the learning effects following the principles of data triangulation [51,52].

Survey Development

For the quantitative part of the study, a survey was conducted at 3 measurement points (T0, T1, and T2) at the end of stages 1, 2, and 4 of the SEQI education concept. The survey was developed using the TUI, a valid measurement instrument based on the TAM [36,53,54] that was created to evaluate the intention to use specific technologies by identifying factors that play a role in the technology adoption decision. Its basis is the theory of reasoned action, which holds that attitudes and behaviors are closely connected. Therefore, the behavioral intention to use a certain technology is determined by a person’s attitudes and social norms.

According to Kothgassner et al [36], the TUI distinguishes 3 main factors that affect technology acceptance and, thus, predict the actual use of technology: perceived usefulness, perceived ease of use, and attitude toward use. Perceived usefulness is defined as the subjectively perceived likelihood of improving performance by using a technology, whereas perceived ease of use is defined as the degree to which a product, system, or interface is designed and structured in a way that allows users to interact with it comfortably, intuitively, and with minimal effort. It encompasses elements such as user-friendly design, simplicity, and the overall accessibility of the system, contributing to a positive and efficient user experience. Both perceived usefulness and perceived ease of use directly influence an individual’s attitudes toward using a technology, which then directly affect their behavioral intention to use and, thus, their actual use of that technology [36].

The TUI consists of 30 items divided into 8 subscales. Of these subscales, 5 (curiosity, fear of technology, interest, immersion, and skepticism) have 4 items, whereas the accessibility, usability, and usefulness subscales each have 3 items. In this study, we adopted 7 of these subscales for our analysis: curiosity, fear of technology, interest, skepticism, accessibility, usability, and usefulness [36].

Implementation of the SEQI: 5-Day Training Course

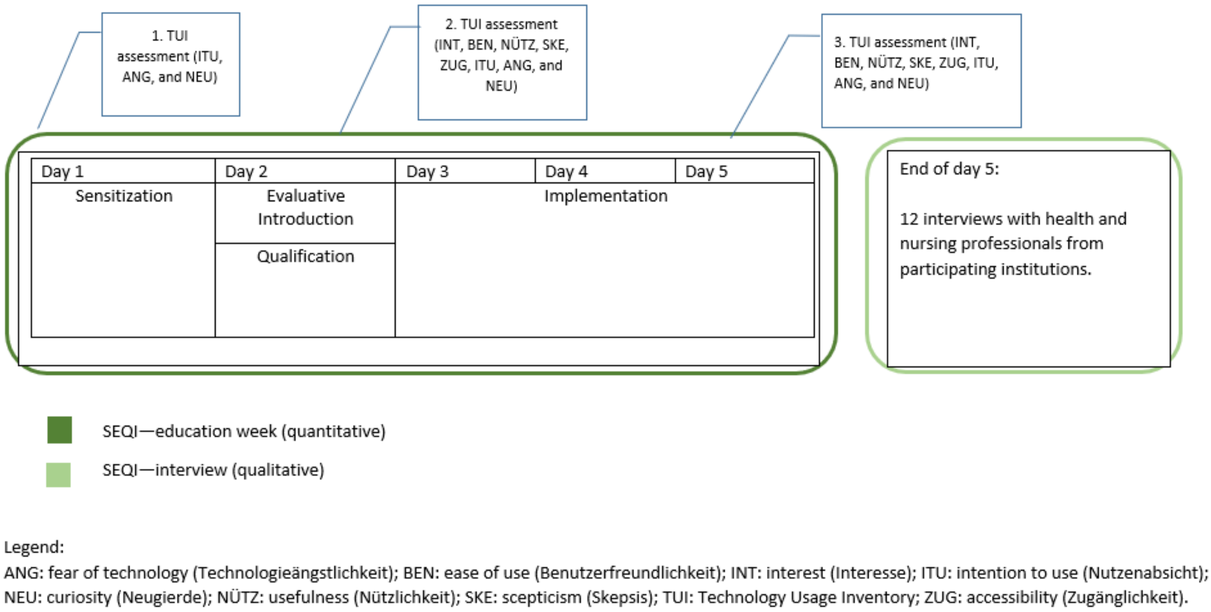
The 4-stage SEQI education concept was implemented in 26 long-term care facilities across the state of Saxony-Anhalt in Germany. For recruitment, a list of all long-term inpatient facilities in Saxony-Anhalt was made available by the discharge management of the University Hospital Halle (Saale). The list contained a total of 446 long-term inpatient facilities distributed across the state of Saxony-Anhalt. All facilities with a minimum occupancy of 50 beds (ie, 225/446, 50.4% of the facilities) were contacted and asked about their willingness to participate. In total, 26 facilities consented to participate in the study and were subsequently recruited. A total of 5 care professionals participated in each of the 26 training courses. On day 1 (sensitization), a workshop was used to sensitize and introduce participants to the digital transformation in health care and DATs. On day 2 (evaluative introduction), case vignettes and real-world nursing situations were discussed, and the care professionals selected a DAT based on the needs assessments of their patients. In addition, on day 2 (qualification), training was given on the proper use of the DATs. On days 3 to 5 (implementation), the selected DAT to be tested in the facilities was introduced.

This multistage structure of the SEQI education approach helped meet the core objective of this study, which was to raise awareness among care professionals of the digital transformation in health care and convey the use of DATs as a possible resource when planning nursing processes as this approach allowed for sufficient time for processing the extensive information provided in the 4 stages. Therefore, this training approach met care professionals’ desire for transparent information and training on new technologies.

In total, 2 researchers from the project team, a technician, and a nursing care researcher led the 5-day training course. Participants received additional informational material about the overall project and each DAT (see the next subsection). They were informed that they could contact the study team by telephone or email at any time during the trial period if they had any questions. Figure 1 shows the time frame of the SEQI intervention.

XSL•FO
RenderX

Figure 1. A schematic timeline of the sensitization, evaluative introduction, qualification, and implementation (SEQI) process.



Included DATs

The Assessment Instrument for Determining Care Dependency (BI) describes 8 modules [55-57]. In this study, these modules were understood as areas of care dependency within which specific nursing problems and a corresponding DAT assigned to achieve a defined nursing goal could be used [57]. On the basis of the assessment of the actual needs of the patients they were caring for, care professionals selected a suitable DAT. Six currently available DATs for use in long-term care facilities were selected:

1. A noninvasive sensor (DFree) that uses an app to determine bladder capacity and informs the user about the right time to go to the toilet [58]
2. A total of 2 robotic technologies belonging to the "social robotics" field (PARO and Pleo) [59,60]
3. A passive exoskeleton for relieving physical strain during demanding care tasks [61]
4. A mobile telepresence system with a self-balancing wheel and display for videoconferences [62,63]
5. Virtual reality applications for stress reduction and mindfulness [64]
6. A total of 2 communication robots for interaction using voice control and speech output (Pepper and Nao robots) [65]

Assessment of the Change in the Intention to Use DATs

The TUI allows for inferences about the intention to use specific technologies [36], whereas other subscales assess the actual use of a technology (electronic supplement in Multimedia Appendix 1 and measurement time and instruments in Multimedia Appendix 2). The 3 measurements were taken at the beginning of the intervention (T0), the end of day 2 (T1), and the end of the intervention (T2). Thereafter, participants evaluated their experience with the DAT and the educational intervention. The data analyses were conducted by health and nursing scientists (MZ and CB) and a colleague from the technical team (DB).

Assessment of the Learning Effects

A total of 12 guided interviews were conducted through theoretical sampling. The interviews were analyzed using systematic text condensation (STC) based on the work by Malterud [66,67]. The STC scheme, as proposed by Malterud [66,67], requires a structured approach to qualitative data analysis. This approach entails a bottom-up categorization process implemented inductively and, thus, hinges on a transparent and systematic methodology. This method begins with a comprehensive reading of the data to gain an overall impression followed by the identification of meaning units that relate to the research question. Subsequently, the meaning units are systematically coded into groups. These groups are then condensed, whereby the content is abstracted into a few comprehensive categories while the integrity of the original data is maintained. In conclusion, the essence of each category is synthesized into a theme or topic, thereby providing a clear and structured understanding of the data. This process guarantees that the analysis remains firmly anchored in the participants' perspectives while enabling the formulation of meaningful interpretations.

Guided Interviews

For the qualitative part of the study, 12 health care professionals participated in guided interviews. A total of 12 interviewees was deemed to be sufficient for facilitating an intensive data analysis, providing deeper insights, and achieving data saturation considering the resource constraints and research objectives. Interviewees were selected using convenience sampling. Statistical representativeness was not sought; rather, the interviews aimed to gather the specific knowledge necessary to understand participants' assessment of the SEQI education concept.

The interviews were conducted after the SEQI intervention in 46% (12/26) of the participating long-term care facilities. In every second facility, the participating care professionals were

XSL•FO
RenderX

asked whether one from the group would volunteer to be interviewed. The interviews were useful for obtaining both practically relevant background knowledge on the implementation of the SEQI education concept and guidance during the practical presentation of the research results [68]. The interviews focused on the meaningful, planned, and systematic integration of DATs into nursing processes from care professionals' perspectives.

The interview guidelines were developed iteratively based on the SPSS method described by Helfferich [69]. The interviews were recorded and transcribed verbatim. The discussions were centered on the pivotal question of the suitability of the SEQI education approach for health care practice. The objective was to examine the concept in the context of the 5-day practical trial, identify potential improvements from a practical perspective, and enhance the content of the concept. The interviews were conducted by a health and nursing scientist (SH and BK) and a colleague from the technical team (DB and CK). The interviews were coded and documented independently by 3 members of the research team (SH, BK, and AW). The objective of data saturation was not initially established. The practical implementation of the saturation principle is contingent upon the availability of a somewhat larger sample and the flexibility to determine the number of interviews conducted. In this case, the sample size was not predetermined; rather, interviews were conducted until no new information or categories were added to the existing information or categories. In the pilot study, data saturation was not reached due to the project team's awareness that data saturation is rarely achievable with regard to the comprehensive subject of the research. Instead, the interviews were to take into account the criterion of "internal representation," which replaces the criterion of representativeness as a quality criterion for samples [69]. Given the inherent difficulties in achieving high concordance between coders, no effort was made to measure it. This is particularly the case in the first round of coding, which often leads to a revision of the category system. Therefore, the resulting data will depend heavily on the extent and differentiation of the category system developed. The data analysis followed the STC method determined by Malterud [66,67] using the MAXQDA analysis software (version 20.0.7; VERBI GmbH).

Ethical Considerations

The ethics committee of the Medical Faculty of Martin Luther University Halle-Wittenberg approved this study on April 14, 2021 (approval 2021-021). The study was registered in the German Clinical Trials Register (DRKS00024967), and the protocol has not been published. Informed consent was obtained from all participants. No reward was given for participation.

Results

Survey Findings

A total of 122 participants were sampled from the 26 participating long-term care facilities. Of the 122 questionnaires returned, 10 (8.2%) were excluded from the analysis because they had incomplete responses or failed to select the specific DAT necessary for evaluating the intention to use DATs. In 41% (50/122) of the questionnaires, missing data were imputed using mean values (individual values were also attributed for various items). This procedure was performed after previous statistical consultation. As a simple random sample was used, the chosen imputation method was sufficient and did not affect the results. However, both simple and multiple imputations were performed to eliminate potential sources of error. The imputations were found to have only a minimal impact on the results. Half (62/112, 55.4%) of the participants were aged >41 years (Table 1). Most of the participants (78/112, 69.6%) had <20 years of professional experience.

Table 1. Sample characteristics (N=112).

Characteristic	Participants, n (%)
Sex	
Female	87 (77.7)
Male	22 (19.6)
No answer	3 (2.7)
Qualification	
3-year duration of training	54 (48.2)
2-year duration of training	3 (2.7)
At least 1-year duration of training	6 (5.4)
Therapist (speech, occupational, or physiotherapy)	13 (11.6)
Social worker	17 (15.2)
Social or welfare worker	3 (2.7)
Other qualification	11 (9.8)
No answer	5 (4.5)

XSL•FO
RenderX

Intention to Use

Intention to use was scored using the selected 7 subscales of the TUI. The total score ranged from 0 to 300. High subscale scores indicated a high level of the corresponding construct, whereas low scores indicated a low level. The mean intention to use score was 232 (SD 55) out of 300 points at the beginning of the intervention (T0; Table 2) and remained at almost the same level throughout the intervention (T1 and T2). No significance tests were conducted. Statements regarding significance were not necessary in this context because the trend was clear. Among the subscales, the participants were curious about DATs (21 out of 28 points), whereas their skepticism about DATs was low (13 out of 28 points; Table 3). In summary, the intention to use DATs and, thus, the predicted actual use of such technology in health care can be considered high.

Table 2. Results for the intention to use digital assistive technologies.

Time point	Scores, mean (SD)	Scores, median (IQR)
T0[a]	232 (55)	241 (203-275)
T1[b]	231 (66)	247 (208-280)
T2[c]	227 (72)	250 (187-288)

[a]T0: beginning of the intervention (n=111).
[b]T1: end of day 2 (n=112).
[c]T2: end of the intervention (n=112).

Table 3. Results for the intention to use digital assistive technologies by subscale.

	Score at T0[a]		Score at T1[b]		Score at T2[c]	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Curiosity[d]	21 (5)	22 (18-24)	21 (5)	21 (18-25)	20 (5)	21 (17-24)
Fear of technology[d]	13 (6)	12 (7-17)	9 (4)	8 (6-12)	10 (5)	8 (6-12)
Interest[d]	—[e]	—	21 (5)	22 (17-25)	21 (5)	22 (18-25)
Ease of use[f]	—	—	16 (4)	17 (14-19)	17 (4)	17 (14-20)
Usability[d]	—	—	20 (5)	21 (17-24)	19 (6)	20 (15-24)
Skepticism[d]	—	—	9 (5)	9 (6-12)	10 (4)	9 (7-11)
Accessibility[f]	—	—	13 (4)	12 (10-15)	12 (4)	12 (9-14)

[a]T0: beginning of the intervention (n=111).
[b]T1: end of day 2 (n=112).
[c]T2: end of the intervention (n=112).
[d]Range 4 to 28.
[e]No measurement at time T0.
[f]Range 3 to 21.

Interview Findings

As detailed in the Methods section, 12 health care professionals participated in guided interviews. The analysis of the interview data produced 14 codes that were condensed into 4 conceptual themes related to the learning effects: evaluation of the education concept (theme 1), effects on work and care structures (theme 2), need for reflection and discussion (theme 3), and improvement potential for health care and nursing care practice (theme 4). The 14 codes are presented in detail in Multimedia Appendix 3. Theme 1 describes the assessment of the SEQI education concept by health care professionals with a focus on the learning effects as SEQI focuses on training under real working conditions and, therefore, links theoretical and practical knowledge. Theme 2 addresses the impact of the education concept on work and care structures when using DATs. SEQI encourages the recapitulation and re-evaluation of work processes and care activities. The possibility to test already available DATs makes this recapitulation more realistic as it allows for the examination of integration possibilities under real working conditions. Theme 3 represents a synthesis of the preceding 2 themes. Care professionals assess application scenarios and determine the interactions among patients, DATs, and health care professionals based on the SEQI education concept. This entails a focus on knowledge and competence acquisition as well as the determination of interactions among patients. The opportunity to learn in a group of colleagues aligns with nursing learning habits and the availability of nursing expertise and previous knowledge. Theme 4 addresses the necessity to expand the scope of education and educate other professional groups on the subject matter. In this context, the dearth of DATs that are specifically tailored to patients'

XSL•FO
RenderX

functional limitations was also discussed. Practical testing experience reduces uncertainties and helps estimate which DATs can be used by other health care–related occupational groups. It is preparation for the possibilities of digitized health care, although it became evidently clear that the currently available DATs lack usefulness and usability for nursing care. This approach appears to be beneficial as it allows for the assessment of DATs using nursing expertise and the redesign of DATs to better align with the specific needs of nursing care.

Discussion

Principal Findings

The qualitative and quantitative findings reflect both the positive and negative aspects of DATs. To facilitate a more nuanced interpretation of the results, we will initially focus on the less complex quantitative survey data before subsequently turning our attention to the qualitative interview findings. This will facilitate the establishment of a clear link between the 2 sets of data. The survey results indicate that the reluctance of care professionals to implement DATs is not the primary cause of the issues that were identified. Rather, the perceived lack of usability and suitability was a significant contributing factor.

This mixed methods study was conducted to investigate the integration of DATs in nursing practices. This was achieved through the use of a structured education concept termed SEQI. Textbox 1 summarizes the key findings and insights derived from the study.

Textbox 1. Representative findings.

- Willingness to integrate digital assistive technologies (DATs) in real-world working conditions. This study revealed that care professionals demonstrated a general willingness to integrate DATs into their nursing processes even before the implementation of the sensitization, evaluative introduction, qualification, and implementation (SEQI) intervention. This suggests that care professionals have a fundamental openness to adopting new technologies.
- Impact of the SEQI intervention. The SEQI intervention furnished care professionals with indispensable background knowledge and sensitized them to the digital transformation. The intervention enabled the participants to evaluate the suitability of DATs in health care settings more effectively, understand the qualifications required for their appropriate application, and identify factors that would facilitate their use and integration.
- Practical and application-oriented learning. This study underscored the significance of practical, application-oriented but structured learning in fostering the long-term integration of DATs. The SEQI approach proved effective in enhancing care professionals' understanding and acceptance of DATs.
- Requirements for technology-related education built on care professionals' expertise, knowledge, and learning habits. Care professionals learn to express specific ideas and requirements for technology-related care as well as education concepts. This feedback is crucial for developing effective training programs that address the practical needs and challenges faced by care professionals due to the digital transformation of health care.
- Active matching of support and needs to be prepared for the digital transformation in the health care sector. The active matching of support and needs is a crucial aspect of this process. This study highlighted the significance of aligning technical support with patients' physical limitations (need to fit) and requirements when selecting and integrating DATs. This need-to-fit approach guarantees that the technologies used are beneficial and appropriate for the patients.

In the statistical methodology, the mean is a widely used measure of the central tendency of a given data set. However, the mean is susceptible to influence from values at the extreme high and low ends of the results. Consequently, the median is a superior measure of central tendency in instances in which a small number of outliers can significantly influence the mean. The median values identified in this study indicate a marginal increase in willingness to use, which is not apparent when interpreting the mean values. This indicates that the care professionals identified shortcomings in the selected DATs during the 3-day practical testing. A survey was conducted with 112 care professionals at 3 distinct measurement points to ascertain their intention to use DATs. The results indicated a positive shift in intention to use DATs following the implementation of the SEQI intervention. No sample size calculation or *P* values were provided. Nevertheless, the overall trend indicated an increase in acceptance, willingness, and intention to use DATs among the participants.

Qualitative interviews with 12 care professionals yielded further insights. The participants expressed appreciation for the SEQI program's structured approach, particularly the evaluative introduction and on-site testing stages, which they found to foster confidence and competence in the use of DATs. In total, 4 primary themes were identified.

In theme 1, "Evaluation of the education concept," the care professionals described the preparatory and theoretical introduction in stage 1 (sensitization), along with stages 2 to 4, as meaningful. Stages 1 and 2 (evaluative introduction) provided knowledge of the digital transformation and served as preparation for the practical experience in stages 3 (qualification) and 4 (implementation). The care professionals noted that a focused discussion of DATs is more effective than a pure knowledge transfer, such as only providing a user manual. Previous theoretical knowledge helps classify new information on DATs both ethically and normatively, and problems are related and supported based on the theoretical background as they arise, which is necessary for integrating DATs into practical work. Furthermore, the interviewees emphasized that practical teaching using case studies and group work is helpful for deepening understanding.

In theme 2, "Effects on work and care structures," the care professionals reported that the use of DATs could enhance the quality of care and optimize work processes. However, it was

XSL•FO
RenderX

also identified that the practical introduction of DATs presents challenges, such as the need for familiarity with new systems and the adaptation of processes.

In theme 3, “Need for reflection and discussion,” the care professionals indicated that the workshops facilitated reflection on the role of DATs in care facilities and on their own attitudes while using DATs. The opportunity to discuss these topics helped them gain a deeper understanding of the digital transformation and learn about different perspectives.

In theme 4, “Improvement potential for health care and nursing care practice,” the care professionals emphasized the importance of careful selection of DATs and the necessity of training on their proper implementation to improve the quality of care. In addition, cooperation between care professionals and technical experts was identified as a key factor in the successful introduction of DATs [40].

Takeaways and Themes

The incorporation of structured educational methodologies such as the SEQI intervention is of paramount importance for the integration of DATs into nursing practices. The SEQI intervention provides care professionals with the requisite knowledge and skills to use these technologies in an effective manner. In addition, a positive attitude toward technology among health care professionals is essential for successful integration.

It is imperative to assess the usability of DATs in long-term care settings. The emphasis on practical application and real-world testing in SEQI proved to be a significant contributing factor to its success. By integrating theoretical knowledge and hands-on experience, SEQI ensures that technologies are tailored to meet specific patient needs, a recurring theme in our discussions. The structured and guided education concept of SEQI serves to enhance care professionals’ capacity to integrate DATs into their practices, facilitating this process through practical, application-oriented training. In the initial phase of the SEQI educational approach (sensitization), the theoretical underpinnings of digital transformation in health care were elucidated and deliberated with care professionals. Given that care professionals constitute the largest group in the health care system, they were encouraged to identify potential applications and areas of use for DATs based on their expertise (evaluative introduction). This process enabled them to evaluate DATs from the perspectives of their patients and the specific care challenges they face. The SEQI education concept can play an instrumental role of the profession as it invites care professionals to assess the usefulness of DATs. The structured theoretical knowledge transfer and the gradual transition from passive knowledge consumers to active users through consecutive training units were well received by the participants. This mutual approach proved conducive to the development of practical applications for DATs.

The qualitative findings provide information on 4 themes.

The SEQI initiative enhanced the comprehension of alternative courses of action and resolutions pertaining to the use of DATs in nursing procedures. The direct interaction during the testing phase enabled care professionals to assess the potential benefits and limitations of DATs, thereby fostering realistic and independent evaluations. The potential for technology to dehumanize care was countered through practical application testing, which demonstrated that care is fundamentally an interpersonal interaction that cannot be replaced by technology. Dehumanization is defined by Biniok [70] as a form of the denial of the human characteristics and qualities of other individuals. For example, social interactions and relationships could become devalued if care professionals were replaced by robots. The opportunity for direct testing highlighted the possibilities and shortcomings of DATs. The importance of the “human factor” for the high-quality provision of services became clear to the participating care professionals, and concerns about care professionals being replaced by DATs were refuted. The care professionals exhibited an awareness that care is fundamentally an interpersonal interaction that cannot be readily substituted by technology [12]. The interview data revealed that the care professionals’ lack of interest in DATs should not be interpreted as a general lack of interest [71-73]. This interpretation has arisen due to a methodological limitation in numerous previous studies in which care professionals were asked about a topic with which they lacked direct experience [29,74,75]. The lack of objective, tangible information on DATs and their potential applications makes it difficult for care professionals to formulate unbiased opinions. Consequently, they may only provide superficial responses to inquiries about their interest in DATs as it has been demonstrated that empirical evidence is essential for making well-informed decisions [76]. The focus was not on the DATs themselves but on the care professionals’ questions, conflicts, and behavioral uncertainties regarding DATs. This process-oriented approach to knowledge transfer, also known as the genetic method, aims not only to convey knowledge but also to acquire practical skills, thereby directing the focus to new or alternative situation interpretations.

The SEQI intervention can have a lasting impact when both cognitive transfer (knowledge, skills, and abilities) and emotional transfer (attitudes, values, motives, and feelings) are achieved [77]. An emotional transfer occurred when the relevance of DATs to specific care problems was perceived, thereby reducing skepticism and increasing willingness to engage with their implementation. SEQI helped defuse the often emotional discussion around robotics in health care because a realistic assessment of DATs was possible for the care professionals. One illustration of the emotional transfer achieved through practice with DATs was when a caregiver described the interaction of an introverted older adult with the PARO device. Upon being stimulated by the device, the older adult relinquished his self-imposed isolation, commenced laughing, and “began to tell stories.” Furthermore, a relatively young man with spinal cord injuries was able to experience a visit to a Rammstein concert through virtual reality immersion. The care professionals were impressed by the absence of any reservations or fears about the virtual environment.

Furthermore, SEQI facilitated the practical transformation of theoretical knowledge, as evidenced by the descriptions of necessary adjustments to workflows and care activities provided by the participating care professionals. Theme 2 (“Effects on work and care structures”) showed that the multiday on-site

XSL•FO
RenderX

training was met with a positive evaluation, with participants noting the rarity of such opportunities for testing DATs in long-term care for older adults [78]. The results of the survey indicated that care professionals, the largest health care group, play a crucial role in the formation of sociotechnical care arrangements. The high scores on the interest and curiosity subscales and the low scores on the fear of technology subscale indicated that the participants were open to testing new technologies provided that they are usable and tailored to patient needs. In line with the theory of transformative learning by Mezirow [44], care professionals developed application ideas through reflective transformation, thereby ensuring long-term applicability in their professional practices. These considerations also extended to their own work processes, such as the integration of exoskeletons into shift planning or the logistics of outpatient care. To accomplish this, nursing professionals would be required to operate a vehicle while wearing the exoskeleton to visit and provide care for their patients and clients in their residences.

The survey results indicated the importance of actively involving nursing care professionals, the largest group of health professionals, in the design of socio-technical care arrangements (as evidenced by the high scores on the interest subscale and the low scores on the fear of technology subscale). In addition, the high scores on the curiosity subscale suggest that caregivers are open to trying new technologies, including DATs, if they are usable and tailored to the needs of target populations [79]. The consistently high scores on the overall intention to use scale were noteworthy as all participants demonstrated a willingness to use DATs at the outset of the study and retained this willingness even when the fit of the DAT to the patient was deemed to be inadequate. This indicates that care professionals perceive the potential for enhancing the suitability of DATs for their needs through participatory development and are willing to use them in the future.

In theme 3 (“Need for reflection and discussion”), the pragmatic aspects of the SEQI, such as sufficient time for questions, reflection during breaks, and mutual exchange with colleagues, were described in a positive light. It is imperative to reflect on DATs for the advancement of one’s professional capabilities, facilitating the expansion of personality traits and meta-competencies. The high score on the overall intention to use scale, despite the low fit of some DATs, indicates a fundamental interest in using technologies and a willingness to participate in their development and adaptation. This indicates that care professionals have a high level of self-responsibility and self-reflection about their professional practice and are willing to continuously educate themselves [80]. Through practical experiences and encounters, participants can easily acquire theoretical knowledge and apply it to their professional practice. If the acquired knowledge yields positive results when applied in practice, it is more likely to be remembered and applied in the long term. Therefore, practical training such as SEQI is often more effective than purely theoretical training as it offers the opportunity to directly implement learned knowledge and gain practical experience [80].

In theme 4 (“Improvement potential for health care and nursing care practice”), care professionals across the 26 facilities recommended an extended testing period on weekends to allow for a realistic assessment of DATs considering increased workloads and reduced staff capacity. Although facility management is ultimately responsible for procurement, care professionals acknowledged the necessity of balancing cost and benefit considerations. It was further proposed that additional health care professionals, such as nursing assistants, be trained in the use of DATs to ensure the maintenance of professional standards and appropriate delegation of caregiving tasks. Care professionals emphasized the importance of careful selection of DATs and the necessity of training on their proper implementation to improve the quality of care. In addition, cooperation between care professionals and technical experts was identified as a key factor in the successful introduction of DATs [40].

In conclusion, the SEQI educational concept effectively integrates structured education with practical application and real-world testing, thereby fostering positive attitudes toward DATs among care professionals. This comprehensive approach guarantees that DATs are adapted to the specific requirements of patients and residents and integrated into nursing practices, thereby improving the quality of care and optimizing work processes.

Limitations and Strengths

The mixed methods approach, which focuses on qualitative, sequential exploration, was a strength of this study. The basic research design was critically reflected upon using the Good Reporting of a Mixed Methods Study checklist [49]. Furthermore, we classified and reflected on the quality criteria following Lincoln et al [81] and Lamnek [82]. The quality criteria of the survey were discussed using the Mixed Methods Appraisal Tool [50]. The openness of the mixed methods approach proved to be effective for this study considering the complexity and range of topics thus far underexplored. In particular, this approach allowed the researchers to provide a flexible response to the problems in the sampling strategy and the design of the empirical study.

This study had certain limitations. Primarily, it was conducted exclusively in long-term care facilities, and thus, the results are applicable only to this area. The recruitment of facilities was carried out exclusively at the management level, which may have introduced a positive selection bias (ie, only technology-friendly facilities were willing to participate). In addition, because all the visited facilities were informed about the observation element for ethical research reasons, the possibility of biased responses should also be considered. One of the key strengths of this study is that it marks the first time that caregivers have had access to DATs on a large scale and been able to test them in practice over a longer period in the context of nursing care. In preliminary work, the validity of the data would need to be critically questioned due to skepticism, which is particularly pronounced among care professionals in cases of lack of access to DATs or predefined media ignorance. However, research on DATs presents a complex picture. Although the self-image of care professionals indicates a high degree of conformity with traditional nursing practices [81],

XSL•FO
RenderX

these practices have expanded, changed, and transformed the professional self-concept [33].

The question of whether it is reasonable to assume that nursing professionals are unable to manage intervention and interview situations independently is open to debate. This suggests that societal trends are employed to delineate uncertainties pertaining to the utilization of DAT in relation to the fundamental tenets of nursing practice, the incorporation of DAT into nursing processes, and the associated protection claims of individuals requiring care. In this context, societal trends are employed to describe nursing and healthcare professionals as occupying a subordinate position within a hierarchical healthcare system. This line of reasoning relies on a pervasive trope that portrays the purported uncertainty of nursing professionals with regard to the utilization of DAT. This gives rise to a debate about the possibility of technological reductionism undermining professional claims to individualized services and the expertise of nurses and healthcare professionals, particularly in the event that automated decisions are made by DATs in the future.

Moreover, there is a concern that the physical aspect of the interaction between care professionals and patients could be substituted by technology in the long term [83,84]. Within the group of participating care professionals, the fear of being replaced—a serious concern from the early days of discussion on DATs—is gradually giving way to the recognition of DATs as a useful complementary support tool [5,9]. Health care professionals are increasingly recognizing that digital technologies and robotics can complement and support their work rather than replace them. This indicates that the attitudes toward and perceptions of the role of technology in care have changed over time. Care professionals are increasingly seeing the interaction between technology and human care as improving caregiving for people [37]. This shift in attitudes suggests a growing recognition of the potential benefits of technology in nursing practices.

The primary challenge remains the development of methods to make DATs accessible and useful for care professionals in long-term care facilities. DATs have the potential to reduce the necessity for formal health, support, and long-term care services by assuming care tasks that are required when limitations prevent a patient from completing self-care independently. This, in turn, could result in a reduction in the workload of care professionals. Consequently, DATs are being developed in nursing to support patients and improve their compliance, which benefits care professionals. Thus, in conjunction with sufficient staffing, DATs can diminish the continuous work performed under time constraints and effectively digitize nursing care [5]. However, the impact of DATs on patients was not a primary focus of this study. Therefore, future studies should assess the extent to which the potential and effective alleviation of care professionals' workload using DATs affects overall care quality.

Ultimately, the best DATs do not serve their purpose if they do not benefit the people being cared for. A strength of this study was that care professionals were granted broad access to DATs for the first time and, thus, experienced the actual usability of DATs in a real-world context. This indicates the conditions and situations in which care professionals accept or are skeptical about the use of DATs in long-term care facilities.

Conclusions, Outlook, and Implications for Practice

This study revealed that care professionals are open to using DATs; however, they need information and knowledge on how to reflect on DATs critically. This study also highlights the effectiveness of the SEQI education concept for transferring theory into practice. Information, real-world practice, and learning are essential for reducing barriers and promoting an understanding of potential application fields and limitations. Our participants' critical reflections revealed that currently available DATs only offer limited relief for care professionals as they merely support social care and daily life management. Hence, the targeted and patient-oriented use of DATs is necessary to promote critical reflection on the suitability of such technologies for nursing processes. The SEQI education concept can be used to strengthen care professionals' competencies in dealing with DATs and enable a realistic assessment.

To achieve the long-term implementation of DATs in practice, practical and economic factors such as the creation and expansion of a comprehensive digital infrastructure must be considered. This study's results underline the important role that care professionals play in the interprofessional team of health care professionals when using DATs as part of nursing processes. In this paper, the structured approach of SEQI was presented as a useful way to integrate DATs into nursing processes. The approach was positively received, and interest in DATs remained high throughout the study. The SEQI approach can help evaluate the suitability of DATs for identified functional nursing problems and integrate DATs into nursing processes. However, because of time restrictions for care professionals, it may be beneficial to involve other professional groups, such as social workers, in structured technology education. In addition, the exclusion of other health care professionals (eg, nursing assistants and station assistants) as potential users seems excessively limiting and should be avoided. In addition, while SEQI supports a more realistic understanding of DATs by care professionals, it can be used to stimulate discourse about DATs. As an education approach, SEQI has strong practical relevance, and its stages can be transferred to different nursing processes. This approach is easily implemented in long-term care facilities and could be included in the professional education of health care workers. Care professionals' lack of experience with DATs highlights the importance of answering questions about the actual approaches regarding the sustainability of implementing these technologies in nursing practices. It is also essential to determine who qualifies to use these technologies.

The participants' high willingness to use DATs in practice should be supported through holistic and application-oriented concepts that also consider ethical and normative aspects. SEQI, accompanied by the creation and expansion of a comprehensive digital infrastructure, can be implemented to build competencies to create the necessary conditions for the long-term implementation of DATs in practice. Regarding the use of DATs in nursing processes, care professionals play a central role among health care professionals.

XSL•FO
RenderX

Acknowledgments

The authors would like to thank the staff of the institutions involved in the implementation of this study for their excellent collaboration, without which this study would not have been possible. This work was created as part of the research project Forschungsbasierte Entwicklung einer beschleunigten praktischen Implementierung assistiver & digitaler Technik in der pflegerischen Versorgung älterer Menschen in Sachsen-Anhalt (FORMAT)-Continuum (Autonomy in Old Age research association, Europäischer Fonds für regionale Entwicklung [EFRE] funds; project duration: September 1, 2019-December 31, 2022; funding reference: ZS/2019/02/97281). We acknowledge the financial support of the Open Access Publication Fund (Publikationsfond) of the Martin-Luther-University Halle-Wittenberg.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of selected technologies tested during study period.
[DOCX File , 14 KB - mededu_v10i1e54083_app1.docx]

Multimedia Appendix 2
Supplement of measurement time during SEQI process.
[DOCX File , 16 KB - mededu_v10i1e54083_app2.docx]

Multimedia Appendix 3
Supplement Learning effects in detail.
[DOCX File , 29 KB - mededu_v10i1e54083_app3.docx]

References

1. Bräseke G, Nägele G, Lingot N. Einsatz von robotischen Systemen in der Pflege in Japan mit Blick auf den steigenden Fachkräftebedarf. IGES Institut. Ein Unternehmen der IGES Gruppe. 2019. URL: https://tinyurl.com/3tynfje8 [accessed 2024-04-29]
2. Kuhn S. Wie revolutioniert die digitale Transformation die Bildung der Berufe im Gesundheitswesen? Bern Open Repository and Information System. 2019. URL: https://boris.unibe.ch/132747/ [accessed 2024-04-29]
3. Alves-Oliveira P, Petisca S, Correia F, Paiva A. Social robots for older adults: framework of activities for aging in place with robots. In: Proceedings of the 7th International Conference on Social Robotics. 2015 Presented at: ICSR '15; October 26-30, 2015; Paris, France p. 11-20 URL: https://link.springer.com/chapter/10.1007/978-3-319-25554-5_2 [doi: 10.1007/978-3-319-25554-5_2]
4. Krick T, Huter K, Seibert K, Domhoff D, Wolf-Ostermann K. Measuring the effectiveness of digital nursing technologies: development of a comprehensive digital nursing technology outcome framework based on a scoping review. BMC Health Serv Res 2020 Mar 24;20(1):243 [FREE Full text] [doi: 10.1186/s12913-020-05106-8] [Medline: 32209099]
5. Hofstetter S, Kraft B, Jahn P. Roboter – die neuen Kollegen im Team? Gesundheits- und Sozialpolitik (G&S). 2023. URL: https://tinyurl.com/9d4n36jb [accessed 2024-04-29]
6. Heinemann S, Matusiewicz D. Rethink Healthcare: Crisis as an Opportunity. Heidelberg, Germany: Medhochzwei Verlag; 2021.
7. Klie T. Im Rückblick 25 Jahre Pflegeversicherung. In: Storm A, editor. 25 Jahre Pflegeversicherung: Kosten der Pflege. Hamburg, Germany: medhochzwei Verlag; 2019:5-8.
8. Paulicke D. Assistive Technologien für pflegende Angehörige von Menschen mit Demenz : beschreibende Studie zu einem transformativen Informations- und Beratungsverständnis. Martin-Luther-Universität Halle-Wittenberg. 2021. URL: https://tinyurl.com/kv93phy2 [accessed 2024-04-29]
9. Geist L, Immenschuh U, Jahn P, Paulicke D, Zilezinski M, Buhtz C, et al. Identifikation von lernfördernden Maßnahmen zur Einführung von Digitalen und assistiven Technologien (DAT) in Prozesse der pflegerischen Versorgung: eine qualitative Studie. HeilberufeScience 2022 Jun 14;13(3-4):152-161 [FREE Full text] [doi: 10.1007/s16024-022-00372-4] [Medline: 35730048]
10. Hofstetter S, Buhtz C, Paulicke D, Jahn P. Lernen in bewegten Zeiten. Pflegez 2020 Oct 20;73(11):42-45. [doi: 10.1007/S41906-020-0927-0]
11. Buhtz C, Paulicke D, Hofstetter S, Jahn P. Technikaffinität und Fortbildungsinteresse von Auszubildenden der Pflegefachberufe: eine Onlinebefragung. HBScience 2020 Jan 30;11(1-2):3-12 [FREE Full text] [doi: 10.1007/S16024-020-00337-5]

XSL•FO
RenderX

12. Robotik für gute Pflege: stellungnahme. Deutscher Ethikrat. URL: https://tinyurl.com/57mk6mzt [accessed 2024-04-29]
13. Assistive technology. World Health Organization (WHO). 2018. URL: https://tinyurl.com/54337fu9 [accessed 2024-04-29]
14. Competencies for care professionals working in primary health care. World Health Organization (WHO). URL: https://ccoms.esenfc.pt/pub/Competencies-care [accessed 2024-04-29]
15. Janda V. Usability ist keine Eigenschaft von Technik. In: Inchubert C, Schulz-Schaeffer I, editors. Berliner Schlüssel zur Techniksoziologie. Cham, Switzerland: Springer; 2019:347-374.
16. Servaty R, Kersten A, Brukamp K, Möhler R, Mueller M. Implementation of robotic devices in nursing care. Barriers and facilitators: an integrative review. BMJ Open 2020 Sep 21;10(9):e038650 [FREE Full text] [doi: 10.1136/bmjopen-2020-038650] [Medline: 32958491]
17. Schüßler I. Zur (Un-)Möglichkeit einer Wirkungsforschung in der Erwachsenenbildung: Kritische Analysen und empirische Befunde. Deutsches Institut für Erwachsenenbildung. 2012. URL: https://tinyurl.com/47wj4bhu [accessed 2024-04-29]
18. Kaihlanen A, Gluschkoff K, Kinnunen UM, Saranto K, Ahonen O, Heponiemi T. Nursing informatics competences of Finnish registered nurses after national educational initiatives: a cross-sectional study. Nurse Educ Today 2021 Nov;106:105060 [FREE Full text] [doi: 10.1016/j.nedt.2021.105060] [Medline: 34315050]
19. Nadav J, Kaihlanen AM, Kujala S, Laukka E, Hilama P, Koivisto J, et al. How to implement digital services in a way that they integrate into routine work: qualitative interview study among health and social care professionals. J Med Internet Res 2021 Dec 01;23(12):e31668 [FREE Full text] [doi: 10.2196/31668] [Medline: 34855610]
20. Albrecht UV, Behrends M, Schmeer R, Matthies HK, von Jan U. Metadata correction: usage of multilingual mobile translation applications in clinical settings. JMIR Mhealth Uhealth 2013 Aug 07;1(2):e19. [doi: 10.2196/mhealth.2866] [Medline: 25098738]
21. Merda M, Schmidt K, Löchert B. Pflege 4.0 – Einsatz moderner Technologien aus der Sicht professionell Pflegender. Berufsgenossenschaft für Gesundheitsdienst und Wohlfahrtspflege (BGW). 2017. URL: https://tinyurl.com/4f9fxu7k [accessed 2024-04-29]
22. Buhtz C, Paulicke D, Schwarz K, Jahn P, Stoevesandt D, Frese T. Receptiveness of GPs in the south of Saxony-Anhalt, Germany to obtaining training on technical assistance systems for caregiving: a cross-sectional study. Clin Interv Aging 2019;14:1649-1656 [FREE Full text] [doi: 10.2147/CIA.S218367] [Medline: 31571844]
23. Hung SY, Ku YC, Chien JC. Understanding physicians' acceptance of the Medline system for practicing evidence-based medicine: a decomposed TPB model. Int J Med Inform 2012 Feb;81(2):130-142. [doi: 10.1016/j.ijmedinf.2011.09.009] [Medline: 22047627]
24. Pynoo B, Devolder P, Duyck W, van Braak J, Sijnave B, Duyck P. Do hospital physicians' attitudes change during PACS implementation? A cross-sectional acceptance study. Int J Med Inform 2012 Feb;81(2):88-97 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.10.007] [Medline: 22071012]
25. Frommeld D, Scorna U, Haug S. Gute Technik für ein gutes Leben?!. In: Frommeld D, Scorna U, Haug S, Weber K, editors. Gute Technik für ein gutes Leben im Alter?: Akzeptanz, Chancen und Herausforderungen altersgerechter Assistenzsysteme. Berlin, Germany: Transcript Verlag; 2021:11-26.
26. Hofstetter S, Richey V, Jahn P. Digitale Revolution: survey zur Akzeptanz sozial assistiver Technologie in der Pflege. Die Schwester Pfleger. 2019. URL: https://www.bibliomed-pflege.de/sp/artikel/38742-digitale-revolution [accessed 2024-04-29]
27. Holden RJ, Brown RL, Scanlon MC, Karsh B. Modeling nurses' acceptance of bar coded medication administration technology at a pediatric hospital. J Am Med Inform Assoc 2012 Nov 01;19(6):1050-1058 [FREE Full text] [doi: 10.1136/amiajnl-2011-000754] [Medline: 22661559]
28. Kowitlawakul Y. The technology acceptance model: predicting nurses' intention to use telemedicine technology (eICU). Comput Inform Nurs 2011 Jul;29(7):411-418. [doi: 10.1097/NCN.0b013e3181f9dd4a] [Medline: 20975536]
29. Zöllick JC, Kuhlmey A, Suhr R, Eggert S, Nordheim J, Blüher S. Akzeptanz von Technikeinsatz in der Pflege. In: Jacobson K, Kuhlmey A, Greß S, Klauber J, Schwinger A, editors. Pflege-Report 2019: Mehr Personal in der Langzeitpflege - aber woher?. Cham, Switzerland: Springer; 2020:211-218.
30. Sun SL, Hwang HG, Dutta B, Peng MH. Exploring critical factors influencing nurses' intention to use tablet PC in Patients' care using an integrated theoretical model. Libyan J Med 2019 Dec;14(1):1648963 [FREE Full text] [doi: 10.1080/19932820.2019.1648963] [Medline: 31357919]
31. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res 2019 Jun 20;19(1):400 [FREE Full text] [doi: 10.1186/s12913-019-4238-3] [Medline: 31221133]
32. Paulicke D, Buhtz C, Voigt J. Aufgeschlossenheit und Fortbildungsinteresse von PflegeschülerInnen zu technischen und digitalen Assistenzsystemen. Konferenzband Zukunft der Pflege. 2018. URL: https://tinyurl.com/58ykzt4t [accessed 2024-04-29]
33. Damschroder LJ, Reardon CM, Widerquist MA, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]

XSL•FO
RenderX

34. Gallivan MJ, Spitlers VK, Koufaris M. Does information technology training really matter? A social information processing analysis of coworkers' influence on IT usage in the workplace. J Manag Inf Syst 2014 Dec 08;22(1):153-192. [doi: 10.1080/07421222.2003.11045830]
35. Lee SM, Kim YR, Lee J. An empirical study of the relationships among end-user information systems acceptance, training, and effectiveness. J Manag Inf Syst 2015 Dec 11;12(2):189-202. [doi: 10.1080/07421222.1995.11518086]
36. Kothgassner O, Felnhofer A, Hauk N. TUI technology usage inventory. Information- and Communication technology Applications: Research on User-oriented Solutinso. 2013. URL: https://tinyurl.com/4vrtvrkj [accessed 2024-04-29]
37. Wilson ML, Rebecca F. Are care professionals able to lead in the digital health evolution? Developing an informatics competent and capable nursing workforce. Healthcare Information and Management Systems Society. URL: https://www.himss.org/resources/are-care [accessed 2024-04-29]
38. Maier I, Möller T. Pflege Mit Masterplan in die Zukunft. Bibliomed. 2021. URL: https://tinyurl.com/537w95z4 [accessed 2024-04-29]
39. Hofstetter S, Lehmann L, Zilezinski M, Steindorff J, Jahn P, Paulicke D. Vermittlung digitaler Kompetenzen in der Pflegeausbildung – eine Vergleichsanalyse der Rahmenpläne von Bund und Ländern. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2022 Sep 09;65(9):891-899 [FREE Full text] [doi: 10.1007/s00103-022-03575-2] [Medline: 35943547]
40. Kaap-Fröhlich S, Ulrich G, Wershofen B, Ahles J, Behrend R, Handgraaf M, et al. Position paper of the GMA committee interprofessional education in the health professions - current status and outlook. GMS J Med Educ 2022;39(2):Doc17 [FREE Full text] [doi: 10.3205/zma001538] [Medline: 35692364]
41. Belliger A, Krieger DJ. The digital transformation of healthcare. In: North K, Maier R, Haas O, editors. Knowledge Management in Digital Change: New Findings and Practical Cases. Cham, Switzerland: Springer; 2018:311-326.
42. Bergen I, Belliger A. über kulturelle Interoperabilität, Future Skills und warum Netzwerke Plattformen ablösen. Visionäre Der Gesundheit. URL: https://visionaere-gesundheit.de/andrea-belliger/ [accessed 2022-05-11]
43. Hasseler M, Lietz AL, Krebs S. Delegation im Krisenfall - Entscheidungen erleichtern: Flexibilität und Kommunikation sind unabdingbar. Procare 2020 Nov 13;25(9):46-49 [FREE Full text] [doi: 10.1007/s00735-020-1268-3] [Medline: 33250584]
44. Mezirow J. Perspective transformation. Adult Educ 2016 Sep 16;28(2):100-110. [doi: 10.1177/074171367802800202]
45. Mezirow J. Transformative Erwachsenenbildung. The German Education Portal. URL: https://tinyurl.com/y464jk5d [accessed 2024-04-29]
46. Hanft A, Brinkmann K. Offene Hochschulen: Die Neuausrichtung der Hochschulen auf Lebenslanges Lernen. Berlin, Germany: Waxmann; 2013.
47. Singer-Brodowski: Transformatives Lernen als neue Theorie-Perspektive in der BNE. Die Kernidee transformativen Lernens und seine Bedeutung für informelles Lernen. PolIBNT. 2016. URL: https://tinyurl.com/3esys5ah [accessed 2024-04-29]
48. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research. Thousand Oaks, CA: Sage Publications; 2017.
49. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008 Apr;13(2):92-98. [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
50. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
51. Halcomb EJ, Andrew S. Triangulation as a method for contemporary nursing research. Nurse Res 2005 Oct;13(2):71-82. [doi: 10.7748/nr.13.2.71.s8] [Medline: 16416981]
52. Östlund U, Kidd L, Wengström Y, Rowa-Dewar N. Combining qualitative and quantitative research within mixed method research designs: a methodological review. Int J Nurs Stud 2011 Mar;48(3):369-383 [FREE Full text] [doi: 10.1016/j.ijnurstu.2010.10.005] [Medline: 21084086]
53. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319. [doi: 10.2307/249008]
54. Davis FD. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. Int J Man Mach Stud 1993 Mar;38(3):475-487. [doi: 10.1006/imms.1993.1022]
55. Wingenfeld K, Büscher A, Gansweid B. Das neue Begutachtungsinstrument zur Feststellung von Pflegebedürftigkeit. GKV-Spitzenverband. 2011. URL: https://tinyurl.com/5n6ckm45 [accessed 2024-04-29]
56. Wingenfeld K, Schaeffer D. Die Weiterentwicklung des Pflegebedürftigkeitsbegriffs in der Pflegeversicherung. Heft 2011;11:7-13 [FREE Full text]
57. Kiefer G, Pick P. Richtlinien des GKV-Spitzenverbandes zur Feststellung der Pflegebedürftigkeit nach dem XI. Buch des Sozialgesetzbuches. Medizinischer Dienst des Spitzenverbandes Bund. 2021. URL: https://tinyurl.com/2hndfbcz [accessed 2024-04-29]
58. Hofstetter S, Zilezinski M, Wolf A, Behr D, Paulicke D, Stoevesandt D, et al. Dfree ultrasonic sensor in supporting quality of life and patient satisfaction with bladder dysfunction. Int J of Uro Nursing 2022 Nov;17(1):62-69. [doi: 10.1111/ijun.12334]

XSL•FO
RenderX

59. Kelly PA, Cox LA, Petersen SF, Gilder RE, Blann A, Autrey AE, et al. The effect of PARO robotic seals for hospitalized patients with dementia: a feasibility study. Geriatr Nurs 2021 Jan;42(1):37-45. [doi: 10.1016/j.gerinurse.2020.11.003] [Medline: 33221556]
60. Pu L, Moyle W, Jones C, Todorovic M. The effect of using PARO for people living with dementia and chronic pain: a pilot randomized controlled trial. J Am Med Dir Assoc 2020 Aug;21(8):1079-1085. [doi: 10.1016/j.jamda.2020.01.014] [Medline: 32122797]
61. Zelik KE, Nurse CA, Schall MC, Sesek RF, Marino MC, Gallagher S. An ergonomic assessment tool for evaluating the effect of back exoskeletons on injury risk. Appl Ergon 2022 Feb;99:103619 [FREE Full text] [doi: 10.1016/j.apergo.2021.103619] [Medline: 34740072]
62. Geier J, Mauch M, Patsch M, Paulicke D. Wie Pflegekräfte im ambulanten Bereich den Einsatz von Telepräsenzsystemen einschätzen - Eine qualitative Studie. Pflege 2020 Feb;33(1):43-51. [doi: 10.1024/1012-5302/a000709] [Medline: 31691626]
63. Hung L, Wong J, Smith C, Berndt A, Gregorio M, Horne N, et al. Facilitators and barriers to using telepresence robots in aged care settings: a scoping review. J Rehabil Assist Technol Eng 2022 Jan 21;9:20556683211072385 [FREE Full text] [doi: 10.1177/20556683211072385] [Medline: 35083063]
64. Planert J, Machulska A, Hildebrand AS, Roesmann K, Otto E, Klucken T. Self-guided digital treatment with virtual reality for panic disorder and agoraphobia: a study protocol for a randomized controlled trial. Trials 2022 May 21;23(1):426 [FREE Full text] [doi: 10.1186/s13063-022-06366-x] [Medline: 35597959]
65. Stoevesandt D, Jahn P, Watzke S, Wohlgemuth WA, Behr D, Buhtz C, et al. Comparison of acceptance and knowledge transfer in patient information before an MRI exam administered by humanoid robot versus a tablet computer: a randomized controlled study. Rofo 2021 Aug 10;193(8):947-954 [FREE Full text] [doi: 10.1055/a-1382-8482] [Medline: 34111898]
66. Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet 2001 Aug 11;358(9280):483-488. [doi: 10.1016/S0140-6736(01)05627-6] [Medline: 11513933]
67. Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scand J Public Health 2012 Dec 04;40(8):795-805. [doi: 10.1177/1403494812465030] [Medline: 23221918]
68. Kaiser R. Konzeptionelle und methodologische Grundlagen qualitativer Experteninterviews. In: Kaiser R, editor. Qualitative Experteninterviews: Konzeptionelle Grundlagen und praktische Durchführung. Cham, Switzerland: Springer; 2014:21-49.
69. Helfferich C. Die Qualität qualitativer Daten: manual für die Durchführung qualitativer Interviews. Cham, Switzerland: Springer; 2011.
70. Biniok P. Assistenz-Triaden. Abwägungen zu Versorgungssicherheit und Entmenschlichung durch assistive Technologien. In: Luthe EW, Müller SV, Müller I, editors. Assistive Technologien im Sozial- und Gesundheitssektor. Cham, Switzerland: Springer; 2022:599-621.
71. Krings BJ, Weinberger N. Assistant without Master? Some conceptual implications of assistive robotics in health care. Technol 2018 Jan 18;6(1):13. [doi: 10.3390/technologies6010013]
72. Maibaum A, Bischof A, Hergesell J, Lipp B. A critique of robotics in health care. AI Soc 2021 Apr 16;37(2):467-477. [doi: 10.1007/S00146-021-01206-Z]
73. Smarr CA, Prakash A, Beer JM, Mitzner TL, Kemp CC, Rogers WA. Older adults' preferences for and acceptance of robot assistance for everyday living tasks. Proc Hum Factors Ergon Soc Annu Meet 2012 Sep;56(1):153-157 [FREE Full text] [doi: 10.1177/1071181312561009] [Medline: 25284971]
74. Kuhlmey A, Blüher S, Nordheim J. Ressource oder Risiko. Wie professionell Pflegende den Einsatz digitaler Technik in der Pflege sehen Zentrum für Qualität. In: Pflege und digitale Technik. Berlin, Germany: Zentrum für Qualität in der Pflege; 2019.
75. Wolbring G, Yumakulov S. Social robots: views of staff of a disability service organization. Int J of Soc Robotics 2014 Mar 28;6(3):457-468. [doi: 10.1007/s12369-014-0229-z]
76. Steckelberg A, Haastert B, Hülfenhaus C, Mühlhauser I. Effekt einer evidenzbasierten Verbraucherinformation zur Entscheidungsfindung beim kolorektalen Screening. Gesundheitswesen 2015 Sep 3;77 Suppl 1(S 01):S93-S94. [doi: 10.1055/s-0032-1329999] [Medline: 23553186]
77. Schüßler I. Lernwirkungen Neuer Lernformen. Arbeitsgemeinschaft Betriebliche Weiterbildungsforschung. 2004. URL: https://abwf.de/content/main/publik/materialien/materialien55.pdf [accessed 2023-11-15]
78. Borutta M, Giesler C. Karriereverläufe von Frauen und Männern in der Altenpflege: Eine sozialpsychologische und systemtheoretische Analyse. Cham, Switzerland: Springer; 2006.
79. Hasseler M. Digitalization and new technologies in care – concepts and potentials for nursing care provision. In: Rubeis G, Hartmann KV, Primc N, editors. Digitalisierung der Pflege: Interdisziplinäre Perspektiven auf digitale Transformationen in der pflegerischen Praxis. Berlin, Germany: V&R unipress; 2019:1033-1016.
80. Weeks KW, Coben D, O'Neill D, Jones A, Weeks A, Brown M, et al. Developing and integrating nursing competence through authentic technology-enhanced clinical simulation education: Pedagogies for reconceptualising the theory-practice gap. Nurse Educ Pract 2019 May;37:29-38. [doi: 10.1016/j.nepr.2019.04.010] [Medline: 31060016]
81. Lincoln YS, Guba EG, Pilotta JJ. Naturalistic inquiry. Int J Intercult Relat 1985 Jan;9(4):438-439. [doi: 10.1016/0147-1767(85)90062-8]
82. Lamnek S. Qualitative Sozialforschung. Beltz Verlag. 2010. URL: https://tinyurl.com/5n6bx92a [accessed 2024-04-29]

XSL•FO
RenderX

83. Remmers H. Altern und Verletzlichkeit: Gero-Technologien als Bestandteil einer therapeutisch-rehabilitativen Dyade? In: Frommeld D, Scorna U, Haug S, Weber K, editors. Gute Technik für ein gutes Leben im Alter? Akzeptanz, Chancen und Herausforderungen altersgerechter Assistenzsysteme. Berlin, Germany: Transcript Verlag; 2019:129-158.
84. Hülsken-Giesler M, Remmers H. Robotische Systeme für die Pflege. Hamburg, Germany: V&R Unipress; 2019.

Abbreviations

ADL: activity of daily living
DAT: digital assistive technology
SEQI: sensitization, evaluative introduction, qualification, and implementation
STC: systematic text condensation
TAM: technology acceptance model
TUI: Technology Usage Inventory

Edited by B Lesselroth; submitted 01.11.23; peer-reviewed by S Hinder, A Kaunnil, K Trainum; comments to author 24.04.24; revised version received 30.04.24; accepted 15.08.24; published 09.10.24.

Please cite as:
Hofstetter S, Zilezinski M, Behr D, Kraft B, Buhtz C, Paulicke D, Wolf A, Klus C, Stoevesandt D, Schwarz K, Jahn P
Integrating Digital Assistive Technologies Into Care Processes: Mixed Methods Study
JMIR Med Educ 2024;10:e54083
URL: https://mededu.jmir.org/2024/1/e54083
doi:10.2196/54083
PMID:

©Sebastian Hofstetter, Max Zilezinski, Dominik Behr, Bernhard Kraft, Christian Buhtz, Denny Paulicke, Anja Wolf, Christina Klus, Dietrich Stoevesandt, Karsten Schwarz, Patrick Jahn. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.10.2024. 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.

XSL•FO
RenderX

Original Paper

Development and Implementation of a Safety Incident Report System for Health Care Discipline Students During Clinical Internships: Observational Study

Eva Gil-Hernández[1], MSc; Irene Carrillo[2], PhD; Mercedes Guilabert[2], PhD; Elena Bohomol[3], PhD; Piedad C Serpa[4], PhD; Vanessa Ribeiro Neves[3], PhD; Maria Maluenda Martínez[5], PhD; Jimmy Martin-Delgado[6,7], PhD; Clara Pérez-Esteve[1], MSc; César Fernández[2], PhD; José Joaquín Mira[1,2], PhD

[1]Fundación para el Fomento de la Investigación Sanitaria y Biomédica (FISABIO), Alicante, Spain

[2]Universidad Miguel Hernández, Elche, Spain

[3]Escola Paulista de Enfermagem, Universidade Federal de São Paulo, São Paulo, Brazil

[4]Clinical Management and Patient Safety Department, Universidad de Santander, Bucaramanga, Colombia

[5]Biomedical Sciences Faculty, Universidad Austral, Pilar, Argentina

[6]Instituto de Investigación e Innovación en Salud Integral, Facultad de Ciencias de la Salud, Universidad Católica de Santiago de Guayaquil, Guayaquil, Ecuador

[7]Hospital de Especialidades Alfredo Paulson, Junta de Beneficencia de Guayaquil, Guayaquil, Ecuador

Corresponding Author:
José Joaquín Mira, PhD
Universidad Miguel Hernández
Avenida Universidad s/n
Elche, 03202
Spain
Phone: 34 966658984
Email: jose.mira@umh.es

Abstract

Background: Patient safety is a fundamental aspect of health care practice across global health systems. Safe practices, which include incident reporting systems, have proven valuable in preventing the recurrence of safety incidents. However, the accessibility of this tool for health care discipline students is not consistent, limiting their acquisition of competencies. In addition, there is no tools to familiarize students with analyzing safety incidents. Gamification has emerged as an effective strategy in health care education.

Objective: This study aims to develop an incident reporting system tailored to the specific needs of health care discipline students, named Safety Incident Report System for Students. Secondary objectives included studying the performance of different groups of students in the use of the platform and training them on the correct procedures for reporting.

Methods: This was an observational study carried out in 3 phases. Phase 1 consisted of the development of the web-based platform and the incident registration form. For this purpose, systems already developed and in use in Spain were taken as a basis. During phase 2, a total of 223 students in medicine and nursing with clinical internships from universities in Argentina, Brazil, Colombia, Ecuador, and Spain received an introductory seminar and were given access to the platform. Phase 3 ran in parallel and involved evaluation and feedback of the reports received as well as the opportunity to submit the students' opinion on the process. Descriptive statistics were obtained to gain information about the incidents, and mean comparisons by groups were performed to analyze the scores obtained.

Results: The final form was divided into 9 sections and consisted of 48 questions that allowed for introducing data about the incident, its causes, and proposals for an improvement plan. The platform included a personal dashboard displaying submitted reports, average scores, progression, and score rankings. A total of 105 students participated, submitting 147 reports. Incidents were mainly reported in the hospital setting, with complications of care (87/346, 25.1%) and effects of medication or medical products (82/346, 23.7%) being predominant. The most repeated causes were related confusion, oversight, or distractions (49/147, 33.3%) and absence of process verification (44/147, 29.9%). Statistically significant differences were observed between the mean

XSL•FO
RenderX

final scores received by country (*P*<.001) and sex (*P*=.006) but not by studies (*P*=.47). Overall, participants rated the experience of using the Safety Incident Report System for Students positively.

Conclusions: This study presents an initial adaptation of reporting systems to suit the needs of students, introducing a guided and inspiring framework that has garnered positive acceptance among students. Through this endeavor, a pathway toward a safety culture within the faculty is established. A long-term follow-up would be desirable to check the real benefits of using the tool during education.

Trial Registration: Trial Registration: ClinicalTrials.gov NCT05350345; https://clinicaltrials.gov/study/NCT05350345

(JMIR Med Educ 2024;10:e56879) doi:10.2196/56879

KEYWORDS

reporting systems; education; medical; nursing; undergraduate; patient safety

Introduction

Background

Patient safety is an objective of health care practice in the health systems of all countries. However, the complexity and uncertainty that accompany health care makes this a practice not without risks. The World Health Organization leads the World Alliance for Patient Safety with the purpose of implementing safe practices and other actions with which to generate a safer environment in all health centers [1].

The information available regarding safety incidents focuses primarily on adverse events (AEs), which are incidents that result in harm to a patient. Slightly more than half of these AEs could have been prevented [2]. The results of research studies show that, in high-income countries, approximately 10% of patients admitted to hospitals experience an AE [3]. In primary and outpatient care, approximately 3% to 10% of patients experience an AE over the course of a year [4]. In 80% of cases, the damages are avoidable. In low- and middle-income countries, there are higher rates of AEs due to deficiencies in infrastructure, facilities, and accessibility [2]. So-called safe practices aim to reduce these figures and have proliferated across all countries [5]. Among them, incident reporting systems (IRSs) have emerged as a valuable tool to prevent safety incidents stemming from the same cause from recurring [2].

Studies indicate that up to 30% of students are involved in an AE during an academic year [6]. Moreover, during their internships, students observe decisions and procedures that may lead to errors or cause harm (AEs) to patients [7]. While access to IRSs is widespread in all health care systems, students of health care disciplines are often not adequately trained on how to use and benefit from these tools to create safer environments for patients. This lack of training restricts students' acquisition of crucial competencies in several ways.

The familiarization of students with incident reporting addresses a significant educational practice gap. First, the absence of IRS exposure hinders students' ability to understand what an incident report is, how to complete it, the extent of the information required, and how it functions to promote safer environments. This exposition to IRSs not only enhances their capability to effectively report incidents in future real-world contexts but also helps reduce the initial reluctance toward reporting. Second, reporting unsafe events can enhance practice and prevent future safety incidents. This active learning helps students identify and avoid recurring incidents by raising awareness of their causes. Third, providing students with access to IRSs raises awareness among future professionals of the critical importance of patient safety. It serves as a vital learning resource and offers an opportunity to change attitudes and foster the development of a proactive safety culture [8].

Despite this, the interventions designed and validated to achieve the goal of promoting incident reporting among health care discipline students are scarce [9]. There are also no tools to introduce these students to the analysis of the remote and immediate causes of safety incidents and the identification of barriers to prevent them from recurring. However, there are digital tools that are starting to be used to increase patient safety, particularly those based on gamification [10,11].

The effectiveness of gamification in health care education has been analyzed in several studies [12,13], showing improvements in knowledge, skills, satisfaction, behavior change, and attitudes compared to control groups. However, the usefulness of engaging health care discipline students in patient safety has not been assessed.

Objectives

The primary objective of this study was to develop a patient safety IRS tailored to the needs of health care discipline students. The secondary objectives were to study the performance of different groups of students in the use of the platform and train them on the correct procedures for reporting.

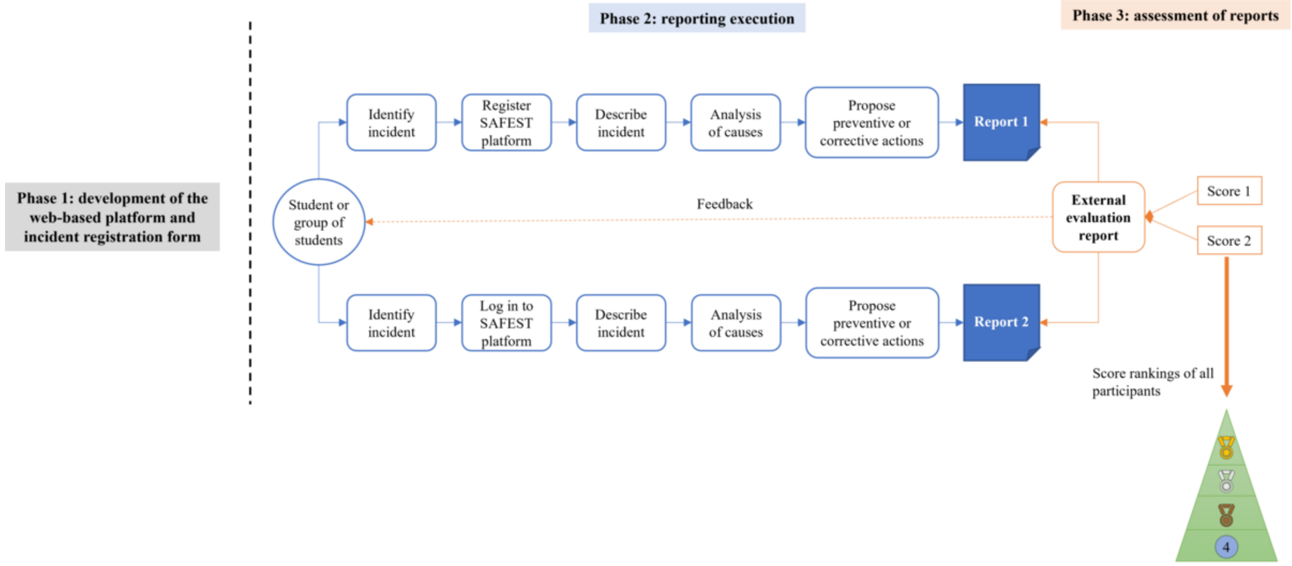
Methods

Study Design

This was an observational study developed in 3 phases (Figure 1), in which safety incident reports made by final-year students in medicine and nursing during their clinical internships were analyzed. The students were enrolled in universities from Argentina, Brazil, Colombia, Ecuador, and Spain once they had gained experience from their clinical placements. All these universities are members of the European Researchers' Network Working on Second Victims Consortium, with the Latin American ones as third-party or observer countries and Spain as the promoter of the network.

XSL•FO
RenderX

Figure 1. Workflow of the study divided into the 3 phases. SAFEST: Safety Incident Report System for Students.



In the participating countries, medical studies are typically completed over 6 years, with the last 3 years progressively incorporating more clinical practice. However, nursing studies exhibit greater variability and can range from 4 to 6 years in duration. In these programs, the final year is usually dedicated to clinical practice. Nursing curricula also show more diversity in their content, with some programs focusing more on hospital-based activities whereas others emphasize community health practice. Nonetheless, due to international guidelines on required competencies and clinical practice hours, these programs are standardized to ensure consistency in training.

In countries such as Argentina, Brazil, Chile, Ecuador, and Spain, medical and nursing programs follow this general structure but with some national variations. For example, in Spain, medical students undergo a rigorous 6-year program with a strong emphasis on clinical rotations in the later years. Nursing programs in Spain typically last 4 years, with the final year focused on intensive clinical practice. In Brazil and Argentina, similar patterns are observed, although the specifics of the curriculum and clinical exposure may differ slightly due to local health care needs and educational frameworks. As in other places, teaching patient safety is limited, representing one of the gaps highlighted in various studies [14].

This study is reported according to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies ([15]; see Multimedia Appendix 1).

Phase 1: Development of the Web-Based Platform and Incident Registration Form

This phase consisted of the development of the web-based platform and the incident registration form, named Safety Incident Report System for Students (SAFEST).

To design the content for the incident registration form, various existing systems at different levels were used as references. The existing systems in health care centers require users to be part of the center's staff, making them inaccessible to other groups, including students. In addition, while these systems collect the reporter's assessment of potential causes, they do not advance to propose alternatives for preventing future safety incidents. This aspect of our educational initiative is crucial to influencing students' attitudes toward safety reporting. Due to these limitations, the available safety systems in the health care facilities were not suitable for student practice and, therefore, deemed inappropriate for our purposes. Consequently, we decided to design a new system that closely mimics the systems that students will encounter in their professional practice. This approach ensures that students receive relevant and practical training, enhancing their ability to effectively report and analyze safety incidents in the future.

A database was then constructed incorporating fields gathered from the Patient Safety Reporting and Learning System of the Spanish National Health System [16]; the Adverse Event Reporting and Registration System of the Valencia Health Agency [17]; and Based on Root Cause Analysis (BACRA) [18,19], a web-based application based on root cause analysis and failure mode and effects analysis.

One of the key features of SAFEST is that each section and registration field (eg, center type, care complications, damage type, or care received after the incident) offered an extensive range of response options in different formats (single-select drop-down menu or multiple-choice answer). This design facilitated the reporting task for the students as they rarely needed to use natural language to describe a situation. This approach aligns with the latest advancements in reporting systems, minimizing errors in subsequent coding while providing a comprehensive catalog of options. However, in some cases in which the preset options may limit the recording, students can add a qualitative description to complement the recording. For example, when describing the incident, the student should characterize the event according to the classic typification of its nature, that is, whether the origin of the incident was related to complications of care, care-related infection, effects of medication or medical devices, complications of a procedure, or other situations not covered by the previous categories (eg, unexpected death of the patient). All these categories are detailed in a list of possibilities in a

XSL•FO
RenderX

multiple-choice format. However, in all categories (including "Other"), the student may choose a final option as "Other," in which case they should describe in words the situation in question. The first version of the database was created in Spanish.

From this database, common and specific aspects of each form were identified, and a preliminary draft of the proposal was developed accordingly. This draft underwent review by 3 subject matter experts from different Latin American countries, and the resulting changes and suggestions were incorporated to produce a high-quality form. This latest version of the tool was translated into English by EB and VRN, both of whom use the 2 languages regularly in the academic setting, ensuring the equivalence of the versions through back translation. The necessary modifications to ensure the adequacy of the system were made. Simultaneously, the visual identity and acronym for the platform were developed (Multimedia Appendix 2).

Phase 2: Introduction Seminar and Incident Reporting Execution

Overview

During this phase, students received an introductory seminar on patient safety and reporting and were given access to the platform. These introductory seminars were conducted by the responsible coordinators from the 5 universities (1 from each country; see Multimedia Appendix 3 for the educational materials used during the seminars). During these seminars, the project and the platform were presented, and attendees were given the opportunity to ask any questions they had.

The seminars contributed to the recruitment of participants based on voluntary participation without offering any academic grade advantages. To incentivize student engagement, they were provided with the opportunity to obtain a Miguel Hernández University nanocourse certificate. Moreover, the highest scores qualified for a draw with 4 new smartwatches as the prize, thus incorporating classic elements of gamification strategies. Of the smartwatches, 2 were assigned to the people with the highest and second-best scores and who had also submitted their feedback, whereas the other 2 were drawn among all reports with a score of >3.0 and who had also completed their feedback.

The specific instruction given to the students was to report any safety incidents that had occurred in their training health care center and of which they were aware, either because they had been involved or because they were witnesses. To introduce students to this exercise and standardize explanations and instructions on how to respond, concise use instructions were created along with video tutorials on navigating the website and submitting reports and a schematic diagram of the operation (Multimedia Appendix 4). The same presentation was used in all countries. In accordance with the academic calendars of the participating countries, the report submission period spanned from September 14, 2022, the day when the first seminar was held, to November 8, 2023.

Participants

Medicine (n=176) and nursing (n=47) students who had completed more than half of their educational program and were performing clinical internships were invited to participate. Recruitment was conducted by the professor in charge in each country with students in the corresponding academic years who met the selection criteria.

Study Size

According to existing literature, in pilot studies, if a problem exists with a 5% probability in a potential study participant, a sample size of 59 participants will almost certainly identify the problem with 95% confidence [20].

Phase 3: Assessment of Reports and Feedback on the Experience

Feedback

This phase ran in parallel to the previous one and involved external evaluation and feedback on the reports received that could prove useful for the students improvement in continuing to send reports. In total, 2 independent assessments were conducted for each incident report by members of the platform's promoting team. As a final exercise, students who had submitted at least 1 report were invited 1 month after this activity was over to fill out a satisfaction questionnaire.

Data Sources

The information provided in this study stems from the firsthand experiences of each student.

Variables

The outcomes we aimed to assess were the students' performance in reporting using the platform, which includes an estimation of potential causes to raise awareness of the inherent risks in health care activities, and their satisfaction with the experience.

To study their ability in reporting, a rubric (Table 1) was followed, in which the 2 reviewers independently rated the information provided about the incident, the analysis of immediate and latent causes, and the corrective or preventive plan proposed by the student using a scale of 1 to 5 points for each one, where the higher the score, the better the assessment. In addition, strengths and areas for improvement were included in the evaluation as an open-text field. The individual score from each evaluator was obtained by calculating the arithmetic mean of these 3 points. The final score for that report was the average of the 2 scores obtained from each evaluator.

XSL•FO
RenderX

Table 1. Rubric designed to assess the correctness of the reports made by students.

Points	Description
To what extent is the information complete and descriptive enough?	
2 points	The provided information allows for the understanding of the events.
2 points	The information is consistent throughout the entire report.
1 point	All fields are properly filled out.
To what extent is the analysis of immediate and latent causes complete and adequate?	
2 points	The provided information is comprehensive, and reasons with a high probability of influence are not overlooked.
2 points	The provided information offers sufficient details to envision the scenario of what happened.
1 point	The selected information is logical and does not appear to have been chosen merely for completion.
To what extent the corrective or preventive plan proposed is realistic and responds to the problem?	
1 point	The plan has a corrective or preventive nature.
1 point	The proposed plan is realistic.
1 point	The proposed plan is understandable.
1 point	Details are addressed to implement the proposed plan.
1 point	Language and spelling are appropriate.

To analyze their satisfaction with the experience, they were asked to complete a questionnaire with 3 aspects to rate on a scale of 1 to 5, with 1 being *not at all* and 5 being *very much*: “Do you believe that after this experience you would be capable of generating reports accurately?” (question 1), “Has viewing the assessments and comments you received on your reports been beneficial for your learning?” (question 2), and “Have you felt confident in terms of the privacy and anonymity of your reports?” (question 3). In addition, they had a text field available to input any suggestions that could contribute to improving the platform (question 4).

The independent variables used included the country from which the report was made, the sex of the reporter, their ongoing studies, and the number of internship hours completed up to the moment of reporting. All these data were incorporated into the incident registration form itself.

Bias

When the form was sent to the partners for review, a language check was also requested to address any idiomatic barriers that may have existed to allow for cross-cultural conclusions of the study and avoid possible biases related to linguistic nuances. Cultural differences were also considered, ensuring that items were comparable across countries.

Statistical Methods

To gain a comprehensive understanding of the reported incidents, descriptive analyses were conducted. To obtain the results of the phase of assessment, various statistical analyses were conducted. Descriptive statistics were computed for each of the 3 dimensions under analysis as well as for the overall score, with stratification by country, sex, and educational background. The weighted Cohen κ was computed to evaluate the agreement among the scores assigned by different pairs of evaluators for each dimension. Before proceeding with the analysis of the final scores of each report, the normality of the sample was assessed using *Q-Q* plots and the Shapiro-Wilk normality test. The relationship between the number of internship hours and the final score was examined using the Spearman correlation coefficient. Finally, differences in scores among countries, sexes, and educational backgrounds were investigated using the nonparametric Kruskal-Wallis test and the Mann-Whitney *U* test. The *P* value significance was set at .05. Data analyses were performed using SPSS Statistics (version 28.0.0; IBM Corp).

Ethical Considerations

This study was authorized by the Research Ethics Committee of Sant Joan d’Alacant University Hospital (22/027) and registered on ClinicalTrials.gov (NCT05350345).

Informed consent for study participation was obtained at the time of registration on the platform, whereby individuals were required to select the corresponding checkbox with instructions provided regarding the process for revoking their participation. After reporting, the report was automatically encoded with a numerical identifier by the platform. Throughout the assessment process, participant sociodemographic data were concealed to ensure evaluator objectivity.

No form of financial compensation was provided for participation or recruitment.

Results

Phase 1: Development of the Web-Based Platform and Incident Registration Form

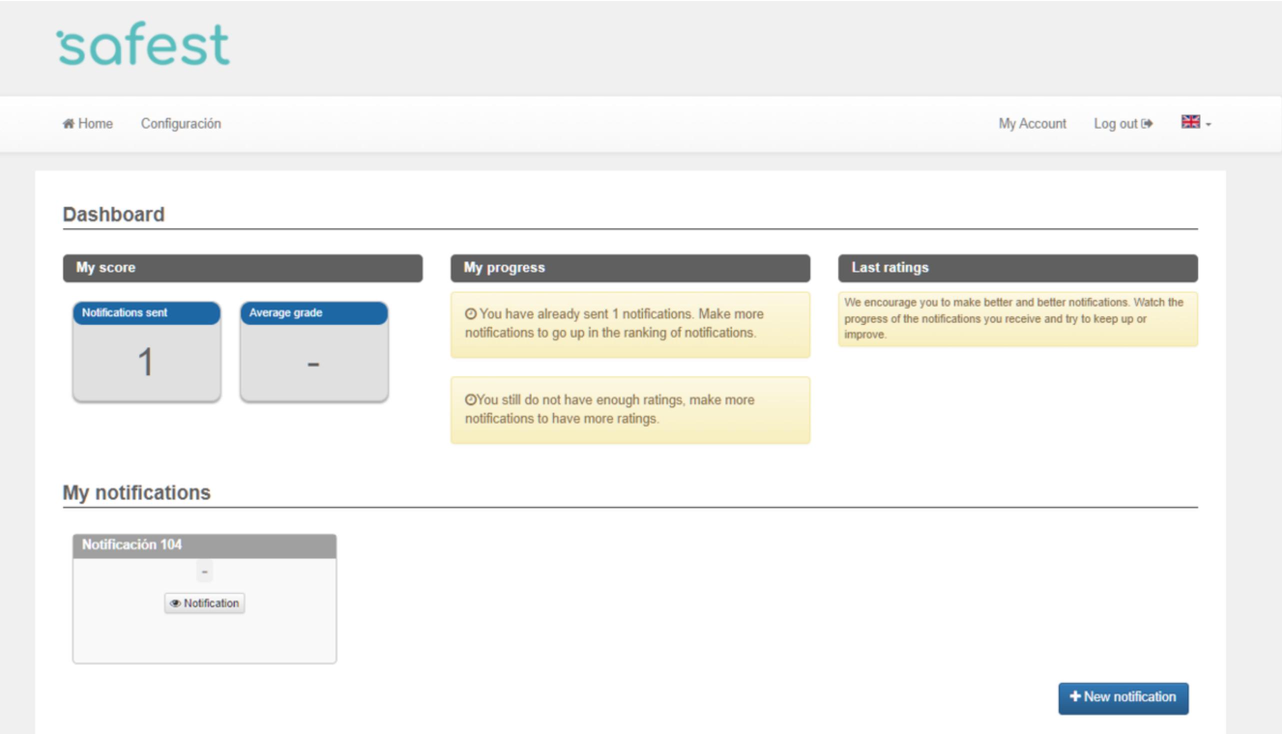
SAFEST [21] and the servers were located in Miguel Hernández of Elche University (Spain). Participation was allowed both individually and in groups of 2 to 3 students.

When accessing the page, users could find an explanatory text about the project, logos of collaborators, and buttons to access the platform or register. Upon initial access, the user was

XSL•FO
RenderX

required to provide consent to participate in the study. The platform was available in both Spanish and English. Once logged in, the dashboard was shown (Figure 2), where they could view the total of submitted reports, their average score, and the progression of their results, as well as their position in the score ranking at any time. In addition, they had access to previously submitted reports, as well as the button to access the incident registration form.

Figure 2. Appearance of the Safety Incident Report System for Students (SAFEST) platform dashboard.



The final form was divided into 9 sections: data of the reporting center, patient data, notifier data, incident data, description of the incident, damage assessment, factors influencing the incident, care received after the incident, and reflections. In total, it consisted of 48 questions (distributed as depicted in Textbox 1) that allowed for obtaining the necessary information about the incident, conducting an analysis of the causes, and proposing corrective or preventive actions. The complete form can be found in Multimedia Appendix 5.

XSL•FO
RenderX

Textbox 1. Questions asked on the reporting form and types of responses.

Data of the reporting center

- Center type (drop-down menu)

Patient data

- Patient's age (drop-down menu)
- Patient's sex (drop-down menu)
- Patient's risk factors (multiple choice)

Notifier data

- Notifier's sex (closed-ended question)
- Country from which the notification was made (drop-down menu)
- Studies in the course (drop-down menu)
- Institution (drop-down menu)
- Year (closed-ended question)
- Internship hours carried out so far in that department (open-ended question)

Incident data

- Date of the incident (date)
- Time of the incident (time)
- Date of the notification (date)
- Time of the notification (time)
- Where it took place (drop-down menu)
- Number of people related to the incident (open-ended question)
- Position or positions of the person or people involved (multiple choice)
- Frequency or probability of recurrence (drop-down menu)
- Participation in the incident (drop-down menu)

Description of the incident

- Care complications (multiple choice)
- Care-related infection (multiple choice)
- Effects of medication or medical products (multiple choice)
- Complications of a procedure (multiple choice)
- Other (multiple choice)

Damage assessment

- Damage type (drop-down menu)
- Severity (drop-down menu)
- Patient autonomy (drop-down menu)
- Estimation of the damage duration (drop-down menu)

Factors that conditioned the incident

- Patient or family factors (multiple choice)
- Equipment and resource factors (multiple choice)
- Individual factors of the health care professional or professionals (multiple choice)
- Work environment factors (multiple choice)
- Oral and written communication between professionals factors (multiple choice)

XSL•FO
RenderX

- Patient communication factors (multiple choice)
- Teamwork and leadership factors (multiple choice)
- Task-related factors (multiple choice)
- Organizational and management factors (multiple choice)
- Other factors (open-ended question)

Care received after the incident

- Care received after the incident (multiple choice)

Reflections

- Has the center been notified? (drop-down menu)
- Could the incident have been prevented? (drop-down menu)
- How could it have been prevented? (open-ended question)
- How could the probability of occurrence or the severity of this event be reduced? (open-ended question)
- To what extent was all the information necessary to analyze the causes of the event available? (open-ended question)
- Have measures been put in place to prevent it from happening in the future? (drop-down menu)
- What measures have been put in place to prevent it from happening in the future? (open-ended question)
- Do you consider that the analysis could have been different if you had had access to another source of information? (open-ended question)
- Write here any other comments you may have (open-ended question)

To streamline the use of the system, selection questions and drop-down menus were used to report incidents. Both the *Description of the incident* and *Factors that conditioned the incident* blocks allowed for more than one option to be selected. Written input was only necessary in the *Reflections* part.

Phase 2: Introduction Seminar and Incident Reporting Execution

A total of 105 students from the 5 countries participated voluntarily and actively by submitting at least 1 report (participation rate: 105/223, 47.1%). By country, this corresponds to 16.2% (17/105) of students from Argentina, 12.4% (13/105) of students from Brazil, 10.5% (11/105) of students from Colombia, 32.4% (34/105) of students from Ecuador, and 28.6% (30/105) of students from Spain. Of the 105 participants, 35 (33.3%) were male, 68 (64.8%) were female, and 2 (1.9%) specified their sex as *other*. Only 1.9% (2/105) of them formed a team. Regarding their studies, 66.7% (70/105) were pursuing a degree in medicine, 28.6% (30/105) were enrolled in nursing studies, 1.9% (2/105) were part of the pediatric specialization program, 1.9% (2/105) were students from the radiology and diagnostic imaging specialization program, and 1% (1/105) belonged to the orthopedics and traumatology specialization program.

A total of 147 reports were submitted as 14 users provided >1 report. Of the 147 received reports, 18 (12.2%) were from Argentina, 13 (8.8%) were from Brazil, 44 (29.9%) were from Colombia, 35 (23.8%) were from Ecuador, and 37 (25.2%) were from Spain.

Of the 147 safety incident reports, a substantial majority, specifically, 144 (98%) reports, occurred in a health care setting, with most occurring in a hospital context (n=132, 89.8%). Within this hospital-centric subset, most incidents were concentrated in hospitalization units (45/132, 34.1%). Other noteworthy locations included surgical block areas (21/132, 15.9%), emergency departments (17/132, 12.9%), support services (14/132, 10.6%), day hospitals (12/132, 9.1%), and intensive care units (10/132, 7.6%).

Regarding the nature of the incidents, Table 2 illustrates the frequency with which each major classification category was selected. On most occasions, events from different blocks were registered in the same report.

Table 2. Nature of the reported safety incidents (n=346).

Type of incident	Reports, n (%)
Care complications	87 (25.1)
Effects of medication or medical products	82 (23.7)
Complications of a procedure	67 (19.4)
Other	60 (17.3)
Care-related infection	50 (14.5)

XSL•FO
RenderX

Specifically, from the available list of the most common safety events included in SAFEST (drop-down list), students' reports were related to "Worse evolutionary course of the main pathology" (27/147, 18.4%), "No harm" (24/147, 16.3%), "Ineffective analgesia-related pain" (19/147, 12.9%), "Falls and consequent fractures" (18/147, 12.2%), "Surgical site or traumatic wound infection" (15/147, 10.2%), "Contusion" (14/147, 9.5%), "Unexpected death" (14/147, 9.5%), "Headache" (13/147, 8.8%), and "Prescription error" (13/147, 8.8%).

The reported causes are shown in Table 3. According to the number of reports in which they appear, we established the following categories: "Patient or family factors" (112/147, 76.2% of reports), "Equipment and resource factors" (71/147, 48.3% of reports), "Individual factors of the healthcare professional(s)" (118/147, 80.3% of reports), "Work environment factors" (103/147, 70.1% of reports), "Oral and written communication between professionals factors" (76/147, 51.7% of reports), "Patient communication factors" (64/147, 43.5% of reports), "Teamwork and leadership factors" (93/147, 63.3% of reports), "Task-related factors" (91/147, 61.9% of reports), and "Organizational and management factors" (81/147, 55.1% of reports).

XSL•FO
RenderX

Table 3. Causes and contributing factors of the incidents grouped by category (n=147).

Factor	Reports, n (%)
Patient or family factors	
Comorbidity or complexity of the condition	33 (22.4)
Low economic level	30 (20.4)
Noncooperative attitude (noncompliance)	35 (23.8)
Lack of family or support networks	40 (27.2)
Poor communication with relatives	18 (12.2)
Altered cognitive status	8 (5.4)
Does not provide correct or enough information	14 (9.5)
Recent surgery	15 (10.2)
Educational and social factors to consider	20 (13.6)
Other patient factors	18 (12.2)
Mental disorder	3 (2)
Equipment and resource factors	
Improper storage or accessibility	19 (12.9)
Malfunctions	14 (9.5)
Equipment maintenance issues	12 (8.2)
Lack of alternative materials	11 (7.5)
Incorrect labeling	11 (7.5)
Equipment deficit (including nonsterile material)	10 (6.8)
Inadequate resource design (eg, bell)	10 (6.8)
Product or drug unavailability	10 (6.8)
Improper calibration	9 (6.1)
Nonstandard equipment	7 (4.8)
New equipment or resource	5 (3.4)
Failure to access or unavailability of the digital medical record	4 (2.7)
Expiration	2 (1.4)
Similar container or name	2 (1.4)
Individual factors of the health care professional or professionals	
Confusion, oversight, or distractions	49 (33.3)
Overload or work pressure	37 (25.2)
Lack of knowledge of regulations or protocols of performance	25 (17)
Uncooperative attitude	24 (16.3)
Inadequate or insufficient anamnesis, examination, or tests	23 (15.6)
Medication error (prescription or dispensing)	21 (14.3)
Inadequate or insufficient knowledge or skills	20 (13.6)
Low motivation	15 (10.2)
Diagnostic error	13 (8.8)
Inadequate or insufficient training	13 (8.8)
Not verifying the treatment that the patient is currently taking	12 (8.2)
Little experience in the workplace	10 (6.8)
Inadequate timetable	8 (5.4)
Inappropriate interpretation of analytical or test results	5 (3.4)

XSL•FO
RenderX

Factor	Reports, n (%)
Work environment factors	
Distractions in the environment	38 (25.9)
Shift-related fatigue	36 (24.5)
High care pressure	29 (19.7)
Inadequate environment—cleaning, beds, or space	26 (17.7)
Inadequate environment—noise, light, or temperature	19 (12.9)
Performance of outside tasks	14 (9.5)
Excessive staff turnover or inexperience	12 (8.2)
Inadequate staff-to-patient ratio	11 (7.5)
Security and access to restricted areas	2 (1.4)
Oral and written communication between professionals factors	
The information does not reach the entire team	30 (20.4)
Ambiguous verbal indications	23 (15.6)
Insufficient or inadequate records	20 (13.6)
Using an inappropriate channel	19 (12.9)
Inappropriate body language	15 (10.2)
Incorrect use of language	11 (7.5)
Patient communication factors	
Insufficient or inadequate records	23 (15.6)
Ambiguous verbal indications	17 (11.6)
Using an inappropriate channel	16 (10.9)
Incorrect use of language	14 (9.5)
Inappropriate body language	12 (8.2)
Language barrier	6 (4.1)
Teamwork and leadership factors	
Lack of coordination in the team	40 (27.2)
Inadequate supervision	40 (27.2)
Low risk awareness	31 (21.1)
Inaccurate assignment of tasks	20 (13.6)
Conflict between team members	12 (8.2)
No effective leadership	9 (6.1)
Task-related factors	
Absence of process verification	44 (29.9)
Unknown protocol or noncompliance	30 (20.4)
Absence of guidelines or protocols	19 (12.9)
Inadequate or outdated protocol	18 (12.2)
Too complex task	7 (4.8)
Organizational and management factors	
Absence of evaluation systems	16 (10.9)
Error in health information	16 (10.9)
Nonexistent or inadequate risk management	16 (10.9)
Error in medical documentation	12 (8.2)
Insufficient organizational structure	12 (8.2)

XSL•FO
RenderX

Factor	Reports, n (%)
Absence of support mechanisms in a risk situation	11 (7.5)
Incorrect patient identification	11 (7.5)
Insufficient deployment of a proactive security culture	11 (7.5)
Delays in the performance of tests or interconsultations	9 (6.1)
Insufficient care structure	9 (6.1)
Wrong appointment or scheduling	7 (4.8)
Gaps or failures in the information system	6 (4.1)
Inadequate or nonexistent treatment plan	5 (3.4)
Long waiting list	5 (3.4)

When asked about whether the event had been reported at the center, in 41.5% (61/147) of the reports the answer was “Yes”; in 34.7% (51/147) of the reports, the answer was “I don’t know”; and, in 23.8% (35/147) of the reports, the answer was “No.” Finally, 93.9% (138/147) of the reported events were classified as preventable compared to 6.1% (9/147) that were categorized as nonpreventable.

Phase 3: Assessment of Reports and Feedback on the Experience

Considering the 147 reports received, the mean final score obtained was 3.40 (SD 0.92) out of 5, and 111 (75.5%) reports had a final score of ≥3.0. For each of the 3 aspects studied, an average score of 3.38 (SD 1.29) was obtained for the section on giving information about the incident, an average score of 3.54 (SD 1.21) was obtained for the analysis of causes, and an average score of 3.30 (SD 1.30) was obtained for the proposal of a corrective or preventive plan. Table 4 shows the means of the final scores segregated by category.

Table 4. Mean final scores segregated by country, sex, and studies.

Variables	Values, mean (SD)
Country	
Argentina	3.66 (0.89)
Brazil	3.65 (0.75)
Colombia	3.28 (0.88)
Ecuador	2.81 (0.88)
Spain	3.89 (0.73)
Sex	
Male	3.06 (0.99)
Female	3.64 (0.84)
Team	3.46 (0.60)
Other	3.58 (0.42)
Studies	
Medicine	3.34 (0.95)
Nursing	3.66 (0.82)
Pediatric specialization	3.25 (0.92)
Radiology and diagnostic imaging specialization	3.55 (0.59)
Orthopedics and traumatology specialization	2.75 (1.30)

Significant differences were found in the final scores based on country (P<.001) and sex (P=.006). However, no significant results were obtained when comparing scores based on studies (P=.47). Similarly, when focusing on the 2 main groups (medicine and nursing), there were no significant differences (P=.11). Comparisons by groups for the significant variables are presented in Tables 5 and 6.

XSL•FO
RenderX

Table 5. *P* values for final score mean comparisons (country).

Country	Argentina	Brazil	Colombia	Ecuador	Spain
Argentina	—[a]	.92	.15	.004	.40
Brazil	.92	—	.19	.004	.36
Colombia	.15	.19	—	.01	.003
Ecuador	.004	.004	.01	—	<.001
Spain	.40	.36	.003	<.001	—

[a]Not applicable.

Table 6. *P* values for final score mean comparisons (sex).

Sex	Male	Female	Other	Team
Male	—[a]	.001	.44	.25
Female	.001	—	.82	.33
Other	.44	.82	—	.69
Team	.25	.33	.69	—

[a]Not applicable.

Regarding the internship hours carried out in that department, no correlation was found with the score obtained on each report (–0.079; *P*=.18). Finally, the interrater agreement analyses revealed consistency between each pair of evaluators across all cases (Table 7).

Table 7. Weighted Cohen κ values obtained for each pair of evaluators.

	Pair 1		Pair 2	
	Cohen κ	*P* value	Cohen κ	*P* value
Complete and descriptive information	0.324	<.001	0.304	<.001
Analysis of immediate and latent causes	0.420	<.001	0.195	.009
Corrective or preventive plan	0.344	<.001	0.258	<.001

A total of 15 students participated in discussing the experience through the satisfaction questionnaire, providing an average rating of 4.06 (SD 1.00) for question 1, an average rating of 4.18 (SD 1.22) for question 2, and an average rating of 4.56 (SD 1.09) for question 3. Moreover, they provided the following feedback in the improvement suggestion section: "The platform has been very useful to me, and I suggest that similar projects continue to be conducted virtually to encourage widespread participation," "It would be beneficial if the platform allows the upload of images as evidence for each incident," and "Consider incorporating a text comment box, allowing individuals involved in the incident to narrate the events rather than solely selecting an option. This would prevent overlooking crucial details that might be of interest for subsequent management and error prevention."

Discussion

Principal Findings

A platform named SAFEST was developed, allowing students to submit reports and receive feedback regarding provided information, causal analysis, and proposed improvements. The educational practice simulates the environment they will encounter in their professional practice regarding the reporting of safety incidents. Students from 5 countries participated in this initiative, sending reports in which most incidents occurred within hospital settings and involved complications related to care and medication.

The identified causes of safety events reported by students using SAFEST included confusion, oversight, or distractions and absence of process verification. Across all countries, the average score exceeded 2.5, although significant differences in average scores among some countries are observed. Overall, the experience was highly regarded.

This paper delves into a comprehensive portrayal of the SAFEST platform, focusing on its inception, development, and implementation. The SAFEST platform stands as a pivotal reporting system strategically crafted to initiate health care discipline students into the realm of identifying and reporting current safety lapses within health care environments. In addition, this paper describes the perception that medical and nursing students had regarding incidents impacting patient safety, their attributed causes, and potential preventive or corrective measures. This reflection on what they identified as incidents can provide professors with feedback for planning their teaching.

XSL•FO
RenderX

The platform and designed materials allowed medical and nursing students to be introduced to safety incident reporting. SAFEST recreated the natural context in which reporting occurs, providing students with an experience close to reality but facilitating the process by allowing for step-by-step guided reporting. The IRS form follows the same structure and covers the same fields as those available for health care professionals as it was developed based on 3 existing systems. However, there are differences in terms of approach or responsibility. Students report situations that they observe or are involved in but are not the authors of. In addition, they encounter disparities regarding the accessibility of information for cause analysis as they do not have full access to the patient's medical history. The gathered information is not disseminated, nor does it bring consequences for the involved parties. This type of active learning can also facilitate a better understanding of the impact of safety incidents and their causes. By doing this, students can grasp the basic concepts of fostering a proactive safety culture for patients. Once in clinical settings as professionals, they will be able to overcome the natural barrier hindering reporting through their participation in SAFEST. They will also have gained experience in analyzing both the immediate and remote causes of safety incidents and identifying preventive or corrective measures. These types of educational interventions prompt reflection on how errors occur in clinical practice, aiding in distinguishing between honest mistakes and intentional errors. If specific patient safety content is taught during regular classes while conducting the reporting practice, one can expect a greater impact of this practice on instrumental and attitudinal competencies. This aspect should be verified in the future.

Notably, the existing literature predominantly reflects studies conducted within the field of nursing [7,8,22], leaving a noteworthy void in the examination of reporting mechanisms across broader health sciences education. Breaking away from this convention, our research introduces a groundbreaking element not only by incorporating medical students into its purview but also by providing a system that can be extended to other disciplines that develop their practices in the clinical field.

Moreover, the provision of feedback helps students learn and improve their skills. Similarly, the integration of a gamified environment adds an element of engagement and motivation to the learning process. By incorporating elements of game design such as rewards and progression systems, the learning experience is transformed into a user-friendly practice.

Reporting systems constitute one of the fundamental tools for creating increasingly safe environments for patients [2]. It has been demonstrated that they also have a positive impact on the safety culture within health care institutions. However, students in health care disciplines typically become familiar with this tool once they are in health care settings either as residents or professionals. Simulation, as portrayed in this case, stands as one of the most used approaches in teaching-learning methods [23]. The approach of this exercise ensures active student engagement in reporting. The feedback provided to the students facilitated the enhancement of their proficiency and enabled them to report accurately. This aspect has been highlighted as significant in other studies [24].

The data from this study suggest that introducing a practice on how to report and why it is important was well received by the participants in this academic exercise, resulting in reports of suitable quality. Previous studies [25] have suggested that students demonstrate enhanced proficiency in detecting and analyzing incidents when they are not involved in them. Therefore, incorporating a reporting exercise during their internship period would contribute to cultivating a patient safety culture among students. This approach facilitates experiential learning, enabling students to comprehend the intricacies of incidents, empowering them to identify and mitigate such occurrences in their future professional endeavors.

The incident reporting by students has 2 strengths: the firsthand experience in clinical risk management within a health care institution and the provision of specific information that can contribute to enhancing comprehensive patient safety education among students. When delving into the results obtained in terms of scores, we found congruence with the results of other studies [25] in that the analysis of causes emerged as the strongest aspect, whereas the proposal of an improvement plan proved to be the weakest. This is particularly evident in cases in which patient safety content was integrated into the curriculum, where greater familiarity with patient safety was correlated with higher-quality reporting. The variations in the scores obtained can be explained by the curricular differences between each country. The Argentinean university involved offers 2 subjects on patient safety during the 5 years of the degree. It also has a patient safety program in which theoretical, simulation, and practical modules on patient safety (international goals and risk management) are offered so that students receive training throughout their degree, from first to fifth year, in all subjects that involve field practice. In contrast, the Ecuadorian university involved lacks any specific courses on the subject during the 6 years. Spanish students receive specific lectures on patient safety in 4 subjects starting in the second year before entering the internship in the sixth year. One of these subjects also incorporates specific topics on AEs and their communication. In the Colombian university involved, there is no specific subject in the curriculum dedicated to this matter in the first 5 years of study. However, before engaging in clinical internships in the final year, students are required to complete a course on clinical management and health, which delves into introductory topics related to patient safety. In the case of the Brazilian university, the term "patient safety" is explicitly referenced in the curriculum of 6 subjects, spread out from the second to the fourth year of studies. Moreover, in another 17 subjects, while there may not be an explicit mention, faculty members address the subject matter throughout the duration of the academic term.

Similarly, female students achieved higher scores in the evaluation of their reports. Nevertheless, there is no existing literature to substantiate this observation, prompting the need to consider the influence of other factors that could account for it. In our case, these outcomes might be influenced by the sample distribution as reports submitted by male students were predominantly concentrated in Colombia (23/147, 15.6%) and Ecuador (21/147, 14.3%), the 2 countries exhibiting the lowest mean scores. Therefore, these differences might not be explained solely by sex but rather by the background in patient safety.

XSL•FO
RenderX

In this study, 47.1% (105/223) of students participated submitting at least one report. This percentage contrasted with the findings of other studies, which reported participation rates of approximately 12% [26]. This increase opens the door to a more in-depth exploration of the factors that may be influencing this elevated level of student participation. One potential line of inquiry focuses on student motivation and how it may be linked to the design or implementation of the reporting system. Examining the effectiveness of strategies used to encourage participation could shed light on the dynamics that lead to more active engagement by students in this particular context.

It is not surprising that many incidents took place in hospitals. First, students from both disciplines undertook most of their practical training in this environment, and this clinical exposure increases the likelihood of witnessing or being involved in safety incidents. Moreover, hospitals typically handle more complex and critical cases compared to other health care settings as well as conducting a greater quantity and variety of procedures. However, this figure may be influenced by the students' risk perception. It is plausible that primary care settings are perceived as less prone to safety incidents, leading students to pay less attention to their surroundings in such environments. In analogous studies, the most frequently reported type of incident was associated with medication administration [7]. However, in our case, what emerged most frequently throughout the reports were incidents related to caregiving. This outcome is likely related to the information more readily accessible to students, explaining why they witness fewer medication errors than expected during their practice [22].

During their clinical placements in health care settings, students frequently witness safety incidents of different severities, triggering conflicting emotions—from fear of speaking up to guilt for remaining silent. Studies suggest that approximately 4 out of every 10 students in training admit to having made at least one medical error during their training period [27]. Most of these errors involve lapses in clinical judgment (7 out of 10 cases). The primary causes of these errors have been associated with deficiencies in supervision and in the students' own technical competencies [28]. They often feel that the causes of these events are not adequately addressed. Once the practice session ends, they are left without information on whether the incident was reported, whether its causes were analyzed, or whether any subsequent actions were taken, all of which they might be unaware of. The attitudes and coping strategies of nursing students following the recognition of a medical error have been explored [29-31]. On the basis of our understanding, students who become implicated in an AE or near-miss situation tend to manifest symptoms aligned with the experience of second victims [27,32]. Familiarizing themselves with reporting and analyzing incident causes offers them a new perspective that we can also expect to aid them emotionally.

Finally, following the suggestions provided by the students and, thus, incorporating user-centered design principles, we found it highly beneficial to incorporate a text box for a brief narrative of the events. We believe that the optimal approach would involve presenting a comprehensive set of options to encourage reflection, prompting individuals to consider aspects they might not have otherwise. Subsequently, a field will be provided for participants to describe the unfolding of events in their own words. This aligns with the findings of King et al [33], who advocate for a balanced approach in future patient reporting systems, integrating closed-ended questions for cause analysis and classification alongside open-ended narratives to accommodate patients' potential limitations in understanding terminology.

Implications of Findings

By providing a guided process, students are aided in considering a variety of factors that could pose potential risks, ranging from material resource deficiencies to patient attitudes or workload overload. Moreover, they learn to analyze different variables, weigh consequences, and make informed decisions based on available information. Consequently, students acquire skills and experience that they are expected to be able to apply in similar situations in the future. Similarly, by increasing awareness of risks and sources of mistakes and empowering students to identify them, the likelihood of involvement in dangerous or problematic situations is expected to be reduced [34], thereby contributing to the creation of safer environments.

The apprehension surrounding potential negative outcomes of reporting has been present since the initial implementation of reporting systems in Australia in 1993 [35]. Introducing students from health-related disciplines to the reporting process, emphasizing the understanding of why, how, and for what purpose they should contribute, aims to foster a safety culture among the forthcoming generations of health care professionals. Encouraging students to view errors as valuable learning opportunities rather than indicators of incompetence is highly necessary. Embracing mistakes as integral components of the learning process can foster a growth mindset where challenges become stepping stones to improvement. This positive approach not only cultivates resilience but also promotes a more constructive and proactive attitude toward learning.

Since digital systems offer a more enduring record-keeping mechanism and facilitate a higher volume of reports than their paper counterparts [9], approaches such as this one can increase the correctness and impact on the future rate of reporting. In addition, this educational practice should help overcome the initial reluctance that discourages reporting safety incidents. To know and have used an incident reporting tool, describing a safety incident and reflecting on its potential causes and the measures that could actively and thoughtfully prevent it, should have an impact on attitudes toward reporting [36,37].

Future Research

Several scales have been developed to assess students' knowledge and the information they receive, aiming to model their safety culture. Among these, the scales proposed by Flin et al [29] and Mira et al [38] are remarkable. However, we need to identify which mechanisms are most effective in integrating curriculum content that matches the students' knowledge levels and attitudes, fostering a cross-disciplinary education in patient safety.

In addition, although the students scored 4.0 out of 5 regarding the fact that after this experience, they would be capable of making reports properly, a follow-up over time is required to

XSL•FO
RenderX

really verify the benefits brought by this experience. Furthermore, it would be interesting to consider the use of this tool in students of earlier courses, provided they undergo some period of their training in clinical settings, to analyze its utility in earlier stages of education. This would also facilitate the development of longitudinal studies to monitor the impact in terms of reporting.

Limitations

The aim of this experience was not to detect the safety incidents themselves but rather to train students to make correct reports in their future professional practice. Thus, the frequencies and features described in this paper did not necessarily represent the actual safety incidents occurring and what students could witness in their countries.

Recognizing an error is not straightforward. Students in training may consider it risky for their future to report an incident, leading to a restriction in the information they provide to the system. If they end up working in an environment where psychological safety is at risk, despite actively participating in this educational practice, they might choose silence, and fear of potential negative consequences could undo what was gained from this practice. The same can happen with other organizational factors that may hinder and make reporting difficult for the group of professionals in a center. This practice does not prevent this from happening in some contexts.

It cannot be guaranteed that the reports accurately reflect incidents that actually occurred. A convenience sample was used, which restricts the generalizability of the results. The medical and nursing curricula in the different participating countries were not identical. Although participation in the study was offered in the context of subjects related to patient safety, it was not possible to control for students' baseline knowledge of incident reporting. These differences may have influenced the quality of the reports. It will be necessary to delve into the safety culture in the course of subjects with patient safety content in training programs. The constant technological evolution requires timely updating of the proposal, adapting it to possible technological solutions. Student involvement should be facilitated by the participation of academics in the project. However, the project schedule may be affected by the academic obligations of this group (eg, exams, vacations, and internship periods).

With this exercise, students become familiar with a fundamental tool in patient safety that they will encounter at the beginning of their professional careers and often approach with some hesitation, particularly in the countries where the study was conducted. However, the reports are based on observations made during their placements, and the analysis of the proposed improvement plan was conducted without accessing all the clinical information necessary for a precise analysis of root and immediate causes. In this case, the remote causes could not be determined during the exercise.

The sample size and the study's cross-sectional nature did not allow for assessing the impact of evaluators' feedback on students' learning and the quality of their subsequent reports. In the future, longitudinal studies with repeated measures over time would make it possible to establish the effect of feedback.

Finally, we would have liked to establish a user-centered platform from the outset. However, due to the lack of previous information from students regarding the subject matter, it was not feasible to conduct a consultation to determine which elements to consider. We have endeavored to compensate for this by incorporating the feedback provided subsequently.

Conclusions

In Europe, only a handful of medical or nursing schools have incorporated curriculum plans addressing patient safety [14]. Studies examining the nature of patient safety training received by students in health care disciplines are limited [39,40]. Faculties and schools might consider these reflections and data, incorporating reporting as a practical exercise into their curriculum. This study presents an initial adaptation of reporting systems to suit the needs of students, introducing a guided and inspiring framework that has garnered positive acceptance and evaluation among students. Through this endeavor, a pathway toward a safety culture within the faculty is established.

Acknowledgments

The students participated voluntarily, making this educational practice possible. The authors disclose receipt of the following financial support for the research, authorship, and publication of this paper. This work was supported by the Grants for the Development of Research Projects for Consolidated Groups 2021 provided by the Foundation for the Promotion of Health and Biomedical Research of the Valencia Community with reference UGP-21-215. Throughout the execution of this study and composition of this manuscript, JJM benefited from an augmented research activity contract granted by the Carlos III Health Institute (reference INT22/00012). In addition, during the execution of this study and composition of this manuscript, EG-H received funding through a Predoctoral Fellowship for Research Training in Health from the Carlos III Health Institute supported by the European Union NextGenerationEU and the Recovery, Transformation, and Resilience Plan from the Spanish Government (reference FI22/00277).

Data Availability

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

XSL•FO
RenderX

Authors' Contributions

All the authors meet the International Committee of Medical Journal Editors criteria for authorship. JJM, IC, MG, and CF were responsible for the design of the study. IC and EGH were responsible for obtaining ethics approval and registration in ClinicalTrials.gov. IC and MG conducted the review and compilation of the existing tools, and JJM and EGH drafted the first version of the form. PCS, JMD, and MMM reviewed the form and made the necessary clarifications. EB and VRN translated the contents into English, and finally, all authors approved the final form. Regarding the platform, CF developed the acronym and visual identity; EGH, IC, MG, and VRN prepared the instructional and introductory materials; and JJM, PCS, JMD, MMM, and EB were in charge of the recruitment and introductory seminar in each country. Finally, EGH, IC, and MG carried out the evaluation and feedback of the reports, and CPE performed the statistical analysis and interpretation of the results. JJM and EGH developed the first version of the manuscript. All authors revised the paper critically for important intellectual content and read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist for cross-sectional studies.
[PDF File (Adobe PDF File), 210 KB - mededu_v10i1e56879_app1.pdf]

Multimedia Appendix 2
Acronym and visual identity developed for the platform.
[PDF File (Adobe PDF File), 5499 KB - mededu_v10i1e56879_app2.pdf]

Multimedia Appendix 3
Educational materials used during the seminars.
[PDF File (Adobe PDF File), 905 KB - mededu_v10i1e56879_app3.pdf]

Multimedia Appendix 4
Schematic diagram of the Safety Incident Report System for Students platform operation.
[PDF File (Adobe PDF File), 81 KB - mededu_v10i1e56879_app4.pdf]

Multimedia Appendix 5
Safety Incident Report System for Students reporting form.
[DOCX File , 98 KB - mededu_v10i1e56879_app5.docx]

References

1. World alliance for patient safety: forward programme 2005. World Health Organization. 2004. URL: https://iris.who.int/handle/10665/43072 [accessed 2024-04-29]
2. Patient safety incident reporting and learning systems: technical report and guidance. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240010338 [accessed 2024-04-29]
3. Global patient safety action plan 2021-2030: towards eliminating avoidable harm in health care. World Health Organization. URL: https://www.who.int/teams/integrated-health-services/patient-safety/policy/global-patient-safety-action-plan [accessed 2024-04-29]
4. Panagioti M, Khan K, Keers RN, Abuzour A, Phipps D, Kontopantelis E, et al. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. BMJ 2019 Jul 17;366:l4185 [FREE Full text] [doi: 10.1136/bmj.l4185] [Medline: 31315828]
5. Kizer KW, Blum LN. Safe practices for better health care. In: Henriksen K, Battles JB, Marks ES, editors. Advances in Patient Safety: From Research to Implementation (Volume 4: Programs, Tools, and Products). New York, NY: Agency for Healthcare Research and Quality; 2005.
6. Asensi-Vicente J, Jiménez-Ruiz I, Vizcaya-Moreno MF. Medication errors involving nursing students: a systematic review. Nurse Educ 2018;43(5):E1-E5. [doi: 10.1097/NNE.0000000000000481] [Medline: 29210898]
7. Stevanin S, Causero G, Zanini A, Bulfone G, Bressan V, Palese A. Adverse events witnessed by nursing students during clinical learning experiences: findings from a longitudinal study. Nurs Health Sci 2018 Dec;20(4):438-444. [doi: 10.1111/nhs.12430] [Medline: 29771463]
8. Dennison S, Freeman M, Giannotti N, Ravi P. Benefits of reporting and analyzing nursing students' near-miss medication incidents. Nurse Educ 2022;47(4):202-207. [doi: 10.1097/NNE.0000000000001164] [Medline: 35113065]

XSL•FO
RenderX

9. Chiou SF, Huang EW, Chuang JH. The development of an incident event reporting system for nursing students. Stud Health Technol Inform 2009;146:598-602. [Medline: 19592912]
10. Ruiz Colón G, Evans K, Kanzawa M, Phadke A, Katznelson L, Shieh L. How many lives will you save? A mixed methods evaluation of a novel, online game for patient safety and quality improvement education. Am J Med Qual 2023 Nov 01;38(6):306-313 [FREE Full text] [doi: 10.1097/JMQ.0000000000000153] [Medline: 37882817]
11. Vestal ME, Matthias AD, Thompson CE. Engaging students with patient safety in an online escape room. J Nurs Educ 2021 Aug;60(8):466-469. [doi: 10.3928/01484834-20210722-10] [Medline: 34346812]
12. van Gaalen AE, Brouwer J, Schönrock-Adema J, Bouwkamp-Timmer T, Jaarsma AD, Georgiadis JR. Gamification of health professions education: a systematic review. Adv Health Sci Educ Theory Pract 2021 May;26(2):683-711 [FREE Full text] [doi: 10.1007/s10459-020-10000-3] [Medline: 33128662]
13. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, Wortley D, Lilienthal A, Tudor Car LT, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res 2019 Mar 28;21(3):e12994 [FREE Full text] [doi: 10.2196/12994] [Medline: 30920375]
14. Sánchez-García A, Saurín-Morán PJ, Carrillo I, Tella S, Põlluste K, Srulovici E, et al. Patient safety topics, especially the second victim phenomenon, are neglected in undergraduate medical and nursing curricula in Europe: an online observational study. BMC Nurs 2023 Aug 24;22(1):283 [FREE Full text] [doi: 10.1186/s12912-023-01448-w] [Medline: 37620803]
15. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
16. Sistema de notificación y aprendizaje para la seguridad del paciente (SiNASP). Ministerio de Sanidad. 2020. URL: https://sinasp.es/ [accessed 2022-03-14]
17. Cuidados 2.0. manual SINEA. Generalitat Valenciana. 2013. URL: https://cuidados20.san.gva.es/web/calidad-y-seguridad-de-cuidados/bienvenida/-/blogs/manual-sinea [accessed 2022-02-04]
18. Herramienta BACRA v1.2. AppAndAbout. URL: https://appandabout.es/bacra/ [accessed 2022-02-04]
19. Carrillo I, Mira JJ, Vicente MA, Fernandez C, Guilabert M, Ferrús L, et al. Design and testing of BACRA, a web-based tool for middle managers at health care facilities to lead the search for solutions to patient safety incidents. J Med Internet Res 2016 Sep 27;18(9):e257 [FREE Full text] [doi: 10.2196/jmir.5942] [Medline: 27678308]
20. Viechtbauer W, Smits L, Kotz D, Budé L, Spigt M, Serroyen J, et al. A simple formula for the calculation of sample size in pilot studies. J Clin Epidemiol 2015 Nov;68(11):1375-1379. [doi: 10.1016/j.jclinepi.2015.04.014] [Medline: 26146089]
21. Safety incident reporting system for students during their clinical internship. safest. URL: https://calite.umh.es/valoracion/en/ [accessed 2024-04-29]
22. Walker D, Barkell N, Dodd C. Error and near miss reporting in nursing education: the journey of two programs. Teach Learn Nurs 2023 Jan;18(1):197-203. [doi: 10.1016/J.TELN.2022.10.001]
23. Song MO, Yun SY, Jang A. Patient safety error reporting education for undergraduate nursing students: a scoping review. J Nurs Educ 2023 Sep;62(9):489-494. [doi: 10.3928/01484834-20230712-04] [Medline: 37672496]
24. Morey S, Magnusson C, Steven A. Exploration of student nurses' experiences in practice of patient safety events, reporting and patient involvement. Nurse Educ Today 2021 May;100:104831. [doi: 10.1016/j.nedt.2021.104831] [Medline: 33676347]
25. Lee S, Roh HR, Kim M, Park JK. Evaluating medical students' ability to identify and report errors: finding gaps in patient safety education. Med Educ Online 2022 Dec;27(1):2011604 [FREE Full text] [doi: 10.1080/10872981.2021.2011604] [Medline: 35129092]
26. Walker D, Altmiller G, Hromadik L, Barkell N, Barker N, Boyd T, et al. Nursing students' perceptions of just culture in nursing programs: a multisite study. Nurse Educ 2020;45(3):133-138. [doi: 10.1097/NNE.0000000000000739] [Medline: 32310625]
27. Van Slambrouck L, Verschueren R, Seys D, Bruyneel L, Panella M, Vanhaecht K. Second victims among baccalaureate nursing students in the aftermath of a patient safety incident: an exploratory cross-sectional study. J Prof Nurs 2021;37(4):765-770. [doi: 10.1016/j.profnurs.2021.04.010] [Medline: 34187676]
28. Singh H, Thomas EJ, Petersen LA, Studdert DM. Medical errors involving trainees: a study of closed malpractice claims from 5 insurers. Arch Intern Med 2007 Oct 22;167(19):2030-2036. [doi: 10.1001/archinte.167.19.2030] [Medline: 17954795]
29. Flin R, Patey R, Jackson J, Mearns K, Dissanayaka U. Year 1 medical undergraduates' knowledge of and attitudes to medical error. Med Educ 2009 Dec;43(12):1147-1155. [doi: 10.1111/j.1365-2923.2009.03499.x] [Medline: 19930505]
30. Gómez Ramírez O, Arenas Gutiérrez W, González Vega L, Garzón Salamanca J, Mateus Galeano E, Soto Gámez A. Cultura de seguridad del paciente por personal de enfermería en Bogotá, Colombia. Cienc Enferm 2011 Dec;17(3):97-111. [doi: 10.4067/S0717-95532011000300009]
31. Hayes AJ, Roberts P, Figgins A, Pool R, Reilly S, Roughley C, et al. Improving awareness of patient safety in a peer-led pilot educational programme for undergraduate medical students. Educ Health (Abingdon) 2014;27(2):213-216. [doi: 10.4103/1357-6283.143775] [Medline: 25420988]
32. Hobgood C, Hevia A, Tamayo-Sarver JH, Weiner B, Riviello R. The influence of the causes and contexts of medical errors on emergency medicine residents' responses to their errors: an exploration. Acad Med 2005 Aug;80(8):758-764. [doi: 10.1097/00001888-200508000-00012] [Medline: 16043533]

XSL•FO
RenderX

33. King A, Daniels J, Lim J, Cochrane DD, Taylor A, Ansermino JM. Time to listen: a review of methods to solicit patient reports of adverse events. Qual Saf Health Care 2010 Apr;19(2):148-157. [doi: 10.1136/qshc.2008.030114] [Medline: 20351164]
34. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003 Aug;78(8):775-780. [doi: 10.1097/00001888-200308000-00003] [Medline: 12915363]
35. Webb RK, Currie M, Morgan CA, Williamson JA, Mackay P, Russell WJ, et al. The Australian incident monitoring study: an analysis of 2000 incident reports. Anaesth Intensive Care 1993 Oct;21(5):520-528. [doi: 10.1177/0310057X9302100507] [Medline: 8273871]
36. Soydemir D, Seren Intepeler S, Mert H. Barriers to medical error reporting for physicians and nurses. West J Nurs Res 2017 Oct;39(10):1348-1363. [doi: 10.1177/0193945916671934] [Medline: 27694427]
37. Aljabari S, Kadhim Z. Common barriers to reporting medical errors. ScientificWorldJournal 2021 Jun 10;2021:6494889-6494888 [FREE Full text] [doi: 10.1155/2021/6494889] [Medline: 34220366]
38. Mira JJ, Navarro IM, Guilabert M, Poblete R, Franco AL, Jiménez P, et al. A Spanish-language patient safety questionnaire to measure medical and nursing students' attitudes and knowledge. Rev Panam Salud Publica 2015 Aug;38(2):110-119. [Medline: 26581051]
39. Mira JJ, Guilabert M, Vitaller J, Ignacio E. Formación en seguridad del paciente en las escuelas de medicina y enfermería en España. Rev Calid Asist 2016;31(3):141-145. [doi: 10.1016/j.cali.2015.08.008] [Medline: 26611250]
40. Kirwan M, Riklikiene O, Gotlib J, Fuster P, Borta M. Regulation and current status of patient safety content in pre-registration nurse education in 27 countries: findings from the Rationing - Missed nursing care (RANCARE) COST Action project. Nurse Educ Pract 2019 May;37:132-140. [doi: 10.1016/j.nepr.2019.04.013] [Medline: 31153130]

Abbreviations

AE: adverse event
BACRA: Based on Root Cause Analysis
IRS: incident reporting system
SAFEST: Safety Incident Report System for Students
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by B Lesselroth; submitted 29.01.24; peer-reviewed by D Chrimes, PCI Pang, Y Gong; comments to author 20.02.24; revised version received 01.03.24; accepted 27.06.24; published 18.07.24.

Please cite as:
Gil-Hernández E, Carrillo I, Guilabert M, Bohomol E, Serpa PC, Ribeiro Neves V, Maluenda Martínez M, Martin-Delgado J, Pérez-Esteve C, Fernández C, Mira JJ
Development and Implementation of a Safety Incident Report System for Health Care Discipline Students During Clinical Internships: Observational Study
JMIR Med Educ 2024;10:e56879
URL: https://mededu.jmir.org/2024/1/e56879
doi:10.2196/56879
PMID:39024005

©Eva Gil-Hernández, Irene Carrillo, Mercedes Guilabert, Elena Bohomol, Piedad C Serpa, Vanessa Ribeiro Neves, Maria Maluenda Martínez, Jimmy Martin-Delgado, Clara Pérez-Esteve, César Fernández, José Joaquín Mira. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.07.2024. 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.

XSL•FO
RenderX

Global Trends in mHealth and Medical Education Research: Bibliometrics and Knowledge Graph Analysis

Yuanhang He[1,2,*], MM; Zhihong Xie[1,2,*], MM; Jiachen Li[1,2], MM; Ziang Meng[1,2], MM; Dongbo Xue[1,2], PhD; Chenjun Hao[1,2], PhD

1

2

*these authors contributed equally

Corresponding Author:
Chenjun Hao, PhD

Abstract

Background: Mobile health (mHealth) is an emerging mobile communication and networking technology for health care systems. The integration of mHealth in medical education is growing extremely rapidly, bringing new changes to the field. However, no study has analyzed the publication and research trends occurring in both mHealth and medical education.

Objective: The aim of this study was to summarize the current application and development trends of mHealth in medical education by searching and analyzing published articles related to both mHealth and medical education.

Methods: The literature related to mHealth and medical education published from 2003 to 2023 was searched in the Web of Science core database, and 790 articles were screened according to the search strategy. The HistCite Pro 2.0 tool was used to analyze bibliometric indicators. VOSviewer, Pajek64, and SCImago Graphica software were used to visualize research trends and identify hot spots in the field.

Results: In the past two decades, the number of published papers on mHealth in medical education has gradually increased, from only 3 papers in 2003 to 130 in 2022; this increase became particularly evident in 2007. The global citation score was determined to be 10,600, with an average of 13.42 citations per article. The local citation score was 96. The United States is the country with the most widespread application of mHealth in medical education, and most of the institutions conducting in-depth research in this field are also located in the United States, closely followed by China and the United Kingdom. Based on current trends, global coauthorship and research exchange will likely continue to expand. Among the research journals publishing in this joint field, journals published by JMIR Publications have an absolute advantage. A total of 105 keywords were identified, which were divided into five categories pointing to different research directions.

Conclusions: Under the influence of COVID-19, along with the popularization of smartphones and modern communication technology, the field of combining mHealth and medical education has become a more popular research direction. The concept and application of digital health will be promoted in future developments of medical education.

(JMIR Med Educ 2024;10:e52461) doi:10.2196/52461

KEYWORDS

mHealth; mobile health; medical education; bibliometric; knowledge map; VOSviewer

Introduction

The rapid development of information and communication technologies in recent years has enabled greater connections to the mobile internet to access any information desired at any time and place, providing favorable conditions for the development of mobile health (mHealth). mHealth offers a full range of health care and medical education services, transcending geographical, time, language, and even organizational barriers [1,2].

mHealth was first defined as "unwired e-med" by Laxminarayan and Istepanian [3] in 2000. In 2003, mHealth was defined as an emerging mobile communication and networking technology for health care systems [4]. mHealth can provide diagnostic and treatment support services through mobile communication devices such as mobile phones, iPads, and personal digital assistants. An mHealth system and its associated app functions have a significant impact on typical health care, clinical data collection, record maintenance, health care information awareness, detection and prevention systems, drug counterfeiting, and theft. Thus, mHealth services have a powerful impact on all health services, including hospitals, care centers, and acute care, and are designed to significantly improve the lives of patients, especially older adults, individuals with physical disabilities, and patients with chronic conditions [4].

XSL•FO
RenderX

Currently, medical resources are extremely unevenly distributed among populations. In many developing countries, medical services have not yet been updated to incorporate current technological capabilities and the level of medical education often lags far behind that of developed countries. The integration and development of mHealth and medical education can help to address this situation. Through mHealth, doctors can provide basic health care and concepts to people living in areas where health services are lacking, and researchers who are experts in the field can share their clinical experience and theoretical knowledge with their peers through mobile communication technologies such as mobile phones. Thus, the widespread adoption of mHealth can not only rapidly raise the level of medical services in a region but can also help to somewhat reduce the gap in health services between different regions of the world and promote the progress of the global health care industry. From the perspective of medical education, medical students have traditionally only been able to acquire theoretical knowledge in the classroom and obtain hands-on experience through clinical practice. With the development and promotion of various medical-related mobile apps, medical education is no longer limited to face-to-face interactions, and more advanced and quality teaching resources can be disseminated through mobile software and other digital means. The combination of mHealth and medical education has provided more access to educational resources for medical students and physician groups at different levels [4].

Bibliometric analysis is a quantitative analysis method combining mathematics and statistics that focuses on the bibliometric characteristics of a research field to help researchers better understand the development trends in the field for guiding more in-depth research [5-7]. As research on mHealth continues to deepen, there have been an increasing number of articles published in the field. However, to date, there has been no bibliometric analysis of research related to the applications of mHealth in medical education. Therefore, in this study, we summarized the literature related to mHealth and medical education to help deepen our understanding of mHealth and identify future directions for its in-depth research in the context of developing medical education.

Methods

Ethical Considerations

All of the data collected and analyzed in this study were obtained from online public databases and did not involve any human or animal; thus, ethical approval was not required.

Data Sources

The Web of Science (WoS) literature database was selected to search, export, and analyze the relevant literature linking mHealth and medical education. Although the concept of mHealth was first proposed in 2000, since it was only officially defined in 2003, we set the start date for the search to 2003 [3,4,8]. We searched the WoS platform on April 2, 2023, selecting the WoS Core Collection, which contains articles included in the SCI (Science Citation Index)-EXPANDED, SSCI (Social Science Citation Index), AHCI (Arts & Humanities Citation Index), CPCI-S (Conference Proceedings Citation Index-Science), CPCI-SSH (Conference Proceedings Citation Index-Social Science & Humanities), BKCI-S (Book Citation Index-Science), BKCI-SSH (Book Citation Index-Social Science & Humanities), ESCI (Emerging Sources Citation Index), CCR (Current Chemical Reactions)-EXPANDED, and IC (Index Chemicus) databases.

Search Strategy

The search in the WoS Core Collection was performed in advanced search mode and the search option was set to "exact search." The search terms included a combination of "mHealth," "mobile health," and "medical education" as follows: "TS=[(mobile health) OR (mHealth)] AND [medical education]." The time span was from January 1, 2003, to March 31, 2023; the document type was limited to "Articles"; and English was selected as the only language of publication. The first output of the articles retrieved was obtained according to this strategy without setting any other inclusion criteria.

Data Analysis and Visualization

The literature retrieved based on the search strategy outlined above was exported in both plain-text (txt) and tab-delimited (txt) file formats. Descriptive statistics were obtained using HistCite Pro 2.1 [9]. Microsoft Excel 2021 was used to summarize the results from the HistCite Pro 2.1 analysis quantitatively and present the data graphically. VOSviewer (version 1.6.17) was used for cocitation correlation analysis and knowledge mapping [10]. VOSviewer (version 1.6.17) [11] and Pajek64 (version 5.16) were used jointly to analyze the current state of research and time trends. Visualization of country/region coauthorship trends was achieved using the combined powerful mapping capabilities of VOSviewer (version 1.6.17) and SCImago Graphica (version 1.0.34).

VOSviewer Software Settings

We used VOSviewer to perform a keyword co-occurrence analysis on the exported documents, setting the unit of analysis to "all keywords" and the counting method to "full counting"; the minimum number of occurrences was set to 10. For the overlay visualization, we utilized Pajek software for classification assistance. In the national and regional coauthorship trends analysis, we set the minimum number of coauthors for each country to 5 in VOSviewer. In the cocited references analysis, we set the minimum number of citations to 10. In the cocited journal sources analysis, we set the minimum cocitation count to 35. In the cocited authors analysis, we set the minimum number of citations to 20.

Results

Search Results and Publication Trends

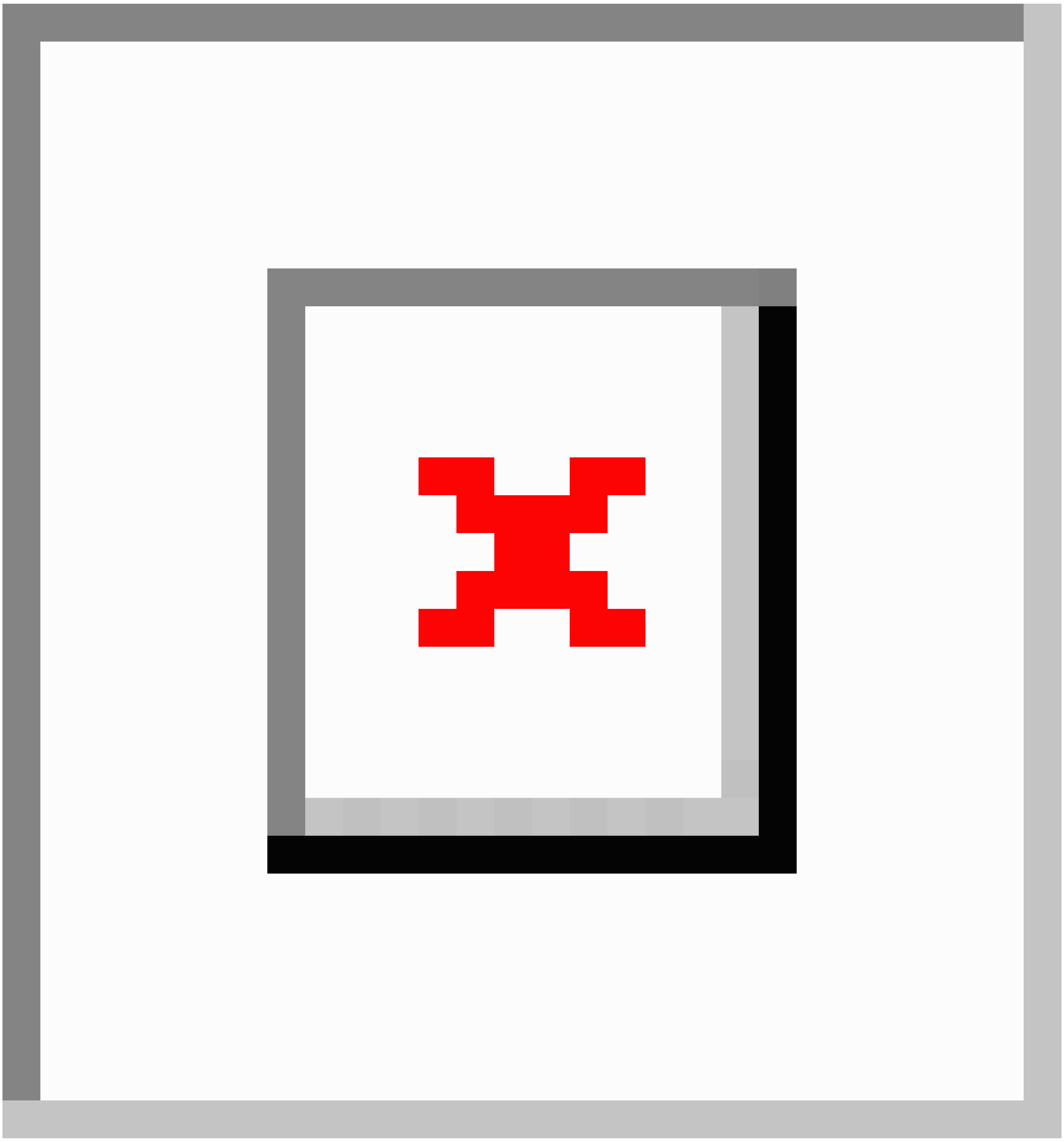
A total of 790 publications related to mHealth and medical education were retrieved based on the search strategy outlined in the Methods, which were analyzed by HistCite Pro 2.1. The local citation score (LCS) and global citation score (GCS) were calculated by the HistCite Pro software based on the information provided in the documents. The LCS refers to the number of times a document is cited within a given topic, reflecting the extent of recognition of research findings within the peer

XSL•FO
RenderX

community. The GCS represents the number of times a document is cited across all fields globally, serving as a significant indicator of the interdisciplinary and cross-domain impact of research outcomes. The GCS for the 790 articles was 10,600, with an average of 13.42 citations per article, and the LCS was 96.

Figure 1 shows the number of mHealth and medical education–related publications and the associated changes in the LCS over time. In the last two decades, especially since 2007, the annual number of publications has been steadily increasing year by year. Since 2020, the annual number of publications has exceeded 100, rising to 130 in 2022. However, data for 2023 only include publications from the first 3 months and are thus incomplete, making it difficult to determine the publication trend for that year. In terms of the LCS, the highest value was 15 in 2016, indicating a significant reference value for research in mHealth and medical education in that year. Additionally, there were peaks in the LCS detected in 2008 (7), 2013 (14), 2014 (14), and 2016 (15), indicating that studies in these years had large contributions to the research published in this field in the subsequent years. However, due to limitations of the search time frame, articles submitted in 2022 and 2023 may still be under review and not yet been published (and therefore not yet cited), resulting in an incomplete calculation of LCS values for the past 3 years.

Figure 1. Annual trend in the number of publications and local citation score (LCS) in the field of mobile health and medical education from 2003 to 2023.

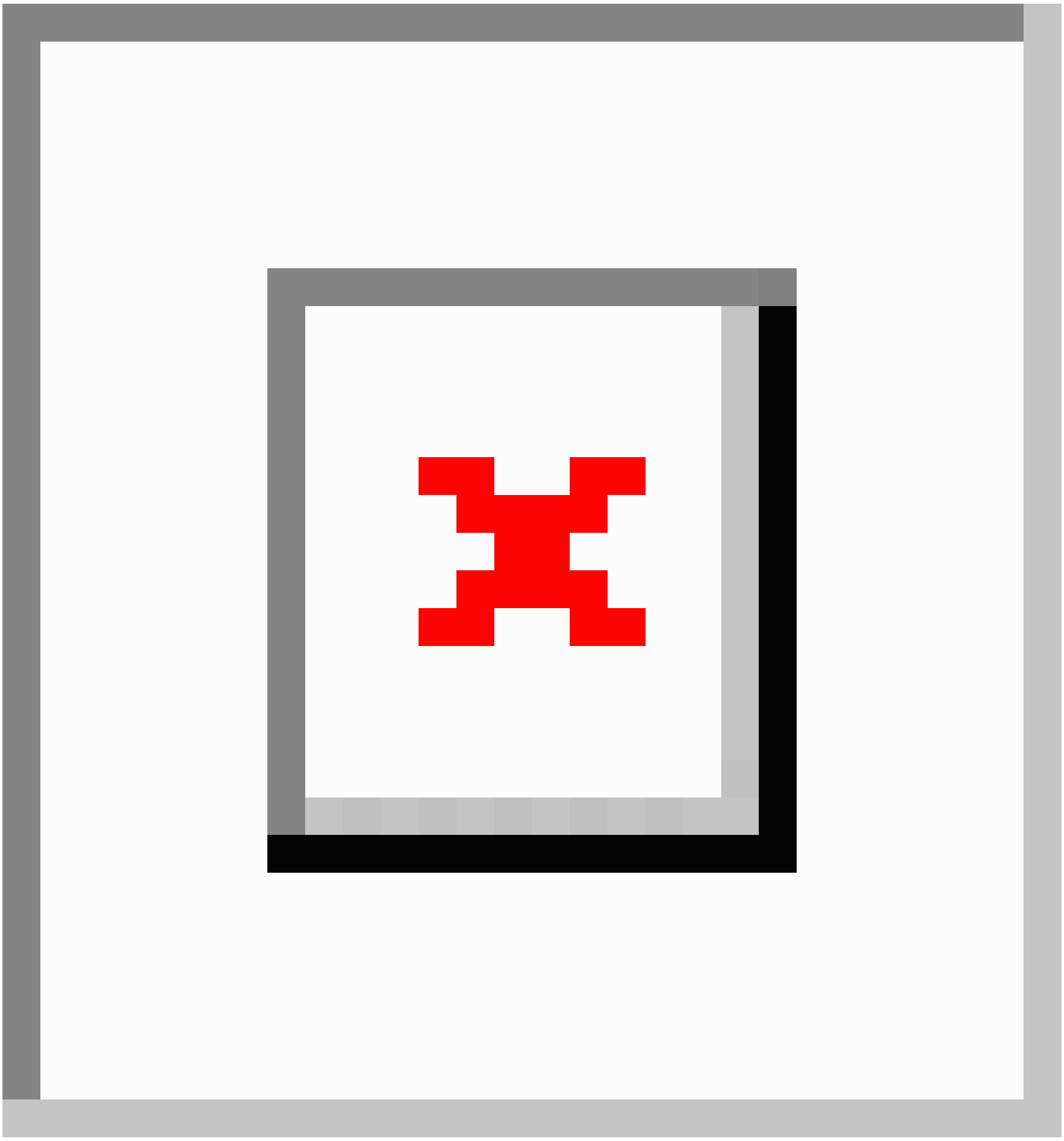


XSL•FO
RenderX

Contributions of Countries and Institutions

We analyzed the top countries and institutions that have published research related to mHealth and medical education. Figure 2A shows the top 10 countries in terms of publication volume, with each country publishing over 20 articles. The United States ranked first with a total of 318 articles, accounting for 40.25% of the total publication volume, representing a contribution far greater than that of other countries. China (n=70) and the United Kingdom (n=62) ranked second and third, respectively. The top 5 countries with respect to the LCS are presented in Figure 2B, with the LCS for the United States reaching 47, which was much higher than that for any other country. Figure 2C lists the top 5 countries in terms of the article H-index, with the United States again ranking first with an H-index of 32; followed by the United Kingdom (17) in second; and China, Canada, and Australia tying for third with an H-index of 14 each. Therefore, the United States leads in both the quantity and quality of publications related to mHealth and medical education, while China and the United Kingdom also rank in the top three for all indicators.

Figure 2. Ranking of top publishing countries and institutions in the field of mobile health and medical education. (A) The top 10 countries with the largest number of publications and their proportions. (B) The top 5 countries with the largest LCS and their proportions. (C) The top 5 countries with the largest H-index values. (D) Institutions with more than 10 publications. (E) The top 6 institutions with the largest LCS. (F) The top 5 institutions with the largest H-index values. Hlth Bur Gansu Prov: Health Bureau of Gansu Province; LCS: local citation score; MCPHS: Massachusetts College of Pharmacy and Health Sciences; Minist Hlth: Ministry of Health; Peoples R China: People’s Republic of China.



XSL•FO
RenderX

We subsequently analyzed the institutions that published the retrieved articles in this field. Figure 2D shows the institutions with more than 10 publications on the topic, with The University of Sydney ranking first with 18 articles, followed by The University of Toronto (n=16), and Harvard Medical School and Johns Hopkins University tied for third place with 15 relevant publications each. In terms of the LCS, The University of Pennsylvania ranked first with a score of 10 (Figure 2E). Figure 2F compares the top 5 institutions in terms of the H-index, with The University of Pennsylvania and The University of Sydney having the highest H-index of 9 each. Thus, overall, the world's leading universities such as The University of Pennsylvania and The University of Sydney are producing relatively advanced research in mHealth and medical education, and this institutional-based analysis is largely consistent with the country-based analysis.

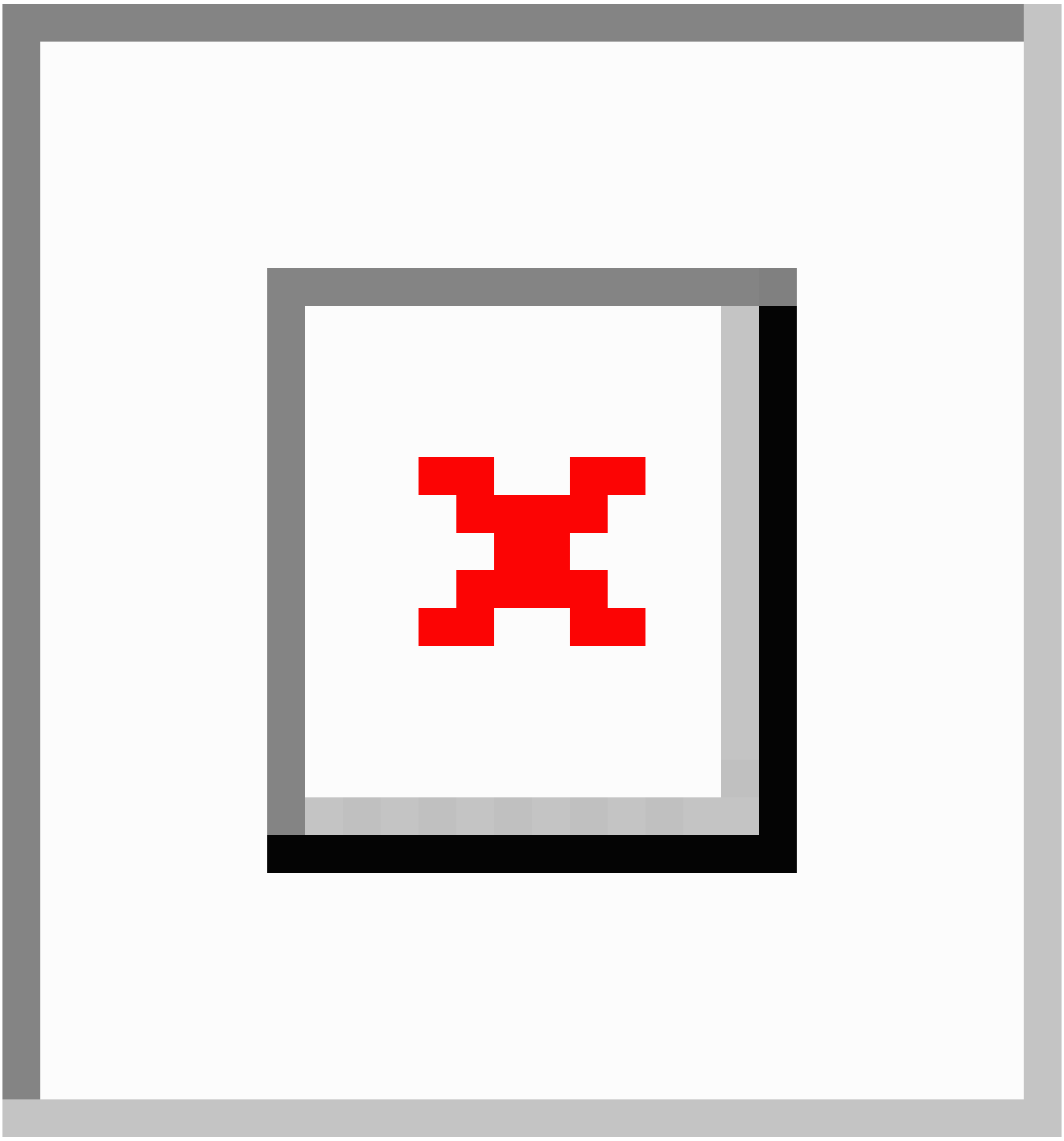
Journal of Publication and Authors

A total of 420 journals were involved in publishing mHealth and medical education–related articles according to statistics compiled with HistCite Pro 2.1. *JMIR mHealth and uHealth* ranked first with 67 related publications, *Journal of Medical Internet Research* ranked second with 35 articles, and *JMIR Formative Research* and *BMJ Open* ranked third with 19 articles each. Among the journals with more than 10 publications, five are from JMIR Publications (Figure 3A). In terms of the H-index, *JMIR mHealth and uHealth* again ranked first with an H-index of 20, *Journal of Medical Internet Research* ranked second (15), and *Telemedicine and e-Health* ranked third (9) (Figure 3B). This finding demonstrates the comprehensiveness and authority of the JMIR Publications journal series in the field of mHealth and medical education.

The authors with the highest number of publications published 5 articles each, and since most of these authors are repeated coauthors, this field appears to be dominated by a relatively small set of researchers. Table 1 lists the authors with more than 4 articles published along with their LCS and GCS; among them, Littman-Quinn R, Aungst TD, and Kovarik CL are at the top of the list in terms of both the quantity and quality of publications.

XSL•FO
RenderX

Figure 3. Contributions of journals to the field of mobile health and medical education. (A) Journals with more than 10 publications. (B) Top 5 journals with the largest H-index values.



XSL•FO
RenderX

Table . Authors with more than 4 publications and their associated local citation score (LCS) and global citation score (GCS).

Author	Number of publications	LCS	GCS
Deng N	5	0	23
Gill CJ	5	4	33
Halim N	5	4	33
Littman-Quinn R	5	9	163
Schooley B	5	1	60
Aungst TD	4	8	103
Barteit S	4	0	53
Briz-Ponce L	4	0	305
Duan HL	4	0	21
Kim J	4	0	11
Kovarik CL	4	8	136
Li Y	4	0	86
Neuhann F	4	0	53
Scott KM	4	2	101
Williams AL	4	4	33

Keyword Co-Occurrence and Research Trends

We used VOSviewer to conduct keyword co-occurrence analysis on the exported 790 documents. A total of 3288 keywords were extracted, with 105 keywords meeting the threshold. The 105 keywords were then plotted using VOSviewer for density, network, and overlay visualization.

Figure 4 shows the density visualization of the 105 keywords, revealing that the majority of research in this field revolves around mHealth or education, which, to a certain extent, verifies the objectivity and scientificity of our search strategy and analysis.

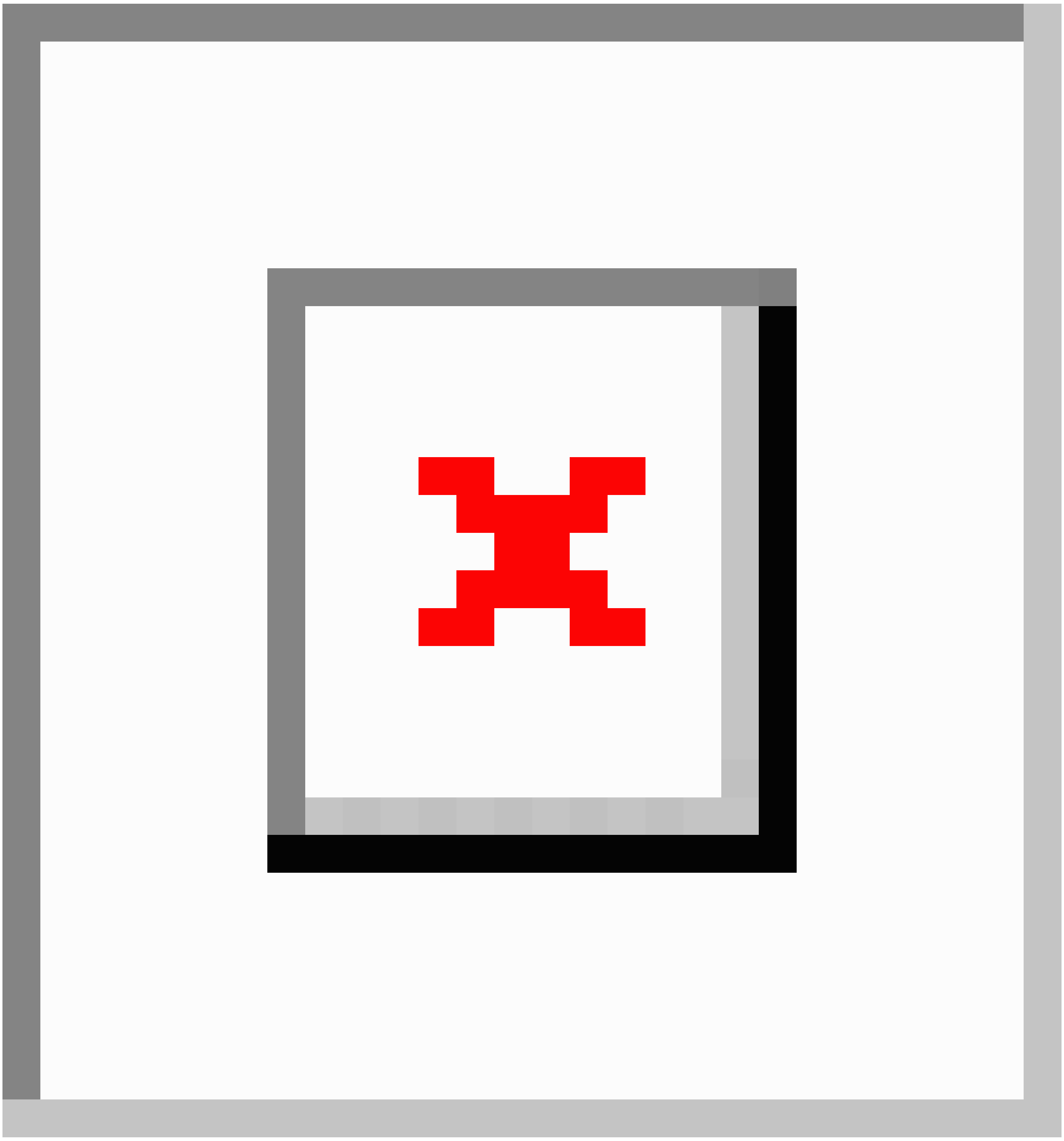
Figure 5A shows a network visualization of the 105 keywords. A color node can roughly represent a research direction and a larger node area typically indicates a more popular keyword. The software divided all keywords into 5 categories. The red cluster consists of 34 keywords, primarily focusing on clinical medical education (including the keywords "education" and "medical education"), demonstrating that the application of mobile medical software (ie, mobile apps) in medical education and knowledge is widely studied. The green cluster comprises 33 keywords, mainly focusing on the management and development of mobile medical devices and software as well as their application in different age groups through the internet and mobile apps (including the keywords "internet," "mobile app," "management," "outcomes," "adults," "children," and "adolescents"). The blue cluster includes 21 keywords, emphasizing the promotion and education of mHealth in public health and epidemiology (including the keywords "public health," "medical informatics," "health education," "epidemiology," "HIV," and "COVID-19"). The yellow cluster contains 16 keywords, primarily investigating the association of mHealth with smartphones, applications in remote diagnosis and treatment, and its role in digital medicine (including the keywords "mHealth," "smartphone," "mobile phone," "telehealth," "telemedicine," and "digital health"). As the purple cluster contains only one keyword, "qualitative research," this serves as a link between various research areas owing to its vague directionality.

Figure 5B presents an overlay visualization of the 109 keywords highlighted in research related to the field of mHealth and medical education. According to the color legend, over time, the main keywords in this research area have gradually shifted from the purple (prior to 2017) to yellow (after 2020) category. This indicates that initially, this field was limited to the understanding and learning of mobile information (including the keywords "information," "mobile," and "mobile learning"). With the development and popularity of the internet and mobile devices, their use in medical education began to be promoted (including the keywords "internet," "mobile devices," "mobile technology," and "medical education"). Further, with the development of mobile phones and mobile software, the application of mHealth in medical education is no longer limited to the teaching of professional knowledge to students but is also oriented toward the general public and the promotion of educational medical health concepts among different groups of people (represented by most keywords in the teal-colored small-sized nodes).

In recent years, mHealth has increasingly shifted into the research spotlight with the continuous support of smartphones and a greater inclination toward public health, along with the implementation of inclusive medical services and health communication (yellow small-sized nodes). In the future, increased promotion and use of mHealth care may push digital health (highlighted as "digital health" in yellow in Figure 5, referring to the application of digital technologies such as the Internet of Things, artificial intelligence, and big data in health management) to a focused area of research.

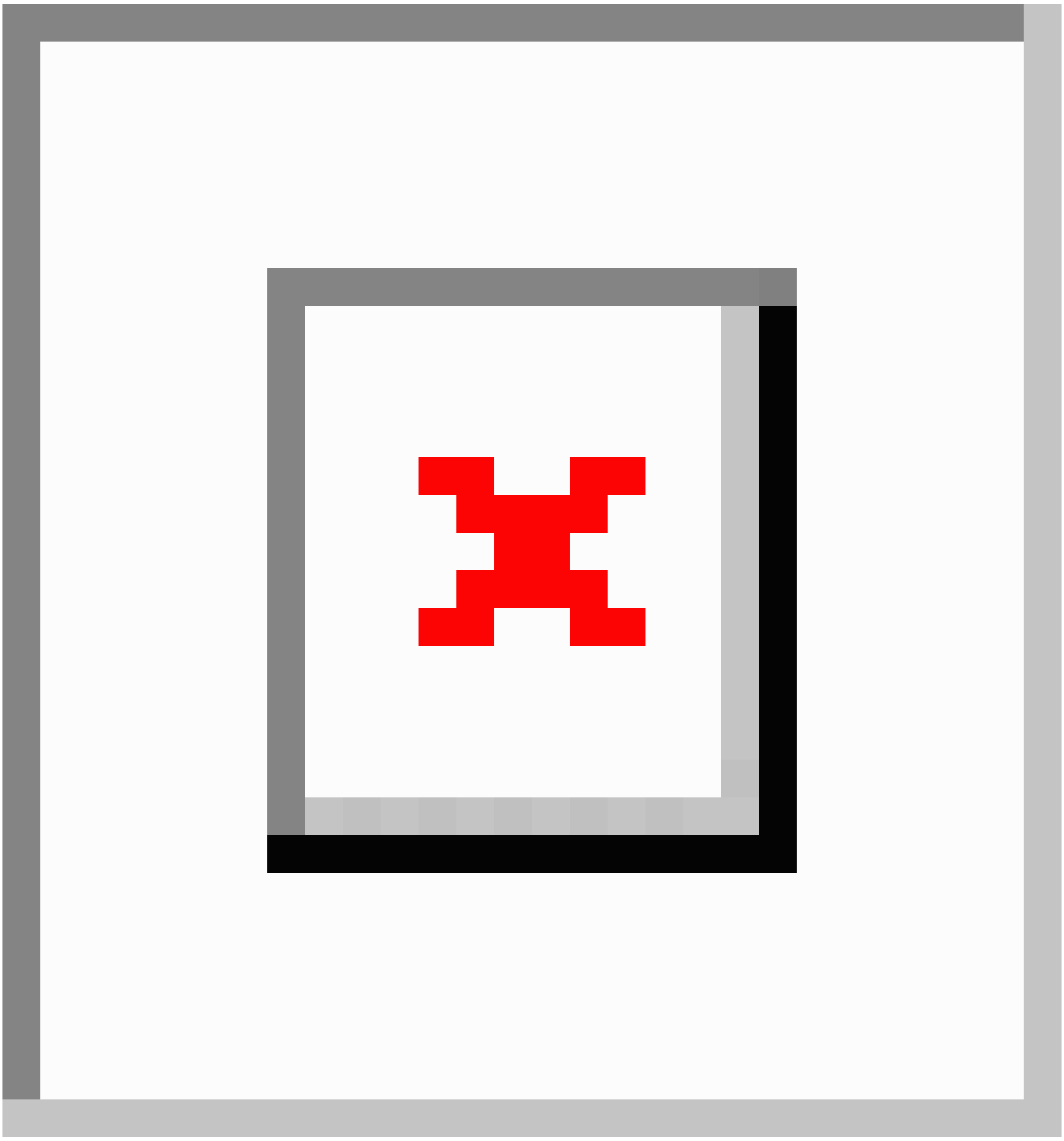
XSL•FO
RenderX

Figure 4. Density visualization of the top 105 keywords. The higher the keyword density, the redder its surrounding color.



XSL•FO
RenderX

Figure 5. (A) Network and (B) overlay visualization of the 105 keywords.

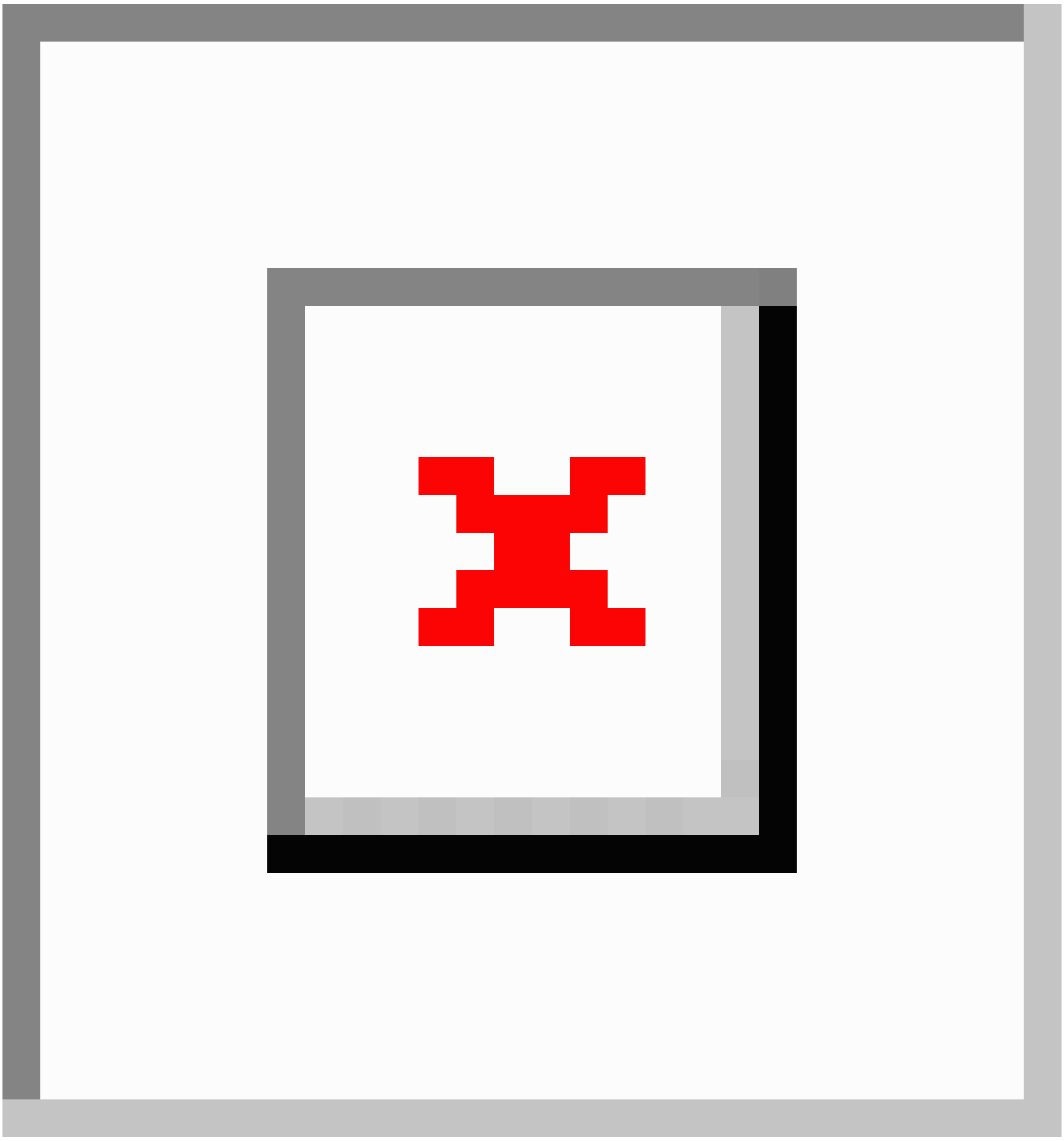


National and Regional Coauthorship Trends

Different countries and regions tend to collaborate on the same research topics rather than working in isolation. VOSviewer identified 37 countries with coauthorship relationships. As shown in Figure 6A, the red clusters (12 countries) have the strongest coauthorship relationships, with the United States (as the country with the highest number of publications) having the most significant coauthorship relationships. We then exported the results of the VOSviewer analysis to SCImago Graphica for further analysis of country coauthorship correlations in a world map (Figure 6B), which provides a clearer visual representation of the strong coauthorship links between countries on all continents, mainly comprising European countries. This map also shows that researchers working in different countries have a large breadth of interactions, even communicating with each other across continents.

XSL•FO
RenderX

Figure 6. National and regional coauthorship trends. (A) Network map of national coauthorship. (B) Country coauthorship correlations in a world map.



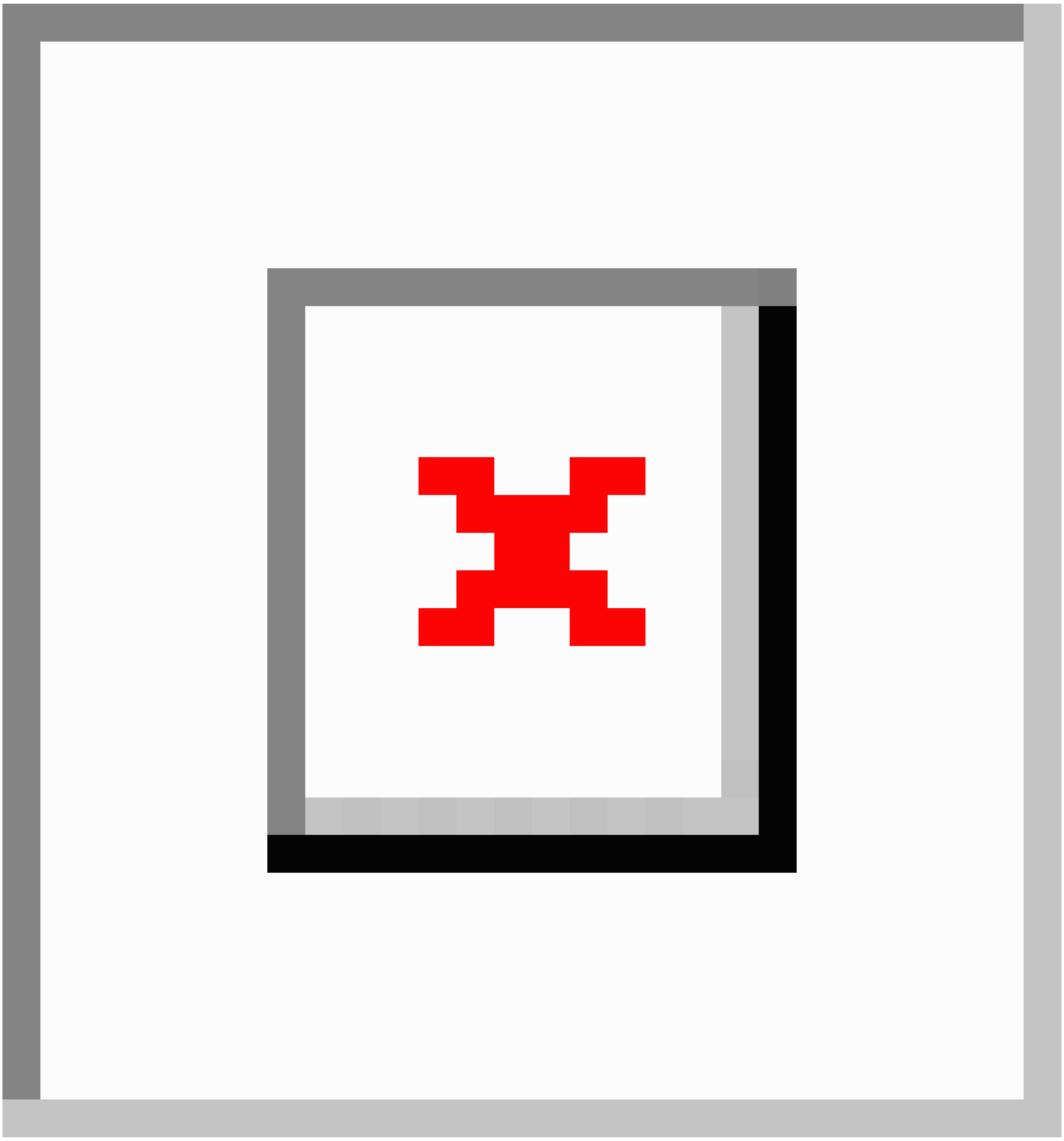
Cocitation Analysis

Cocited References

Cocited references are an important indicator of the extent to which a particular field is linked to different researchers or research areas. A total of 25,986 cited references were considered valid in VOSviewer, with a total of 28 articles meeting the minimum threshold. These 28 references were divided into three interconnected clusters (Figure 7A), with 11 articles in the red cluster, 9 articles in the green cluster, and 8 articles in the blue cluster. The article “Smartphone and medical related app use among medical students and junior doctors in the United Kingdom (UK): a regional survey” by Payne and colleagues [12], published in *BMC Medical Informatics and Decision Making* [12], showed the highest cocitation frequency, with 28 citations.

XSL•FO
RenderX

Figure 7. Cocitation analysis. (A) Network map of cocited references. (B) Network map of cocited journal sources. (C) Network map of cocited authors.



Cocited Journal Sources

The analysis of cocited journal sources demonstrates the extent to which research in the fields of mHealth and medical education is published in journals that have previously published relevant literature on these topics. A total of 83 journal sources met the minimum threshold (Figure 7B), with the *Journal of Medical Internet Research* having the most cited articles at 898. The 83 journals were divided into six clusters, including 23 journals in the red cluster, 15 in the green cluster, 14 in the blue cluster, 13 in the yellow cluster, 11 in the purple cluster, and 7 in the cyan cluster.

Cocited Authors

The number of cocited authors serves as an important indicator for bibliometrics analysis, highlighting the closeness of scholarly relationships and research directions among scholars. In VOSviewer, 22 authors met the minimum threshold of 20,821 citations. According to the network of cocited authors (Figure 7C), a larger area of a color node indicates more citations. The World Health Organization (WHO) had the highest number of citations, reaching 160, reflecting its authority in the field of mHealth combined with medical education. Although each of the four colors represents a different research focus for different authors, the different clusters are not absolutely isolated from each other.

XSL•FO
RenderX

Discussion

Principal Results

In this study, we conducted a search of the literature in the WoS Core Collection and obtained 790 relevant articles on mHealth and medical education published from 2003 to 2023 according to the search strategy. In the past two decades, especially since 2007, the number of published papers in this combined field has gradually increased, reaching 130 published papers in 2022. The GCS is 10,600, with an average of 13.42 citations per article, and the LCS is 96. The United States stands out as the country with the greatest application of mHealth in medical education, and most of the institutions with in-depth research in this field are also located in the United States. The depth of research in China and the United Kingdom followed closely behind. Based on current trends, global coauthorship and research exchange will continue to expand. Among the journals publishing research on this topic, JMIR Publications journals have an absolute advantage in this joint field. The 105 keywords identified were divided into five categories pointing to different research directions.

An important indicator of research trends in a field is the number of relevant articles published each year. The results of this analysis show a general upward trend in research in mHealth and medical education, a field that has received a great deal of attention in recent years. As of 2020, research in this joint field can be divided into two phases: the nascent phase and the stable growth phase. The nascent phase spans from the introduction of mHealth in 2003 to 2007 when the model of combining mHealth and medical education was first proposed and associated research was in its infancy, as represented by the small number of relevant articles published in this period. The period from 2007 to 2020 represents a phase of steady growth, with a gradual increase in the number of relevant research articles. In terms of the LCS, there were four peaks detected in 2008, 2013, 2014, and 2016, respectively. Considering the annual publication volume over the entire period of mHealth research, it can be inferred that the research achievements in 2008 played a crucial role in the development of mHealth applications in medical education.

From Figure 5B, it can be seen that the main keywords representing the direction of mHealth before 2017 were “health care,” “internet,” and “information”; however, after 2019, the main keywords changed to “mobile phone,” “mHealth,” and “education,” indicating that the direction of mHealth development has been changing in recent years. This may be due to the popularity of smartphones, development of mobile software, spread of the internet, and rapid development of communication technology. mHealth has evolved from an initial focus on understanding and learning about mobile information and health care information to a combination of mHealth and mobile devices for research and medical education. On January 9, 2007, Steve Jobs, as the Chief Operating Officer of Apple, presented the iPhone 2G and its operating system iOS to the world. This event triggered the rapid development of smartphones and associated apps, as well as the emergence of new mobile platforms. Likely due to these breakthroughs in smartphones and mobile-related technologies, mHealth began to enter the minds of researchers, attracting the attention of scientists worldwide, and thus the number of annual publications related to mHealth began to rise steadily. In addition, the rapid development of communication technology, increasing popularity of smartphones, and development of mobile software provided a suitable platform for medical schools, hospitals, and research institutions in different regions to collaborate and communicate with each other.

On March 11, 2020, the WHO announced COVID-19 as a global pandemic caused by SARS-CoV-2, which affected the daily lives of billions of people [13-15]. The COVID-19 pandemic not only posed a serious challenge to global medical care systems [16] but also limited access to learning and education, with most students having to access knowledge via the internet using communication devices such as mobile phones, iPads, and computers at home. This led to the rapid development of online teaching and learning software, and ultimately accelerated the integration of mHealth and medical education. Consequently, the number of mHealth-related research articles exceeded 100 in 2020 and rose to 130 in 2022. The development and application of 5G mobile technology and the rapid development of online teaching–related software collectively contributed to the deeper integration of mHealth and medical education [17]. Analysis of keyword clusters (Figures 4 and 5A) showed that mHealth research in the last two decades can be roughly divided into four clusters: a clinical education–related cluster, an mHealth equipment and software–related cluster, a health care and public health mission cluster, and a telemedicine cluster. The development of the discipline requires mutual cooperation with other fields. Promoting the integration and development of mHealth and medical education is extremely important to improve the health care conditions in less developed areas such as developing countries and to promote the common development of the world’s health care standards, which is in line with the WHO’s aim to improve the health of people around the world as much as possible.

The high number of citations in this joint field is somewhat indicative of the quality of the research cited. The study by Payne et al [12] received a particularly high number of citations, indicating its significant impact on medical education and mHealth. This study found that medical students and physician groups enjoy acquiring theoretical knowledge through an mHealth teaching model, which is consistent with the overall findings of this bibliometric analysis. In terms of researchers, the WHO has the highest number of cited articles in the field of mHealth combined with medical education, which not only reflects the authority of the organization but also shows the importance the WHO attaches to mHealth combined with medical education. The top three cited journals for mHealth and medical education research are *JMIR mHealth and uHealth* (impact factor 4.948, Q1), *Journal of Medical Internet Research* (impact factor 7.077, Q1), and *BMJ Open* (impact factor 3.007, Q2). According to their impact factors obtained from Journal Citation Reports 2022 [18], these three journals are considered Q1 and Q2 journals, indicating their strong contributions to their respective fields. The cocitation analysis demonstrated the authority of *Journal of Medical Internet Research* in the field

XSL•FO
RenderX

of mHealth and medical education research, with an annual volume of 318 articles, an LCS of 47, and an H-index of 32. Although the United States is clearly the world leader in mHealth and medical education, making a significant contribution to the field, academic exchanges between different countries are also ongoing.

Limitations

There are limitations of our study that should be acknowledged. First, data completeness may be inadequate; although the WoS database has the most complete coverage of articles, our literature search was limited to the English language, which may have resulted in the omission of some key information for some countries where research was published in other languages. In addition, the search strategy was limited to the string "TS=[(mobile health) OR (mHealth)] AND [medical education]" and therefore may not have been sufficiently comprehensive. Second, we only searched under the category "Articles," which may have also led to missing relevant publications in other formats.

Conclusion

Bibliometric analysis indicates that mHealth-related research has been growing at an accelerating rate over the last two decades. In the area of combining mHealth and medical education, the WHO is playing an important leadership role, with many researchers following suit. With the influence of COVID-19, the spread of smartphones, and constant developments in modern communication technologies, the field of combining mHealth and medical education is becoming increasingly popular, and the concept and application of digital health will be promoted in the future drive for medical education.

Acknowledgments

We thank the Web of Science database for providing all of the data used in this analysis. We are also grateful to Stable Diffusion, a deep learning model used to convert text into images, which was used to generate the table of contents image. This work was supported by the National Natural Science Foundation of China (grants 82170654 and 82100675), Key Research and Development Program of Heilongjiang Province (grant 2022ZX06C06), Excellent Youth Foundation of the First Affiliated Hospital of Harbin Medical University (grant 2021Y12), and Postgraduate Research & Practice Innovation Program of Harbin Medical University (grant YJSCX2023-207HYD).

Data Availability

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

YH designed the study and analyzed the data. YH and ZX wrote the manuscript. JL and ZM prepared the figures and tables. CH reviewed and revised the manuscript. DX and CH supervised the research and thus made equal contributions to the work. All authors approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Akter S, Ray P. mHealth - an ultimate platform to serve the unserved. Yearb Med Inform 2010:94-100. [doi: 10.1055/s-0038-1638697] [Medline: 20938579]
2. Tachakra S, Wang XH, Istepanian RSH, Song YH. Mobile e-health: the unwired evolution of telemedicine. Telemed J E Health 2003;9(3):247-257. [doi: 10.1089/153056203322502632] [Medline: 14611692]
3. Laxminarayan S, Istepanian RS. UNWIRED E-MED: the next generation of wireless and internet telemedicine systems. IEEE Trans Inf Technol Biomed 2000 Sep;4(3):189-193. [doi: 10.1109/titb.2000.5956074] [Medline: 11026588]
4. Silva BMC, Rodrigues JJPC, de la Torre Díez I, López-Coronado M, Saleem K. Mobile-health: a review of current state in 2015. J Biomed Inform 2015 Aug;56:265-272. [doi: 10.1016/j.jbi.2015.06.003] [Medline: 26071682]
5. Ellegaard O, Wallin JA. The bibliometric analysis of scholarly production: how great is the impact? Scientometrics 2015;105(3):1809-1831. [doi: 10.1007/s11192-015-1645-z] [Medline: 26594073]
6. Glanville J, Kendrick T, McNally R, Campbell J, Hobbs FDR. Research output on primary care in Australia, Canada, Germany, the Netherlands, the United Kingdom, and the United States: bibliometric analysis. BMJ 2011 Mar 8;342:d1028. [doi: 10.1136/bmj.d1028] [Medline: 21385804]
7. Chen C, Dubin R, Kim MC. Emerging trends and new developments in regenerative medicine: a scientometric update (2000 – 2014). Expert Opin Biol Ther 2014 Sep;14(9):1295-1317. [doi: 10.1517/14712598.2014.920813] [Medline: 25077605]

XSL•FO
RenderX

8. Istepanian RSH, Lacal JC. Emerging mobile communication technologies for health: some imperative notes on m-health. Presented at: 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Sep 17 to 21, 2023:; Cancun, Mexico p. 1414-1416. [doi: 10.1109/IEMBS.2003.1279581]
9. Chen C. Searching for intellectual turning points: progressive knowledge domain visualization. Proc Natl Acad Sci U S A 2004 Apr 6;101(Suppl 1):5303-5310. [doi: 10.1073/pnas.0307513100] [Medline: 14724295]
10. Chen C, Hu Z, Liu S, Tseng H. Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace. Expert Opin Biol Ther 2012 May;12(5):593-608. [doi: 10.1517/14712598.2012.674507] [Medline: 22443895]
11. Zhang J, Luo Z, Zhang R, et al. The transition of surgical simulation training and its learning curve: a bibliometric analysis from 2000 to 2023. Int J Surg 2024 May 9. [doi: 10.1097/JS9.0000000000001579] [Medline: 38729115]
12. Payne KFB, Wharrad H, Watts K. Smartphone and medical related app use among medical students and junior doctors in the United Kingdom (UK): a regional survey. BMC Med Inform Decis Mak 2012 Oct 30;12:121. [doi: 10.1186/1472-6947-12-121] [Medline: 23110712]
13. Mahase E. Covid-19: WHO declares pandemic because of “alarming levels” of spread, severity, and inaction. BMJ 2020 Mar 12;368:m1036. [doi: 10.1136/bmj.m1036] [Medline: 32165426]
14. Polack FP, Thomas SJ, Kitchin N, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. N Engl J Med 2020 Dec 31;383(27):2603-2615. [doi: 10.1056/NEJMoa2034577] [Medline: 33301246]
15. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet 2020 Feb 15;395(10223):470-473. [doi: 10.1016/S0140-6736(20)30185-9] [Medline: 31986257]
16. Muntz MD, Franco J, Ferguson CC, Ark TK, Kalet A. Telehealth and medical student education in the time of COVID-19-and beyond. Acad Med 2021 Dec 1;96(12):1655-1659. [doi: 10.1097/ACM.0000000000004014] [Medline: 35134026]
17. Simkó M, Mattsson MO. 5G wireless communication and health effects–a pragmatic review based on available studies regarding 6 to 100 GHz. Int J Environ Res Public Health 2019 Sep 13;16(18):3406. [doi: 10.3390/ijerph16183406] [Medline: 31540320]
18. Journal citation reports. Clarivate. URL: https://clarivate.com/products/scientific-and-academic-research/research-analytics-evaluation-and-management-solutions/journal-citation-reports/ [accessed 2024-05-24]

Abbreviations

AHCI: Arts & Humanities Citation Index
BKCI-S: Book Citation Index-Science
BKCI-SSH: Book Citation Index-Social Science & Humanities
CCR: Current Chemical Reactions
CPCI-S: Conference Proceedings Citation Index-Science
CPCI-SSH: Conference Proceedings Citation Index-Social Science & Humanities
ESCI: Emerging Sources Citation Index
GCS: global citation score
IC: Index Chemicus
LCS: local citation score
mHealth: mobile health
SCI: Science Citation Index
WHO: World Health Organization
WoS: Web of Science

Edited by TDA Cardoso; submitted 04.09.23; peer-reviewed by R Poss-Doering; revised version received 17.03.24; accepted 14.05.24; published 04.06.24.

Please cite as:
He Y, Xie Z, Li J, Meng Z, Xue D, Hao C
Global Trends in mHealth and Medical Education Research: Bibliometrics and Knowledge Graph Analysis
JMIR Med Educ 2024;10:e52461
URL: https://mededu.jmir.org/2024/1/e52461
doi:10.2196/52461

© Yuanhang He, Zhihong Xie, Jiachen Li, Ziang Meng, Dongbo Xue, Chenjun Hao. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.6.2024. 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

XSL•FO
RenderX

information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Development of Web-Based Education Modules to Improve Carer Engagement in Cancer Care: Design and User Experience Evaluation of the e-Triadic Oncology (eTRIO) Modules for Clinicians, Patients, and Carers

Rebekah Laidsaar-Powell[1,2], PhD; Sarah Giunta[1], BPsych (Hons); Phyllis Butow[1,2], PhD; Rachael Keast[1], MClinPsych; Bogda Koczwara[3,4], BM, BS; Judy Kay[5], PhD; Michael Jefford[6], MBBS, MPH, PhD; Sandra Turner[7,8], MBBS, PhD; Christobel Saunders[9], MBBS; Penelope Schofield[6,10,11], PhD; Frances Boyle[8,12], MBBS, PhD; Patsy Yates[13], RN, PhD; Kate White[14], RN, PhD; Annie Miller[15], Dip Life Coaching, Dip Management, Dip Business; Zoe Butt[1], MClinPsych; Melanie Bonnaudet[5,16], MIT; Ilona Juraskova[1,2], PhD

[1]Centre for Medical Psychology & Evidence-based Decision-making, School of Psychology, The University of Sydney, Sydney, Australia
[2]Psycho-Oncology Co-operative Research Group, The University of Sydney, Sydney, Australia
[3]Flinders Medical Centre, Adelaide, Australia
[4]College of Medicine and Public Health, Flinders University, Adelaide, Australia
[5]School of Computer Science, The University of Sydney, Sydney, Australia
[6]Health Services Research and Implementation Science, Peter MacCallum Cancer Centre, Melbourne, Australia
[7]Department of Radiation Oncology, Westmead Hospital, Westmead, Australia
[8]Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
[9]Department of Surgery, Royal Melbourne Hospital, University of Melbourne, Melbourne, Australia
[10]Department of Psychology and Iverson Health Innovation Research Institute, Swinburne University, Melbourne, Australia
[11]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Australia
[12]Patricia Ritchie Centre for Cancer Care & Research, Mater Hospital, Sydney, Australia
[13]Faculty of Health, Queensland University of Technology, Brisbane, Australia
[14]Susan Wakil School of Nursing, The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council New South Wales, Sydney, Australia
[15]Cancer Council New South Wales, Sydney, Australia
[16]School of Electrical Engineering and Computer Science, Kungliga Tekniska högskolan Royal Institute of Technology, Stockholm, Sweden

Corresponding Author:
Rebekah Laidsaar-Powell, PhD
Centre for Medical Psychology & Evidence-based Decision-making
School of Psychology
The University of Sydney
Room 310, Level 3, Griffith Taylor Building (A19)
Manning Road
Sydney, 2006
Australia
Phone: 61 2 9351 6811
Email: rebekah.laidsaar-powell@sydney.edu.au

Abstract

Background: Carers often assume key roles in cancer care. However, many carers report feeling disempowered and ill‐equipped to support patients. Our group published evidence-based guidelines (the Triadic Oncology [TRIO] Guidelines) to improve oncology clinician engagement with carers and the management of challenging situations involving carers.

Objective: To facilitate implementation of the TRIO Guidelines in clinical practice, we aimed to develop, iteratively refine, and conduct user testing of a suite of evidence-based and interactive web-based education modules for oncology clinicians

XSL•FO
RenderX

(e-Triadic Oncology [eTRIO]), patients with cancer, and carers (eTRIO for Patients and Carers [eTRIO - pc]). These were designed to improve carer involvement, communication, and shared decision-making in the cancer management setting.

Methods: The eTRIO education modules were based on extensive research, including systematic reviews, qualitative interviews, and consultation analyses. Guided by the person-based approach, module content and design were reviewed by an expert advisory group comprising academic and clinical experts (n=13) and consumers (n=5); content and design were continuously and iteratively refined. User experience testing (including "think-aloud" interviews and administration of the System Usability Scale [SUS]) of the modules was completed by additional clinicians (n=5), patients (n=3), and carers (n=3).

Results: The final clinician module comprises 14 sections, requires approximately 1.5 to 2 hours to complete, and covers topics such as carer-inclusive communication and practices; supporting carer needs; and managing carer dominance, anger, and conflicting patient-carer wishes. The usability of the module was rated by 5 clinicians, with a mean SUS score of 75 (SD 5.3), which is interpreted as good. Clinicians often desired information in a concise format, divided into small "snackable" sections that could be easily recommenced if they were interrupted. The carer module features 11 sections; requires approximately 1.5 hours to complete; and includes topics such as the importance of carers, carer roles during consultations, and advocating for the patient. The patient module is an adaptation of the relevant carer module sections, comprising 7 sections and requiring 1 hour to complete. The average SUS score as rated by 6 patients and carers was 78 (SD 16.2), which is interpreted as good. Interactive activities, clinical vignette videos, and reflective learning exercises are incorporated into all modules. Patient and carer consumer advisers advocated for empathetic content and tone throughout their modules, with an easy-to-read and navigable module interface.

Conclusions: The eTRIO suite of modules were rigorously developed using a person-based design methodology to meet the unique information needs and learning requirements of clinicians, patients, and carers, with the goal of improving effective and supportive carer involvement in cancer consultations and cancer care.

(JMIR Med Educ 2024;10:e50118) doi:10.2196/50118

KEYWORDS

family carers; patient education; health professional education; web-based intervention; mobile phone

Introduction

Background

Carers (including but not limited to spouses, partners, adult children, siblings, parents, or friends [1]) of adults with cancer assume many responsibilities in supporting and caring for their loved one [2]. Carers can experience many challenges in this demanding role and often report high distress [3,4], poor physical health, low quality of life, and unmet needs [5,6]. As carer burden increases, carers may neglect their own needs, which can also impact their ability to support and care for their loved one [7,8].

While issues faced by carers are well recognized by health care professionals [9], many clinicians report that they do not know how to appropriately engage with carers or address their unique challenges [9,10]. Oncologists have reported a lack of education about communicating with carers [10], and suboptimal carer-clinician communication is common [11]. Some carers report being overlooked in medical consultations and feeling disempowered and unprepared in their caregiving role [12]. Clinician inclusion and support of carers have been reported as highly valued by both carers and patients [12].

Improving carer engagement and support needs to be addressed from multiple perspectives. Not only are clinicians uncertain about how to include carers in consultations [9] but also many carers often lack confidence and skills in caregiving [12,13], and some patients are unsure about what role their carer should assume in medical consultations and decision-making [14]. Therefore, interventions targeting *all* members of the clinician-patient-carer trio are needed.

Web-based delivery of education offers efficacy, efficiency, ability to undertake training in discrete periods, lower cost, flexibility, and greater reach than traditional face-to-face formats [15]. A systematic review of web-based health education by George et al [16] found web-based training for health professionals to be as effective as or better than face-to-face formats on outcomes such as knowledge, skills, and attitudes. Web-based communication skills interventions have been found to be effective in improving self-rated clinician confidence, communication skills, and knowledge among cancer clinicians [15]. For example, a web-based module developed by our group to educate nurses about managing conflict involving patients and carers (the Triadic Oncology [TRIO]–Conflict module) was found to improve cancer nurses' attitudes and confidence in interacting with carers [17].

Patients and carers can also benefit from web-based resources and educational tools [18]. A systematic review of digital psychosocial interventions for patients with cancer and carers found web-based interventions to be both feasible and acceptable [19]. Digital interventions for carers have been shown to improve carer outcomes, knowledge, and skills, with the additional benefit of being accessible from home, thus minimizing the demands on carers' time [20]. For example, a web-based psychosocial intervention for patients with cancer, Stress-Aktiv-Mindern (STREAM), has demonstrated beneficial patient outcomes including reduced stress and improved quality of life [21]. Similarly, the psychoeducational platform, Comprehensive Health Enhancement Support System (CHESS), has demonstrated favorable outcomes among carers such as significant reduction in negative mood and carer burden [22]. These beneficial effects were comparable to those of traditional psychoeducation interventions [23,24]. While STREAM and

XSL•FO
RenderX

CHESS demonstrate the efficacy of web-based patient and carer support, their focus is on *psychosocial* support. To date, there have been no web-based education modules dedicated to empowering and upskilling patients and carers in *carer-relevantcommunication and engagement* with cancer clinicians and in carer participation in cancer treatment decision-making. Therefore, we aimed to develop and evaluate a web-based learning tool to address these needs.

Interventions to support cancer carers are often difficult to implement in clinical practice and face barriers to implementation including problems with design, feasibility, acceptability, and cost [25]. One way to improve the acceptability and sustainability of an intervention is to use a co-design approach with the target population as stakeholders, to ensure that the program targets user needs and preferences. The person-based approach [26] ensures that intervention development is grounded in the perspectives and psychosocial context of end users via iterative, qualitative research with relevant stakeholders. This approach has been effectively used in the development of web-based health care interventions [27,28].

Objectives

This paper describes the development, iterative refinement, and user testing of evidence-based and interactive web-based interventions designed to improve engagement and communication with carers in cancer care. We have published the study protocol for a randomized controlled trial to test the efficacy of the e-Triadic Oncology (eTRIO) modules elsewhere [29]. However, necessary amendments to the planned randomized controlled trial due to the COVID-19 pandemic were made after publication of the protocol. The evaluation approach was revised to hybrid effectiveness and implementation studies using a pre-post, single-arm intervention design.

In this paper, we have reported about the development of web-based education modules for all 3 relevant stakeholder groups, including oncology health professionals and patients with cancer and carers (eTRIO for patients and carers [eTRIO-pc]).

Methods

Overview

The person-based co-design approach by Yardley et al [26] underpinned the module design. Development and user experience testing of the clinician (eTRIO) and patient-carer (eTRIO-pc) modules was undertaken in multiple cyclical phases of data collection, analysis, and integration, in a process of iterative refinement [30]. Consistent with the approach by Yardley et al [26], this involved (1) *planning*: development of module content based on evidence, qualitative interviews with stakeholders, and input from our expert advisory group; (2) *design*: iterative review and refinement based on advisory group feedback; and (3) *development and evaluation of acceptability and feasibility*: formal heuristic evaluation, System Usability Scale (SUS) questionnaire, and think-aloud review of the eTRIO modules by stakeholders (Figures 1 and 2). The final phase of implementation and trialing is currently being conducted in a separate pre-post evaluation study, which will be reported elsewhere.

XSL•FO
RenderX

Figure 1. e-Triadic Oncology (eTRIO; clinician) module development process. SUS: System Usability Scale.

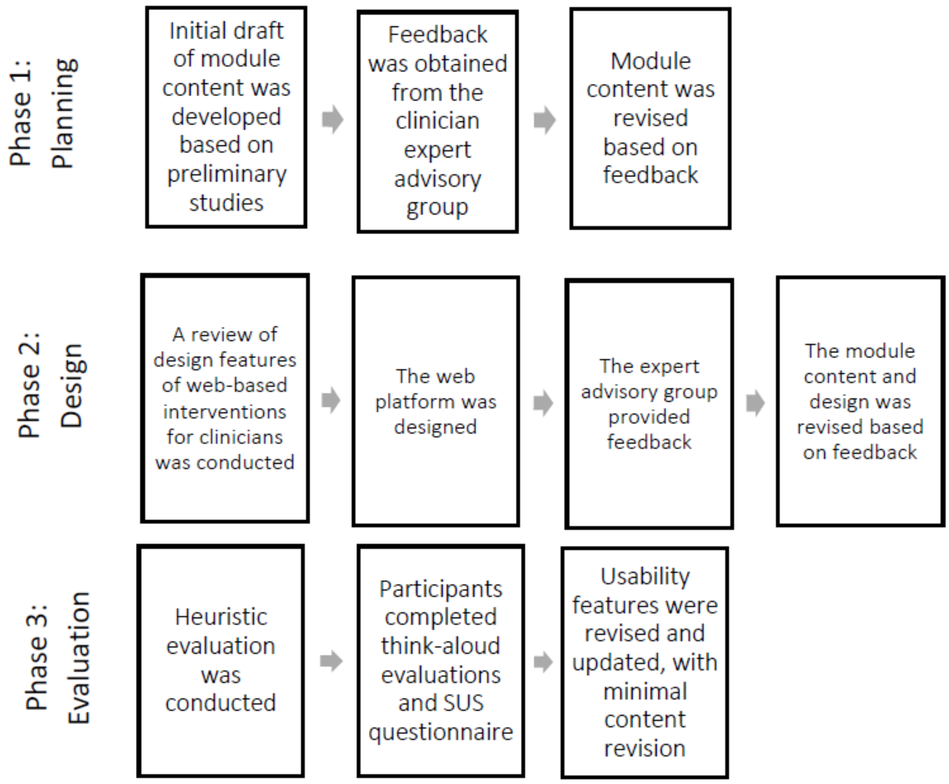
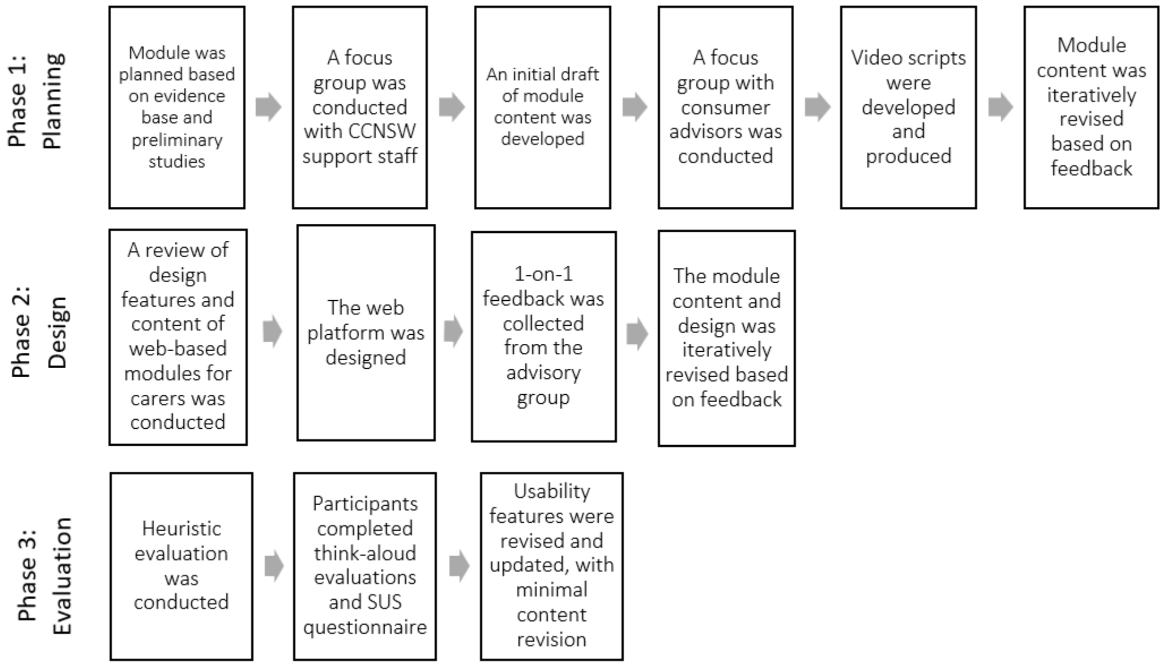


Figure 2. e-Triadic Oncology for patients and carers (eTRIO-pc) module development process. CCNSW: Cancer Council New South Wales; SUS: System Usability Scale.



XSL•FO
RenderX

Phase 1: Development of eTRIO and eTRIO-pc Module Content

Development of the eTRIO Clinician Module

The content of the modules was informed by our extensive Triadic Oncology (TRIO) research program, which includes a systematic review of carer involvement in consultations [31]; qualitative interviews with oncology clinicians, patients, and carers [9,14,32]; quantitative and qualitative analyses of audiotaped oncology consultations [11]; a conceptual framework of carer involvement in medical decisions [33]; and carer communication guidelines for clinicians (TRIO Guidelines) developed via a Delphi consensus process [34,35]. Key clinician training needs, strategies, and behaviors relevant to the module were ascertained through this extensive research program.

On the basis of this prior research, we developed an initial draft of the eTRIO content. The draft module comprised 14 study sections (1 introductory section and 13 strategy areas covered in the TRIO Guidelines [34,35]). A clinician expert advisory group was formed to provide feedback about the module content, comprising medical oncologists (3/13, 23%), oncology nurses (2/13, 15%), psychologists (2/13, 15%), a radiation oncologist (1/13, 8%), an oncology surgeon (1/13, 8%), and the research team comprising psycho-oncologists (4/13, 31%). Each member of the clinician expert advisory group reviewed a text-based draft of the module content and provided written feedback about each module section, including interactive activities, reflective exercises, and wording of strategies. Multiple teleconferences were conducted, where group members provided feedback about the content and structure of each section. Major changes were discussed with the group until consensus was reached. Feedback from the advisory group was collated, and the module content was iteratively refined.

Development of the eTRIO-pc Patient-Carer Module

The eTRIO-pc module content was drafted based on a review of current web-based guidance for carers about involvement in medical consultations [18], qualitative studies of patients and carers [9,14,32], and analyses of audiotaped consultations [11]. A meeting with the staff at a leading nongovernment cancer support and advocacy organization (n=5) was also conducted to inform the content of the eTRIO-pc initial draft. The staff members were asked to describe the key content that should be included in the eTRIO-pc module, based on their experience in supporting patients and carers via a telephone information and support service.

Consumer advisers (3/5, 60% cancer carers and 2/5, 40% patients with cancer) also provided iterative feedback about the module content during a half-day workshop and via email. Consumer advisers were asked to comment about whether the module content was understandable, the relevance of the module content and feasibility of the suggested strategies, the language, and tone of the module. All feedback from the Cancer Council New South Wales support staff and consumer advisers was collated and discussed with the project team until consensus was reached through revisions.

After the development and iterative revision of the module content was complete, video vignettes modeling key carer communication skills were developed to supplement the written content. Video vignettes have been demonstrated as an effective educational tool for patients and carers and can improve accessibility for those with low literacy [36,37]. We engaged a professional medical education and communication production company to develop a script covering key learning areas for carers, as determined by the consumer advisory groups. The script was iteratively reviewed by the research team, consumer advisers, and a physician to ensure that the videos aligned with the TRIO communication guidelines [34,35] and were clinically relevant and feasible.

Phase 2: Iterative Design, Review, and Refinement of eTRIO and eTRIO-pc Web-Based Modules

Design and Refinement of eTRIO Clinician Module

As shown in Figure 1, phase 2 involved consumer input and refinement of the modules. To translate the text-based content into an interactive web-based educational module, we studied the best practice principles for the delivery of e-learning to health professionals [16,38-40]. This included a review by de Leeuw et al [38] about e-learning features targeted at postgraduate medical students and health professionals completing ongoing professional development, which identified 6 domains of important elements for e-learning quality (preparation, design, communication, content, assessment, and maintenance). Informed by a previous review [38], we developed a base design and catalog of potential design features.

A prototype web platform was developed by a professional web development company. In 2 sessions conducted via Zoom (Zoom Video Communications), the clinician advisory group completed a walk-through of the module and provided comprehensive feedback. Their verbal and written feedback was collated and integrated into a revised web-based module.

Design and Refinement of the eTRIO-pc Patient-Carer Module

Similarly, as displayed in Figure 2, phase 2 involved the conversion of the text-based module content for patients and carers into an interactive web-based platform. We conducted a review of the content and design features of other available evidence-based web-based platforms for carers [18], drew on the evidence base surrounding education for carers [41-43], and received input from the consumer advisory group. To inform the website design, we reviewed the publicly available web-based resources for carers.

The final design features of eTRIO and eTRIO-pc were implemented by a professional web development company and included interactive activities, video vignettes, and text-based content. The clinician and consumer advisory groups were given access to the draft module, and its content and format were revised based on their extensive feedback. An expert in human-centered IT design was involved in all stages of development of the clinician and patient-carer modules.

XSL•FO
RenderX

Phase 3: Heuristic Evaluation and "Think Aloud" User Experience Evaluation of eTRIO and eTRIO-pc Web-Based Modules

As shown in Figures 1 and 2, phase 3 involved usability evaluations of the developed web-based module. We conducted a heuristic evaluation to discover technical and usability issues [44]. The modules were examined by the researchers to identify problems that did not comply with the usability principles recognized by Nielsen [45], which include consistency and standards, error prevention, and aesthetic and minimalist design. The severity and prevalence of the issues were ranked from 1 to 5, with a high rank indicating that the problem was a priority to fix, and the web platform was updated accordingly.

Usability and user experience testing for the penultimate versions of eTRIO and eTRIO-pc were conducted using think-aloud methodology with 11 participants, including clinicians, patients, and carers, all of whom were naïve to the TRIO Guidelines and modules. Think aloud is an effective evaluation method in which participants are provided with an interface and asked to verbalize their thoughts as they work through it [46,47]. Potential participants were identified through the research team's professional networks and via social media advertisements.

The consenting participants completed a demographic questionnaire and a 4-item self-report measure of health literacy [48]. Participants were provided access to the relevant eTRIO module and asked to speak aloud their thoughts and impressions as they were completing the module (think-aloud). These sessions were conducted face to face or via videoconferencing. After working through the module, participants completed the SUS [49]. Think-aloud evaluations were audio recorded and transcribed verbatim. Transcripts were qualitatively analyzed using thematic analysis [50], which involved familiarization with the transcripts, coding of salient initial ideas as codes, identification of patterns in the codes to generate themes and subthemes, and iterative review of the themes and subthemes to ensure a coherent and comprehensive thematic structure. This process was conducted collaboratively and through iterative discussion by RLP, PB, ZB, MB, and IJ. Themes were related to the following: usability and technical issues, positive aspects of design and function, attitudes toward the content of the program, and perspectives about the impact or implementation of the program. All transcripts were analyzed based on the established thematic framework and were grounded in illustrative quotations. Subsequently, the modules were iteratively refined based on this feedback.

Ethical Considerations

Ethics approval was obtained from the University of Sydney Human Research Ethics Committee (protocol 2015/468). Participants provided informed consent and were given the opportunity to opt out at any point in time. Participant data were deidentified. Participants were provided a gift card worth Aus $20 (US $13.22) as compensation for their time.

Results

This section describes the clinician, patient, and carer feedback; iterative revisions made; and lessons learned in the design and development of the eTRIO and eTRIO-pc modules.

Phase 1: Development of eTRIO and eTRIO-pc Module Content

eTRIO Clinician Module

The clinician advisory group members (n=13) emphasized the importance of the module being concise. They suggested more content for the introductory section such as including a broad and inclusive definition for "carers," content about culturally diverse carers, and more information about the legal and ethical aspects of involving carers. Clinicians also suggested the inclusion of self-reflections about one's own attitudes and potential biases toward carers. Additional suggestions included addressing the diversity of settings in which family or carer interactions can occur (eg, outside traditional outpatient consultations such as at the patient's bedside or via the telephone). Several clinicians stressed the importance of including clear learning outcomes and summaries for each of the 14 sections.

eTRIO-pc Patient-Carer Module

Cancer support staff (n=5) suggested a clear definition of the role of carers, tailoring based on the cultural backgrounds of patients and carers, and consideration of power imbalances that may exist in patient-carer relationships. They emphasized checking in on patient and carer emotions such as grief and distress, suggested that modules could include opportunities for self-reflection, and highlighted the need to include information about available support for carers.

The overall impression of the consumer advisory group (n=5) was that the language and tone of the draft module was very formal and academic; they wanted the tone to be more "personal," "empathetic," and "softer" and the language to be less prescriptive. They suggested additional strategies for patients with newly diagnosed cancer and carers, such as making notes during medical consultations, and suggested including quotes and stories from actual carers to illustrate examples.

Phase 2: Iterative Design, Review, and Refinement of eTRIO and eTRIO-pc Web-Based Modules

Overview

Table 1 describes the results from phase 2 using the e-learning design features by de Leeuw et al [38] applied to the eTRIO and eTRIO-pc modules.

XSL•FO
RenderX

Table 1. e-Learning design features identified by de Leeuw et al, as applied to the e-Triadic Oncology (eTRIO) and e-Triadic Oncology for patients and carers (eTRIO-pc) modules.

Elements of e-learning	Description	Use in eTRIO and eTRIO-pc
Preparation	Identifying the needs of the target audience	• Th research team conducted an extensive program of previous studies on the needs of carers • Stakeholder input, feedback, and evaluation
Design	Including elements of accessibility, reliability, user-friendly navigation, and visual appeal	• Web-based program, simple layout, and designed for easy use • Font, color, size, and layout are optimized for accessibility • User progress is saved when users log out • Website is designed and tested on various software and hardware
Communication	Communication with users and program facilitators	• Landing page introduces users to the learning objectives and goals (ie, communication skills and strategies, understanding carer roles, and benefits of carer involvement in cancer care) • Clear information about program use and navigation is included

Content	Including words, images, videos, interactive activities, summaries, and so on	• All modules include multimedia content such as several clinical vignette videos, audios, text, images, and interactive features. Interactive activities were designed, including the following:
		  • *eTRIO (clinician)*: sorting and drag-and-drop activities, true-or-false exercises, open-text written responses, click-to-expand sections, and identifying behaviors in a vignette video
		  • *eTRIO-pc (patient-carer)*: resources that can be individually tailored (eg, assembling a caregiving team, building a question prompt list, and checklist for patients and carers to discuss carer role), click-to-expand sections, and open-text written responses
		• Downloadable summaries are provided to allow access after completing or outside the module

Assessment	Assessing learning and acquiring feedback	• Each section of each module contains clear learning objectives, displayed on the first page of each section. For example, section 9 of eTRIO (clinician), related to the use of interpreters, states the following: "In this section you will explore reasons why patients/carers might resist professional language interpreters, and understand strategies to overcome these issues. You will learn practical strategies to engage and use formal interpretation services."
		• All modules include assessment activities to facilitate learning and reflection:
		  • *eTRIO (clinician)*: self-reflection and assessment of own attitudes and practices, true-or-false assessment of content with correct answers and explanations, multiple choice questions asking users to reflect about how they would navigate a clinical scenario, and open-ended responses
		  • *eTRIO-pc (patient-carer)*: self-assessment of emotions, opportunities to reflect about own preferences and attitudes and to plan future actions or behaviors, and open-text reflections about video vignettes modeling key skills

Maintenance	Providing long-term access and updating information and links	• Website is regularly maintained and updated • All users will have access to the program after completion of the training

eTRIO Clinician Module

During the transformation of content to a web-based module, features of e-learning [38] were applied as described in Table 1. The design features of other web-based clinician training modules were examined, revealing display, navigation, and interactive activity styles (eg, minimal use of text, prominent navigation buttons, and clickable and expandable content). Our team worked closely with graphic and web designers to develop a consistent color scheme and intuitive navigation system and aimed to minimize visual noise on each page. The refined content and design features were transformed into a web-based web platform.

All members of the clinician advisory group (13/13, 100%) commented that there was excessive content and that there would not be clinician appetite for web-based training that extended beyond 2 hours in total. The content was subsequently condensed, with the core content displayed with the option of more extensive content, which could be expanded for clinicians interested in deeper learning regarding an issue.

The final eTRIO clinician module comprises 14 sections (submodules), of which clinicians must complete a minimum of 8. The sections range between 3 and 15 minutes in duration. The following 4 sections were deemed to be mandatory by the clinician advisory group, based on their critical relevance to all clinicians: section 1—introduction, section 4—building rapport

XSL•FO
RenderX

with carers, section 7—supporting carers' emotional and informational needs, and section 10—managing conflicting patient-carer treatment preferences. Clinicians could select additional 4 sections based on their interest and preference. The eTRIO module requires approximately 1.5 to 2 hours to complete, as determined by multiple stakeholders working through the content and documenting the amount of time each section required to complete.

eTRIO-pc Patient-Carer Module

Consistent with the principles of computer-based teaching for adult learners by Lau [51], the web-based eTRIO-pc module was created by transforming the written content into interactive, engaging learning activities. Our review of carer resources demonstrated several useful stylistic, formatting, and usability features, for example, the use of bullet points to convey written information, 1-page displays eliminating the need to scroll, and use of simple navigation buttons. These features and principles of web-based education were collated and discussed with the team's academic IT expert and web developers to select and finalize the most appropriate features to be included. The resultant module prototype included video vignettes that could easily be played and paused, interactive activities such as "drag-and-drop" and "click to reveal" exercises, and type-your-response activities (Multimedia Appendix 1). We maintained consistency in design and formatting across the clinician, patient, and carer modules.

We sent the prototype to the members of the consumer advisory group (n=5), and they provided written feedback via email and offered additional personal quotes that could be included in the module to personalize the content. They re-emphasized the need for content that was empathetic and offered practical advice. The final eTRIO-pc modules contain 7 sections for patients and 11 sections for carers and requires approximately 1 to 1.5 hours to complete.

Phase 3: "Think Aloud" Usability Evaluation of eTRIO and eTRIO-pc Web-Based Modules

Heuristic Evaluation

Using the heuristic evaluation method [44], we identified 37 usability issues across the draft eTRIO and eTRIO-pc modules, and each was rated for severity. The main areas of the identified problems were as follows: (1) inconsistency of icons and redundancy in buttons (5/37, 14% of the issues; eg, inconsistent use of star and book icons to indicate the bookmark function), (2) buttons and interactions were not working (16/37, 43% of the issues; eg, nothing happens when the print button is clicked), (3) layout problems (6/37, 16% of the issues; eg, text is not aligned with the textbox), and (4) presentation of content (10/37, 27% of the issues; eg, color selection in the bar-slider activity may be confusing) [52]. Following this evaluation, problems with high severity and prevalence were prioritized, and all issues that could be corrected were fixed before conducting the think-aloud user evaluations.

Think-Aloud User Experience Evaluations

Overall, 11 individuals (n=5, 45% health professionals; n=3, 27% patients; and n=3, 27% carers) participated in the think-aloud evaluations in individual sessions lasting between 40 and 60 minutes. Participant characteristics are displayed in Table 2.

XSL•FO
RenderX

Table 2. Characteristics of participants of the think-aloud evaluations.

Participant category and characteristics	Values
Health professionals (n=5)	
Age (y), mean (SD; range)	47 (10.3; 35-58)
Sex, n (%)	
Female	4 (80)
Male	1 (20)
Profession, n (%)	
Physician	2 (40)
Nurse	3 (60)
Clinical expertise, n (%)	
Oncology	2 (40)
Palliative care	2 (40)
Geriatrics	1 (20)
Experience (years), mean (SD; range)	22 (9.8; 12-37)
Patients (n=3)	
Age (y), mean (SD; range)	65 (13.7; 50-77)
Sex (female), n (%)	3 (100)
Diagnosis, n (%)	
Kidney cancer	1 (33)
Colorectal cancer	1 (33)
Non-Hodgkins lymphoma	1 (33)
Cancer stage, n (%)	
Local	2 (67)
Advanced	1 (33)
Health literacy, n (%)	
Low	1 (33)
Medium	1 (33)
High	1 (33)
Carers (n=3)	
Age (y), mean (SD; range)	65 (8.7; 58-75)
Sex, n (%)	
Female	2 (67)
Male	1 (33)
Relationship with care recipient, n (%)	
Spouse or partner	2 (67)
Mother	1 (33)
Diagnosis of care recipient, n (%)	
Lung cancer	1 (33)
Multiple myeloma	1 (33)
Non-Hodgkins lymphoma	1 (33)
Cancer stage of care recipient, n (%)	
Local	1 (33)
Advanced	2 (67)

XSL•FO
RenderX

Participant category and characteristics	Values
Health literacy, n (%)	
Medium	1 (33)
High	2 (67)

eTRIO Clinician Module

The usability of the module was rated by 5 clinicians, with a mean SUS score of 75 (range 68-80), which is interpreted as good [49]. All clinicians gave high ratings to their ability to use the module independently without technical assistance. Clinicians identified technical and navigation issues, which were subsequently rectified (such as the side scroll bar not appearing, text appearing outside the text bubble, and a sliding bar not working responsively). For some, the use of specific web browsers corrected these issues. Clinicians described the overall navigation through the module as "straightforward." Formatting issues with font size and background color were highlighted. Clinicians commented that the ability to easily navigate back to certain sections to "refer back to later" was valued.

Content analysis of think-aloud evaluations revealed 7 categories related to clinicians' attitudes toward the design and formatting of eTRIO. Clinicians appreciated that the modules could be completed in small "snackable" periods in any order, that they could keep track of what sections were completed (*trackable*), and that they were able to refer back to any module at any time. Clinicians enjoyed the "clickable" activities where they interacted with the content. Despite attempts to make the sections as short as possible (average 5-10 min/section), a few clinicians still perceived them as "too long," with some stating that the videos were "slow" at times. They highlighted a preference for material that is brief, uses simple language, is easy to digest, and "skimmable." A few clinicians reported "glossing over" or "tuning out" when sections were perceived as very long. They suggested simplifying the language and formatting the text to highlight important information (eg, use of bullet points and bold and italic style). Revisions were made to the text to further improve conciseness, including rephrasing the core content, moving some content to the expandable 'additional information' section, and greater use of bullet points and bold text. Where possible, videos were edited to remove nonessential scenes. Most participants appreciated that the content and activities were relevant and "relatable" to them as clinicians, that claims were "supported" by evidence, and that the activities and media were "diverse" and varied to facilitate engagement and interest. Illustrative quotes are provided in Table 3.

XSL•FO
RenderX

Table 3. Illustrative quotes from think-aloud evaluations by clinicians.

Usability and content feature	Description	Illustrative quotes
Snackable	Ability to complete the module in small segments	• "So, you're saying you don't have to do it all in one go...oh, I think that's really important because you do get called away and the phone is ringing...because I know even with our mandatory online training in the past, you just [had to] forfeit [all progress] if you couldn't finish." [Nurse 2]
Trackable	Ability to know what has been completed and refer to the content later	• "It's nice to have things you can refer back to because this might trigger things that make you think oh yeah, I did read about that." [Physician 2]
Clickable	Importance of interactive content	• "I like this section - it's really good. I like that activity. I've never done one of those before - that's really good. [Interactive activity clicking points of rapport building throughout a video vignette]. You definitely engage a thousand percent more with the activities." [Physician 2] • "I think [the activities] are quite good because at least you are giving people a little bit more of themselves...I think it's good to have that interaction rather than just reading...that gets a bit boring. And then, that you ask people to actually write something is good." [Nurse 2]
Skimmable	Importance of simple, concise language	• "After reading articles all day I don't want to read something that has too much jargon in it...Go back and simplify the language...when I read something apart from patients notes, I skim it. So, it's got to be something that I can get the message with a glance." [Physician 1] • "Uhm why I am I finding it difficult to understand? I think it could be worded more simply." [Physician 2] • "Yeah. I hate the time pressure...It's so built into our working day, it's like get on, get it done, that you gloss over so much. I actually didn't realize before doing this how much I gloss over...I probably would watch [the video] to the end but there's a part of me thinking yeah it's going on a little bit." [Physician 2]
Relatable	Relevance of content to the user	• "Yeah, I like that there is the suggestions of things to say. That makes it really relatable - I think those are good." [Physician 2] • "I like scenarios...Just sort of triggers you to think a little bit more rather than just reading through something. I think the scenario allows me to put it into practice or put it into place a little bit more." [Nurse 3]
Supported	Evidence-based content	• "I like the use of the quotes. It gives a bit of a supportive evidence to it, as nurse I like that...It has got some stats [statistics] there...When you hover over it...it gives the reference." [Nurse 3]
Diverse	Importance of variety in media and activities	• "Oh a video, that's interesting, it's sort of mixing it up, it's nice to have the different things." [Physician 2]

eTRIO-pc Patient and Carer Module

The average SUS score as rated by 6 patients and carers was 78 (SD 16.2; range 55-97.5), which is interpreted as good [49]. Patients and carers were generally happy with the content and usability of the eTRIO-pc module. They commented that the content was relatable and were pleased by the emphasis placed on carers. Overall, they found the web platform easy to navigate and enjoyed the interactive activities; however, 1 (17%) of the 6 patients found the interface to be "overwhelming." A major critique of the formatting and layout was that the pages were "too busy" and contained excessive information. Illustrative quotes are provided in Table 4.

The final eTRIO and eTRIO-pc modules were updated based on this feedback. All technical and navigation issues were addressed by the web developers.

For both modules, the text was condensed and reformatted with the use of bold and italic style to highlight the important points and allow for easier reading and a more streamlined user interface.

XSL•FO
RenderX

Table 4. Illustrative quotes from think-aloud evaluations by patients and carers.

Usability and content feature	Description	Illustrative quotes
Snackable	Ability to complete the module in small segments	• "Looking at this dashboard I like it that it tells you how long each part is going to take just so you know in advance. You're busy and maybe you just have time to do half of it and then you can sort of plan how you're going to tackle it." [Carer 1]
Clickable	Importance of interactive content	• "Some of the activities like the questions, I really liked. The ones where you wrote down what you thought the carer might do for you if you then use it as a communication tool, really good as well." [Patient 1]
Usable	Ease of navigation	• "I think [navigation] is pretty easy and straight forward. I think anybody who's used to doing online training, modules and so on will probably find it really easy." [Carer 1]
Relatable	Relevance of content to the user	• "I think this is a very useful slide. When we went in to our first meeting we were just there, me and my son did this." [Carer 1]
Visually simple	Cleanness of layout, formatting, and images	• "I think you are making this page very busy with text and it's a bit confronting." [Carer 2] • "It is pretty text-heavy and I guess that I am more of a visual learner so it might be nice to have some more pictures, icons, to make it a little bit more visually appealing." [Patient 1]

Final Web Platform Design and Content Summary

The eTRIO modules reflect the reported informational needs of health professionals, patients with cancer, and carers. A full description of the module content has been published elsewhere [29]. The eTRIO modules have been rigorously designed to be easy to use, require minimal time commitment, and be flexible in terms of when and how the platform can be used. The modules are optimized for use on a computer but can also be used on a smartphone or tablet. Some notable features include the following: navigation buttons and a progress bar along the bottom of the page, expandable content for those who want deeper information about a specific topic, and downloadable summaries and lists. Notable interactive activities include the following: testing of knowledge through true-or-false exercises, identifying specific behaviors in a short video vignette, and building a question prompt list. Refer to Table 5 for descriptions and images of key features; full explanations of the interactive activities are provided in Multimedia Appendices 1 and 2.

Table 5. Key features of the e-Triadic Oncology (eTRIO) modules.

Feature	Description	Images
Interactive activities	Includes self-reflection, knowledge tests, and free-text responses	
Learning outcomes	The eTRIO clinician module features sign-posted learning outcomes at the beginning of each section	
Downloadable content	Includes materials and personalized checklists for patients and carers and downloadable summaries for clinicians	and
Video vignettes	Realistic scenarios modeling communication skills	
Intuitive navigation features	Navigation buttons are explained in the module's introduction	

XSL•FO
RenderX

Discussion

Principal Findings

The web-based modules described in this paper represent a crucial step in the development and design of education for clinicians, patients, and carers that is evidence based, practical, and interactive and can be easily disseminated. Drawing on the evidence for best practice web-based learning design [38,51], we sought input from a variety of stakeholders to develop a unique learning experience strongly informed by the needs of the target populations. Rigor was ensured via 3 stages of development in which module content and design were continually revised and refined. Overall, participants were positive about the content and interface. The final prototype was appraised as highly acceptable, relevant, and feasible among the small sample of users; however, more studies are needed to confirm this and to ascertain the effectiveness of the intervention. We are currently conducting a pre-post evaluation of these modules to explore their potential effectiveness in improving communication within the patient-carer-clinician trio.

Lessons Learned

Throughout the development and design of these modules, we observed the specific needs and preferences of end users. The person-based approach to developing eTRIO and eTRIO-pc was highly dynamic, and the modules underwent numerous iterations throughout all phases of the design process, which included the involvement of consumers and user-driven evaluations. While there are multiple approaches to developing health interventions, the benefits of the person-based approach include grounding the design in user contexts and lived experiences, integrating feedback based on the actual use of an intervention, and investigating user needs and perspectives beyond just the usability of the intervention [26]. The utility of the person-based approach has been extolled in recent studies [53-55] and is supported by the findings of this study. The eTRIO development process (Figures 1 and 2) provided the necessary building blocks to revise and refine the module for effective use in the real world. Consistent with other studies [56,57], we found that the collaborative co-design process led to positive evaluations of acceptability and usability and high levels of end-user satisfaction.

As highlighted in the person-based approach, the 3 user groups (clinicians, patient, and carers) demonstrated diverse learning preferences and needs. This was accommodated via tailoring the formatting or content to the strengths and contextual demands of different user groups and differentiating the content based on user needs. We found that clinicians had a strong desire for content that was written in simple, concise, and “sharp” language; could be “skim read”; and could be completed in brief, “snack-sized” sections. For example, clinicians in our advisory group often stressed that they lacked time and that training needed to be short, precise, and able to be stopped and restarted due to interruptions. On the other hand, the structure and time demands of training appeared to be less important to patients and carers. Instead, these groups emphasized the need for the module to be easy to use and navigate and for the content to be more conversational, empathetic, and in plain language (in contrast to the preferences of clinicians). Clinicians in our study valued the integration of academic literature and referencing, whereas some carers advocated for greater inclusion of carer experiences and quotes. The preferences of carers in our study are consistent with previous studies, which have similarly found that carers often prefer web-based education to have an empathetic and supportive tone, the web program to be easy to navigate, and the integration of other carers’ experiences into the content [58-60]. While several differences were identified between the clinician and carer user groups, there were also several similarities across all user groups in how the web-based modules should be structured and delivered. This is reflected in the evidence base, where health professionals, patients, and carers alike report that they prefer flexible, self-paced delivery of web-based programs that are interactive and include a variety of activities across media (visual, written, and auditory) [19,38]. These detailed insights are valuable in designing future training modules to facilitate their acceptability among users in each specific group.

The final interface used design principles to ensure engaging and interactive content. There is robust empirical evidence suggesting that interactivity in e-learning improves quality, efficacy, and learning outcomes [38,61]. For example, users of a web-based public health program had better learning outcomes when they used a gamified, interactive version featuring responsive design, learning challenges, visible progress, and rapid feedback compared to those using a minimally interactive, survey-based program [62]. Such interactivity was also demonstrated as important for users of the eTRIO modules. For example, in the initial design phases, when content was largely text based, the advisory committee members noted how dense the information appeared. While this was never intended to be the final format of the educational intervention, comments obtained from users in phase 1 highlighted the limitations of passive, didactic, text-heavy information. There is evidence suggesting that people do not learn effectively when information is given without any opportunity to reflect on, test, or demonstrate their knowledge and views [63]. Interactive activities, including assessments of learning and personal reflection activities, offer users the opportunity to reflect and reinforce their learning and become active participants in their education rather than passive consumers of information. Multimedia Appendices 1 and 2 display the engaging interactive activities that were acceptable to eTRIO users, which may be used in other web-based learning interventions and resources.

For both the clinician and patient-carer modules, we also incorporated a variety of media (text, audio, video, graphics, and images to cater to different learning styles and preferences. There is evidence suggesting that the use of multimedia may increase user satisfaction, acceptability, and engagement [64,65] and thus may improve adherence and broad implementation. The modules were designed such that users could navigate through them at their own pace and read, view, and explore the sections in a self-directed manner based on how they like to engage with and process content. For example, we found that users had mixed responses to the videos embedded in the training module. Some users commented that the videos were

XSL•FO
RenderX

very long and that they would mentally switch off or skip them. Others claimed to be “visual learners” and thoroughly enjoyed the opportunity to observe scenarios in this format, especially because the videos included interactive “trigger” questions such as “What would you do next?” where they were required to apply some of their learning to a scenario. This approach has been used in other web-based health interventions [66,67], which include complementary text, images, videos, audios, and interactive content to convey the educational content and cater to these diverse user preferences.

Strengths and Limitations

A thoughtful process of iterative design was conducted over a 2-year period, ultimately producing a suite of web-based interventions intended to improve communication between cancer clinicians, patients, and carers. However, important limitations should be noted. While extensive end-user feedback was collected through iterative feedback from clinician and consumer advisory groups, the sample size of participants (patients with cancer and carers) naive to the modules in phase 3 was small, and there was limited diversity among consumer advisers and participants. In addition, we did not measure the computer literacy of the participants, which may have impacted their views about the program’s usability. Thus, the attitudes and preferences of participants may not be reflective of the wider population. For example, we were unable to recruit a carer with low health literacy, and there was an overrepresentation of women.

Further usability and acceptability testing is currently underway in a larger study with a more diverse sample of patients and carers. Recruitment of participants in phase 3 was conducted through professional networks and social media, and therefore, the participants may have had a strong interest web-based learning or carer communication, which could have biased their views. This study focused only on development and user testing, and therefore, no assessment of the effectiveness or uptake of the modules has been conducted. Larger evaluation studies of the modules are currently being conducted, which will provide insight into the utility of the eTRIO modules in improving carer-related communication and inclusion.

Finally, while most patients with cancer have a carer or support person, some patients do not. Further studies are required to better understand the needs of people without a carer, which is beyond the scope of this study.

Future Directions

The eTRIO and eTRIO-pc modules are now undergoing pre-post evaluation with additional qualitative learner feedback to inform the broad implementation and uptake of these educational resources.

Conclusions

By including and being receptive to the needs of our user groups throughout the design process, we were able to create interventions that end users are likely to be more engaged and satisfied with.

Acknowledgments

This study was supported by Cancer Australia and Cancer Council New South Wales Grant, through the Priority-Driven Collaborative Cancer Research Scheme (project 1146383).

Conflicts of Interest

None declared.

Multimedia Appendix 1
e-Triadic Oncology patient and carer module features.
[DOCX File , 1304 KB - mededu_v10i1e50118_app1.docx]

Multimedia Appendix 2
e-Triadic Oncology clinician module features.
[DOCX File , 1069 KB - mededu_v10i1e50118_app2.docx]

References

1. Castro A, Arnaert A, Moffatt K, Kildea J, Bitzas V, Tsimicalis A. "Informal caregiver" in nursing: an evolutionary concept analysis. ANS Adv Nurs Sci 2023;46(1):E29-E42. [doi: 10.1097/ANS.0000000000000439] [Medline: 36006014]
2. Molassiotis A, Wang M. Understanding and supporting informal cancer caregivers. Curr Treat Options Oncol 2022 Apr;23(4):494-513 [FREE Full text] [doi: 10.1007/s11864-022-00955-3] [Medline: 35286571]
3. Geng HM, Chuang DM, Yang F, Yang Y, Liu WM, Liu LH, et al. Prevalence and determinants of depression in caregivers of cancer patients: a systematic review and meta-analysis. Medicine (Baltimore) 2018 Sep;97(39):e11863 [FREE Full text] [doi: 10.1097/MD.0000000000011863] [Medline: 30278483]
4. Bedaso A, Dejenu G, Duko B. Depression among caregivers of cancer patients: updated systematic review and meta-analysis. Psychooncology 2022 Nov;31(11):1809-1820 [FREE Full text] [doi: 10.1002/pon.6045] [Medline: 36209385]

XSL•FO
RenderX

5. Lambert S, Girgis A, Descallar J, Levesque JV, Jones B. Trajectories of mental and physical functioning among spouse caregivers of cancer survivors over the first five years following the diagnosis. Patient Educ Couns 2017 Jun;100(6):1213-1221. [doi: 10.1016/j.pec.2016.12.031] [Medline: 28089132]
6. Lambert SD, Girgis A. Unmet supportive care needs among informal caregivers of patients with cancer: opportunities and challenges in informing the development of interventions. Asia Pac J Oncol Nurs 2017 Apr;4(2):136-139 [FREE Full text] [doi: 10.4103/2347-5625.204485] [Medline: 28503646]
7. Høeg BL, Frederiksen MH, Andersen EA, Saltbæk L, Friberg AS, Karlsen RV, et al. Is the health literacy of informal caregivers associated with the psychological outcomes of breast cancer survivors? J Cancer Surviv 2021 Oct;15(5):729-737. [doi: 10.1007/s11764-020-00964-x] [Medline: 33169190]
8. Kershaw T, Ellis KR, Yoon H, Schafenacker A, Katapodi M, Northouse L. The interdependence of advanced cancer patients' and their family caregivers' mental health, physical health, and self-efficacy over time. Ann Behav Med 2015 Dec;49(6):901-911 [FREE Full text] [doi: 10.1007/s12160-015-9743-y] [Medline: 26489843]
9. Laidsaar-Powell R, Butow P, Bu S, Fisher A, Juraskova I. Oncologists' and oncology nurses' attitudes and practices towards family involvement in cancer consultations. Eur J Cancer Care (Engl) 2017 Jan 01;26(1):e12470. [doi: 10.1111/ecc.12470] [Medline: 26931469]
10. Røen I, Stifoss-Hanssen H, Grande G, Kaasa S, Sand K, Knudsen AK. Supporting carers: health care professionals in need of system improvements and education - a qualitative study. BMC Palliat Care 2019 Jul 16;18(1):58 [FREE Full text] [doi: 10.1186/s12904-019-0444-3] [Medline: 31311536]
11. Laidsaar-Powell R, Butow P, Bu S, Dear R, Fisher A, Coll J, et al. Exploring the communication of oncologists, patients and family members in cancer consultations: development and application of a coding system capturing family-relevant behaviours (KINcode). Psychooncology 2016 Jul 30;25(7):787-794. [doi: 10.1002/pon.4003] [Medline: 26514374]
12. McCarthy B. Family members of patients with cancer: what they know, how they know and what they want to know. Eur J Oncol Nurs 2011 Dec;15(5):428-441. [doi: 10.1016/j.ejon.2010.10.009] [Medline: 21094087]
13. Morris S, Thomas C. The carer's place in the cancer situation: where does the carer stand in the medical setting? Eur J Cancer Care (Engl) 2001 Jun;10(2):87-95. [doi: 10.1046/j.1365-2354.2001.00249.x] [Medline: 11829054]
14. Laidsaar-Powell R, Butow P, Bu S, Charles C, Gafni A, Fisher A, et al. Family involvement in cancer treatment decision-making: a qualitative study of patient, family, and clinician attitudes and experiences. Patient Educ Couns 2016 Jul;99(7):1146-1155. [doi: 10.1016/j.pec.2016.01.014] [Medline: 26873544]
15. Berg M, Ngune I, Schofield P, Grech L, Juraskova I, Strasser M, et al. Effectiveness of online communication skills training for cancer and palliative care health professionals: a systematic review. Psychooncology 2021 Sep;30(9):1405-1419 [FREE Full text] [doi: 10.1002/pon.5702] [Medline: 33909328]
16. George PP, Zhabenko O, Kyaw BM, Antoniou P, Posadzki P, Saxena N, et al. Online digital education for postregistration training of medical doctors: systematic review by the digital health education collaboration. J Med Internet Res 2019 Feb 25;21(2):e13269 [FREE Full text] [doi: 10.2196/13269] [Medline: 30801252]
17. Laidsaar-Powell R, Keast R, Butow P, Mahony J, Hagerty F, Townsend J, et al. Improving breast cancer nurses' management of challenging situations involving family carers: pilot evaluation of a brief targeted online education module (TRIO-Conflict). Patient Educ Couns 2021 Dec;104(12):3023-3031. [doi: 10.1016/j.pec.2021.04.003] [Medline: 33941422]
18. Keast R, Butow PN, Juraskova I, Laidsaar-Powell R. Online resources for family caregivers of cognitively competent patients: a review of user-driven reputable health website content on caregiver communication with health professionals. Patient Educ Couns 2020 Dec;103(12):2408-2419. [doi: 10.1016/j.pec.2020.04.026]
19. Heynsbergh N, Heckel L, Botti M, Livingston PM. Feasibility, useability and acceptability of technology-based interventions for informal cancer carers: a systematic review. BMC Cancer 2018 Mar 02;18(1):244 [FREE Full text] [doi: 10.1186/s12885-018-4160-9] [Medline: 29499663]
20. Zhai S, Chu F, Tan M, Chi NC, Ward T, Yuwen W. Digital health interventions to support family caregivers: an updated systematic review. Digit Health 2023 May 19;9:20552076231171967 [FREE Full text] [doi: 10.1177/20552076231171967] [Medline: 37223775]
21. Grossert A, Urech C, Alder J, Gaab J, Berger T, Hess V. Web-based stress management for newly diagnosed cancer patients (STREAM-1): a randomized, wait-list controlled intervention study. BMC Cancer 2016 Nov 03;16(1):838 [FREE Full text] [doi: 10.1186/s12885-016-2866-0] [Medline: 27809796]
22. DuBenske LL, Gustafson DH, Namkoong K, Hawkins RP, Atwood AK, Brown RL, et al. CHESS improves cancer caregivers' burden and mood: results of an eHealth RCT. Health Psychol 2014 Oct;33(10):1261-1272 [FREE Full text] [doi: 10.1037/a0034216] [Medline: 24245838]
23. Kaltenbaugh D, Klem M, Hu L, Turi E, Haines AJ, Hagerty Lingler J. Using Web-based interventions to support caregivers of patients with cancer: a systematic review. Oncol Nurs Forum 2015 Mar;42(2):156-164. [doi: 10.1188/15.ONF.156-164] [Medline: 25806882]
24. Northouse LL, Katapodi MC, Song L, Zhang L, Mood DW. Interventions with family caregivers of cancer patients: meta-analysis of randomized trials. CA Cancer J Clin 2010 Aug 16;60(5):317-339 [FREE Full text] [doi: 10.3322/caac.20081] [Medline: 20709946]

XSL•FO
RenderX

25. Ugalde A, Gaskin CJ, Rankin NM, Schofield P, Boltong A, Aranda S, et al. A systematic review of cancer caregiver interventions: appraising the potential for implementation of evidence into practice. Psychooncology 2019 Apr 07;28(4):687-701 [FREE Full text] [doi: 10.1002/pon.5018] [Medline: 30716183]
26. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
27. Beatty L, Koczwara B, Butow P, Turner J, Girgis A, Schofield P, et al. Development and usability testing of a web-based psychosocial intervention for women living with metastatic breast cancer: finding my way-advanced. J Cancer Surviv 2021 Jun 15;15(3):403-409. [doi: 10.1007/s11764-021-01019-5] [Medline: 33723741]
28. Wagner LI, Duffecy J, Begale M, Victorson D, Golden SL, Smith ML, et al. Development and refinement of FoRtitude: an eHealth intervention to reduce fear of recurrence among breast cancer survivors. Psychooncology 2020 Jan 08;29(1):227-231 [FREE Full text] [doi: 10.1002/pon.5297] [Medline: 31760667]
29. Juraskova I, Laidsaar-Powell R, Keast R, Schofield P, Costa DS, Kay J, et al. eTRIO trial: study protocol of a randomised controlled trial of online education modules to facilitate effective family caregiver involvement in oncology. BMJ Open 2021 May 28;11(5):e043224 [FREE Full text] [doi: 10.1136/bmjopen-2020-043224] [Medline: 34049902]
30. Gould JD, Lewis C. Designing for usability: key principles and what designers think. Commun ACM 1985;28(3):300-311. [doi: 10.1145/3166.3170]
31. Laidsaar-Powell R, Butow P, Bu S, Charles C, Gafni A, Lam W, et al. Physician-patient-companion communication and decision-making: a systematic review of triadic medical consultations. Patient Educ Couns 2013 Apr;91(1):3-13. [doi: 10.1016/j.pec.2012.11.007] [Medline: 23332193]
32. Laidsaar-Powell R, Butow P, Bu S, Fisher A, Juraskova I. Attitudes and experiences of family involvement in cancer consultations: a qualitative exploration of patient and family member perspectives. Support Care Cancer 2016 Oct 30;24(10):4131-4140. [doi: 10.1007/s00520-016-3237-8] [Medline: 27137213]
33. Laidsaar-Powell R, Butow P, Charles C, Gafni A, Entwistle V, Epstein R, et al. The TRIO framework: conceptual insights into family caregiver involvement and influence throughout cancer treatment decision-making. Patient Educ Couns 2017 Nov;100(11):2035-2046. [doi: 10.1016/j.pec.2017.05.014] [Medline: 28552193]
34. Laidsaar-Powell R, Butow P, Boyle F, Juraskova I. Facilitating collaborative and effective family involvement in the cancer setting: guidelines for clinicians (TRIO Guidelines-1). Patient Educ Couns 2018 Jun;101(6):970-982. [doi: 10.1016/j.pec.2018.01.019] [Medline: 29526389]
35. Laidsaar-Powell R, Butow P, Boyle F, Juraskova I. Managing challenging interactions with family caregivers in the cancer setting: guidelines for clinicians (TRIO Guidelines-2). Patient Educ Couns 2018 Jun;101(6):983-994. [doi: 10.1016/j.pec.2018.01.020] [Medline: 29526388]
36. Cruz-Oliver DM, Pacheco Rueda A, Viera-Ortiz L, Washington KT, Oliver DP. The evidence supporting educational videos for patients and caregivers receiving hospice and palliative care: a systematic review. Patient Educ Couns 2020 Sep;103(9):1677-1691 [FREE Full text] [doi: 10.1016/j.pec.2020.03.014] [Medline: 32241583]
37. Gysels M, Higginson IJ. Interactive technologies and videotapes for patient education in cancer care: systematic review and meta-analysis of randomised trials. Support Care Cancer 2007 Jan 23;15(1):7-20. [doi: 10.1007/s00520-006-0112-z] [Medline: 17024500]
38. de Leeuw RA, Westerman M, Nelson E, Ket JC, Scheele F. Quality specifications in postgraduate medical e-learning: an integrative literature review leading to a postgraduate medical e-learning model. BMC Med Educ 2016 Jul 08;16(1):168 [FREE Full text] [doi: 10.1186/s12909-016-0700-7] [Medline: 27390843]
39. Hillen MA, van Vliet LM, de Haes HC, Smets EM. Developing and administering scripted video vignettes for experimental research of patient-provider communication. Patient Educ Couns 2013 Jun;91(3):295-309. [doi: 10.1016/j.pec.2013.01.020] [Medline: 23433778]
40. Scott KM, Baur L, Barrett J. Evidence-based principles for using technology-enhanced learning in the continuing professional development of health professionals. J Contin Educ Health Prof 2017;37(1):61-66. [doi: 10.1097/CEH.0000000000000146] [Medline: 28252469]
41. Berry LL, Dalwadi SM, Jacobson JO. Supporting the supporters: what family caregivers need to care for a loved one with cancer. J Oncol Pract 2017 Jan;13(1):35-41. [doi: 10.1200/JOP.2016.017913] [Medline: 27997304]
42. Sherwood PR, Given BA, Given CW. Caregiver Knowledge and Skills. In: Bellizzi KM, Gosney M, editors. Cancer and Aging Handbook: Research and Practice. Hoboken, NJ: Wiley-Blackwell; 2012:445-458.
43. Connor KI, Siebens HC, Chodosh J. Person-centered approaches to caregiving. In: Gaugler JE, Kane RL, editors. Family Caregiving in the New Normal. San Diego, CA: Academic Press; 2015:251-268.
44. Alonso-Ríos D, Mosqueira-Rey E, Moret-Bonillo V. A systematic and generalizable approach to the heuristic evaluation of user interfaces. Int J Hum Comput Interact 2018 Jan 24;34(12):1169-1182. [doi: 10.1080/10447318.2018.1424101]
45. Nielsen J. Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1994 Presented at: CHI '94; April 24-28, 1994; Boston, MA p. 152-158 URL: https://dl.acm.org/doi/10.1145/191666.191729 [doi: 10.1145/191666.191729]
46. Duncker K, Lees LS. On problem-solving. Psychol Monogr 1945;58(5):1-113. [doi: 10.1037/h0093599]

XSL•FO
RenderX

47. Nielsen J. Usability Engineering. Cambridge, MA: Morgan Kaufmann; 1994.
48. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. Fam Med 2004 Sep;36(8):588-594 [FREE Full text] [Medline: 15343421]
49. Brooke J. SUS: A 'Quick and Dirty' Usability Scale. Boca Raton, FL: CRC Press; 1995.
50. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Jun 13;11(4):589-597. [doi: 10.1080/2159676x.2019.1628806]
51. Lau KH. Computer-based teaching module design: principles derived from learning theories. Med Educ 2014 Mar;48(3):247-254. [doi: 10.1111/medu.12357] [Medline: 24528459]
52. Bonnaudet M. A usability evaluation of TRIO's e-learning modules enhancing the communication between cancer patients, clinicians and carers. KTH Royal Institute of Technology. 2020. URL: https://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1466296&dswid=-4538 [accessed 2024-02-29]
53. Murfield J, Moyle W, O'Donovan A. Planning and designing a self-compassion intervention for family carers of people living with dementia: a person-based and co-design approach. BMC Geriatr 2022 Jan 14;22(1):53 [FREE Full text] [doi: 10.1186/s12877-022-02754-9] [Medline: 35031015]
54. O'Cathain A, Croot L, Sworn K, Duncan E, Rousseau N, Turner K, et al. Taxonomy of approaches to developing interventions to improve health: a systematic methods overview. Pilot Feasibility Stud 2019 Mar 12;5(1):41 [FREE Full text] [doi: 10.1186/s40814-019-0425-6] [Medline: 30923626]
55. Band R, Bradbury K, Morton K, May C, Michie S, Mair FS, et al. Intervention planning for a digital intervention for self-management of hypertension: a theory-, evidence- and person-based approach. Implement Sci 2017 Feb 23;12(1):25 [FREE Full text] [doi: 10.1186/s13012-017-0553-4] [Medline: 28231840]
56. Yardley L, Ainsworth B, Arden-Close E, Muller I. The person-based approach to enhancing the acceptability and feasibility of interventions. Pilot Feasibility Stud 2015 Oct 26;1(1):37 [FREE Full text] [doi: 10.1186/s40814-015-0033-z] [Medline: 27965815]
57. Bradbury K, Morton K, Band R, van Woezik A, Grist R, McManus RJ, et al. Using the person-based approach to optimise a digital intervention for the management of hypertension. PLoS One 2018;13(5):e0196868 [FREE Full text] [doi: 10.1371/journal.pone.0196868] [Medline: 29723262]
58. Köhle N, Drossaert CH, Oosterik S, Schreurs KM, Hagedoorn M, van Uden-Kraan CF, et al. Needs and preferences of partners of cancer patients regarding a web-based psychological intervention: a qualitative study. JMIR Cancer 2015 Dec 29;1(2):e13 [FREE Full text] [doi: 10.2196/cancer.4631] [Medline: 28410157]
59. Vaughan C, Trail TE, Mahmud A, Dellva S, Tanielian T, Friedman E. Informal caregivers' experiences and perceptions of a web-based peer support network: mixed-methods study. J Med Internet Res 2018 Aug 28;20(8):e257 [FREE Full text] [doi: 10.2196/jmir.9895] [Medline: 30154074]
60. Teles S, Ferreira A, Paúl C. Attitudes and preferences of digitally skilled dementia caregivers towards online psychoeducation: a cross-sectional study. Behav Inf Technol 2022 Jan 09;42(4):345-359. [doi: 10.1080/0144929x.2021.2021285]
61. Chang V. Review and discussion: E-learning for academia and industry. Int J Inf Manag 2016 Jun;36(3):476-485. [doi: 10.1016/j.ijinfomgt.2015.12.007]
62. Trevors G, Ladhani F. It's contagious! examining gamified refutation texts, emotions, and knowledge retention in a real-world public health education campaign. Discourse Process 2022 Jun 27;59(5-6):401-416. [doi: 10.1080/0163853x.2022.2085477]
63. Karpicke JD, Roediger HL. Repeated retrieval during learning is the key to long-term retention. J Mem Lang 2007 Aug;57(2):151-162. [doi: 10.1016/j.jml.2006.09.004]
64. Zhang D. Interactive multimedia-based e-learning: a study of effectiveness. Am J Distance Educ 2005 Sep;19(3):149-162. [doi: 10.1207/s15389286ajde1903_3]
65. Violante MG, Vezzetti E. Virtual interactive e-learning application: an evaluation of the student satisfaction. Comput Appl Eng Educ 2013 Aug 13;23(1):72-91. [doi: 10.1002/cae.21580]
66. Scott AF, Ayers S, Pluye P, Grad R, Sztramko R, Marr S, et al. Impact and perceived value of iGeriCare e-learning among dementia care partners and others: pilot evaluation using the IAM4all questionnaire. JMIR Aging 2022 Dec 22;5(4):e40357 [FREE Full text] [doi: 10.2196/40357] [Medline: 36150051]
67. Heynsbergh N, Heckel L, Botti M, Livingston PM. A smartphone app to support carers of people living with cancer: a feasibility and usability study. JMIR Cancer 2019 Jan 31;5(1):e11779 [FREE Full text] [doi: 10.2196/11779] [Medline: 30702432]

Abbreviations

CHESS: Comprehensive Health Enhancement Support System
eTRIO: e-Triadic Oncology
eTRIO-pc: e-Triadic Oncology for patients and carers
STREAM: Stress-Aktiv-Mindern
SUS: System Usability Scale
TRIO: Triadic Oncology

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 20.06.23; peer-reviewed by A Castro, Y Asada; comments to author 29.09.23; revised version received 27.11.23; accepted 31.01.24; published 17.04.24.

Please cite as:
Laidsaar-Powell R, Giunta S, Butow P, Keast R, Koczwara B, Kay J, Jefford M, Turner S, Saunders C, Schofield P, Boyle F, Yates P, White K, Miller A, Butt Z, Bonnaudet M, Juraskova I
Development of Web-Based Education Modules to Improve Carer Engagement in Cancer Care: Design and User Experience Evaluation of the e-Triadic Oncology (eTRIO) Modules for Clinicians, Patients, and Carers
JMIR Med Educ 2024;10:e50118
URL: https://mededu.jmir.org/2024/1/e50118
doi:10.2196/50118
PMID:38630531

©Rebekah Laidsaar-Powell, Sarah Giunta, Phyllis Butow, Rachael Keast, Bogda Koczwara, Judy Kay, Michael Jefford, Sandra Turner, Christobel Saunders, Penelope Schofield, Frances Boyle, Patsy Yates, Kate White, Annie Miller, Zoe Butt, Melanie Bonnaudet, Ilona Juraskova. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.04.2024. 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.

XSL•FO
RenderX

Impact of a New Gynecologic Oncology Hashtag During Virtual-Only ASCO Annual Meetings: An X (Twitter) Social Network Analysis

Geetu Bhandoria[1], MS; Esra Bilir[2,3], MSc, MD; Christina Uwins[4], MRCS, MRCOG; Josep Vidal-Alaball[5,6,7], MPH, MD, PhD; Aïna Fuster-Casanovas[7,8], RPh, MSc; Wasim Ahmed[9], BA, MSc, PhD

1
2
3
4
5
6
7
8
9

Corresponding Author:
Wasim Ahmed, BA, MSc, PhD

Abstract

Background: Official conference hashtags are commonly used to promote tweeting and social media engagement. The reach and impact of introducing a new hashtag during an oncology conference have yet to be studied. The American Society of Clinical Oncology (ASCO) conducts an annual global meeting, which was entirely virtual due to the COVID-19 pandemic in 2020 and 2021.

Objective: This study aimed to assess the reach and impact (in the form of vertices and edges generated) and X (formerly Twitter) activity of the new hashtags #goASCO20 and #goASCO21 in the ASCO 2020 and 2021 virtual conferences.

Methods: New hashtags (#goASCO20 and #goASCO21) were created for the ASCO virtual conferences in 2020 and 2021 to help focus gynecologic oncology discussion at the ASCO meetings. Data were retrieved using these hashtags (#goASCO20 for 2020 and #goASCO21 for 2021). A social network analysis was performed using the NodeXL software application.

Results: The hashtags #goASCO20 and #goASCO21 had similar impacts on the social network. Analysis of the reach and impact of the individual hashtags found #goASCO20 to have 150 vertices and 2519 total edges and #goASCO20 to have 174 vertices and 2062 total edges. Mentions and tweets between 2020 and 2021 were also similar. The circles representing different users were spatially arranged in a more balanced way in 2021. Tweets using the #goASCO21 hashtag received significantly more responses than tweets using #goASCO20 (75 times in 2020 vs 360 times in 2021; *z* value=16.63 and *P*<.001). This indicates increased engagement in the subsequent year.

Conclusions: Introducing a gynecologic oncology specialty–specific hashtag (#goASCO20 and #goASCO21) that is related but different from the official conference hashtag (#ASCO20 and #ASCO21) helped facilitate discussion on topics of interest to gynecologic oncologists during a virtual pan-oncology meeting. This impact was visible in the social network analysis.

(JMIR Med Educ 2024;10:e45291) doi:10.2196/45291

KEYWORDS

social media; academic tweeting; hashtag; gynecologic oncology; Twitter; ASCO; gynecology; oncology; virtual; engagement; software application; users; cancer; social network; health promotion

Introduction

X (formerly Twitter) has emerged as one of the social media platforms most frequently used by health care professionals [1]. In addition to individuals sharing information and networking, several academic groups, scientific societies, medical journals, and conference organizers use Twitter for educational purposes [2-4]. The reach and impact of conference hashtags have been studied previously [5-7]. Scientific conferences and academic meetings promote dedicated "conference hashtags" and encourage attendees to share their insights, experiences, and learning on the web through social media. Similarly, a study

XSL•FO
RenderX

demonstrated the significant impact of a social media ambassador program during the European Society of Gynaecological Oncology (ESGO) congresses on Twitter, highlighting substantial increases in engagement metrics and follower growth, thus advocating for the efficacy of such initiatives in enhancing congress-related engagement and visibility [8]. Furthermore, another study assessed the impact and reach of the 2020 World Gynecologic Oncology Day Twitter campaign, revealing significant participation from health care professionals and the effectiveness of the #WorldGODay hashtag in raising awareness for gynecologic cancers [9].

The official hashtag is announced in advance and widely disseminated on various social media channels [3]. These hashtags are also displayed across conference venues, and some conferences even display live tweeting during designated scientific sessions or plenaries. The aim is to disseminate meeting information and learning to attendees as well as the wider scientific community.

The COVID-19 pandemic has had a profound impact on scientific conferences. Many meetings were canceled, and others became virtual. Going virtual has affected the use of Twitter during meetings. Beste et al [10] found that the number of tweets and Twitter users at a virtual conference compared to the previous year's in-person meeting reflected the decline in the number of registrations between the 2 years.

The American Society of Clinical Oncology (ASCO) annual meeting has used its official hashtag, #ASCO, since 2011 [11]. ASCO meetings are one of the largest gatherings of oncology professionals globally. Conversations on the web and offline center around particular topics of interest, subspecialties, and the latest evidence. The COVID-19 pandemic forced both the 2020 and 2021 ASCO meetings to be held virtually. New hashtags (#goASCO20 and #goASCO21) were created for the ASCO virtual conferences in 2020 and 2021 to encourage focused gynecologic oncology discussions at the ASCO meetings. As ASCO meetings cover all oncology topics, subspecialties conversations relating to particular tumor types or subspecialties could get lost in the general discussion. Our study aimed to investigate the impact of virtualization on Twitter engagement during virtual-only ASCO annual meetings, with a focus on gynecologic oncology, and explore strategies for enhancing focused discussions and knowledge dissemination through dedicated conference hashtags.

Methods

Data Collection

Twitter data were retrieved using the hashtags #goASCO20 and #goASCO21 for 2020 and 2021, respectively. Data from the whole year were retrieved from the year each conference took place (from January to December) for each meeting using the Academic Track Twitter application programming interface, which provides access to all tweets [12].

Data Analysis

Data (influential users, topics, web sources, and social network analysis) were analyzed using social network analysis in the NodeXL software application (Social Media Research Foundation) [10], allowing an understanding of the shape of the conversation. Both graphs' vertices were clustered using the Clauset-Newman-Moore cluster algorithm to generate network visuals. The graphs were then laid out using the Harel-Koren Fast Multiscale layout algorithm. Authors in previous publications have used this methodology successfully [13-15]. Circles with lines between them represent individual Twitter users or accounts: the "mentions" and "replies." The size of the circles means how influential the user is, with bigger circles representing more influential users. The visuals presented illustrate the interactions between Twitter users. Multimedia Appendix 1 provides a compiled list of terms related to social media research for readers' ease of understanding. We also applied a 2-proportion *z* test to determine whether the change in response rates between 2020 and 2021 were statistically significant. This allowed an understanding of the shape of connections resulting from conversations to be visualized.

Visuals were created to provide an overview of the resulting social networks. Dots represent users. The green lines shown between users are known as "edges." Edges indicate both the presence and strength of a relationship between a user. There is an edge for each "reply" and "mention" and a "self-loop edge" for each tweet that is neither a "reply" nor a "mention." The "betweenness centrality" score was used to rank the size of the nodes. This score measures the influence of an individual "vertex" (an individual Twitter user, also referred to as a "node") on the flow of information between all other "vertices." This score assumes that information flows along the shortest paths between vertices. In each group, various color dots are bigger than others, indicating that these users are more influential. In addition, green lines from these groups indicate a serious relationship with other users and highlight how they have a strong influence.

Ethical Considerations

This study gained ethical approval from Newcastle University (Ref: 26055/2022). Twitter users who have been named in the study were personally contacted by the authors and provided their consent before their names or Twitter handles were published.

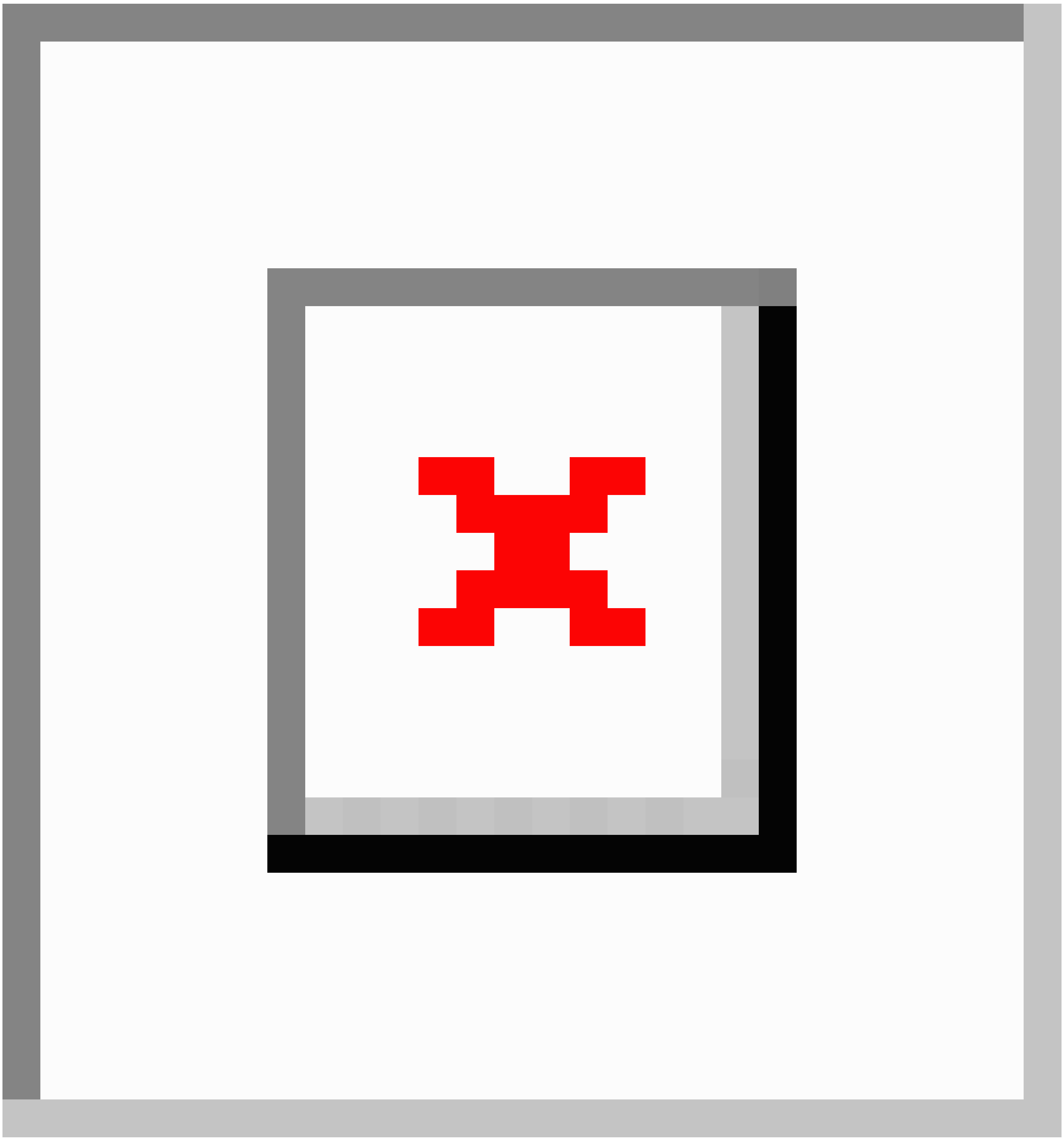
Results

Overview of the Social Networks

The most frequently used words or hashtags are highlighted in each group in Figure 1. At the top right of each group, the most used hashtags in order of interaction can be seen. For example, in group 1, the hashtag used the most was #ASCO20, while in groups 2, 3, and 4, it was #goASCO20. It is evident from the figure that different groups discussed varied topics, as depicted by other hashtags apart from #ASCO20 and #goASCO20. Figure 1 illustrates how the various communities of users shared and tweeted the #goASCO20 hashtag. Groups 1, 2, 3 and 4 have an increased number of green lines between them, indicating that their users were tweeting and mentioning one another frequently. Group 3 additionally has red lines connecting itself to groups 2 and 4. The red lines indicate stronger connections in social networks.

XSL•FO
RenderX

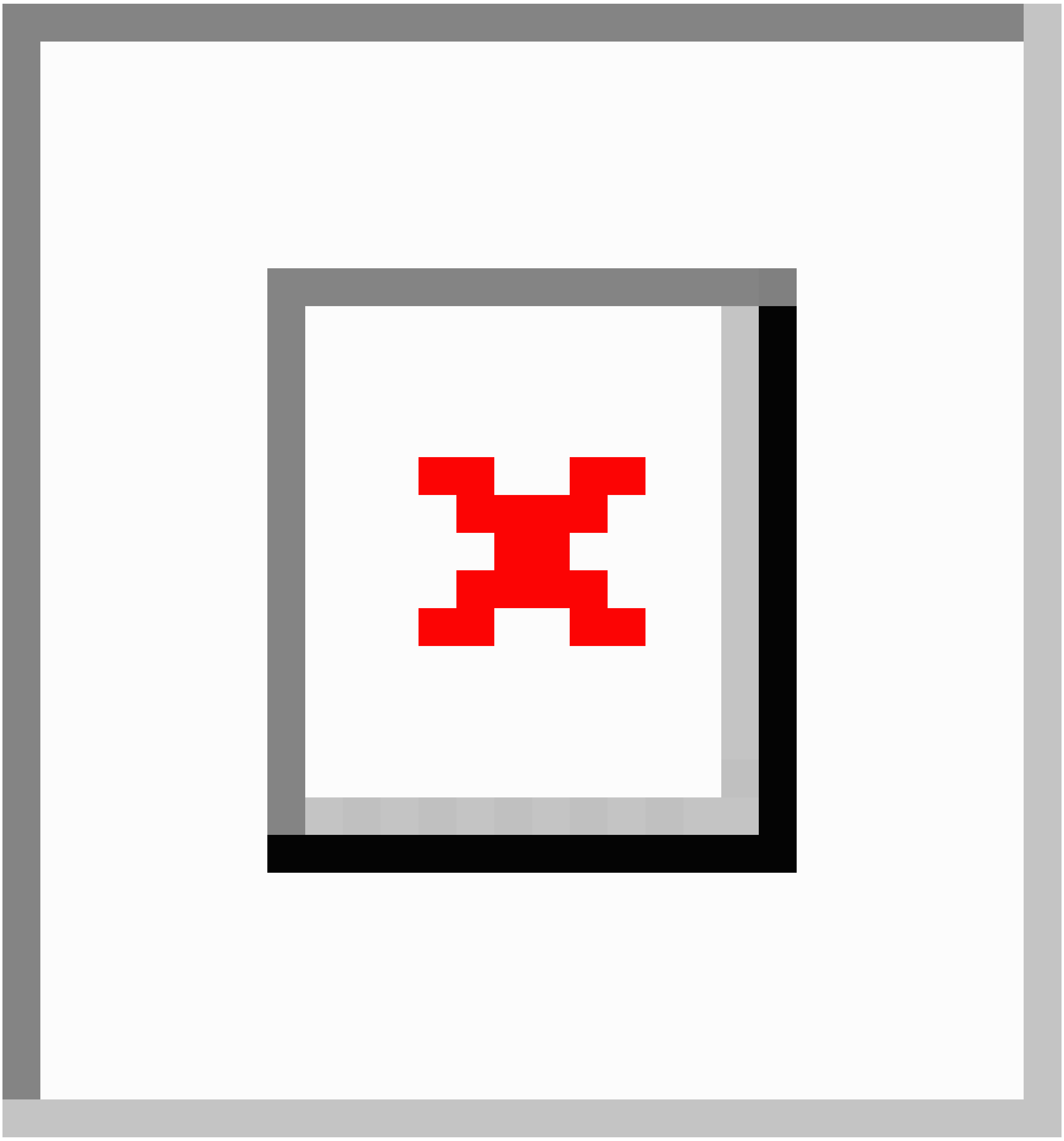
Figure 1. A visual overview of the #goASCO20 Twitter network. ASCO: American Society of Clinical Oncology; G: group.



The most used hashtag was #goASCO21. Different groups of users talked about various topics using the same hashtags. Green lines between groups indicate their relationship and influence on other users. Figure 2 illustrates the various communities of users who shared and tweeted #goASCO21, and all the groups have many green lines between them, indicating that the users were tweeting and mentioning one another. In addition, group 3 strongly influences other groups (red lines), especially group 2. Only 1 circle is more prominent than others in groups 1 and 3, indicating that these users were more influential. Other groups have circles of variable size, showing no clear influential user. In Figure 2, group 3 has more relationships (edges) with other groups than in Figure 1. The most promoted hashtags by group 3 in 2020 were #goASCO20, #ASCO20, #ASCO, and #gyncsm and in 2021 were #goASCO21, #ASCO21, #ASCO, and #ovariancancer.

XSL•FO
RenderX

Figure 2. A visual overview of the #goASCO21 Twitter network. ASCO: American Society of Clinical Oncology; G: group.



In 2021, the circles representing Twitter users were spatially arranged in a more balanced way, indicating that there were more users among the different groups in 2021. The increased lines between them illustrate an increase in cross-group discussion.

Overview of Network Metrics

Table 1 summarizes the network metrics for #goASCO20 and #goASCO21. The 240 tweets using #goASCO20 had 150 unique users and 2519 total edges. The 243 tweets using #goASCO21 had 174 unique users and 2062 total edges. A 16% (174 vs 150 unique users) increase in Twitter use was observed between 2020 and 2021. In 2020, the tweets formed 5 types of edges (mentions, retweets, replies, mention in retweets, and quote tweets) in which #goASCO20 was tagged. These tweets were mentioned 903 times, replied to 75 times, retweeted 367 times, and mentioned in retweets 934 times. In 2021, the tweets also formed 5 types of edges in which #goASCO21 was tagged: these tweets were mentioned 870 times, replied to 360 times, retweeted 33 times, and mentioned in retweets 556 times. To determine if the increase in responses to tweets using the #goASCO21 hashtag compared to the #goASCO20 hashtag was statistically significant, a 2-proportion z test was conducted. We compared the proportion of responses for each hashtag (360/2062, 17.5% for #goASCO21 and 75/2519, 3% for #goASCO20). The test resulted in a z value of approximately 16.63 and a P value <.001, indicating that the difference in response rates is statistically significant.

XSL•FO
RenderX

Table . Overview of network metrics (#goASCO20 and #goASCO21).

Graph metric	#goASCO20, n	#goASCO21, n	Change, n (%)[a]
Graph types	Directed	Directed	—[b]
Vertices (unique users)	150	174	24 (16)
Unique edges	505	505	0 (0)
Edges with duplicates	2014	1557	−457 (−22.7)
Total edges	2519	2062	−457 (−18)
Edge types	5	5	0 (0)
Mentions	903	870	−33 (−3.6)
Mentions in retweet	934	556	−378 (−40.5)
Replies	75	360	285 (380)
Retweets	367	33	−334 (−91)
Tweets	240	243	3 (1.2)

[a]The denominator (N) is the #goASCO20 value.

[b]Not applicable.

Table 2 presents an overview of the top 10 users promoting #goASCO20 and #goASCO21. This study identified 10 influential users based on their location in the network and their "betweenness centrality" score. The rank column orders the users by their "betweenness centrality" score, which reports the influence a user exerts on other users. The "in-degree" value depicts the number of times other users have mentioned an account in their tweets. Users having a high "in-degree" value means that other Twitter users consider them to have high levels of trustworthiness. For example, the user who ranked first in 2020 (@esragbilir) has been mentioned 30 times by other users. The "out-degree" value measures the number of times users mention other users in their tweets. The user who ranked first in 2020 had mentioned other users 90 times in her tweets. The top 3 users in the 2020 ranking (@esragbilir, @Bhandoria, and @ChristinaUwins) belong to the accounts of 3 authors of this study. They have a similar level of trustworthiness, and the first in the 2020 ranking is the user who has mentioned other users the most. The fourth rank in 2020 belongs to @ASCO, the user with the highest level of trustworthiness because of its high "in-degree" value.

XSL•FO
RenderX

Table . Overview of top users (#goASCO20 and #goASCO21).

Rank	#goASCO20					#goASCO21				
	User	In-degree value	Out-degree value	Betweenness centrality score	Followers, n	User	In-degree value	Out-degree value	Betweenness centrality score	Followers, n
1	@esragbilir[a]	30	90	7688.582	1355	@esragbilir[a]	17	110	13315.076	1355
2	@Bhandoria[a]	30	58	4150.223	1174	@Bhandoria[a]	37	70	11371.626	1174
3	@ChristinaUwins[a]	30	52	2704.416	888	@BatistaTP	10	52	2103.704	727
4	@ASCO	43	1	1686.453	125,888	@DrFMartinelli	14	15	1474.510	709
5	@XXXXX[b]	4	53	1459.960	1329	@gyncsm	14	16	1403.206	5726
6	@GOG	9	5	876.668	824	@AinhoaMada	19	6	1391.266	351
7	@RossFH	18	7	846.868	836	@drmnevsmne2	8	4	1104.949	428
8	@AinhoaMada	7	1	840.000	351	@PayamKashMD	11	35	1040.534	2523
9	@BatistaTP	12	12	729.352	727	@was3210[a]	11	4	806.254	9943
10	@gyncsm	9	9	703.734	5726	@dsmgyo	12	9	659.031	1246

[a]Project team members.

[b]Twitter handle anonymized.

The user who ranked first in 2021 (@esragbilir) was mentioned 17 times by other users in their tweets and mentioned other users 110 times in her tweets. The top 2 users in the ranking in 2021 belong to the accounts of 2 authors of this study, as in the previous year. @Bhandoria had a higher level of trustworthiness than the first user in the ranking, and @esragbilir mentioned other users more than @Bhandoria. The third in rank is @BatistaTP, a gynecologic oncology surgeon. The fourth place in the ranking in 2021 belongs to @DrFMartinelli, a gynecologist specializing in oncology. The fifth place in the ranking belongs to @gyncsm, a community for those impacted by gynecologic cancers.

Table 3 provides an overview of the top 20 cowords used with #goASCO20 and #goASCO21.

XSL•FO
RenderX

Table . Overview of the top 20 cowords used with hashtags #goASCO20 and #goASCO21.

Rank	#goASCO20			#goASCO21		
	Word 1	Word 2	Count, n	Word 1	Word 2	Count, n
1	#gyncsm	#some4gynonc	130	#asco21	asco	88
2	bhandoria	christinauwins	106	#goasco21	#eva_asco2021	72
3	#womeninstem	#gyncsm	100	#goasco21	#asco21	72
4	christinauwins	ilkerselcukmd	94	sbco_oficial	br_gynoncgroup	60
5	asco	#asco20	92	#eva_asco2021	#sbco	60
6	gynaecological	ncology	89	#sbco	#asco21	60
7	use	#goasco20	88	sgo_org	gog	59
8	#goasco20	#asco20	83	asco	#gynecologiconcol-ogy	58
9	during	#asco20	80	gog	esgo_society	56
10	follow	use	79	esgo_society	essonews	56
11	#some4gynonc	#somedocs	69	essonews	sbco_oficial	56
12	#somedocs	#medtwitter	69	br_gynoncgroup	ijgconline	56
13	#asco20	#gyncsm	66	ijgconline	igcsociety	56
14	promote	raiseawareness	63	igcsociety	gyncsm	54
15	raiseawareness	#gynecologiconcol-ogy	63	#asco21	#goasco21	42
16	shared	photos	58	#asco21	#gyncsm	31
17	photos	app	58	christinauwins	was3210	25
18	app	photo	58	#cervicalcancer	#endometrialcancer	25
19	esragbilir	bhandoria	55	#cervicalcancer	#goasco21	24
20	#goasco20	promote	55	#goasco21	clin	24

In 2020, the cowords used the most with the studied hashtag were #gyncsm and #some4gynonc (130 times). #Gyncsm is a community for those impacted by gynecologic cancers. #Some4gynonc is a social media group promoting the goal of curing gynecologic cancer globally. In second place, 2 users were mentioned 106 times with #goASCO20: @Bhandoria, a gyneoncologist and obstetrician, and @ChristinaUwins, a surgeon and senior research fellow in robotic gynecologic oncology. In third place, 2 hashtags (#womeninstem and #gynscm) were used 100 times. The hashtag #womeninstem promotes women and gender equality in science, technology, engineering, and mathematics. In fourth place, there were 2 users, both of whom were mentioned 94 times with #goASCO20. Finally, the fifth place belongs to the hashtags #asco and #goasco20, and both were mentioned 92 times with #goASCO20.

In 2021, the cowords used the most with #goASCO21 were #asco21 and #asco (88 times). In second place, #asco21 and #eva_asco2021 were used 72 times. The first refers to the ASCO, and the second (#eva_asco2021) refers to a group focused on gynecologic tumors from Brazil. The third most used cowords (72 times) were hashtags that promoted the spread of clinical knowledge (#goASCO21 and #asco21). The fourth-ranked cowords (60 times) were sbco_oficial, a Brazilian society of oncologic surgery, and br_gynoncgroup, a Brazilian gynecologic oncology group. The last word pairs in the top 5 most used cowords were #eva_asco2021, a group focused on gynecologic tumors from Brazil, and #sbco, a hashtag used to refer to the Brazilian Society of Oncologic Surgery.

Discussion

Principal Findings

This study hypothesized that introducing a new hashtag specific to gynecologic oncology could provide a focus for tweeting about gynecologic cancers. A new hashtag, #goASCO20, was presented on Twitter during the ASCO 2020 virtual conference and was replaced with #goASCO21 in 2021. Conference attendees were encouraged to use these new hashtags when discussing anything related to gynecologic cancers. The use of these new hashtags was actively encouraged. Users who promoted the hashtag in 2020 did not tend to respond to tweets but, in 2021, increased their response rate (75 times in 2020 vs 360 times in 2021). This shows that the gynecologic oncology community started engaging better in the second virtual congress. Consistent use of hashtags has enhanced Twitter engagement, as evident in the study by Morgan et al [15]. The cumulative number of impressions for #ASCO16 was 468.2 million compared with approximately 1.12 billion for #ASCO20

XSL•FO
RenderX

[15]. We predict a similar growth of #goASCO if its use is continued.

COVID-19 played a crucial role in social media use among the oncology community. It forced the annual meeting to go entirely virtual. As evidenced by our study, the conference attendees used social media channels more to interact.

The 2 users who promoted the hashtags the most were the same in 2020 and 2021. It should be noted that @esragbilir and @Bhandoria significantly increased their “betweenness centrality” score, indicating that their location in the network became more influential among the users. Establishing a core social media team that actively promotes it is essential.

Strengths and Weaknesses

This is the first study where a new hashtag was introduced and social media interaction was measured. This study contributes to the literature on this topic, highlighting how networks can be used to spread trustworthy information and share relevant information among the scientific community on Twitter.

A limitation of this study is that it was not designed to assess the validity of any tweets but to evaluate the success of promoting the use of a gynecologic oncology–specific hashtag in increasing interaction between individual Twitter users and organizations. Misinformation on Twitter is a recognized phenomenon; future studies should investigate whether the quality and quantity of discussion are affected [16]. Since the inception of oncology hashtags, we acknowledge the existence of the gynecology-specific hashtag #gyncsm [17]. We created the #goASCO hashtags to study its impact as #gyncsm is used more by patients with gynecologic cancer and their advocates [18]. We should have examined the effect of #gyncsm during these virtual meetings, and this may be seen as a weakness, with no comparator group being available. Lastly, some of the “influential Twitter users” named in the results included a few authors. However, this is not aimed at self-promotion but is part of the results’ description.

Conclusion

The use of a gynecologic cancer–specific hashtag helped facilitate discussion on topics in gynecologic oncology on Twitter during the 2020 and 2021 ASCO virtual meetings. This impact was visible in the social network analysis.

Data Availability

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

Authors' Contributions

EB, GB, and CU contributed to conceptualization. WA, JV-A, and AF-C contributed to data curation and analysis. GB and EB contributed to project administration. All authors contributed to writing–original draft and writing–review and editing.

Multimedia Appendix 1
Terms related to social media research.
[DOCX File, 13 KB - mededu_v10i1e45291_app1.docx]

References

1. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med 2018 May 28;7(6):121. [doi: 10.3390/jcm7060121] [Medline: 29843360]
2. Mishori R, Levy B, Donvan B. Twitter use at a family medicine conference: analyzing #STFM13. Fam Med 2014 Sep;46(8):608-614. [Medline: 25163039]
3. Pemmaraju N, Mesa RA, Majhail NS, Thompson MA. The use and impact of Twitter at medical conferences: best practices and Twitter etiquette. Semin Hematol 2017 Oct;54(4):184-188. [doi: 10.1053/j.seminhematol.2017.08.003] [Medline: 29153078]
4. Cohen D, Allen TC, Balci S, et al. #InSituPathologists: how the #USCAP2015 meeting went viral on Twitter and founded the social media movement for the United States and Canadian Academy of Pathology. Mod Pathol 2017 Feb;30(2):160-168. [doi: 10.1038/modpathol.2016.223] [Medline: 28084341]
5. Wilkinson SE, Basto MY, Perovic G, Lawrentschuk N, Murphy DG. The social media revolution is changing the conference experience: analytics and trends from eight international meetings. BJU Int 2015 May;115(5):839-846. [doi: 10.1111/bju.12910] [Medline: 25130687]
6. Mackenzie G, Søreide K, Polom K, et al. Beyond the hashtag - an exploration of tweeting and replies at the European Society of Surgical Oncology 39th clinical conference (ESSO39). Eur J Surg Oncol 2020 Jul;46(7):1377-1383. [doi: org/10.1016/j.ejso.2020.02.018] [Medline: 32127248]
7. Chaudhry A, Glodé LM, Gillman M, Miller RS. Trends in Twitter use by physicians at the American Society of Clinical Oncology annual meeting, 2010 and 2011. J Oncol Pract 2012 May;8(3):173-178. [doi: 10.1200/JOP.2011.000483] [Medline: 22942812]

XSL•FO
RenderX

8. Bilir E, Ahmed W, Kacperczyk-Bartnik J, et al. Social media ambassadors and collaboration with OncoAlert: a European Network of Young Gynae Oncologists study of comparative Twitter analysis of #ESGO2021 and #ESGO2022. Int J Gynecol Cancer 2023 Jun 5;33(6):964-970. [doi: 10.1136/ijgc-2023-004371] [Medline: 37130625]
9. Uwins C, Yilmaz Y, Bilir E, Bhandoria GP. World Gynecologic Oncology Day: the use of Twitter to raise awareness of gynecologic cancers. AJOG Glob Rep 2022 Jul 21;2(3):100079. [doi: 10.1016/j.xagr.2022.100079] [Medline: 36276802]
10. Beste NC, Davis X, Kloeckner R, et al. Comprehensive analysis of Twitter usage during a major medical conference held virtually versus in-person. Insights Imaging 2022 Jan 20;13(1):8. [doi: 10.1186/s13244-021-01140-0] [Medline: 35050426]
11. Pemmaraju N, Thompson MA, Mesa RA, Desai T. Analysis of the use and impact of Twitter during American Society of Clinical Oncology annual meetings from 2011 to 2016: focus on advanced metrics and user trends. J Oncol Pract 2017 Jul;13(7):e623-e631. [doi: org/10.1200/JOP.2017.021634] [Medline: 28514195]
12. Ahmed W, Lugovic S. Social media analytics: analysis and visualisation of news diffusion using NodeXL. Online Inf Rev 2019 Feb 11;43(1):149-160. [doi: 10.1108/OIR-03-2018-0093]
13. Ahmed W, Marin-Gomez X, Vidal-Alaball J. Contextualising the 2019 e-cigarette health scare: insights from Twitter. Int J Environ Res Public Health 2020 Mar 26;17(7):2236. [doi: 10.3390/ijerph17072236] [Medline: 32225020]
14. Ahmed W, Vidal-Alaball J, Lopez Segui F, Moreno-Sánchez PA. A social network analysis of tweets related to masks during the COVID-19 pandemic. Int J Environ Res Public Health 2020 Nov 7;17(21):8235. [doi: 10.3390/ijerph17218235] [Medline: 33171843]
15. Morgan G, Choueiri TK, Patel R, Balaji K, Subbiah V. Impact of #ASCO Twitter impressions on the oncology community. J Clin Oncol 2021 May 28;39(15_suppl):11039. [doi: 10.1200/JCO.2021.39.15_suppl.11039]
16. Kreps S, George J, Watson N, Cai G, Ding K. (Mis)information on digital platforms: quantitative and qualitative analysis of content from Twitter and Sina Weibo in the COVID-19 pandemic. JMIR Infodemiol 2022 Feb 24;2(1):e31793. [doi: 10.2196/31793] [Medline: 36406147]
17. Katz MS, Utengen A, Anderson PF, et al. Disease-specific hashtags for online communication about cancer care. JAMA Oncol 2016 Mar;2(3):392-394. [doi: 10.1001/jamaoncol.2015.3960] [Medline: 26539640]
18. Monuszko KA, Fish LJ, Sparacio D, et al. Understanding the needs and perspectives of ovarian cancer patients when considering PARP inhibitor maintenance therapy: findings from two online community events. Gynecol Oncol Rep 2022 Jul;43:101050. [doi: 10.1016/j.gore.2022.101050] [Medline: 35942110]

Abbreviations

ASCO: American Society of Clinical Oncology
ESGO: European Society of Gynaecological Oncology

Edited by TDA Cardoso; submitted 09.02.23; peer-reviewed by Z Zhang; revised version received 03.07.24; accepted 10.07.24; published 14.08.24.

Please cite as:
Bhandoria G, Bilir E, Uwins C, Vidal-Alaball J, Fuster-Casanovas A, Ahmed W
Impact of a New Gynecologic Oncology Hashtag During Virtual-Only ASCO Annual Meetings: An X (Twitter) Social Network Analysis
JMIR Med Educ 2024;10:e45291
URL: https://mededu.jmir.org/2024/1/e45291
doi:10.2196/45291

© Geetu Bhandoria, Esra Bilir, Christina Uwins, Josep Vidal-Alaball, Aïna Fuster-Casanovas, Wasim Ahmed. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.8.2024. 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.

XSL•FO
RenderX

Social Media Usage for Medical Education and Smartphone Addiction Among Medical Students: National Web-Based Survey

Thomas Clavier[1,2], MD, PhD; Emma Chevalier[3], MD; Zoé Demailly[1], MSc, MD; Benoit Veber[1,3], MD, PhD; Imad-Abdelkader Messaadi[3], MD; Benjamin Popoff[1], MSc, MD

1

2

3

Corresponding Author:
Benjamin Popoff, MSc, MD

Abstract

Background: Social media (SoMe) have taken a major place in the medical field, and younger generations are increasingly using them as their primary source to find information.

Objective: This study aimed to describe the use of SoMe for medical education among French medical students and assess the prevalence of smartphone addiction in this population.

Methods: A cross-sectional web-based survey was conducted among French medical students (second to sixth year of study). The questionnaire collected information on SoMe use for medical education and professional behavior. Smartphone addiction was assessed using the Smartphone Addiction Scale Short-Version (SAS-SV) score.

Results: A total of 762 medical students responded to the survey. Of these, 762 (100%) were SoMe users, spending a median of 120 (IQR 60 - 150) minutes per day on SoMe; 656 (86.1%) used SoMe for medical education, with YouTube, Instagram, and Facebook being the most popular platforms. The misuse of SoMe in a professional context was also identified; 27.2% (207/762) of students posted hospital internship content, and 10.8% (82/762) searched for a patient's name on SoMe. Smartphone addiction was prevalent among 29.1% (222/762) of respondents, with a significant correlation between increased SoMe use and SAS-SV score (*r*=0.39, 95% CI 0.33 - 0.45; *P*<.001). Smartphone-addicted students reported a higher impact on study time (211/222, 95% vs 344/540, 63.6%; *P*<.001) and a greater tendency to share hospital internship content on social networks (78/222, 35.1% vs 129/540, 23.8%; *P*=.002).

Conclusions: Our findings reveal the extensive use of SoMe for medical education among French medical students, alongside a notable prevalence of smartphone addiction. These results highlight the need for medical schools and educators to address the responsible use of SoMe and develop strategies to mitigate the risks associated with excessive use and addiction.

(JMIR Med Educ 2024;10:e55149) doi:10.2196/55149

KEYWORDS
medical student; social network; social media; smartphone addiction; medical education; mobile addiction; social networks

Introduction

In today's globalized world, social media (SoMe) have taken a significant role in the medical field, serving as essential tools for promoting research, medical innovations, and updates from various specialties (eg, techniques and congresses). With the explosion in the number of platforms and their evolution, several definitions of SoMe have been proposed and gradually amended [1,2]. Recent definitions agree that true SoMe are defined as "web-based technologies that facilitate multi-user interaction that goes beyond fact sharing, centering around content creation, curation, and community engagement, placing user interaction at their art" [3]. These platforms include, for example, major actors like Facebook, Twitter (now X), Instagram, YouTube, TikTok, Snapchat, LinkedIn, or WhatsApp, and exclude websites or blogs with comment sections and podcasts due to their primarily unidirectional nature [3]. SoMe are now widely used by health care professionals for numerous purposes, such as education, patient communication, and colleague discussions [4,5]. SoMe provide a platform for the rapid dissemination of research findings and facilitate networking and collaboration among researchers and clinicians worldwide [6].

Younger generations increasingly rely on SoMe as their primary source of information about brands or organizations, with this usage even surpassing that of internet search engines among 16 - to 24-year-olds [7]. The main reason for using SoMe is to "stay up-to-date with news and current events" [8]. Time spent on SoMe has consistently grown, rising from 1 hour and 51 minutes per day in 2015 to 2 hours and 24 minutes per day in 2023 [8]. Furthermore, several studies have documented the

XSL•FO
RenderX

benefits of using SoMe for medical education [9,10]. Consequently, a growing number of educators and medical societies are leveraging SoMe to showcase their educational content [11,12]. The COVID-19 pandemic has acted as an amplifier of the trend toward distance learning, with SoMe playing a significant role in this regard [13,14]. However, there is a lack of comprehensive data on medical students' use of these educational resources from SoMe for knowledge acquisition.

Several studies have identified significant risks associated with prolonged SoMe use. Notably, smartphone addiction correlates with the intensity of SoMe usage [15]. This addiction, in turn, can negatively impact students' quality of life, leading to sleep disorders, musculoskeletal disorders, severe social withdrawal, decreased physical activity, and hypertension [16-19]. Finally, all available data on smartphone addiction among medical students originate from Asia, with no data from Western countries.

The purpose of this study was to address the gap in knowledge regarding the use of SoMe by medical students in Western countries, specifically in France, for medical education. We aimed to describe, on a nationwide scale, how medical students use SoMe for medical learning, their motivations and preferences, and the extent to which they rely on these platforms for educational purposes. Additionally, we sought to determine the prevalence of smartphone addiction in this population and explore its potential impact on academic performance and professional behavior. We hypothesized that a significant proportion of medical students use SoMe for educational purposes; that this usage correlates with specific patterns of SoMe behavior, including misuse such as breaches of patient confidentiality; and that high levels of SoMe use are associated with increased rates of smartphone addiction.

Methods

Objectives

The primary objective was to describe how medical students use SoMe to learn about medicine. Secondary objectives were to evaluate their use of these platforms for choosing a medical specialty, analyze the prevalence of smartphone addiction in this population, and describe their potential misuse of SoMe. SoMe misuse was defined as the disclosure of information about hospital internships (text, photo, or video) that may breach patient confidentiality and the active internet-based search for private information [20].

Ethical Considerations

The study was approved by the Ethics and Evaluation Committee for Non-Interventional Research of Rouen University Hospital (E2023-06). Participation was entirely voluntary. Participants were informed about the study's objectives and provided their consent before completing the survey. The survey was conducted anonymously, and no identifying information was collected or attempted to be gathered at any stage, ensuring participants' privacy and confidentiality. No compensation was offered to participants, and they had the right to withdraw from the survey at any time without any consequences. According to institutional guidelines, as this was a noninterventional, anonymous survey with no personal health data collected, the study did not require further ethical exemptions or waivers beyond the initial approval.

Population Selection

We conducted a prospective study in France using a declarative survey. The link to an open Google Form internet survey, consisting of 32 items on 1 web page, was emailed to the board of the French medical students' association (Association des Etudiants en Médecine de France). This board forwarded the questionnaire to the association's representatives at each of the 35 medical schools in France. These 35 representatives were instructed to share the link with their respective faculty's students via email. All contacted students were asked to forward the survey link to their colleagues. All participants received information about the survey objectives, which were reiterated in the questionnaire's introduction. There was no incentive to answer the questionnaire. As responses were anonymous, no information was collected to prevent multiple entries by the same individual. Also, as the questionnaire was open-ended, anyone with the link could answer it. The distribution via medical student representatives was intended to restrict the questionnaire's visibility to the target audience only.

This survey was developed according to available guidelines for self-administered surveys [21]. Responses were submitted on a single web page with 1 "submit" button, which only allowed submissions via these unique links, making uninvited responses extremely unlikely. The request was sent to 35 representatives in France; however, as we were unable to determine how many students the request was forwarded to, we do not know the overall number of students who received the request to participate in the survey. We did not organize any specific follow-up on the distribution of the questionnaire with the contacted representatives. The survey was conducted in accordance with the Checklist for Reporting Results of Internet E-Surveys (CHERRIES; Checklist 1) [22].

The participants included in the analysis were medical students in their first (second to third year) and second (fourth to sixth year) cycles of medical study. The medical curriculum in France consists of 6 years of study before residency, although the first year of medical school is a competitive year with a success rate of about 15%; students in this year of study were not included in this analysis, which concerned only students who were certain to become physicians after their studies.

Survey Design

The survey was developed by a team consisting of medical students, medical educators, and an expert in SoMe to ensure comprehensive coverage of relevant topics. The questions in the "use of social network" sections were designed based on a thorough review of existing literature on SoMe usage in medical education and consultations with subject matter experts [1,3,10]. The initial draft was reviewed by a panel of medical educators and students to assess face validity and to validate the list of social networks that met our definition. To further ensure the validity and reliability of the survey, a pilot test was conducted with the participation of the board of the French Medical

XSL•FO
RenderX

Students' Association to ensure the comprehensibility and relevance of the questions. Feedback from the pilot test was used to refine the survey questions for clarity and relevance. The psychometric properties of the survey were not formally assessed.

The survey consisted of 3 sections. The "demographic data" section collected information on city, age, gender, year of study, and whether the student had already retaken an exam (catch-up exam). The "use of social networks" section gathered data on personal and professional use of WhatsApp, types of social networks consulted at least once a week, average daily time spent on SoMe, usage of SoMe for learning about medicine and choosing a medical specialty, and the misuse of SoMe (searching for a patient's name and spreading information from hospital internships).

We defined the sharing of content from hospital internships on SoMe as a misuse based on professional standards. It is generally considered unprofessional for students to share details about their clinical experiences on SoMe, as this behavior can undermine patient confidentiality and trust. The Health Insurance Portability and Accountability Act (HIPAA) regulations in the United States clearly state that sharing any patient information on SoMe is unacceptable [23] and medical societies have issued strict recommendations against such practices [24]. Since most student internships involve direct patient contact, sharing content related to these internships often involves discussing patient care experiences, which is inappropriate even if no identifiable patient information is shared.

Students also rated, using a 6-point Likert scale (1 point="completely irrelevant" and 6 points="completely relevant"), whether they thought it was appropriate to offer a teaching module on the professional or educational use of SoMe in medical school.

Finally, the "assessment of smartphone addiction" section evaluated smartphone addiction using the short version of the Smartphone Addiction Scale Short-Version (SAS-SV) developed by Kwon et al [25]. The SAS-SV was already translated into French, demonstrating the validity and reliability of the translated version adapted to French [26]. The scale comprises 10 positive 6-point Likert questions describing smartphone usage, with the total score ranging from 10 to 60 and higher scores indicating a greater risk of addiction. According to the threshold recommended for student populations, and given previous data concerning the absence of gender difference for the cutoff value of the SAS-SV among French-speaking students, the cutoff point was determined as superior or equal to 32 points to identify smartphone addiction [25,26]. The SAS-SV scale covers 6 addictive symptoms and they are loss of control, disruption of family or schooling, disregard for consequences, withdrawal, preoccupation, and tolerance. Each item is associated with an addictive symptom, except for 4 item clusters that are items 1 and 8 (both assessing "loss of control"), items 2 and 10 ("disruptions"), items 3 and 7 ("disregard for consequences"), and items 4 and 5 ("withdrawal") [26]. As previously described, a rating of 4 or higher for each symptom was considered to signify the presence of this specific symptom [26]. Participants had to answer all questions to validate the questionnaire but could go back at any time to change their answers before the final validation. The original version of the questionnaire and its English translation can be found in Multimedia Appendices 1 and 2.

Statistical Analysis

The values are presented as the number and percentage (n, %) for qualitative variables and as the median (IQR) for quantitative variables. Statistical analyses were performed in complete case analysis on fully completed questionnaires [27]. After ensuring the abnormal distribution of the data via a Shapiro-Wilk test, quantitative variables were compared using a Mann-Whitney test. Qualitative variables were analyzed using a Fischer exact test. The Spearman correlation test was used to assess the strength of the association between 2 quantitative variables. All statistical tests were 2-sided, and the $P<.05$ probability threshold was used to establish statistical significance. All statistical analyses were performed using R (version 4.1.3; R Core Team).

Results

Demographic Data

The compilation of responses took place from May 22 to October 26, 2021. A total of 762 medical students responded to the survey. Among them, the median age was 22 (IQR 21 - 24) years, and the gender ratio was 0.39 (212 males and 547 females; 3 students identified themselves as "gender neutral/non-gendered"). Respondents came from all the French metropolitan regions (Table 1). The participants were distributed as 149 (19.5%) second-year students, 121 (15.9%) third-year students, 139 (18.2%) fourth-year students, 119 (15.6%) fifth-year students, and 234 (30.7%) sixth-year students. Among the 762 respondents, 287 (37.7%) had retaken an exam at least once during their medical curriculum.

XSL•FO
RenderX

Table . Distribution of respondents by region.

Region	Respondents, n (%)
Auvergne-Rhône-Alpes	105 (13.8)
Bourgogne-Franche-Comté	106 (13.9)
Bretagne	21 (2.8)
Centre-Val de Loire	29 (3.8)
Grand Est	25 (3.3)
Hauts-de-France	24 (3.2)
Normandie	134 (17.6)
Nouvelle-Aquitaine	24 (3.2)
Occitanie	88 (11.5)
Pays de la Loire	61 (8.0)
Provence-Alpes-Côte d’Azur	90 (11.8)
Île-de-France	55 (7.2)

Use of SoMe

Among the 762 respondents included, 624 (81.8%) were WhatsApp users and 762 (100%) were SoMe users, spending a median time of 120 (IQR 60 - 150) minutes per day on them. A total of 555 (72.8%) students felt that their time spent on SoMe impacted their study time and 656 (86.1%) used SoMe to learn about medicine. The 3 most used SoMe for this purpose were YouTube (504/762, 66.1%), Instagram (433/762, 56.8%), and Facebook (320/762, 42%; Table 2). A total of 115 (15.1%) students used WhatsApp for professional purposes (internship questions, exchange of night shifts, and discussion about a patient).

Table . Proportion of medical students using specific social networks for medical education, exploration of medical specialties prior to selection, and sharing content related to hospital internship.

SoMe[a]	Usage, n (%)		
	Exploring medical specialties	Medical education	Sharing content about hospital internships
Facebook	235 (30.8)	320 (42.0)	47 (6.2)
Instagram	375 (49.2)	433 (56.8)	157 (20.6)
LinkedIn	2 (0.3)	3 (0.4)	1 (0.1)
Pinterest	1 (0.1)	7 (0.9)	0 (0)
Reddit	3 (0.4)	8 (1.0)	0 (0)
Snapchat	2 (0.3)	9 (1.2)	68 (8.9)
TikTok	9 (1.2)	20 (2.6)	1 (0.1)
Twitch	1 (0.1)	0 (0)	0 (0)
Twitter	70 (9.2)	63 (8.3)	18 (2.4)
YouTube	220 (28.9)	504 (66.1)	0 (0)

[a]SoMe: social media.

On SoMe, 79.1% (604/762) students followed 1 or more physicians whom they knew (resident or senior physician), and 25.9% (197/762) followed 1 or more national medical societies. A total of 522/762 (67.9%) students used SoMe to learn about a medical specialty in anticipation of choosing one. We identified significant misuse of social networks in a professional context, as 27.2% (207/762) of students had already posted content on SoMe (text, photo, and video) related to their hospital internship, and 10.8% (82/762) had ever searched for a patient’s name on a SoMe platform. SoMe that used to post content related to hospital internships are presented in Table 2.

Regarding the interest in teaching modules on the professional or educational use of SoMe in medical school, 61.4% (468/762) students found this proposal relevant (204/762, 26.7%), very relevant (156/762, 20.4%), or completely relevant (108/762, 14.2%). The remaining students (294/762, 38.6%) did not find this teaching relevant.

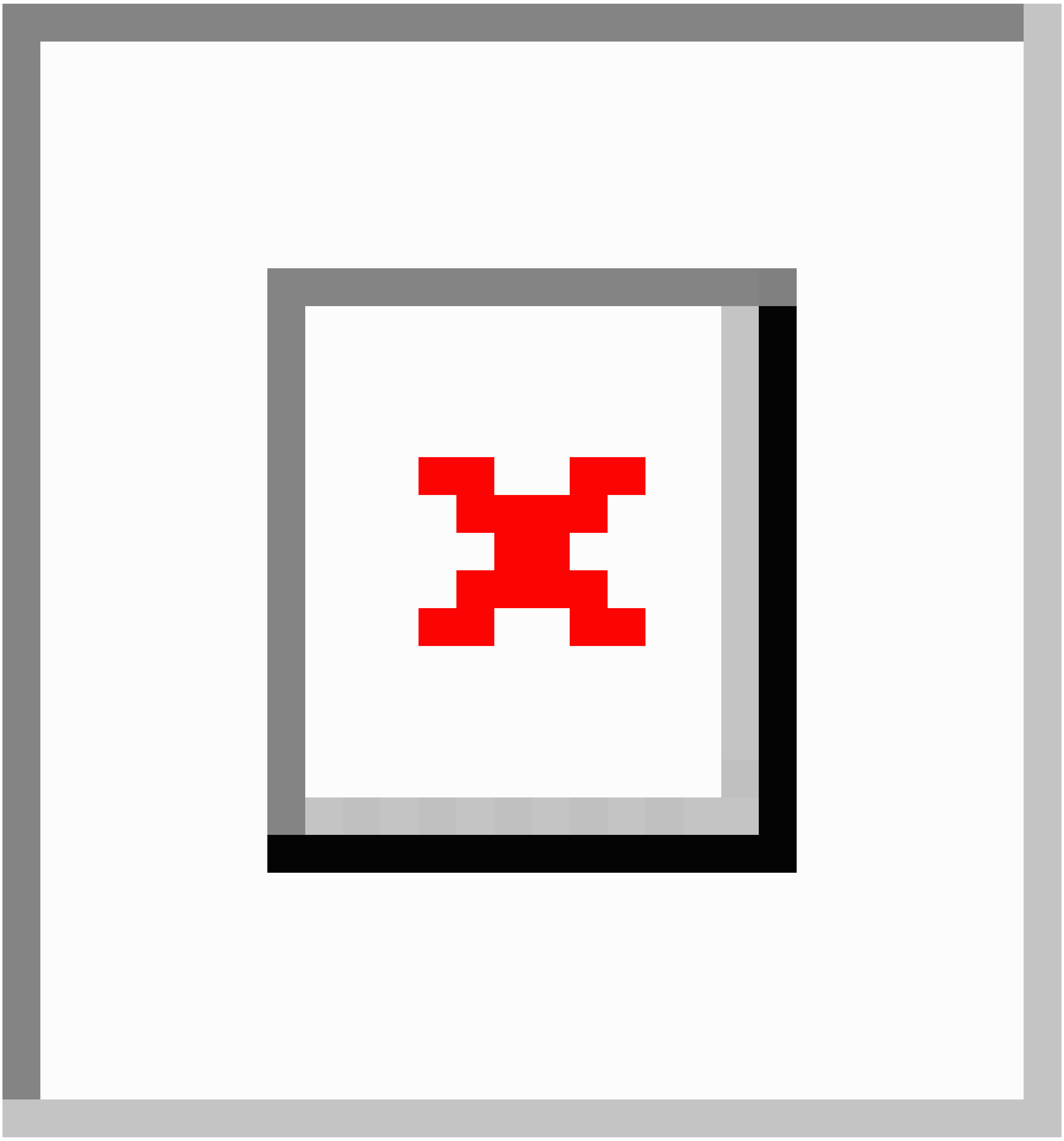
Smartphone Addiction

Among the 762 students analyzed, 222 (29.1%) had an SAS-SV score of at least 32/60, defining smartphone addiction. Nonaddicted students had a median SAS-SV score of 24 (IQR

XSL•FO
RenderX

20 - 27), while addicted students had a median score of 37 (IQR 35 - 42; *P*<.001). There was a significant correlation between the time spent on SoMe and the SAS-SV score (*r*S=0.39, 95% CI 0.33 - 0.45; *P*<.001; Figure 1). There were no demographic differences between smartphone-addicted and nonaddicted students (Table 3). However, addicted students spent more time on SoMe, with a more frequent impact on their study time, and a higher tendency to post content from their hospital internships on SoMe (Table 3). Among the 222 addicted students, the most frequent addiction symptoms were tolerance (207/222, 93.2%), loss of control (166/222, 74.8%), disruption of family or schooling (138/222, 62.2%), and withdrawal (127/222, 57.2%). Only a few students displayed disregard for consequences symptoms (30/222, 13.5%) and none presented preoccupation about their smartphone use.

Figure 1. Correlation between the time spent on social media and the SAS-SV score. SAS-SV: Smartphone Addiction Scale Short-Version.



XSL•FO
RenderX

Table . Characteristics and social media behavior of medical students with or without smartphone addiction.

	Overall (n=762)	Nonaddicted students (n=540)	Addicted students (n=222)	*P* value
Gender, n (%)				.36
Male	212 (27.8)	150 (27.8)	62 (27.9)	
Female	547 (71.8)	389 (72.0)	158 (71.2)	
Other	3 (0.4)	1 (0.2)	2 (0.9)	
Age (years), median (IQR)	22 (21 - 24)	22 (21 - 24)	22 (21 - 23)	.71
Cycle of study, n (%)				.91
First cycle	270 (35.4)	192 (35.7)	78 (35.1)	
Second cycle	492 (64.6)	348 (64.3)	144 (64.8)	
Retook an exam, n (%)				.09
Yes	287 (37.7)	193 (35.7)	94 (42.3)	
No	475 (62.3)	347 (64.3)	128 (57.7)	
Time spent on social media (minutes per day), median (IQR)	120 (60 - 150)	94 (60 - 120)	150 (120 - 200)	<.001
Impact on the study time, n (%)				<.001
Yes	555 (72.8)	344 (63.6)	211 (95.0)	
No	207 (27.2)	196 (36.4)	11 (5.0)	
Social media use to learn about medicine, n (%)				.98
Yes	656 (86.1)	465 (86.0)	191 (86.0)	
No	106 (13.9)	75 (14.0)	31 (14.0)	
Posts related to the hospital internship, n (%)				.002
Yes	207 (27.2)	129 (23.8)	78 (35.1)	
No	555 (72.8)	411 (76.2)	144 (64.9)	
Ever searched a patient's name on social media, n (%)				.07
Yes	82 (10.8)	51 (9.4)	31 (14.0)	
No	680 (89.2)	489 (90.6)	191 (86.0)	

Subgroup Analyses

First- and second-cycle students show different patterns of social network use. Compared to their second-cycle peers, first-cycle students are less likely to use Facebook for their medical education (77/270, 28.5% vs 243/492, 49.3%; *P*<.001), and more inclined to use Instagram or Snapchat for example (Figure S1 in Multimedia Appendix 3). Smartphone addiction is the same between first- and second-cycle students (78/270, 28.9% vs 144/492, 29.3, respectively; *P*=.91), but graduate students are more likely to exhibit SoMe misuse behaviors (Table S1 in Multimedia Appendix 3).

Regarding gender differences, female and male students spent the same amount of time on SoMe (median 120, IQR 60 - 150 minutes per day in both groups; *P*=.80) and had the same prevalence of smartphone addiction (158/547, 28.9% vs 62/212, 29.2% respectively; *P*=.92; Table S2 in Multimedia Appendix 3). They presented differences in the use of different social networks (Figure S2 in Multimedia Appendix 3). Female students were more likely to use SoMe for medical education (482/547, 88.1% vs 173/212, 81.6%; *P*=.04). The 2 genders showed different patterns of misuse, with a greater tendency to post content relating to hospital internships for female students (160/547, 29.1% vs 47/212, 21.8%; *P*=.04) and a greater tendency to search a patient's name on SoMe for male students (33/212, 15.6% vs 49/547, 9%; *P*=.009).

Discussion

Principal Findings

This nationwide web-based survey aimed to describe the use of SoMe by medical students in France for their medical education and to evaluate the prevalence of smartphone addiction in this population. Our findings indicate that among the respondents, a significant majority (656/762, 86.1%) used SoMe to learn about medicine, with YouTube, Instagram, and Facebook being the most popular platforms. Respondents reported spending a substantial portion of their day on SoMe, with a median time of 120 (IQR 60 - 150) minutes per day. However, misuse of SoMe was also reported, with 10.8%

XSL•FO
RenderX

(82/762) of students searching for patients' names on SoMe platforms. Nearly one-third (222/762, 29.1%) of respondents met the criteria for smartphone addiction according to the SAS-SV. Our study is the first nationwide survey to explore the use of SoMe for medical education among medical students in France and to investigate smartphone addiction on such a large scale in this specific population.

Use of SoMe for Medical Education

Our study highlights the extensive use of SoMe among respondents for medical learning mainly by following physicians' accounts (603/762, 79.1%) and, to a lesser extent, by following scientific societies (197/762, 25.9%). The main SoMe platforms used for this purpose were YouTube and Instagram, suggesting a preference for visual and multimedia content over textual information. While our survey did not assess whether SoMe were used more often than school-provided content, it indicates that SoMe was a significant supplementary source to traditional medical education, fostering collaboration among students and health care professionals. The COVID-19 pandemic has likely accelerated this trend [13,14], making it crucial for medical schools and educators to recognize the potential of these platforms and to integrate SoMe effectively into their teaching strategies.

However, it should be noted that 72.8% (555/762) of students surveyed felt that their time spent on SoMe negatively impacted their study time, highlighting the ambivalence of these platforms. Although students use SoMe for educational purposes, they also engage with these platforms for noneducational activities, which can detract from their study time. This point is acknowledged by the respondents, as 61.4% (468/762) of them found it relevant to add a teaching module on the professional or educational use of SoMe in medical school.

It is important to note that the quality and reliability of educational content on SoMe can vary widely [28,29]. There is also a need to consider how to produce and validate medical educational content on these platforms. Some authors proposed their personal guidelines on this topic, but there is still a lack of large consensus on how to provide medical education on SoMe [30,31]. Medical teachers who publish and moderate content on SoMe should probably be valorized at an institutional level, as they play a significant role in the current dissemination of knowledge to students. However, further research is still needed to evaluate the impact of these resources on students' medical knowledge, skills, and professional development.

We describe that 79.1% (603/762) of students follow a physician they already know on SoMe and that 68.5% (522/762) use SoMe to help them choose a specialty. This is an important message, showing that physicians who are active on SoMe are seen (and are likely to be imitated) by students. This reinforces the absolute necessity to maintain perfect professionalism when communicating on SoMe as a health professional or as a teacher. Our results also suggest that SoMe could be an effective communication tool for medical academic societies to promote and present their specialty among students.

Misuse of SoMe in a Professional Context

Our study also identified significant misuse of social networks in a professional context. Posting content related to hospital internship was reported by 27.2% (207/762) of respondents, and 10.8% (82/762) admitted to searching for patients' names on SoMe platforms. Second-cycle students were more likely to present misuse with the risk of a breach of confidentiality, which is consistent with the fact that they spend more time on in-hospital internships than first-cycle students. We also observed a gender difference in misuse behavior, with female respondents more likely to post internship-related content and male respondents more likely to search for patient information. These results have already been reported in other studies [32,33]. These behaviors pose ethical and legal concerns, as they can lead to breaches of confidentiality and compromise patient privacy. Additionally, these actions may have severe consequences for medical students and their future careers, including disciplinary action and damage to their professional reputation. French law strictly forbids sharing any medical information with anyone other than the caring team, and any information apart from the ones strictly necessary to the patient's course of treatment or care, except for some listed exceptions [34].

The platforms that seemed to be more problematic were Snapchat and Instagram, possibly due to the sensationalist nature of sharing photographs or videos, particularly ephemeral ones ("stories"). This is particularly worrying, as we recently showed that, on SoMe, photographs are much more likely to breach medical confidentiality than written posts [35]. In contrast, YouTube, which requires more effort to upload a video, appears much less likely to be used to share such content. Therefore, we suggest for the first time that the risk of SoMe misuse and medical confidentiality breaches seem to vary greatly among the different platforms. This information should probably be used to provide relevant information on the individual risks of each SoMe, which are not a uniform entity.

This problem of unprofessional use of SoMe by medical students is the subject of significant research literature, with which our results are consistent. Some authors report interesting results after implementing "social media and professionalism" course in medical school, with improvement in students' SoMe behavior [36]. It is crucial for all medical schools and health care institutions to address these issues and promote the responsible use of SoMe among their medical students through the development of policies and educational interventions. SoMe misuse, particularly breaches of confidentiality, is not only a problem among students but also among physicians. Ahmed et al [32] identified tweets from 656 health care professionals' Twitter profiles, including 486 (74.1%) doctors. Through these tweets, friends and family were able to identify clinical scenarios in 242 of the 754 (32.1%) tweets. In a study of the profiles of anesthesia and intensive care professionals, 5.3% of doctors' accounts had posted content posing a confidentiality problem [35].

Prevalence and Impact of Smartphone Addiction

Our findings indicate a concerning prevalence of smartphone addiction among medical students, with nearly one-third

XSL•FO
RenderX

(222/762, 29.1%) of respondents meeting the criteria for addiction. This finding is consistent with the literature, which reports a smartphone addiction rate between 15% and 40% among students [19,37-39]. The SAS-SV score for addiction diagnosis is well established, and this scale has been already validated in French, increasing its reliability and reproducibility [26]. Our results showed a significant correlation between increased length of SoMe use and SAS-SV score. It is impossible to establish a causality link here, as students can become addicted to these platforms through overuse, but they may also use them intensively because of their addiction. In a Norwegian student population, Hjetland et al [40] also found a strong association between addiction and daily time spent on SoMe, particularly when this usage occurred in the evening.

The negative psychological, social, and physical effects of smartphone addiction are well described—lower academic performance, sleep disorders, anxiety, musculoskeletal disorders, severe social withdrawal, decreased physical activity, and hypertension [16-19]. It is thus recognized that this addiction has a negative impact at the individual level. However, our results also showed that addicted students reported more impact on their study time and a higher tendency to share hospital internship content on social networks, suggesting that students' smartphone addiction could also have a negative impact on patients. There are no data on the potential link between medical students' smartphone addiction and unprofessional behavior on SoMe, and further works are needed to explore this hypothesis.

Limitations

Despite these interesting results, our work has several limitations. First, the response rate was low. Even if we were unable to determine the total number of medical students who received the survey invitation, the targeted population is approximately 42,000 students (which would give a response rate of 1.8% if we consider that all students received the invitation). As a result, it is not possible to estimate the nonresponse bias, which is probably significant, as students who chose to participate in the study might have different characteristics or behaviors than those who did not. Additionally, the study relied on self-reported data, which may be subject to recall and social desirability biases. Participants might have underreported their SoMe use or smartphone addiction due to concerns about stigma or privacy. Furthermore, the mode of distribution, an online survey distributed via email, may have introduced a selection bias. It is possible that students who are more active on SoMe were more likely to respond to the survey, potentially leading to an overestimation of the prevalence of SoMe use and smartphone addiction among the general population of medical students. Conversely, respondents may have different patterns of SoMe use compared to nonrespondents, which we were unable to assess.

Moreover, the response rate among female students (547/762, 71.8%) was significantly higher compared to male students. According to national statistics, the gender distribution in French medical schools was approximately 66% female and 34% male in 2021 [41]. This overrepresentation of female respondents might have influenced the results, as female students could have different SoMe usage patterns and concerns compared to their male counterparts. This limits the generalizability and interpretation of our findings to the overall population of medical students.

To address these limitations, future research should aim to achieve a higher and balanced response rate, possibly by using multiple distribution methods and follow-ups to reach a more representative sample of the student population. Additionally, qualitative studies could explore the reasons behind nonresponse and differences in SoMe patterns among different student demographics.

The use of the SAS-SV as the sole instrument to assess smartphone addiction has its limitations. While the SAS-SV has demonstrated good validity and reliability, it may not capture the full spectrum of addictive behaviors related to smartphone use. Further research could enable a better understanding of these behaviors in the medical student population and, more specifically, their impact on medical students' results at the national final examination at the end of the sixth year of the medical course. Furthermore, the cross-sectional design of our study does not allow us to establish causal relationships between the use of social networks, smartphone addiction, and the potential consequences on medical students' academic performance and well-being. Longitudinal studies would be required to better understand the directionality of these relationships. In addition, the psychometric properties of the "use of social network" section were not formally assessed, which is acknowledged as a limitation of this study. Future research should include a comprehensive psychometric evaluation to confirm the reliability and validity of the survey instrument.

Conclusions

In conclusion, this study highlights the extensive use of SoMe for medical education among respondents and the concerning prevalence of smartphone addiction. Educators should recognize the potential of these platforms, promote responsible use, and address addiction issues. Further research is needed to optimize SoMe usage for medical education while minimizing risks associated with excessive use and addiction.

Acknowledgments

The authors would like to express their gratitude to the ANEMF (Association des Etudiants en Médecine de France) and all the student representatives for their assistance in disseminating this questionnaire.

XSL•FO
RenderX

Authors' Contributions

TC performed the conceptualization, methodology, investigation, formal analysis, and writing—original draft. EC contributed to the investigation, formal analysis, and writing—review and editing. ZD contributed to the supervision and writing—review and editing. BV performed the supervision and writing—review and editing. IAM performed the investigation, formal analysis, and writing—review and editing. BP contributed to the formal analysis, data curation, visualization, and writing—review and editing. All authors read and approved the final manuscript. The corresponding author had full access to all of the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original survey.
[DOCX File, 26 KB - mededu_v10i1e55149_app1.docx]

Multimedia Appendix 2
Web-based survey in English.
[DOCX File, 22 KB - mededu_v10i1e55149_app2.docx]

Multimedia Appendix 3
Secondary analyses.
[DOCX File, 216 KB - mededu_v10i1e55149_app3.docx]

Checklist 1
Reporting Results of Internet E-Surveys (CHERRIES) checklist.
[DOCX File, 27 KB - mededu_v10i1e55149_app4.docx]

References

1. Cheston CC, Flickinger TE, Chisolm MS. Social media use in medical education: a systematic review. Acad Med J Assoc Am Med Coll 2013 Jun;88(6):893-901. [doi: 10.1097/ACM.0b013e31828ffc23] [Medline: 23619071]
2. Sutherland S, Jalali A. Social media as an open-learning resource in medical education: current perspectives. Adv Med Educ Pract 2017;8:369-375. [doi: 10.2147/AMEP.S112594] [Medline: 28652840]
3. Guckian J, Utukuri M, Asif A, et al. Social media in undergraduate medical education: a systematic review. Med Educ 2021 Nov;55(11):1227-1241. [doi: 10.1111/medu.14567] [Medline: 33988867]
4. von Muhlen M, Ohno-Machado L. Reviewing social media use by clinicians. J Am Med Inform Assoc 2012;19(5):777-781. [doi: 10.1136/amiajnl-2012-000990] [Medline: 22759618]
5. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. Pharm Ther 2014 Jul;39(7):491-520. [Medline: 25083128]
6. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med 2018 May 28;7(6):121. [doi: 10.3390/jcm7060121] [Medline: 29843360]
7. How young people consume news and the implications for mainstream media. : Reuters Institute for the Study of Journalism; 2021 URL: https://reutersinstitute.politics.ox.ac.uk/our-research/how-young-people-consume-news-and-implications-mainstream-media [accessed 2024-10-16]
8. Global social media statistics. DataReportal. URL: https://datareportal.com/social-media-users [accessed 2023-05-07]
9. Thamman R, Gulati M, Narang A, Utengen A, Mamas MA, Bhatt DL. Twitter-based learning for continuing medical education? Eur Heart J 2020 Dec 7;41(46):4376-4379. [doi: 10.1093/eurheartj/ehaa346] [Medline: 32338736]
10. Sterling M, Leung P, Wright D, Bishop TF. The use of social media in graduate medical education: a systematic review. Acad Med 2017;92(7):1043-1056. [doi: 10.1097/ACM.0000000000001617]
11. Hill SS, Dore FJ, Em ST, et al. Twitter use among departments of surgery with general surgery residency programs. J Surg Educ 2021;78(1):35-42. [doi: 10.1016/j.jsurg.2020.06.008] [Medline: 32631768]
12. Kauffman L, Weisberg EM, Zember WF, Fishman EK. #RadEd: how and why to use Twitter for online radiology education. Curr Probl Diagn Radiol 2021;50(3):369-373. [doi: 10.1067/j.cpradiol.2021.02.002] [Medline: 33637393]
13. Dedeilia A, Sotiropoulos MG, Hanrahan JG, Janga D, Dedeilias P, Sideris M. Medical and surgical education challenges and innovations in the COVID-19 era: a systematic review. In Vivo 2020 Jun;34(3 Suppl):1603-1611. [doi: 10.21873/invivo.11950] [Medline: 32503818]
14. Yang S, Jin C, Wang J, Xu X. The use of social media to deliver surgical education in response to the COVID-19 pandemic. J Invest Surg 2022 Jun;35(6):1350-1356. [doi: 10.1080/08941939.2022.2035859] [Medline: 35130457]

XSL•FO
RenderX

15. Fischer-Grote L, Kothgassner OD, Felnhofer A. Risk factors for problematic smartphone use in children and adolescents: a review of existing literature. Neuropsychiatr 2019 Dec;33(4):179-190. [doi: 10.1007/s40211-019-00319-8] [Medline: 31493233]
16. Ren Z, Tan J, Huang B, et al. Association between 24-hour movement behaviors and smartphone addiction among adolescents in Foshan City, Southern China: compositional data analysis. Int J Environ Res Public Health 2022 Aug 12;19(16):9942. [doi: 10.3390/ijerph19169942] [Medline: 36011576]
17. Zou Y, Xia N, Zou Y, Chen Z, Wen Y. Smartphone addiction may be associated with adolescent hypertension: a cross-sectional study among junior school students in China. BMC Pediatr 2019 Sep 4;19(1):310. [doi: 10.1186/s12887-019-1699-9] [Medline: 31484568]
18. Alshahrani A, Samy Abdrabo M, Aly SM, et al. Effect of smartphone usage on neck muscle endurance, hand grip and pinch strength among healthy college students: a cross-sectional study. Int J Environ Res Public Health 2021 Jun 10;18(12):6290. [doi: 10.3390/ijerph18126290] [Medline: 34200762]
19. Tateno M, Teo AR, Ukai W, et al. Internet addiction, smartphone addiction, and hikikomori trait in Japanese young adult: social isolation and social network. Front Psychiatry 2019;10:455. [doi: 10.3389/fpsyt.2019.00455] [Medline: 31354537]
20. Thompson LA, Black E, Duff WP, Paradise Black N, Saliba H, Dawson K. Protected health information on social networking sites: ethical and legal considerations. J Med Internet Res 2011 Jan 19;13(1):e8. [doi: 10.2196/jmir.1590] [Medline: 21247862]
21. Burns KEA, Duffett M, Kho ME, et al. A guide for the design and conduct of self-administered surveys of clinicians. Can Med Assoc J 2008 Jul 29;179(3):245-252. [doi: 10.1503/cmaj.080372]
22. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
23. Alder S. HIPAA and social media guidelines. The HIPAA Journal. 2024. URL: https://www.hipaajournal.com/hipaa-social-media/ [accessed 2024-06-12]
24. Borgmann H, Cooperberg M, Murphy D, et al. Online professionalism-2018 update of European Association of Urology (@Uroweb) recommendations on the appropriate use of social media. Eur Urol 2018 Nov;74(5):644-650. [doi: 10.1016/j.eururo.2018.08.022] [Medline: 30177286]
25. Kwon M, Kim DJ, Cho H, Yang S. The smartphone addiction scale: development and validation of a short version for adolescents. PLoS One 2013;8(12):e83558. [doi: 10.1371/journal.pone.0083558] [Medline: 24391787]
26. Lopez-Fernandez O. Short version of the Smartphone Addiction Scale adapted to Spanish and French: towards a cross-cultural research in problematic mobile phone use. Addict Behav 2017 Jan;64:275-280. [doi: 10.1016/j.addbeh.2015.11.013] [Medline: 26685805]
27. Mirzaei A, Carter SR, Patanwala AE, Schneider CR. Missing data in surveys: key concepts, approaches, and applications. Res Soc Admin Pharm 2022 Feb;18(2):2308-2316. [doi: 10.1016/j.sapharm.2021.03.009]
28. Hillman T, Sherbino J. Social media in medical education: a new pedagogical paradigm? Postgrad Med J 2015 Oct;91(1080):544-545. [doi: 10.1136/postgradmedj-2015-133686] [Medline: 26338982]
29. D'Souza RS, D'Souza S, Sharpe EE. YouTube as a source of medical information about epidural analgesia for labor pain. Int J Obstet Anesth 2021 Feb;45:133-137. [doi: 10.1016/j.ijoa.2020.11.005] [Medline: 33339713]
30. Khalid F, Wu M, Ting DK, et al. Guidelines: the do's, don'ts and don't knows of creating open educational resources. Perspect Med Educ 2023;12(1):25-40. [doi: 10.5334/pme.817] [Medline: 36908747]
31. Lowe-Calverley E, Barton M, Todorovic M. Can we provide quality #MedEd on social media? Trends Mol Med 2022 Dec;28(12):1016-1018. [doi: 10.1016/j.molmed.2022.08.002] [Medline: 36008252]
32. Ahmed W, Jagsi R, Gutheil TG, Katz MS. Public disclosure on social media of identifiable patient information by health professionals: content analysis of Twitter data. J Med Internet Res 2020 Sep 1;22(9):e19746. [doi: 10.2196/19746] [Medline: 32870160]
33. Koohikamali M, Peak DA, Prybutok VR. Beyond self-disclosure: disclosure of information about others in social network sites. Comput Human Behav 2017 Apr;69:29-42. [doi: 10.1016/j.chb.2016.12.012]
34. Article L1110-4 - code de la santé publique. Code de la santé publique. 2021. URL: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000043895798 [accessed 2023-04-30]
35. Pineau I, Pineau M, Selim J, et al. Evaluation of medical confidentiality breaches on Twitter among anesthesiology and intensive care health care workers. Anesth Analg 2023 Aug 1;137(2):418-425. [doi: 10.1213/ANE.0000000000006540] [Medline: 37227950]
36. Guraya SS, Yusoff MSB, Rashid-Doubell F, et al. Changing professional behaviors in the digital world using the Medical Education e-Professionalism (MEeP) framework-a mixed methods multicentre study. Front Med (Lausanne) 2022;9:846971. [doi: 10.3389/fmed.2022.846971] [Medline: 35425778]
37. Rathakrishnan B, Bikar Singh SS, Kamaluddin MR, et al. Smartphone addiction and sleep quality on academic performance of university students: an exploratory research. Int J Environ Res Public Health 2021 Aug 5;18(16):8291. [doi: 10.3390/ijerph18168291] [Medline: 34444042]
38. Tayhan Kartal F, Yabancı Ayhan N. Relationship between eating disorders and internet and smartphone addiction in college students. Eat Weight Disord 2021 Aug;26(6):1853-1862. [doi: 10.1007/s40519-020-01027-x] [Medline: 33034868]

XSL•FO
RenderX

39. Haug S, Castro RP, Kwon M, Filler A, Kowatsch T, Schaub MP. Smartphone use and smartphone addiction among young people in Switzerland. J Behav Addict 2015 Dec;4(4):299-307. [doi: 10.1556/2006.4.2015.037] [Medline: 26690625]
40. Hjetland GJ, Skogen JC, Hysing M, Sivertsen B. The association between self-reported screen time, social media addiction, and sleep among Norwegian university students. Front Public Health 2021;9:794307. [doi: 10.3389/fpubh.2021.794307] [Medline: 34976935]
41. Les effectifs d'étudiants dans le supérieur continuent leur progression en 2021-2022. Ministére de l'Enseignement supérieur et de la Recherche. 2022. URL: https://www.enseignementsup-recherche.gouv.fr/fr/les-effectifs-d-etudiants-dans-le-superieur-continuent-leur-progression-en-2021-2022-88609 [accessed 2024-06-10]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
HIPAA: Health Insurance Portability and Accountability Act
SAS-SV: Smartphone Addiction Scale Short-Version
SoMe: social media

Edited by B Lesselroth; submitted 04.12.23; peer-reviewed by L Park, M Wu; revised version received 12.06.24; accepted 19.08.24; published 22.10.24.

Please cite as:
Clavier T, Chevalier E, Demailly Z, Veber B, Messaadi IA, Popoff B
Social Media Usage for Medical Education and Smartphone Addiction Among Medical Students: National Web-Based Survey
JMIR Med Educ 2024;10:e55149
URL: https://mededu.jmir.org/2024/1/e55149
doi:10.2196/55149

©Thomas Clavier, Emma Chevalier, Zoé Demailly, Benoit Veber, Imad-Abdelkader Messaadi, Benjamin Popoff. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.10.2024. 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.

XSL•FO
RenderX

Original Paper

A SIMBA CoMICs Initiative to Cocreating and Disseminating Evidence-Based, Peer-Reviewed Short Videos on Social Media: Mixed Methods Prospective Study

Maiar Elhariry[1*], MBChB, FHEA; Kashish Malhotra[2,3*], MBBS, SFHEA; Kashish Goyal[4,5], MBBS, MD; Marco Bardus[2], BA, MA, PhD; SIMBA and CoMICs Team[6]; Punith Kempegowda[2,7], MBBS, MSc, MD, MRCP, SFHEA, PhD

[1]Sandwell General Hospital, Sandwell and West NHS Trust, Birmingham, United Kingdom
[2]Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom
[3]Rama Medical College Hospital and Research Centre, Hapur, India
[4]Delhi Heart Institute and Multispeciality Hospital, Bathinda, India
[5]School of Medical Sciences & Research, Sharda University, Greater Noida, India
[6]see Authors' Contributions
[7]Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Punith Kempegowda, MBBS, MSc, MD, MRCP, SFHEA, PhD
Applied Health Sciences, School of Health Sciences
College of Medicine and Health
University of Birmingham
Edgbaston
Birmingham, B15 2TT
United Kingdom
Phone: 44 7721930777
Email: p.kempegowda@bham.ac.uk

Abstract

Background: Social media is a powerful platform for disseminating health information, yet it is often riddled with misinformation. Further, few guidelines exist for producing reliable, peer-reviewed content. This study describes a framework for creating and disseminating evidence-based videos on polycystic ovary syndrome (PCOS) and thyroid conditions to improve health literacy and tackle misinformation.

Objective: The study aims to evaluate the creation, dissemination, and impact of evidence-based, peer-reviewed short videos on PCOS and thyroid disorders across social media. It also explores the experiences of content creators and assesses audience engagement.

Methods: This mixed methods prospective study was conducted between December 2022 and May 2023 and comprised five phases: (1) script generation, (2) video creation, (3) cross-platform publication, (4) process evaluation, and (5) impact evaluation. The SIMBA-CoMICs (Simulation via Instant Messaging for Bedside Application–Combined Medical Information Cines) initiative provides a structured process where medical concepts are simplified and converted to visually engaging videos. The initiative recruited medical students interested in making visually appealing and scientifically accurate videos for social media. The students were then guided to create video scripts based on frequently searched PCOS- and thyroid-related topics. Once experts confirmed the accuracy of the scripts, the medical students produced the videos. The videos were checked by clinical experts and experts with lived experience to ensure clarity and engagement. The SIMBA-CoMICs team then guided the students in editing these videos to fit platform requirements before posting them on TikTok, Instagram, YouTube, and Twitter. Engagement metrics were tracked over 2 months. Content creators were interviewed, and thematic analysis was performed to explore their experiences.

Results: The 20 videos received 718 likes, 120 shares, and 54,686 views across all platforms, with TikTok (19,458 views) and Twitter (19,678 views) being the most popular. Engagement increased significantly, with follower growth ranging from 5% on Twitter to 89% on TikTok. Thematic analysis of interviews with 8 out of 38 participants revealed 4 key themes: views on social

XSL•FO
RenderX

media, advice for using social media, reasons for participating, and reflections on the project. Content creators highlighted the advantages of social media, such as large outreach (12 references), convenience (10 references), and accessibility to opportunities (7 references). Participants appreciated the nonrestrictive participation criteria, convenience (8 references), and the ability to record from home using prewritten scripts (6 references). Further recommendations to improve the content creation experience included awareness of audience demographics (9 references), sharing content on multiple platforms (5 references), and collaborating with organizations (3 references).

Conclusions: This study demonstrates the effectiveness of the SIMBA CoMICs initiative in training medical students to create accurate medical information on PCOS and thyroid disorders for social media dissemination. The model offers a scalable solution to combat misinformation and improve health literacy.

(JMIR Med Educ 2024;10:e52924) doi:10.2196/52924

KEYWORDS

influencers; social media; public engagement; apps; healthcare; medical students; online medical information; simulation; peer-reviewed information

Introduction

In July 2023, there were more than 4.9 billion social media users globally, equating to over 61% of the world's population [1]. Social media usage has increased by 3.7% in the past year, with 173 million new users (5.5 new users every second). Checking one's social media profile has become a predominant activity for 9 out of 10 internet users. Furthermore, 7 of the top 10 most popular social media platforms claim over 1 billion monthly active users. These are Facebook (Meta; 2.989 billion), YouTube (Google; 2.537 billion), WhatsApp (Meta) and Instagram (Meta; 2 billion), WeChat Inc or Weixin (Tencent; 1.319 billion), TikTok (ByteDance; 1.081 billion), and Facebook Messenger (Meta; 1.038 billion). In addition, 4 platforms are owned by the same company, Meta (Facebook, Instagram, WhatsApp, and Facebook Messenger). These are followed by Snapchat (Snap Inc, 750 million users), Douyin (ByteDance, 730 million users), and Telegram (Telegram Messenger Inc, 700 million users). Social media reach extends beyond personal interactions on each platform as users adopt multiple platforms. For example, nearly 78% of Facebook users also use Instagram.

GWI's data from DataReportal reported that 49% of active users aged 16 to 64 (outside China) use social media to keep in touch with friends and family, 37% to fill spare time, 35% to read new stories, 30% to find content (studies and videos). While some platforms are used for passive entertainment (eg, TikTok), Instagram, Facebook, and Snapchat are used for content creation by sharing posts and videos. Given its reach, social media has also emerged as a powerful tool for promoting health and disseminating health-related research findings, surgical education, and medical information [2-4].

However, the role of social media in sharing clinical experiences is complex and carries potential benefits and challenges. Some benefits include rapid and wide dissemination of information at minimal costs to the end user, bridging the gap to health care access, and patient education [5]. However, this unprecedented access to information may also provide a breeding ground for misinformation. This spread of misleading or false information can have dire consequences in health care, where accurate knowledge is crucial for making informed decisions about one's well-being [6]. This risk comes with the exponential growth of short video platforms such as TikTok, mimicked by Instagram reels, Facebook stories, and YouTube shorts, whose algorithms tend to propose similar content based on the users' histories and preferences. Short video platforms are echo chambers [7] that reinforce beliefs, prejudices, fake news, and misinformation. There is a need to address this by producing evidence-based content to ensure the dissemination of accurate information without bias [7,8].

Considering the negative consequences of misinformation, in the last 2 years, the World Health Organization partnered with major technology companies that have leverage on major social networking sites such as Alphabet (Google and YouTube) [9] and Meta (Facebook, Instagram, and WhatsApp) [10] to minimize misinformation. Solutions included semiautomated flagging, labeling, or removing content that violates community guidelines and misinformation policies [11]. Google and YouTube have recently invested US $13.2 million in the International Fact-Checking Network [12] to enhance misinformation response. Video creation platforms, such as YouTube, Twitch, or TikTok, have their community guidelines, which try to regulate the content produced. Nevertheless, there is little to no guidance on creating content before it is uploaded on social media platforms.

Regarding medical or health-related content, social media platforms generally include content that rarely reflects clinical guidelines (eg, low back pain and laparoscopic hysterectomy) [13,14]. Many studies report methods for evaluating medical content on social media, specifically YouTube [15], but few describe medical education videos' development, implementation, and evaluation.

Since, to the best of our knowledge, there are no specific international guidelines to create evidence-based medical and health content related to polycystic ovary syndrome (PCOS) and thyroid disorders, an international medical education initiative was launched to create evidence-based and peer-reviewed bite-sized videos on various medical conditions in collaboration with various patient support groups [16-19]. The initiative named "SIMBA CoMICs" (Simulation via Instant Messaging for Bedside Application–Combined Medical Information Cines) involves medical students, junior doctors, and patient groups who collaborate to create bite-sized videos for different social media platforms. Combined Medical

XSL•FO
RenderX

Information Cines (CoMICs) initiative is a novel approach merging intricate medical concepts with illustrative graphics, presented in video format for swift assimilation. Each CoMIC video meticulously portrays distinct medical conditions, encompassing their presentations, diagnostic tools, subsequent treatment options, and recommended follow-up measures. The content for each presentation aligns with national and international guidelines and undergoes rigorous evaluation by leading experts in the corresponding medical domain.

This study describes generating, creating, disseminating, and evaluating evidence-based, peer-reviewed, short social media videos about PCOS and thyroid diseases. The project followed a collaborative approach with people living with these conditions and health care professionals with a special interest in these conditions. We analyzed the project outreach and audience engagement on social media (specifically TikTok, Instagram reels, YouTube shorts, and Twitter). Furthermore, we gathered participant experience for effective engagement and evidence synthesis. This examination of public discourse through social media platforms provides insights into disseminating and perpetuating viewpoints regarding thyroid conditions and PCOS.

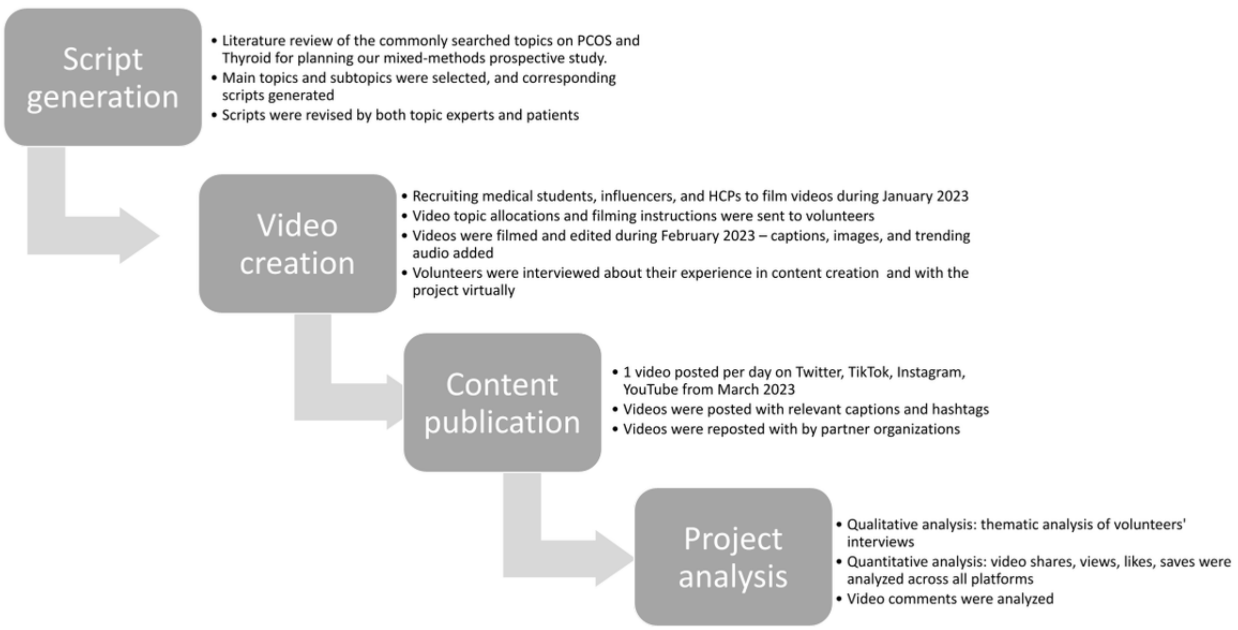
Methods

The SIMBA-CoMICs (Simulation via Instant Messaging for Bedside Application–Combined Medical Information Cines) initiative provides a structured framework for simplifying medical concepts and converting them into visually engaging, evidence-based videos. The initiative actively recruited medical students and resident doctors, who were guided through the script creation and video production process. The SIMBA-CoMICs initiative emphasised on collaboration among health care professionals, content creators, and individuals with lived experience of PCOS and thyroid conditions to ensure that the final content was scientifically accurate and relatable to the public. This structured approach ensured consistency across all phases of the project, from initial script development to final video dissemination on social media platforms.

Design

This mixed methods prospective study was conducted between December 2022 and May 2023 and consisted of 5 phases: script generation, video creation, cross-platform publication, process, and impact evaluation (Figure 1). This study was inspired by the “Knowledge-to-action (KTA)” framework, whereby the content of the videos is based on scientific evidence synthesis, translation, and validation through expert consultation and patient engagement [20]. The content is then produced, and its fruition is monitored, evaluated, and critically appraised for improvement and for generating recommendations.

Figure 1. Overview of the main steps to complete this project. HCP: health care professionals; PCOS: polycystic ovary syndrome.



Script Generation

A literature search of the frequently asked questions in PCOS and thyroid conditions was conducted across various databases, including PubMed, Cochrane Library, and Embase. Key MeSH (Medical Subject Headings) terms, including “frequent,” “questions,” “popular,” “topics,” “Thyroid,” “conditions,” “PCOS,” and “Polycystic Ovary Syndrome,” were explored across all databases in this scoping review. Based on this, a list of the most popular topics was created. The experts and representatives of patient support groups reviewed the topics independently to arrive at 5 main topics for both PCOS and thyroid. Each topic was subdivided into 3-5 subtopics. We then created corresponding scripts and guidance on when and how to seek medical intervention or advice. The scripts’ scientific accuracy and ease of understandability were checked by 2 early

XSL•FO
RenderX

career researchers, a consultant endocrinologist, and members of patient support groups to ensure they align with the relevant international guidelines.

Video Creation

We invited medical students and junior doctors globally to contribute to this project by advertising the role of video creation on our social media handles (Twitter: SIMBAsimulation, TikTok: simbacomic, Instagram: simba.comic, YouTube shorts: @simbasimulation8047). Medical students at the University of Birmingham and other authors' institutes were also invited by email to express interest in creating these videos. Similar invites were sent to members of UK-based and international patient support groups supported by partnering institutions (PCOS Vitality and British Thyroid Foundation). Patient support groups were invited to contribute to the videos. Participants were incentivized by certificates acknowledging their contribution to this project. The invitations included a brief role description and instructions on the format that the volunteers will be asked to film themselves.

Upon expressing interest, each volunteer received an email with the script of the content they were allocated in the video series. The emails also included a sample video for participants to visualize the expected product. Participants were requested to film in portrait orientations in keeping with the main format of short videos on TikTok, Instagram reels, YouTube shorts, and Twitter. In addition, we advised participants to keep the videos under 30 seconds to match the average video duration on social media [21]. Participants who submitted a video were invited to create more videos within an agreed set of personalized deadlines. A total of 5 volunteers agreed to this. After we received all the videos, we added the transcripts to the video as captions, one line at a time, and highlighted keywords, with the relevant pictures and emojis to help the audience better understand and interpret the videos. We edited the videos where necessary to ensure they did not exceed 30 seconds. The finalized videos were shared with early-career endocrinologists and members of patient support groups, who reviewed them and helped finalize the transcripts. Further changes were made to the videos to ensure accuracy and acceptability.

Content Publication

The finalized videos were uploaded to TikTok to be edited using the app's video editing features, followed by sharing them as Instagram reels, YouTube shorts, and Twitter as a series of video posts. A target minimum of one video was posted every day between March 2, 2023, and March 26, 2023, with a 1-day break after each video series. The videos were posted with their subheading as the video's public title, and relevant, popular hashtags were included for each video to stimulate the social media algorithm and help redirect the videos to interested audiences. Volunteers were tagged in the videos they filmed if they had an account on the social media platform used after their consent.

Process Evaluation: Participants' Experience

We invited our content creators to share their views about the project in a semistructured online interview. The interview specifically aimed to explore their experiences, motivation to participate, and interest in developing similar videos for other subjects. Participants were reassured that their responses would be anonymized and would not impact whether they received a certificate for contributing to this project. These interviews were conducted on Zoom (Zoom Video Communications, Inc) and lasted approximately 10 minutes. Each interview was recorded after consent, and participants were allowed to have their cameras switched off during the meeting. All study data were stored in a password-protected folder, with only the study team having access to it. All questions were asked to each participant following the set order, as indicated in the interview guide provided in Multimedia Appendix 1. Interviews with the participants were anonymized and transcribed verbatim. Furthermore, 2 independent authors did the coding of interview transcripts using NVivo (version 12.0). The codes were combined to identify themes using thematic inductive analysis [22,23]. The research team discussed the codes and agreed on the thematic structure proposed.

Impact Evaluation: Quantitative and Thematic Analysis

We evaluated the impact of the produced videos using a combination of quantitative and qualitative data. Quantitative data analyses were based on video analytics on each platform, including the total number of views, the highest number of views on a video, total shares, total number of saved videos, total likes, and change in the number of followers or subscribers were extracted along with other similar variables across all platforms after a month of publishing the video. Considering the proprietary algorithms, recommendations, and search engines followed by each platform, comparative statistical tests were not run across platforms. Qualitative data was based on a content analysis of the comments posted under each video to understand the viewers' perspectives. Comments were inductively coded.

Ethical Considerations

No patient data were collected, and this study was approved by the ethics committee of Delhi Heart Institute and Multispecialty Hospital (DHIMH/IEC/2023-008). Informed consent was taken from each participant, and participation was voluntary.

Results

Script Generation

The topics and subtopics of the videos we generated through our literature searches and validated through expert consultations are presented in Textbox 1.

XSL•FO
RenderX

Textbox 1. Outline of topics and subtopics of the videos.

Thyroid nodules:

- Definition and differentials
- Red flag symptoms
- Investigations
- Management

Hypothyroidism:

- Symptoms and signs
- Investigations
- Risk factors
- Management

Diagnosing polycystic ovary syndrome (PCOS):

- Diagnostic criteria
- Symptoms
- Associated complications
- Risk factors
- Emotional well-being

PCOS implications and associations:

- Impact on menstruation
- Impact on pregnancy and its likelihood
- The link between PCOS and thyroid

Thyroid and pregnancy:

- Conception
- Effect of hypothyroidism
- Effect of hyperthyroidism
- Impact on infants
- Breastfeeding

Study Participants and Content Created

We recruited 38 content creators, mostly students (33/38, 87%), 2 people with 1 or more of the conditions, and 2 social media influencers. In total, 11 students and 1 self-identified influencer created 21 videos.

Process Evaluation: Participants' Experience

Out of the 12 volunteers who filmed the videos, 8 (67%) completed an interview on their experiences and views on the project. Thematic analysis of the anonymized interview transcripts yielded 4 main themes, that are, views on social media, advice when using social media, the reason for taking part in this project, and thoughts on this project (Table 1).

Table 1. Outline of central themes and subthemes from the thematic analysis of participant interviews.

Theme	Subthemes
Views on social media	Disadvantages, advantages, and uses of social media.
Advice on making the most of social media	Factors impacting public engagement, how to improve engagement, and general advice.
Reason for taking part in this project	Barriers, motivation, and previous experience.
Thoughts on this project	Positive aspects, tips to improve the project, and project outreach.

XSL•FO
RenderX

Views on Social Media

Participants highlighted several advantages of social media, including “large outreach” (12 references), “convenience of getting things done from the comfort of their homes” (10 references), and “accessibility to opportunities” in fields of interest (7 references). Participants also noted that social media is beneficial for “finding information” (8 references), advertising and fundraising (5 references), and expanding audience and outreach (7 references). However, several participants shared their concerns about social media. A total of 7 participants referenced “misinformation” as a key threat in social media that should be taken seriously. Some participants expanded to explain the consequences of inaccurate information, such as anxiety (5 references), confusion (3 references), or a false sense of reassurance (3 references). A total of 4 participants commented on the potential of “wasting time” on social media.

Advice on Making the Most of Social Media

The most common themes were being aware of “audience’s demographics,” “sharing on more than one social media platform” (5 references), “collaborating with well-known organizations” (3 references), and “linking the videos to reliable web pages with more information” (2 references). There were also comments on how to maximize positive outcomes and minimize negative ones when using social media generally, including “the importance of monitoring and limiting the unproductive time spent on it” (9 references) and the importance of verifying any information for an unknown source (7 references).

Thoughts on the Project

Participants mentioned “non-restricting participation criteria,” “convenience” (n=8), and “ability to record videos from home with a pre-written script” (6 references), which made it a lot easier to participate. In addition, 3 participants mentioned the time needed to memorize the scripts (3 references), the need to step out of their comfort zone (1 reference), and the dates the volunteers were recruited (1 reference), made their participation challenging.

Tips to improve the project included having a “meetup with other volunteers” (6 references), increasing the “variety of video formats” submitted (4 references), and “extending the period that students had to submit their project” (2 references). Participants believed the project is beneficial for people “impacted by the respective conditions” (16 references), “healthcare professionals” (7 references), and “medical students” (6 references). All participants also stated that they believe this project design is “sustainable” for the long-term (n=8).

Impact Evaluation

Quantitative Assessment

The 21 videos generated a total of 108,210 views across platforms, with Twitter and TikTok generating the most views (n=47,342), followed by Instagram (n=13,207), and YouTube (n=13,773), as of September 19, 2023, as summarized in Table 2 below. The highest number of views and shares on a single video were found in TikTok (4327 views, 57 shares), followed by Twitter (4176 views, 53 shares), whereas YouTube had the lowest views and shares (738 and 0, respectively). TikTok was also the platform with the highest number of likes, saves or bookmarks, and likes and saves for a single video, compared with the other platforms. The videos viewed and shared the most across all platforms were those about thyroid nodules, thyroid and pregnancy, and PCOS and thyroid. Between March 2, 2023, and March 26, 2023, the project’s social media profiles had an overall increase of 259 subscribers or followers (64.8 on average), with 178 new users subscribed on YouTube, followed by TikTok (n=31), Twitter (n=28), and Instagram (n=22). TikTok was the best-performing platform for new users and subscribers.

Table 2. Summary of data from the videos posted across all social media platforms as quantified on May 25, 2023.

Indicators and platforms	TikTok	Instagram	YouTube	Twitter	Total (average, SD)
Total number of views	47,342	13,207	13,773	28,888	108,210 (27,052, 13,935.80)
Highest number of views for a single video	12,140	1440	4118	4252	21,950 (5488, 4001.25)
Total likes	611	124	138	94	967 (241.5, 213.78)
Most likes for a single video	114	20	33	14	211 (53, 40.28)
Total shares	82	20	0	53	155 (39, 31.33)
Highest number of shares for a single video	24	3	0	9	36 (9, 9.25)
Total number of saves or bookmarks of videos	101	9	0	0	110 (27.5, 42.59)
Highest number of saves for a single video	29	3	—[a]	—	32 (16, 13)
Change in number of followers or subscribers (between March 2023 and September 2023)	+38	+22	+356	+128	+544 (136, 133.29)
% change in the number of followers or subscribers (between March 2023 and May 2023)	+950%	+8%	+32%	+26%	—

[a]Not applicable.

XSL•FO
RenderX

Qualitative Evaluation of Comments

A total of 38 comments were posted across all platforms as of September 19, 2023. Out of these, 17 were either emoji-based comments or commended the video without providing further context. In addition, 21 comments were further analyzed. Of these, 8 included praises for the project team, and 3 came from a spammer who promoted their services maliciously and spread misinformation. Furthermore, 5 asked for advice and further elaboration on conditions discussed in the videos. In total, 3 viewers shared their journey, specifically fears linked to the diagnosis and dissatisfaction with the diagnosis process. One of the viewers mentioned it provided helpful academic context as they had an assignment on a related topic. Another user shared that they had no idea about these conditions and thanked the team for publishing this information.

Discussion

Principal Findings

This report describes our experience generating, creating, disseminating, and evaluating evidence-based, peer-reviewed short social media videos. In this experience, we focused on PCOS and thyroid diseases based on our collective expertise. To our knowledge, this is the first study that describes the development and evaluation of videos for multiple social media platforms and discusses the abovementioned topics. Previous research has focused on other conditions (eg, low back pain) or specific surgical procedures (eg, laparoscopic hysterectomy) without guiding medical content creation [13,14].

Several studies have evaluated the content of short videos published on social media networks, highlighting unsatisfactory quality videos and warning about blind reliance on online content [24-26]. These studies acknowledged misinformation and invalidity as major factors compromising video content and negatively influencing the audience. This, paired with the acknowledgment of the increase in reliance on social media for medical information, flags the importance of starting evidence-based awareness campaigns and medical education initiatives for the masses. While previous researchers assessed the quality of online short videos across multiple social media platforms, none directly tackled the issues or attempted to generate evidence-based short videos.

This study builds on the KTA framework and focuses on knowledge dissemination [20]. After noting the rise in popularity and the increase in public reliance on short-video platforms as a source of medical information and acknowledging the issues linked to it, our study decided to use the short-video social media platforms to translate accurate knowledge and evidence-based information to the public. We adapted the information to be translated in lay terms and duration suitable for such platforms. Awareness of the social media algorithms, such as hashtags in promoting videos and attracting the right audience, allowed us to minimize barriers and facilitate knowledge transmission. Assessment of post video release of engagement and content creator feedback helped select and highlight obstacles and facilitators to this process of video dissemination. This methodology can be used to tailor further videos in other specialties. The dissemination of correct information could positively impact health behaviors, encouraging prompt diagnosis, preventing disease prognosis, and allowing an early management plan to be followed. Using the framework for content creation in the project is a way of standardizing and easing this process [27].

Involving medical professionals as content creators on social media brings a unique blend of credibility and educational value to combat misinformation and build trust in public health care. Their expertise allows for accurate and impactful messaging, addressing complex health topics, and debunking myths effectively. Time constraints, oversimplification, and regulatory considerations are some of the challenges that need to be addressed. Alongside creating misinformation-debunking videos, leveraging medical professionals' knowledge can significantly contribute to raising awareness, promoting healthy behaviors, and fostering a culture of informed decision-making among the public.

Process Evaluation

Responses from participants suggest that this project benefits health care professionals, patients, and the public. The experience gained from working on this video series has been noted to help medical students and health care practitioners at different levels of training. It also harnesses their skills and knowledge to ensure that the patient and public get the most up-to-date and accurate information while allowing them to ask questions and suggest further video topics. While some participants emphasized the potential for reaching wider audiences and fostering connections, others expressed concerns about social media engagement's competitive nature and potential negative impacts. These varying viewpoints indicate the complex interplay between the benefits and drawbacks of using social media platforms for content dissemination. Our thematic analysis underscores the multifaceted nature of content creators' motivations, which extend beyond monetary incentives and highlight the intrinsic rewards associated with engaging in creative endeavors. Participants' positive feedback and constructive suggestions indicate ownership and investment in the project's success. These insights contribute to the academic discourse on digital content creation and offer practical implications for educators, marketers, and content creators aiming to navigate the dynamic landscape of social media platforms. This study's findings underscore the need for a holistic approach to understanding content creators' motivations and behaviors within the ever-evolving realm of digital media.

Impact Evaluation

We noticed an incremental engagement with our videos on all social media platforms over time. The videos were shared and saved across all platforms, implying their circulation as the audience uses multiple social media platforms. All social media platforms also witnessed an increase in followers, indicating that the public found our video series helpful. The varied number of views received underscores the diverse audiences on different platforms and highlights the potential for content to resonate strongly within specific communities. TikTok's exceptional performance in likes, shares, and saves further highlights its potential for viral content dissemination and user interaction

XSL•FO
RenderX

while aligning the content per platform-specific trends, formats, and audience preferences.

Our qualitative analysis highlights the dual nature of online engagement, from spammers spreading misinformation to genuine users sharing personal experiences and expressing gratitude. The presence of spammers underscores the need for robust moderation mechanisms to ensure a safe and accurate information-sharing environment with proactive community management. The inquiries and comments seeking further information about specific medical conditions suggest an avenue for generating additional content that addresses viewers' questions and unmet needs. We highlight the diverse nature of engagement on different platforms and reveal the potential for meaningful engagement, education, and community-building through project-specific content. These findings offer insights for future content creation, platform selection, and audience engagement strategies, emphasizing the importance of tailoring content to different platform dynamics and user expectations.

Recommendations

In the absence of evidence-based guidelines to generate medical social media content, this study allowed us to formulate a set of recommendations, which we summarized in a checklist that can also be used in further studies (Table 3).

Table 3. Recommendations for creating videos on medical topics for dissemination in the public domain.

Recommendation	Rationale
Recommendation 1: involving all stakeholders, including health care professionals at all levels, students, patients, and the public.	All stakeholders mutually benefited from the diverse perspectives to deliver easily understandable content for a large audience.
Recommendation 2: using a variety of video formats (ie, interview-style videos, role play, or Q&A[a] format).	Content creators explained that they believe this will prove more engaging and, hence, be more appealing to participants.
Recommendation 3: linking the videos to information pages.	This will join between both the recent reliance on social media videos and the use of accurate research-based findings.
Recommendation 4: signposting the audience to relevant peer-reviewed studies or pages or organizations.	Misinformation spreads from the inability to distinguish reliable and nonreliable information. By signposting credible organizations, you point the audience in the right direction.
Recommendation 5: allowing the content creators to meet beforehand and familiarize themselves with one another and the project leads.	Volunteers suggested that familiarizing themselves with their colleagues would have helped build context and reach out for feedback on their respective videos to optimize video quality.

[a]Q&A: question and answer.

Strengths and Limitations

The strengths of our initiative include using a collaborative approach of involving multiple stakeholders to generate credible peer-reviewed videos with ease of understanding. Feedback from viewers and volunteers who made the video provides essential insights into the public's unmet needs and the importance of disseminating reliable information to debunk misinformation. Furthermore, according to Bloom's Taxonomy [28], allowing students to engage with patients and health care professionals, analyze and evaluate video content, and generate these videos constitutes the highest form of learning. On the other hand, our study was limited by the small sample size of volunteers, videos, and total viewers. The sample size of content creators, consisting primarily of medical students and junior doctors, may not represent a diverse range of perspectives or experiences in content creation. This could limit the generalizability of findings regarding motivations, challenges faced, and overall experiences of content creation on social media platforms. In addition, as the results of online social media platforms are dynamic over time, similar studies conducted at different times or regions may yield different results. While this study focused on creating and disseminating peer-reviewed videos on PCOS and thyroid conditions, it may not encompass the full spectrum of health-related topics relevant to broader public health concerns. Furthermore, the evaluation period of 2 months for video engagement and audience outreach may provide only a snapshot of the long-term impact and sustainability of such initiatives. Future research could address these limitations by including a more diverse range of content creators, expanding the scope of health topics covered, and conducting longer-term evaluations to assess maintained audience engagement and behavior change. Obtaining detailed demographic and socioeconomic profiles of both volunteer and nonvolunteer groups will offer invaluable insights into the barriers hindering participation. These data will not only aid researchers in understanding the dynamics influencing volunteerism but also furnish crucial information for the design and implementation of future projects. Furthermore, uncovering these barriers could shed light on broader societal issues, potentially informing policies and interventions aimed at fostering greater community engagement and participation.

Conclusion

Our study demonstrates how to codesign, disseminate, and evaluate evidence-based, peer-reviewed medical information for short videos to be distributed across social media platforms. This experience focused on PCOS and thyroid diseases and showed how social media could be used to increase awareness and tackle misinformation about these issues. In particular, social media videos can be used to engage the public and stimulate patients who might ask questions. This benefits both the viewers and the video creators, especially if they are medical students or junior doctors.

XSL•FO
RenderX

Acknowledgments

We thank all the health care professionals who participated in this study. Volunteers in this project made valuable contributions to the project delivery and constituted responses for thematic analysis. We thank Ms Anna Woollven from the British Thyroid Foundation (BTF) and Ms Maureen Ann Busby from the PCOSvitality for their continuous support throughout the project. This project received a Public Engagement Grant from the Society for Endocrinology in November 2022.

Data Availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files). Further inquiries can be directed to the corresponding author.

Authors' Contributions

ME and KM are the joint first authors, having made all-round contributions to the study. ME contributed to the study conception, supervised executive aspects of the project, interviewed participants, wrote the first draft, and conducted the thematic analysis. KM contributed to interview coding, qualitative result analysis, fine-tuning the research methods, and writing the first draft. KG contributed to fine-tuning the research methods, writing the first draft, and obtaining ethics approval. PK supervised executive aspects of the project, finalized the research methods, and critically reviewed and revised the manuscript. MB critically reviewed and contributed to the revised manuscript's introduction, methodology, and results. All authors contributed substantially to drafting and approving the final draft of the manuscript. The SIMBA (Simulation via Instant Messaging for Bedside Application) and CoMICs (Combined Medical Information Cines) team included Ms Alexander Browne, Ms Hannah Khan, Ms Anum Chaudry, Ms Shubhi Ratra, Ms Pavithra Sakthivel, Ms Damilola Akande, Mr Matthew Smith, Ms Shruti Attarde, and Ms Kayleigh Harrylal; created the content for all videos. The final version has been reviewed and approved by all the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Public Engagement interview questions asked to volunteers who filmed the videos.
[DOCX File , 15 KB - mededu_v10i1e52924_app1.docx]

References

1. Global social media statistics? Global digital insights.: DataReportal URL: https://datareportal.com/social-media-users [accessed 2023-08-31]
2. Lima DL, Viscarret V, Velasco J, Lima RNCL, Malcher F. Social media as a tool for surgical education: a qualitative systematic review. Surg Endosc 2022;36(7):4674-4684 [FREE Full text] [doi: 10.1007/s00464-022-09150-9] [Medline: 35230534]
3. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
4. Patrick M, Venkatesh RD, Stukus DR. Social media and its impact on health care. Ann Allergy Asthma Immunol 2022;128(2):139-145. [doi: 10.1016/j.anai.2021.09.014] [Medline: 34555532]
5. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. Pharmacy and Therapeutics 2014 Jul;39(7):491-520 [FREE Full text] [Medline: 25083128]
6. Yeung AWK, Tosevska A, Klager E, Eibensteiner F, Tsagkaris C, Parvanov ED, et al. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. J Med Internet Res 2022;24(1):e28152 [FREE Full text] [doi: 10.2196/28152] [Medline: 34951864]
7. Gao Y, Liu F, Gao L. Echo chamber effects on short video platforms. Sci Rep 2023;13(1):6282 [FREE Full text] [doi: 10.1038/s41598-023-33370-1] [Medline: 37072484]
8. Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. The echo chamber effect on social media. Proc Natl Acad Sci USA 2021;118(9):e2023301118 [FREE Full text] [doi: 10.1073/pnas.2023301118] [Medline: 33622786]
9. Google. World Health Organization and Google's collaboration to provide health information. URL: https://blog.google/technology/health/world-health-organization-google-collaboration-health-information/ [accessed 2023-09-16]
10. Meta. Facebook and leading health organizations form alliance for advancing health online. URL: https://about.fb.com/news/2021/06/facebook-leading-health-organizations-form-alliance-for-advancing-health-online/ [accessed 2023-09-16]
11. YouTube. Misinformation Policies - YouTube Help. URL: https://support.google.com/youtube/answer/10834785?hl=en-GB [accessed 2023-09-16]
12. Mashable. Google and YouTube are investing to fight misinformation. URL: https://mashable.com/article/google-youtube-fact-checking-misinformation [accessed 2023-09-16]

XSL•FO
RenderX

13. Maia LB, Silva JP, Souza MB, Henschke N, Oliveira VC. Popular videos related to low back pain on YouTube™ do not reflect current clinical guidelines: a cross-sectional study. Braz J Phys Ther 2021;25(6):803-810 [FREE Full text] [doi: 10.1016/j.bjpt.2021.06.009] [Medline: 34332887]
14. Unal F, Atakul N, Turan H, Yaman Ruhi I. Evaluation of YouTube laparoscopic hysterectomy videos as educational materials during the COVID-19 era using the LAParoscopic surgery video educational guidelines (LAP-VEGaS) and LAP-VEGaS video assessment tool. J Obstet Gynaecol 2022;42(5):1325-1330. [doi: 10.1080/01443615.2021.1962823] [Medline: 34704513]
15. Drozd B, Couvillon E, Suarez A. Medical YouTube videos and methods of evaluation: literature review. JMIR Med Educ 2018;4(1):e3 [FREE Full text] [doi: 10.2196/mededu.8527] [Medline: 29434018]
16. The Thyroid Trust. CoMICs: a combination of medicine and graphic storytelling. URL: https://www.thyroidtrust.org/blog/comics-a-combination-of-medicine-and-graphic-storytelling [accessed 2023-09-16]
17. Davitadze M, Ooi E, Ng CY, Zhou D, Thomas L, Hanania T, et al. SIMBA: using Kolb's learning theory in simulation-based learning to improve participants' confidence. BMC Med Educ 2022;22(1):116 [FREE Full text] [doi: 10.1186/s12909-022-03176-2] [Medline: 35193557]
18. Simba Simulation. URL: https://sites.google.com/view/simbasimulation/home [accessed 2023-09-16]
19. Society for Endocrinology. A place for CoMICs in medical education. URL: https://www.endocrinology.org/endocrinologist/139-spring-2021/features/a-place-for-comics-in-medical-education/ [accessed 2023-09-16]
20. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, et al. Lost in knowledge translation: time for a map? J Contin Educ Health Prof 2006;26(1):13-24. [doi: 10.1002/chp.47] [Medline: 16557505]
21. Terrasse M, Gorin M, Sisti D. Social media, e-Health, and medical ethics. Hastings Cent Rep 2019;49(1):24-33 [FREE Full text] [doi: 10.1002/hast.975] [Medline: 30790306]
22. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115 [FREE Full text] [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
23. Novak M, Drummond K, Kumar A. Healthcare professionals' experiences with education in short term medical missions: an inductive thematic analysis. BMC Public Health 2022;22(1):997 [FREE Full text] [doi: 10.1186/s12889-022-13349-9] [Medline: 35581562]
24. He Z, Wang Z, Song Y, Liu Y, Kang L, Fang X, et al. The reliability and quality of short videos as a source of dietary guidance for inflammatory bowel disease: cross-sectional study. J Med Internet Res 2023;25:e41518 [FREE Full text] [doi: 10.2196/41518] [Medline: 36757757]
25. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-Video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021;23(12):e28318 [FREE Full text] [doi: 10.2196/28318] [Medline: 34931996]
26. Yao L, Li Y, Lian Q, Sun J, Zhao S, Wang P. Health information sharing on social media: quality assessment of short videos about chronic kidney disease. BMC Nephrol 2022;23(1):378 [FREE Full text] [doi: 10.1186/s12882-022-03013-0] [Medline: 36443741]
27. Kong W, Song S, Zhao YC, Zhu Q, Sha L. TikTok as a health information source: assessment of the quality of information in diabetes-related videos. J Med Internet Res 2021;23(9):e30409 [FREE Full text] [doi: 10.2196/30409] [Medline: 34468327]
28. Vanderbilt University Center for Teaching. Bloom's taxonomy. URL: https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/ [accessed 2023-09-16]

Abbreviations

KTA: Knowledge-to-action
MeSH: Medical Subject Headings
PCOS: polycystic ovary syndrome
SIMBA CoMICs: Simulation via Instant Messaging for Bedside Application–Combined Medical Information Cines

Edited by B Lesselroth; submitted 20.09.23; peer-reviewed by C Tong, R Weeks; comments to author 08.04.24; revised version received 12.05.24; accepted 15.08.24; published 30.10.24.

Please cite as:
Elhariry M, Malhotra K, Goyal K, Bardus M, Team SIMBAAC, Kempegowda P
A SIMBA CoMICs Initiative to Cocreating and Disseminating Evidence-Based, Peer-Reviewed Short Videos on Social Media: Mixed Methods Prospective Study
JMIR Med Educ 2024;10:e52924
URL: https://mededu.jmir.org/2024/1/e52924
doi:10.2196/52924
PMID:

XSL•FO
RenderX

©Maiar Elhariry, Kashish Malhotra, Kashish Goyal, Marco Bardus, SIMBA and CoMICs Team, Punith Kempegowda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.10.2024. 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.

XSL•FO
RenderX

Using the Kirkpatrick Model to Evaluate the Effect of a Primary Trauma Care Course on Health Care Workers' Knowledge, Attitude, and Practice in Two Vietnamese Local Hospitals: Prospective Intervention Study

Ba Tuan Nguyen[1], MD; Van Anh Nguyen[2], MD, PhD; Christopher Leigh Blizzard[1], PhD; Andrew Palmer[1], BmedSci, MBBS; Huu Tu Nguyen[3], MD, PhD; Thang Cong Quyet[3], MD, PhD; Viet Tran[1,4,5], MBBS; Marcus Skinner[6], MSc, MBBS, AM; Haydn Perndt[4], MBBS; Mark R Nelson[1], MBBS, MFM, PhD

1

2

3

4

5

6

Corresponding Author:
Ba Tuan Nguyen, MD

Abstract

Background: The Primary Trauma Care (PTC) course was originally developed to instruct health care workers in the management of patients with severe injuries in low- and middle-income countries (LMICs) with limited medical resources. PTC has now been taught for more than 25 years. Many studies have demonstrated that the 2-day PTC workshop is useful and informative to frontline health staff and has helped improve knowledge and confidence in trauma management; however, there is little evidence of the effect of the course on changes in clinical practice. The Kirkpatrick model (KM) and the knowledge, attitude, and practice (KAP) model are effective methods to evaluate this question.

Objective: The aim of this study was to investigate how the 2-day PTC course impacts the satisfaction, knowledge, and skills of health care workers in 2 Vietnamese hospitals using a conceptual framework incorporating the KAP model and the 4-level KM as evaluation tools.

Methods: The PTC course was delivered over 2 days in the emergency departments (EDs) of Thanh Hoa and Ninh Binh hospitals in February and March 2022, respectively. This study followed a prospective pre- and postintervention design. We used validated instruments to assess the participants' satisfaction, knowledge, and skills before, immediately after, and 6 months after course delivery. The Fisher exact test and the Wilcoxon matched-pairs signed rank test were used to compare the percentages and mean scores at the pretest, posttest, and 6-month postcourse follow-up time points among course participants.

Results: A total of 80 health care staff members attended the 2-day PTC course and nearly 100% of the participants were satisfied with the course. At level 2 of the KM (knowledge), the scores on multiple-choice questions and the confidence matrix improved significantly from 60% to 77% and from 59% to 71%, respectively (P<.001), and these improvements were seen in both subgroups (nurses and doctors). The focus of level 3 was on practice, demonstrating a significant incremental change, with scenarios checklist points increasing from a mean of 5.9 (SD 1.9) to 9.0 (SD 0.9) and bedside clinical checklist points increasing from a mean of 5 (SD 1.5) to 8.3 (SD 0.8) (both P<.001). At the 6-month follow-up, the scores for multiple-choice questions, the confidence matrix, and scenarios checklist all remained unchanged, except for the multiple-choice question score in the nurse subgroup (P=.005).

Conclusions: The PTC course undertaken in 2 local hospitals in Vietnam was successful in demonstrating improvements at 3 levels of the KM for ED health care staff. The improvements in the confidence matrix and scenarios checklist were maintained for at least 6 months after the course. PTC courses should be effective in providing and sustaining improvement in knowledge and trauma care practice in other LMICs such as Vietnam.

Trial Registration: Australian New Zealand Clinical Trial Registry (ANZCTR) ACTRN 12621000371897; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380970

(JMIR Med Educ 2024;10:e47127) doi:10.2196/47127

XSL•FO
RenderX

KEYWORDS

trauma care; emergency medicine; primary trauma care course; short course; medical education; trauma; emergency; urgent; professional development; workshop; injury; injured; injuries; primary care

Introduction

Health Care Burden of Road Trauma

Road traffic trauma is a leading cause of morbidity and mortality globally [1]. Road trauma is responsible for 1.3 million deaths and 20-50 million injuries annually with 90% of these occurring in low- and middle-income countries (LMICs) [2,3]. It is predicted that the prevalence of road trauma will increase to become the third leading cause of death by 2030 in LMICs [4]. While there are many contributing factors to the higher impact of road trauma in LMICs than in high-income countries, including infrastructure, vehicle design, underdevelopment of health care systems, and lack of trauma care education [5], the latter factor was highlighted among the 5 key World Health Organization targets for the first decade of action on road safety for LMICs from 2011 to 2020 [6,7]. If this target is achieved, it is estimated that one-third of annual global trauma deaths could be prevented [8,9].

As in other LMICs, road traffic trauma is a major public health problem in Vietnam [10]. In the last 15 years, approximately 10 people per 100,000 population have been killed in road accidents in Vietnam each year, with an equal number hospitalized [11,12]. Road traffic injury is the second most common cause of death for people in Vietnam in the age group of 5 - 14 years, representing the most vulnerable and dependent population, and is the most common cause of death and disability for those in the age group of 15 - 49 years, representing the most productive population [13]. The Vietnamese health care system is built on a "pine tree" model in which a trauma center has responsibility for various "satellite" hospitals and receives patients experiencing severe trauma. In a satellite or frontline hospital, the staff first encountering a trauma patient may be a surgeon, nurse, anesthetist, or general practitioner who may not be trained in a trauma subspecialty. Indeed, the trauma training system in Vietnam is not adapted to this circumstance [14]. There is therefore a need for a training system to solve this education gap.

There are several trauma training courses that have been delivered around the world, such as the Advanced Trauma Life Support (ATLS), Trauma Team Training, and Primary Trauma Care (PTC) courses. Several trauma courses are currently being used in LMICs effectively, resulting in increases in knowledge and at a lower reported cost than the gold-standard ATLS course [15]. Among these, the PTC course is designed and structured to fit within health care systems in LMICs such as Vietnam. In particular, the PTC course requires minimal resources and is therefore sustainable for these countries [16]. For this reason, since 2007, the PTC course has been run in many regional and provincial areas of Vietnam, including Binh Dinh (2007), Hanoi (2008), Ninh Binh (2008), and Ho Chi Minh City (2018) [17-19]. However, none of these Vietnamese courses has been evaluated with respect to the effect on the clinical practice of health care staff.

Education Frameworks

One of the methods commonly used to assess educational programs is the 4-level Kirkpatrick model (KM). This model was developed by Kirkpatrick in 1959 and has since been widely used to evaluate the effectiveness of continuing education in many fields, including medicine [15]. The KM evaluates the training outcomes of a course at 4 levels depending on the amount of time the evaluation is undertaken after the course. Level 1 evaluates trainees' satisfaction toward the instructors and the training program, level 2 assesses trainees' learning of professional knowledge or skills, level 3 measures the changes in trainees' behavior or performance, and level 4 quantifies the improvement of the outcomes closely linked to the training program that will work effectively in the long term. This model has been considered a suitable assessment tool for educational programs as it has a simple process, measures a limited number of variables, and does not depend on individual variables [20].

Moreover, previous studies in education have identified that trainees' knowledge of issues and possession of skills are required for this knowledge to transfer into behavioral change [21], and that positive attitudes and behaviors could lead an individual to be better motivated toward an issue [22]. Since trauma care is a common and vital practice of many health care staff working at the frontline of the Vietnamese health care system, the effectiveness of PTC training courses should be improved and consolidated through knowledge, attitude, and practice (KAP)–based education. As such, the KAP model can be an appropriate approach to help identify knowledge gaps, attitude barriers, and practice patterns that may facilitate understanding the knowledge and practice of health care professionals after attending a PTC training course [23].

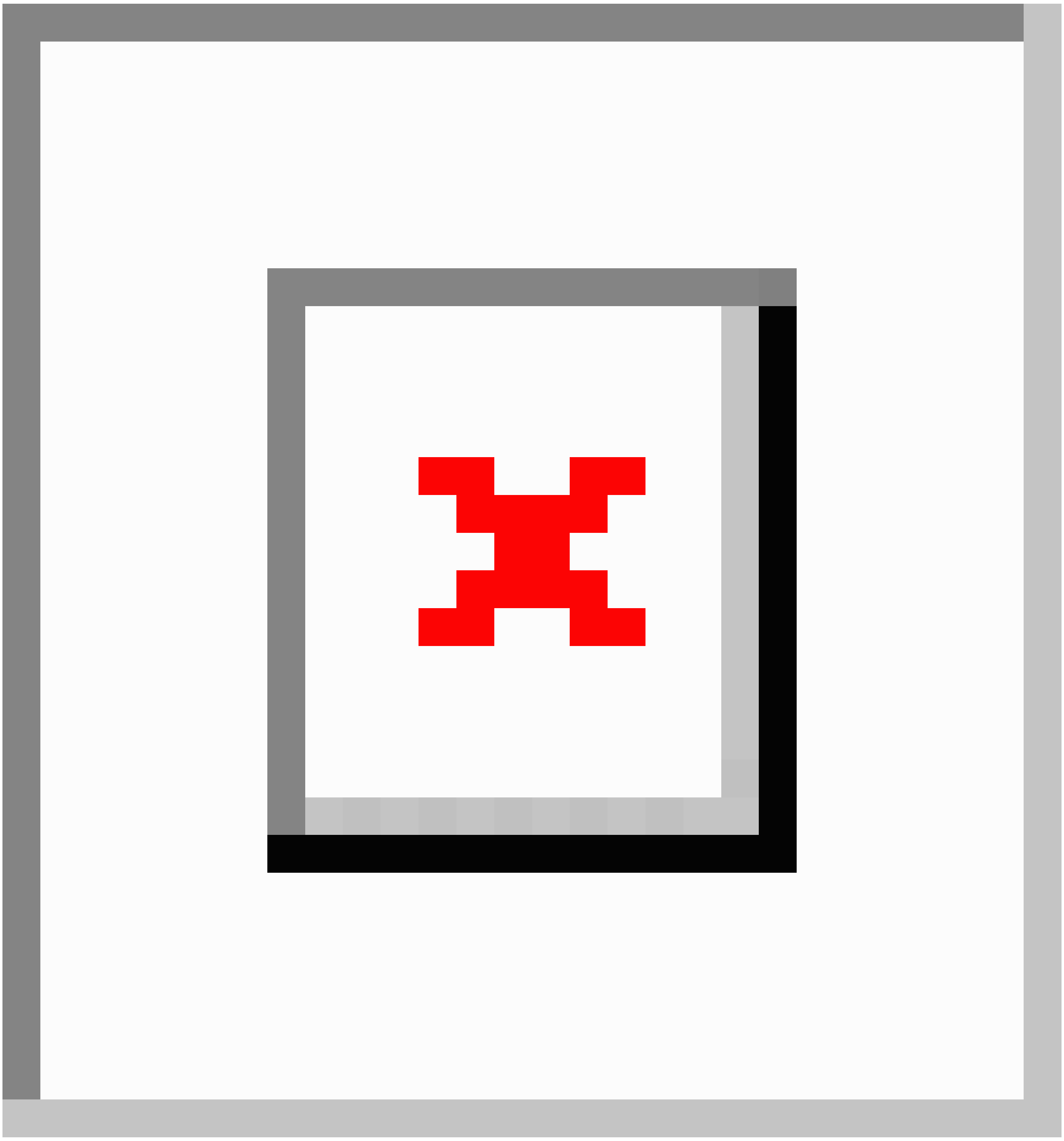
Previous studies have used the KM and KAP model as theoretical frameworks to evaluate the effectiveness of training courses in the health care field [20,24]. In applying these models for this study, we aimed to: (1) measure the outcomes of the PTC course by asking the participants to clarify their reaction to the relevance and usefulness of the course (level 1 of the KM); (2) test the knowledge and level of confidence at the precourse, immediately postcourse, and 6-month postcourse time points (level 2 of the KM); and (3) observe changes in the practice of participants in scenarios and handling trauma patients in their daily work (level 3 of the KM) (Figure 1). Level-4 changes have not yet been assessed but are the subject of an ongoing investigation.

To gain a more comprehensive understanding on the effectiveness of the education of PTC training courses, this study was performed to assess the impact of the PTC course on health care staff by using a conceptual framework involving two theories of evaluation: the KAP model and the 4-level KM . The specific study objectives were to (1) investigate the impact of the PTC training course on the satisfaction, confidence, and change of knowledge and skills of participants in two hospitals in Vietnam; and (2) evaluate the retention of the participants'

XSL•FO
RenderX

knowledge and changes in skills 6 months after the PTC training course.

Figure 1. Theoretical framework of the study based on the KAP and Kirkpatrick models of evaluation. KAP: knowledge, attitude, and practice; PTC: Primary Trauma Care; L1: level 1; L2: level 2; L3: level 3; L4: level 4.



Methods

Setting

This study was carried out in the emergency departments (EDs) of 2 provincial hospitals in Vietnam, the Thanh Hoa and Ninh Binh hospitals. There are 78 beds in both EDs with 86 staff comprising 35 doctors and 51 nurses.

PTC Course

The 2-day PTC courses were delivered by local PTC instructors from Hanoi Medical University following the standard format of the PTC Foundation [25]. The PTC course included lectures, skills workshops, and case scenarios to cover a range of topics and practical skills. The objective of the course was to train participants to approach trauma patients in a sequential manner without missing life-threating symptoms and processes.

All health care staff at the 2 EDs were invited to participate in the course. To maintain an effective clinical workforce in the respective EDs, each site was divided into 2 classes that rotated and ran over 3 days. In Thanh Hoa Hospital, the courses ran from February 17 to February 20, 2021, whereas in Ninh Binh Hospital, the courses ran from March 3 to March 6, 2021. Each

XSL•FO
RenderX

class received the same training and assessment. All material is available on the PTC website [25].

Study Design

This study followed a pre- and postinterventional design. We used validated instruments to assess the participants' satisfaction and confidence prior to the course and immediately after training, as well as to compare their level of trauma knowledge and skills at the precourse, postcourse, and 6-month follow-up time points. We also stratified participants into doctor and nurse groups for further investigation. The trial has been registered in the Australian New Zealand Clinical Trial Registry (ACTRN 12621000371897) [26].

Outcome Measures

KM Level 1: Satisfaction

A self-completed questionnaire was developed to explore participants' reactions to the PTC course. The questionnaire contained 5 items regarding the relevance and usefulness of the course. Trainees indicated their agreement with the corresponding statements using a 5-point Likert scale, with 1 indicating "strongly disagree" and 5 indicating"strongly agree" (Multimedia Appendix 1).

KM Level 2: Knowledge

Twenty multiple-choice questions were used to evaluate the complete teaching content in the course, which have also been used in previous evaluations [16]. Most of the items assessed the trainees' knowledge domain in the knowledge and comprehension categories according to the Bloom taxonomy [27]. The multiple-choice questions focused on knowledge in trauma management, including the areas of head, thoracic, and abdominal trauma, requiring the trainees to remember key fundamental points from serial lectures in the course (Multimedia Appendix 2). The results were calculated as the percentage of correct answers. This multiple-choice question set was developed for the PTC program, translated into Vietnamese by local PTC instructors, and edited by experts and educators from the PTC Foundation.

A confidence matrix was also included with 8 questions assessing the level of confidence while dealing with various circumstances related to patients experiencing trauma. Each question was rated according to 5 levels of confidence, ranging from 1 (the lowest level of confidence) to 5 (the highest level of confidence). This result was also calculated as a percentage for analysis (Multimedia Appendix 3).

KM Level 3: Practice Skills

Scenario checklists, which included various clinical scenarios, were used to evaluate participants' practice skills in a simulation. The participants at the 2 hospitals were divided into small groups of 6 - 7 people. Each group included both doctors and nurses to replicate an emergency team on duty in the ED. The assessments were conducted using an Objective Structured Clinical Examination format [20,24] with 4 stations, each station lasting up to 10 minutes, with one observer who used a standardized checklist to rate the performance of the team. Each scenario checklist comprised 10 key actions. If the examined group achieved this action, they were given 1 point; otherwise, they received 0 points. We chose 4 scenarios for the evaluation and rotated all groups within these 4 scenarios to ensure that the maximum amount of skills were evaluated (Multimedia Appendix 4).

For the bedside clinical checklist, we used an observed checklist with 10 vital points that had been stressed in the course. This checklist was used by experienced clinicians who were local PTC instructors. To minimize observation stress, all participants were informed about the presence of the examiners before and after the course. The examiners were allocated cases randomly in both the pre- and postcourse phases (Multimedia Appendix 5).

Statistical Analysis

We used Stata version 15.1 software for statistical analyses. The Fisher exact and Wilcoxon matched-pairs signed rank tests were used to compare percentages (multiple-choice questions and confidence matrix) and the mean scores (scenario and bedside clinical checklists) among the precourse, postcourse, and 6-month follow-up time points for the participants. A P value <.05 was considered statistically significant.

Ethical Considerations

The University of Tasmania Human Research Ethics Committee approved this study (reference number H0023982) [28]. All participants who volunteered to take part in the study (without compensation) were required to sign the consent form (Multimedia Appendix 6). All data were deidentified and stored online in a password-protected Google drive of the research group's account to ensure privacy and confidentiality.

Results

Participant Characteristics

Among the 86 health care staff in the EDs, 80 participated in the course; the 6 individuals who could not participate in the course were excluded owing to testing positive for COVID-19 at the time of course delivery. All participants completed the pre-and posttest assessments and the scenarios checklist. Only 57 (71%) of the 80 participants completed the joint evaluation of knowledge and scenarios checklist at the 6-month follow-up. The cohort consisted of 34 doctors (mean age 28.0, SD 2.5 years) and 46 nurses (mean age 32.0, SD 6.4 years). Nurses had more general medical work experience than doctors (mean 6.3, SD 5.6 years vs 2.4, SD 1.7 years). Male staff accounted for 68% of the doctors and 65% of the nurses.

KM Level 1: Survey Responses

All participants responded to the survey, with 78 of the 80 participants (98%) indicating satisfaction with the course. Likewise, an equal number of respondents stated that "the course enhanced their knowledge" and 79 of the 80 participants (99%) stated that they would "suggest the course to others." Furthermore, 77 of the 80 participants (96%) agreed that "the course was relevant to ED staff" (Table 1).

XSL•FO
RenderX

Table . Level 1 of the Kirkpatrick model: participant reactions to the Primary Trauma Care (PTC) course (N=80).

Question	Strongly disagree (1), n (%)	Somewhat disagree (2), n (%)	Neither agree or disagree (3), n (%)	Somewhat agree (4), n (%)	Strongly agree (5), n (%)
I was satisfied with the PTC course overall	1 (1)	1 (1)	0 (0)	8 (10)	70 (88)
This course enhanced my knowledge of the subject matter	1 (1)	0 (0)	1 (1)	6 (8)	72 (90)
The course was relevant to what I might be expected to do (to prevent, prepare for/respond to a trauma) in an emergency department[a]	1 (1)	0(0)	1 (1)	9 (11)	68 (86)
This course provided content that is relevant to my daily job	1 (1)	0 (0)	2 (3)	16 (20)	61 (76)
I would recommend this course to others	1 (1)	0 (0)	0 (0)	4 (5)	75 (94)

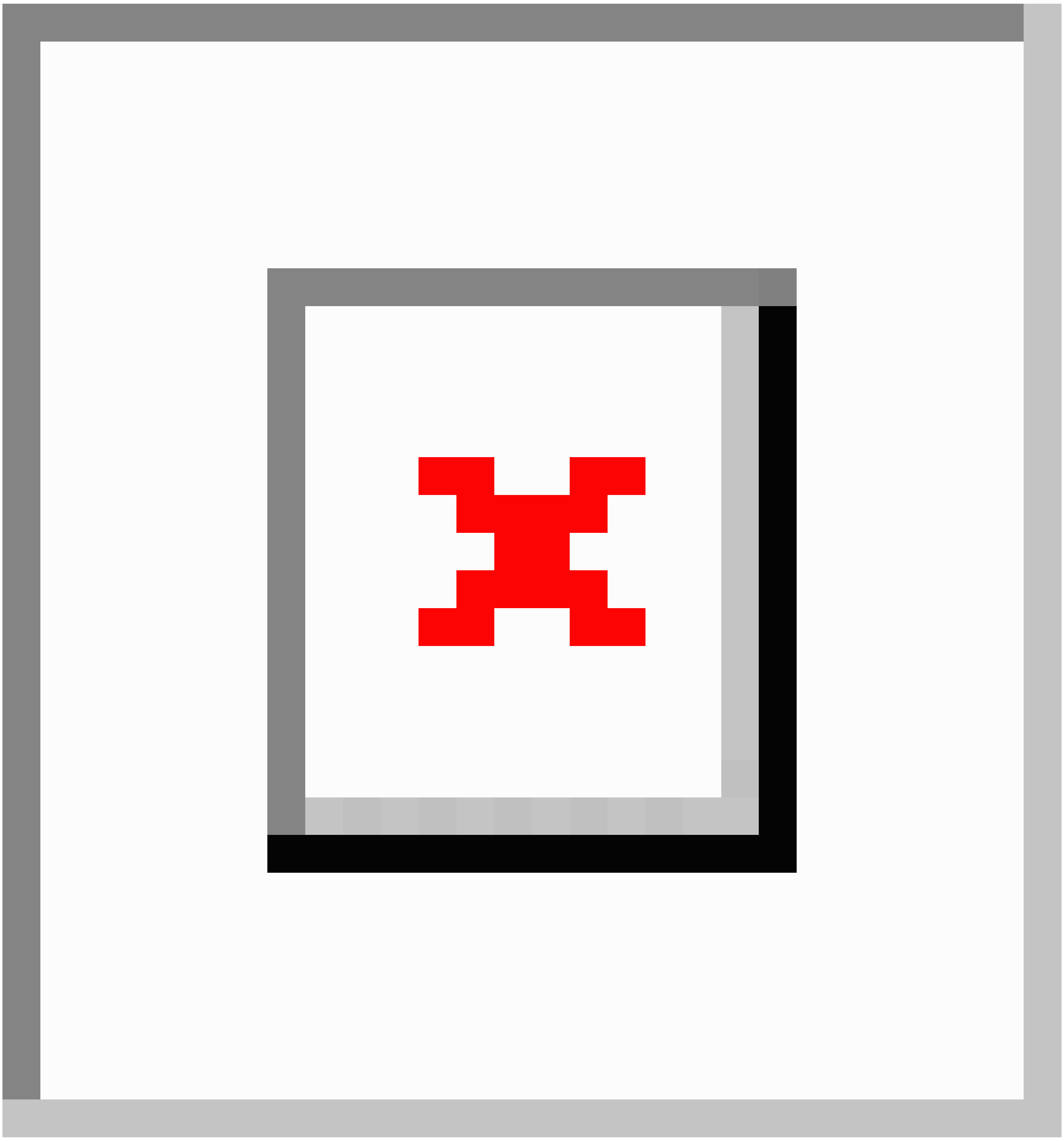
[a]One respondent is missing for this question (N=79).

KM Level 2: Knowledge and Confidence Matrix Assessments

There was a significant improvement in correct multiple-choice question responses between the pre- and postcourse assessments in both the doctor and nurse groups, as these scores increased from 67% and 59% to 82% and 74%, respectively (both *P*<.001). Comparing the immediate postcourse assessment to the 6-month follow-up, there was a significant reduction in correct multiple-choice question responses among nurses (74% vs 67%; *P*=.005) but not among doctors (82% vs 77%; *P*=.31). Compared to the precourse scores, the confidence matrix assessment improved significantly immediately following the course in both the doctors and nurses, from 60% to 76% (*P*<.001) and from 58% to 68% (*P*=.02), respectively. Both figures declined after 6 months, although these changes were not statistically significant for doctors (*P*=.07) or nurses (*P*=.51) (Figure 2).

XSL•FO
RenderX

Figure 2. Level-2 assessment of the Kirkpatrick model. The left y-axis shows the percentage of correct multiple-choice questions (MCQs) and the right y-axis shows the percentage of correct responses to the confidence matrix (CM) for the entire cohort and in the nurse and doctor subgroups over time.



KM Level 3: Practice Skill Evaluations

Scenarios Checklist

In the scenarios section, the average postcourse score in all groups significantly improved ($P<.001$). Six months later, the average scenarios checklist score did not deteriorate ($P=.99$) and this pattern was observed at both hospitals, with $P=.99$ and $P=.81$ for Thanh Hoa and Ninh Binh, respectively (Table 2).

XSL•FO
RenderX

Table . Scenarios checklist results at the 2 hospitals prior to, immediately following, and 6 months after the course.

Site and assessment		Scenario checklist score, mean (SD)
All sites		
	Precourse	5.9 (1.4)
	Post course	9.0 (0.9)
	6-month follow-up	8.5 (0.8)
Thanh Hoa Hospital		
	Precourse	6.1 (1.1)
	Post course	9.1 (0.9)
	6-month follow-up	8.4 (0.7)
Ninh Binh Hospital		
	Precourse	5.8 (1.6)
	Post course	8.9 (0.8)
	6-month follow-up	8.7 (0.8)

Bedside Clinical Checklist

There were 157 possible observed bedside clinical cases precourse and 161 such cases post course at both sites. The bedside clinical scores equally achieved a mean of 5 (SD 1.3) for both sites precourse and increased significantly to 8.4 (SD 0.8) at Ninh Binh and to 8.6 (SD 0.9) at Thanh Hoa, resulting in an overall significant increase of 8.5 (SD 0.8) (P<.001) (Table 3).

The responses to all questions showed a significant improvement, with the most significant effects seen for the questions “Is a primary survey/secondary survey undertaken?”, “Was the cervical spine stabilized (manual/collar)?”, and “Was the patient fully exposed and assessed for other injuries?”with changes in percentage post course of 99%, 92%, and 71%, respectively. By contrast, the questions “Was a log roll performed to evaluate the full length of the spine?” and “After any intervention (eg, insertion of an endotracheal tube, treatment of pneumothorax, rapid infusion of fluids) was the ABC (airway, breathing, circulation) reassessed?” showed the lowest rate of correct responses in both pre- and postcourse observed cases, with a change from 1% to 17% and from 10% to 56%, respectively (Table 4).

Table . Bedside clinical checklist results at the 2 hospitals prior to and immediately following the course.

Site and assessment		Bedside clinical checklist score, mean (SD)
All sites		
	Precourse	5 (1.3)
	Post course	8.5 (0.8)
Ninh Binh Hospital		
	Precourse	5 (1.3)
	Post course	8.4 (0.8)
Thanh Hoa Hospital		
	Precourse	5 (1.3)
	Post course	8.6 (0.9)

XSL•FO
RenderX

Table . Correct percentages of pre- versus postcourse clinical checklist scores of all observed cases.

Question	Precourse (n=157), n (%)	Post course (n=161), n (%)	Change, %	*P* value
Is a primary survey/secondary survey undertaken?	2 (1.3)	161 (100)	98.7	<.001
Was the cervical spine stabilized (manual/collar)?	11 (7)	160 (99.4)	92.4	<.001
Was oxygen administered/a pulse oximeter probe attached?	111 (70.7)	161 (100)	29.3	<.001
Was the airway assessed? (breathing or not, chest moving or not, obstructed sounds or not)?	131 (83.4)	161 (100)	16.6	<.001
Was the breathing clinically assessed by looking (breath count), feeling (palpation of trachea, percussion of chest), and listening (auscultation)?	123 (78.3)	161 (100)	21.7	<.001
Was the circulation assessed by measurement of heart rate and blood pressure? Was there an assessment of the quality of the pulse, capillary return, and temperature of the peripheries?	140 (89.2)	161 (100)	10.8	<.001
Was blood taken for cross match and hemoglobin/hematocrit analysis? Was an intravenous infusion started?	144 (91.7)	161 (100)	8.3	<.001
Was an AVPU/GCS[a] neurological assessment of disability done?	89 (56.7)	160 (99.4)	42.7	<.001
Was the patient fully exposed and assessed for other injuries?	3 (1.9)	117 (72.7)	70.8	<.001
Was a log roll performed to evaluate the full length of the spine?	2 (1.3)	30 (18.6)	17.3	<.001
After any intervention (eg, insertion of an endotracheal tube, treatment of pneumothorax, rapid infusion of fluids) was the ABC[b] reassessed?	15 (9.6)	90 (55.9)	46.3	<.001

[a]AVPU/GCS: Alert, Voice, Pain, Unresponsive/Glasgow Coma Scale.
[b]ABC: airway, breathing, circulation.

Discussion

Principal Findings

This study demonstrated that the PTC course led to improvements at all 3 levels of the KM. This improvement was maintained for at least 6 months post intervention, except for knowledge in the nurse group. These findings suggest that the knowledge and skills acquired in the PTC course are likely to be translated into clinical practice.

Most of the doctors in the ED who joined the course were junior doctors. This is because, in the Vietnamese medical system, ED work is poorly paid and nonspecialized. In addition, the majority of the nurses and doctors were male. These trends are consistently found in ED staff across the Vietnamese medical system, with most of these staff moving out of the ED into a recognized specialty within a few years [29]. Like many LMICs, ED staffing in the Vietnamese medical system is built on the Franco-German model where staff are not trained as a specialty [30]. Therefore, staff turnover requires frequent redelivery of

XSL•FO
RenderX

the PTC course. For this reason, a trauma course such as the PTC course is more suitable for this country. A male predominance among health care staff has also been reported in PTC research of Alwawi (64%), Ologunde (66%), and Nogaro (77%) [31-33]. However, this predominance was not explained in these papers.

Impact of PTC on Level 1 of the KM: Participants' Reactions to the Course

Nearly all participants were satisfied with the course. This is in line with the study of Tolppa et al [34], who found that the majority (56/59) of participants agreed or strongly agreed that trauma services are important and 57/59 of participants would recommend the PTC course to their colleagues. In addition, Jawaid et al [35] organized a PTC course with 20 participants, which received a rating of 100% satisfaction, and 100% of the participants also agreed or strongly agreed that their knowledge and skills were enhanced after the course. The authors of these studies argued that having extremely high postcourse ratings was attributed to the course being well-structured/organized as well as having a local champion, along with two other reasons. First, unlike other medical training programs, this course is free of charge. Second, the course is organized locally; therefore, participants were not required to move to other locations, which avoided travel-related logistic issues. Our results are in line with these previous findings and suggest that future courses in limited-resource settings, if organized, should be held locally.

Impact of the PTC Course on Level 2 of the KM (Knowledge and Confidence Matrix Retention): A Refresh Course on Demand

Our study showed that multiple-choice question scores improved significantly after the course (60% vs 77%; $P<.001$). This finding is similar to those of other studies assessing PTC courses. Amiri et al [36] reported that multiple-choice question scores improved from 63% to 89% for 64 participants comprising physicians and surgeons. Other studies demonstrated significant increases in scores after the course, ranging from 12% to 32% [31,32,35,37-40]. There are 2 multiple-choice questionnaires available in PTC resources (20 and 30 multiple-choice question forms) [25], with comparable results for either form. This reflects the high level of reliability of PTC multiple-choice question tools.

Furthermore, the confidence matrix scores in our study also improved significantly by 17%, which matches the degree of improvement of 20%‑23% reported in previous studies [32,37,39].

However, although multiple-choice question and confidence matrix scores of the doctors and nurses remained unchanged after 6 months, the multiple-choice question scores of nurses declined significantly (74% vs 67%, $P=.005$). This may be explained by the fact that nurses, unlike doctors, are less likely to apply the multiple-choice questions in their routine clinical activities so that there is no reinforcement of these concepts in the workplace after the PTC course. Tolppa et al [34] found that the knowledge gained from a PTC course can be maintained for up to 2 years; this difference from our findings may be related to the fact that Tolppa et al [34] did not report the findings for different subgroups.

Impact of PTC on Level 3 of the KM: Translation From Acquired Learning to Practice

In simulation situations, the simulation check scores improved significantly from 5.9/10 to 9/10 ($P<.001$). Our result matches that of Jawaid et al [35] on assessing the effectiveness of a PTC course with 20 participants, demonstrating an increase in the median simulation check score from 3.5/20 to 9.5/20 ($P<.001$).

Furthermore, improvement was also seen in clinical application after the course. The bedside clinical checklist score increased significantly from 5/10 to 8.5/10 ($P<.001$). This change demonstrates that the knowledge and skills acquired in the course are effectively converted into clinical practice. In contrast, in a study conducted in El Salvador, Cioè-Peña et al [41] found that despite a significant improvement in the median correct response rate of multiple-choice questions from 74% to 86% after the course, there was no significant change in clinical practice in 194 observed cases for both assessment periods ($P=.94$). This difference could be explained by this study's different setting along with the different assessment tools and criteria used [41]. Additionally, in the observed cases, we found a significant improvement of all clinical checklist scores ($P<.001$). In particular, the correct response rate for the question "Is a primary survey/secondary survey undertaken?" increased from 1% in the precourse assessment to 100% in the postcourse assessment. This demonstrates that the primary/secondary survey skill, which is a key point of the PTC and other trauma courses (eg, ATLS), was relatively unknown prior to the PTC course, but was greatly improved post course. By contrast, the questions "Was a log roll performed to evaluate the full length of the spine?" and "After any intervention (eg, insertion of an endotracheal tube, treatment of pneumothorax, rapid infusion of fluids) was the ABC reassessed?" demonstrated the least improvement (1% vs 17% and 10% vs 56%, respectively). In the scenarios checklist evaluation, both participant groups demonstrated an understanding of these fundamental aspects. However, in clinical practice, ED staff might still ignore these concepts due to the excess clinical workload. It is recommended that local hospitals adequately support their ED staff to ensure they can provide care to the best of their knowledge and abilities.

Unlike knowledge assessment, it is a long-standing challenge to evaluate clinical activities due to various barriers and obstacles, including time, high cost, availability of skilled clinical supervisors, and other bias/confounding variables such as evaluation and selection bias [42,43]. As we were cognizant of these difficulties, our study was designed to minimize this bias and the potential impacts of confounding factors. However, because the assessment was direct and intermittent, the presence of examiners may have led to a Hawthorne effect, which causes the alteration of behavior by the presence of examinees due to the awareness of being observed [44,45]. It is worth noting that a previous PTC course had been organized in Ninh Binh Hospital in 2008 [17]. Perhaps some residual education and clinical practice effects may have persisted at the time this study was carried out. However, none of the staff who attended the PTC course had been working at the ED prior to when the 2008

XSL•FO
RenderX

course was conducted; thus, a residual educative effect is likely to be negligible or nonexistent.

Application of the KM and KAP Model in Training

From an educational aspect, among the many factors that affect training program outcomes, the knowledge, attitude, and practice of the trainees are critical components, as they will influence the process of behavioral change, which is the most desired outcome of these courses. In this study, the knowledge, practice, and attitude of health care professionals after the PTC training course were evaluated and showed positive outcomes. When trainees have sufficient and technical knowledge of trauma care, they have positive attitudes and good clinical practice when dealing with trauma patients in the ED. This result is similar to the work of other authors who also applied the KAP approach in their course evaluation [46,47]

Using the KM to evaluate the effectiveness of education/training intervention is not a novel approach [48-50] and it is considered to be more appropriate than other models [51,52]. This study's results confirmed the effectiveness of PTC training courses at the first 3 levels of the KM, similar to the findings of previous studies [16,34,35]. However, the simple 4-level KM does not help to explain the impact of individual or contextual factors in the evaluation. In the situation of PTC training, contextual factors such as the hospital's or ED unit's goals, values, and work environment would impact the application of trained skills on the job of trainees after the courses. In our study, the overloaded situation of the provincial hospitals may be assumed to prevent health care staff from performing the learned procedure to the full capacity when handling trauma patients. In addition, the nature of the tasks of nurses in an ED team may affect their long-term knowledge retention, which could be an expression of the impact of contextual factors on the effectiveness of the training courses. These assumptions warrant deeper investigation in future study, which should consider several organizational factors such as staff turnover, relationships among professionals, and the gender distribution of ED staff. Moreover, in our study, although nearly 100% of the participants were satisfied with the course, which indicates the effectiveness of PTC training courses at the first level of the KM, we do not have evidence to link this positive reaction of the participants to their knowledge transfer and absolute positive postcourse results. A qualitative approach such as in-depth interviews with participants would be useful in detecting the hidden factors that may influence the effectiveness of both the process and outcomes of training, including the organizational aspect of the course, teaching methods, or adequacy of material resources in the courses.

Limitations

This study has some limitations. To minimize evaluation bias, we informed all examinees of the assessment process. Level 4 of KM is considered a primary endpoint of medical intervention; but, this was not assessed in our study. This will be reported in the subsequent papers by the corresponding author. The study had only a 6 months follow-up and thus lacked a longer evaluation and did not evaluate directly how the actual trauma system changed post intervention. The trauma system includes components such as leadership, professional resources, and financial budget, etc, and therefore may require multiple efforts to be improved [53,54]. Future studies which include these components are required to clarify these issues.

Conclusions

The PTC course undertaken in 2 provincial hospitals of Vietnam was successful in improving 3 levels of the KM for ED health care staff. This improvement was maintained for at least 6 months after the course. The PTC courses are effective in providing sustained improvement over 3 levels of the KM for LMICs such as Vietnam.

Acknowledgments

This work was funded by the Elphinstone Group and Broadreach Holdings Pty Ltd via an unconditional grant. BTN holds a University of Tasmania scholarship. The Primary Trauma Care course has been completed and fully evaluated with the great assistance from the lectures team of the Anaesthesia Department of Hanoi Medical University, including the Thanh Hoa Hospital leaders Dr Van Sy Le, Dr Van Cuong Le, Dr Tien Tung Lam, Thi Nga Linh Luong, and the Ninh Binh Hospital leaders Dr Chinh Chuyen Le, Dr Tu Vu Ngoc Dinh, Dr Thanh Nam Phan, Sy Thuoc Phan, and Thi Hoa Do. The courses also received support from Dr Huu Hoang Nguyen, anesthetist, Vietnamese French Hospital, Hanoi, Vietnam.

Authors' Contributions

BTN, VATN, and MN conceived the idea for the study. BTN, TCQ, and HTN collected the data. BTN, VATN, and VT drafted the manuscript. BTN and CLB were responsible for the statistical analyses. AP, CLB, TCQ, MS, HP, and HTN revised the manuscript. MN contributed to the critical revision of the manuscript for important intellectual content and approved the final version. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Level of satisfaction questionnaire.
[DOCX File, 17 KB - mededu_v10i1e47127_app1.docx]

XSL•FO
RenderX

Multimedia Appendix 2
Multiple-choice question test.
[DOCX File, 20 KB - mededu_v10i1e47127_app2.docx]

Multimedia Appendix 3
Confidence matrix.
[DOCX File, 17 KB - mededu_v10i1e47127_app3.docx]

Multimedia Appendix 4
Scenarios.
[DOCX File, 59 KB - mededu_v10i1e47127_app4.docx]

Multimedia Appendix 5
Bedside clinical checklist.
[DOCX File, 17 KB - mededu_v10i1e47127_app5.docx]

Multimedia Appendix 6
Informed consent form.
[DOCX File, 17 KB - mededu_v10i1e47127_app6.docx]

References

1. Road traffic injuries. World Health Organization. URL: https://www.who.int/health-topics/road-safety#tab=tab_1 [accessed 2023-03-20]
2. Social Determinants of Health (SDH) WHO Team. Global status report on road safety 2018. : World Health Organization; 2018 URL: https://www.who.int/publications/i/item/9789241565684 [accessed 2023-03-20]
3. Khalaf MK, Rosen HE, Mitra S, et al. Estimating the burden of disability from road traffic injuries in 5 low- and middle-income countries: protocol for a prospective observational study. JMIR Res Protoc 2023 Feb 1;12:e40985. [doi: 10.2196/40985] [Medline: 36723997]
4. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 2006 Nov;3(11):e442. [doi: 10.1371/journal.pmed.0030442] [Medline: 17132052]
5. WHO South-East Asia. Accelerating actions for implementation of decade of action for road safety. : World Health Organization; 2019 URL: https://www.who.int/publications-detail-redirect/9789290226246 [accessed 2023-03-20]
6. Social Determinants of Health (SDH) WHO Team. Global plan for the decade of action for road safety 2011–2020. : World Health Organization; 2011 URL: https://www.who.int/publications/m/item/global-plan-for-the-decade-of-action-for-road-safety-2011-2020 [accessed 2023-03-20]
7. The Global Health Observatory. Health topics: sustainable development goals (SDGs). World Health Organization. URL: https://www.who.int/sdg/targets/en/ [accessed 2022-11-21]
8. Mock C, Joshipura M, Arreola-Risa C, Quansah R. An estimate of the number of lives that could be saved through improvements in trauma care globally. World J Surg 2012 May;36(5):959-963. [doi: 10.1007/s00268-012-1459-6] [Medline: 22419411]
9. Mock CN, Jurkovich GJ, nii-Amon-Kotei D, Arreola-Risa C, Maier RV. Trauma mortality patterns in three nations at different economic levels: implications for global trauma system development. J Trauma 1998 May;44(5):804-812. [doi: 10.1097/00005373-199805000-00011] [Medline: 9603081]
10. Rossiter ND. Trauma-the forgotten pandemic? Int Orthop 2022 Jan;46(1):3-11. [doi: 10.1007/s00264-021-05213-z] [Medline: 34519840]
11. United Nations Economic and Social Commission for Asia and the Pacific, United Nations Economic Commission for Europe, United Nations Economic Commission for Latin America and the Caribbean. Road safety performance review Viet Nam. : United Nations; 2018 URL: https://unece.org/DAM/trans/roadsafe/unda/RSPR_Viet_Nam_FULL_e.pdf [accessed 2023-03-20]
12. Le Van D. Vietnam national road safety goals and action plan: opportunities and challenges. Presented at: Regional Meeting on Renewing Regional Road Safety Goals and Targets for Asia and the Pacific; Dec 5-7, 2023; Manila, Philippines URL: https://www.unescap.org/sites/default/files/3.%20%20Vietnam%20National%20Road%20Safety%20Goals%20and%20Action%20Plan%20Opportunities%20and%20Challenges.pdf [accessed 2024-07-05]
13. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. Lancet 2020 Oct 17;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]

XSL•FO
RenderX

14. Nguyen BT, Phung TL, Khuc THH, et al. Trauma care training in Vietnam: narrative scoping review. JMIR Med Educ 2022 Jan 24;8(1):e34369. [doi: 10.2196/34369] [Medline: 34967756]
15. Brown HA, Tidwell C, Prest P. Trauma training in low- and middle-income countries: a scoping review of ATLS alternatives. Afr J Emerg Med 2022 Mar;12(1):53-60. [doi: 10.1016/j.afjem.2021.11.004] [Medline: 35070655]
16. Kadhum M, Sinclair P, Lavy C. Are primary trauma care (PTC) courses beneficial in low- and middle-income countries - a systematic review. Injury 2020 Feb;51(2):136-141. [doi: 10.1016/j.injury.2019.10.084] [Medline: 31679834]
17. Primary trauma care courses in Binh Dinh 2006. Primary Trauma Care Foundation. URL: https://www.primarytraumacare.org/wp-content/uploads/2011/09/PTC-Report-Vietnam-May-2006.pdf [accessed 2022-01-17]
18. Primary trauma care courses in Binh Dinh 2007. Primary Trauma Care Foundation. URL: https://www.primarytraumacare.org/wp-content/uploads/2011/09/PTC-Summary-Report-Vietnam-June-2007.pdf [accessed 2022-01-17]
19. McDougall R, Skinner M. Report on primary trauma care program at Cho Ray Hospital, Ho Chi Minh City, April 2018. Primary Trauma Care Foundation. 2018. URL: https://www.primarytraumacare.org/wp-content/uploads/2018/06/Summary-report-for-PTCF-Cho-Ray-HCMC-April-2018Vietnam-1-1.pdf [accessed 2022-01-17]
20. Bates R. A critical analysis of evaluation practice: the Kirkpatrick model and the principle of beneficence. Eval Program Plan 2004 Aug;27(3):341-347. [doi: 10.1016/j.evalprogplan.2004.04.011]
21. Hungerford HR, Volk TL. Changing learner behavior through environmental education. J Environ Educ 1990 Mar;21(3):8-21. [doi: 10.1080/00958964.1990.10753743]
22. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
23. The World Health Report 2008 - primary health care: now more than ever. : World Health Organization; 2008 URL: https://reliefweb.int/report/world/world-health-report-2008-primary-health-care-now-more-ever [accessed 2023-03-20]
24. Wang J, Chen L, Yu M, He J. Impact of knowledge, attitude, and practice (KAP)-based rehabilitation education on the KAP of patients with intervertebral disc herniation. Ann Palliat Med 2020 Mar;9(2):388-393. [doi: 10.21037/apm.2020.03.01]
25. PTC resources. Primary Trauma Care Foundation. URL: https://www.primarytraumacare.org/get-involved/download-resources/ [accessed 2023-03-20]
26. Australian New Zealand Clinical Trials Registry (ANZCTR). URL: https://anzctr.org.au/trial/MyTrial.aspx [accessed 2024-02-28]
27. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theory Pract 2002 Nov 1;41(4):212-218. [doi: 10.1207/s15430421tip4104_2]
28. Human research ethics application - research ethics applications. University of Tasmania. URL: https://ethics.utas.edu.au/Project/Index/72931 [accessed 2024-02-28]
29. Human resources for health country profiles: Viet Nam. : World Health Organization. Regional Office for the Western Pacific; 2016 URL: https://iris.who.int/handle/10665/259990 [accessed 2024-08-07]
30. Sakr M, Wardrope J. Casualty, accident and emergency, or emergency medicine, the evolution. J Accid Emerg Med 2000 Sep;17(5):314-319. [doi: 10.1136/emj.17.5.314] [Medline: 11005398]
31. Alwawi˙ A, Amro NRN, Inkaya BV. The effectiveness of the primary trauma care courses in West Bank, Palestine: are the outcomes acceptable? J Educ Pract 2019. [doi: 10.7176/JEP/10-9-12]
32. Nogaro MC, Pandit H, Peter N, et al. How useful are primary trauma care courses in sub-Saharan Africa? Injury 2015 Jul;46(7):1293-1298. [doi: 10.1016/j.injury.2015.04.010] [Medline: 25907403]
33. Ologunde R, Le G, Turner J, et al. Do trauma courses change practice? A qualitative review of 20 courses in East, Central and Southern Africa. Injury 2017 Sep;48(9):2010-2016. [doi: 10.1016/j.injury.2017.06.007] [Medline: 28625562]
34. Tolppa T, Vangu AM, Balu HC, Matondo P, Tissingh E. Impact of the primary trauma care course in the Kongo central province of the Democratic Republic of Congo over two years. Injury 2020 Feb;51(2):235-242. [doi: 10.1016/j.injury.2019.12.013] [Medline: 31864671]
35. Jawaid M, Ahmed Memon A, Masood Z, Nadeem Alam S. Effectiveness of the primary trauma care course: is the outcome satisfactory? Pak J Med Sci 2013 Sep;29(5):1265-1268. [doi: 10.12669/pjms.295.4002] [Medline: 24353733]
36. Amiri H, Gholipour C, Mokhtarpour M, Shams Vahdati S, Hashemi Aghdam Y, Bakhshayeshi M. Two-day primary trauma care workshop: early and late evaluation of knowledge and practice. Eur J Emerg Med 2013 Apr;20(2):130-132. [doi: 10.1097/MEJ.0b013e32835608c6] [Medline: 22717774]
37. Peter NA, Pandit H, Le G, Nduhiu M, Moro E, Lavy C. Delivering a sustainable trauma management training programme tailored for low-resource settings in East, Central and Southern African countries using a cascading course model. Injury 2016 May;47(5):1128-1134. [doi: 10.1016/j.injury.2015.11.042] [Medline: 26725708]
38. Sadiq MA, Rehman KU, Tariq N, Bashir EA. Impact of primary trauma care workshop on the cognitive domain of final year medical students. J Surg Pak 2018;23(2):64-67. [doi: 10.21699/jsp.23.2.6]
39. Muzzammil M, Minhas MS, Ramzan Ali SAA, Jooma R, Minhas MO, Jabbar S. Primary trauma care course: alternative basic trauma course in developing countries. "The Need Of The Hour". Int J Clin Pract 2021 Aug;75(8):e14327. [doi: 10.1111/ijcp.14327] [Medline: 33982374]
40. Uma K, Harshad D, Dhanashree D. Impact of trauma workshop on knowledge, attitude and practice conducted on undergraduate MBBS students. Perspect Med Res 2020;8(3):81-85. [doi: 10.47799/pimr.0803.17]

XSL•FO
RenderX

41. Cioè-Peña E, Granados J, Wrightsmith L, Henriquez-Vigil A, Moresky R. Development and implementation of a hospital-based trauma response system in an urban hospital in San Salvador, El Salvador. Trauma 2017 Apr;19(2):118-126. [doi: 10.1177/1460408616672491]
42. Torabizadeh C, Ghodsbin F, Javanmardifard S, Shirazi F, Amirkhani M, Bijani M. The barriers and challenges of applying new strategies in the clinical evaluation of nursing students from the viewpoints of clinical teachers. Iran J Nurs Midwifery Res 2018;23(4):305-310. [doi: 10.4103/ijnmr.IJNMR_17_17] [Medline: 30034492]
43. Carley S, Driscoll P. Trauma education. Resuscitation 2001 Jan;48(1):47-56. [doi: 10.1016/s0300-9572(00)00317-8] [Medline: 11162882]
44. Mayo E. The Human Problems of an Industrial Civilization: Routledge; 2003. [doi: 10.4324/9780203487273]
45. Roethlisberger FJ, William JD. Management and the Worker: Routledge; 2003.
46. Alzghoul BI, Abdullah NAC. Pain management practices by nurses: an application of the knowledge, attitude and practices (KAP) model. Glob J Health Sci 2015 Oct 26;8(6):154-160. [doi: 10.5539/gjhs.v8n6p154] [Medline: 26755474]
47. Al Mansour MA, Al-Bedah AM, AlRukban MO, et al. Medical students' knowledge, attitude, and practice of complementary and alternative medicine: a pre- and post-exposure survey in Majmaah University, Saudi Arabia. Adv Med Educ Pract 2015;6:407-420. [doi: 10.2147/AMEP.S82306] [Medline: 26082671]
48. Ragsdale JW, Berry A, Gibson JW, et al. Evaluating the effectiveness of undergraduate clinical education programs. Med Educ 2020 Dec;25(1):1757883. [doi: 10.1080/10872981.2020.1757883] [Medline: 32352355]
49. Firooznia M, Hamta A, Shakerian S. The effectiveness of in-service training "pharmacopeia home health" based on Kirkpatrick's model: a quasi-experimental study. J Educ Health Promot 2020;9:218. [doi: 10.4103/jehp.jehp_170_20] [Medline: 33062751]
50. Heydari MR, Taghva F, Amini M, Delavari S. Using Kirkpatrick's model to measure the effect of a new teaching and learning methods workshop for health care staff. BMC Res Notes 2019 Jul 10;12(1):388. [doi: 10.1186/s13104-019-4421-y] [Medline: 31292006]
51. Rouse DN. Employing Kirkpatrick's evaluation framework to determine the effectiveness of health information management courses and programs. Perspect Health Inf Manag 2011 Apr 1;8(Spring):1c. [Medline: 21464860]
52. Smidt A, Balandin S, Sigafoos J, Reed VA. The Kirkpatrick model: a useful tool for evaluating training outcomes. J Intellect Dev Disabil 2009 Sep;34(3):266-274. [doi: 10.1080/13668250903093125]
53. Carter P, Blanch A. A trauma lens for systems change. Stanf Soc Innov Rev 2019;17(3):48-54. [doi: 10.48558/ESG7-3823]
54. Soto JM, Zhang Y, Huang JH, Feng DX. An overview of the American trauma system. Chin J Traumatol 2018 Apr;21(2):77-79. [doi: 10.1016/j.cjtee.2018.01.003] [Medline: 29605432]

Abbreviations

ABC: airway, breathing, circulation
ATLS: Advanced Trauma Life Support
ED: emergency department
KAP: knowledge, attitude, and practice
KM: Kirkpatrick model
LMIC: low- and middle-income country
PTC: Primary Trauma Care

Edited by TDA Cardoso; submitted 05.04.23; peer-reviewed by A Nguyen, D Ikwuka, E Power, F Hussain, K Marijke, S Afzal; revised version received 12.06.24; accepted 15.06.24; published 23.07.24.

Please cite as:
Nguyen BT, Nguyen VA, Blizzard CL, Palmer A, Nguyen HT, Quyet TC, Tran V, Skinner M, Perndt H, Nelson MR
Using the Kirkpatrick Model to Evaluate the Effect of a Primary Trauma Care Course on Health Care Workers' Knowledge, Attitude, and Practice in Two Vietnamese Local Hospitals: Prospective Intervention Study
JMIR Med Educ 2024;10:e47127
URL: https://mededu.jmir.org/2024/1/e47127
doi:10.2196/47127

© Ba Tuan Nguyen, Van Anh Nguyen, Christopher Leigh Blizzard, Andrew Palmer, Huu Tu Nguyen, Thang Cong Quyet, Viet Tran, Marcus Skinner, Haydn Perndt, Mark R Nelson. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.7.2024. 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,

XSL•FO
RenderX

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.

XSL•FO
RenderX

Original Paper

The Use of a Novel Virtual Reality Training Tool for Peritoneal Dialysis: Qualitative Assessment Among Health Care Professionals

Caterina Lonati[1*], PhD; Marie Wellhausen[2*]; Stefan Pennig[3], PhD; Thomas Röhrßen[3]; Fatih Kircelli[2], Prof Dr; Svenja Arendt[2]; Ulrich Tschulena[2], Dr rer nat

[1]Center for Preclinical Research, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

[2]Fresenius Medical Care, Bad Homburg, Germany

[3]context, Essen, Germany

[*]these authors contributed equally

Corresponding Author:
Ulrich Tschulena, Dr rer nat
Fresenius Medical Care
Else-Kröner-Str. 1
Bad Homburg, 61352
Germany
Phone: 49 61722688932
Email: Ulrich.Tschulena@freseniusmedicalcare.com

Abstract

Background: Effective peritoneal dialysis (PD) training is essential for performing dialysis at home and reducing the risk of peritonitis and other PD-related infections. Virtual reality (VR) is an innovative learning tool that is able to combine theoretical information, interactivity, and behavioral instructions while offering a playful learning environment. To improve patient training for PD, Fresenius Medical Care launched the *stay•safe* MyTraining VR, a novel educational program based on the use of a VR headset and a handheld controller.

Objective: This qualitative assessment aims to investigate opinions toward the new tool among the health care professionals (HCPs) who were responsible for implementing the VR application.

Methods: We recruited nursing staff and nephrologists who have gained practical experience with the *stay•safe* MyTraining VR within pilot dialysis centers. Predetermined open-ended questions were administered during individual and group video interviews.

Results: We interviewed 7 HCPs who have 2 to 20 years of experience in PD training. The number of patients trained with the *stay•safe* MyTraining VR ranged from 2 to 5 for each professional. The *stay•safe* MyTraining VR was well accepted and perceived as a valuable supplementary tool for PD training. From the respondents' perspective, the technology improved patients' learning experience by facilitating the internalization of both medical information and procedural skills. HCPs highlighted that the opportunity offered by VR to reiterate training activities in a positive and safe learning environment, according to each patient's needs, can facilitate error correction and implement a standardized training curriculum. However, VR had limited use in the final phase of the patient PD training program, where learners need to get familiar with the handling of the materials. Moreover, the traditional PD training was still considered essential to manage the emotional and motivational aspects and address any patient-specific application-oriented questions. In addition to its use within PD training, VR was perceived as a useful tool to support the decision-making process of patients and train other HCPs. Moreover, VR introduction was associated with increased efficiency and productivity of HCPs because it enabled them to perform other activities while the patient was practicing with the device. As for patients' acceptance of the new tool, interviewees reported positive feedback, including that of older adults. Limited use with patients experiencing dementia or severe visual impairment or lacking sensomotoric competence was mentioned.

Conclusions: The *stay•safe* MyTraining VR is suggested to improve training efficiency and efficacy and thus could have a positive impact in the PD training scenario. Our study offers a process proposal that can serve as a guide to the implementation of a VR-based PD training program within other dialysis centers. Dedicated research is needed to assess the operational benefits and the consequences on patient management.

XSL•FO
RenderX

(*JMIR Med Educ 2024;10:e46220*) doi:10.2196/46220

KEYWORDS

peritoneal dialysis; virtual reality; patient education; patient training; chronic kidney disease; nursing; qualitative assessment

Introduction

Background

Compared with in-center dialysis, peritoneal dialysis (PD) confers significant benefits to patients with chronic kidney disease (CKD), including better preservation of residual renal function and higher treatment flexibility [1,2]. As a result, kidney health organizations recommend facilitating and increasing patients' access to home dialysis [3]. The SARS-CoV-2 pandemic emphasized the importance of expanding PD use [4]. However, only 11% of patients on dialysis are currently treated with PD [5]. Poor patient education and inadequate training were identified as significant factors contributing to the global underuse of PD [6,7]. In line with this, patients who received structured training more likely chose home dialysis over in-center treatment [8]. In addition, in a telemedicine patient education study, patients with CKD stages 4 and 5 receiving telemedicine predialysis education had increased health literacy and increased home modality choice [9].

Besides learning theoretical concepts, patients with CKD need to acquire physical skills to perform the PD procedure by themselves. Therefore, PD training programs mainly involve individual sessions between patients and nephrology health care professionals (HCPs) [10,11], with most of them represented by nurses [11]. According to an international survey, a successful PD training program typically requires an average training time of 30 hours or 6 days per patient and is predominantly conducted in a one-to-one setting with both the patient and the nurse [11,12]. Increasing evidence demonstrated that efficient patient training can also lead to improved patient outcomes [13-15]. In fact, longer training time was associated with lower PD-related peritonitis rates [13,15], while frequent patient retraining reduced the risk of exit-site infections [14]. Consistently, the International Society for Peritoneal Dialysis guidelines recommended PD training standardization and optimization to reduce peritonitis rates [16]. On the basis of these observations, the International Home Dialysis Roundtable prompted the adoption of new strategies to improve education programs and boost patient engagement in training activities [3]. Of interest, the use of visual or audio aids and computer-assisted instructions was proposed to enhance patient learning [17].

To meet these requirements, Fresenius Medical Care (Bad Homburg, Germany) launched a novel training program for continuous ambulatory PD (CAPD) based on the use of virtual reality (VR), referred to as *stay•safe* MyTraining VR [18]. VR is an innovative technology with a huge potential in medical training [19-25] and patient education [26-32]. VR uses head-mounted displays to create an immersive, computer-generated, 3D, and interactive environment. *Stay•safe* MyTraining VR equipment includes a VR headset and a handheld controller designed to support the training of patients on hygiene procedures, preparation and posttreatment steps, and bag exchange and operation. The novel digital training tool was introduced to support the classical PD training at 3 dialysis centers in Germany. As for the traditional PD training, nurses and nephrologists continue to be responsible for patient education when being supported by the *stay•safe* MyTraining VR.

This Study

This qualitative research investigated, for the first time, HCPs' perspectives, preferences, and attitudes toward the implementation of the *stay•safe* MyTraining VR within CAPD training. We sought to describe how and when the tool was integrated in the traditional training framework and to evaluate whether VR introduction had an impact on patient learning. Moreover, HCPs' opinions about the potential benefits and limitations related to VR were explored. Professionals' and patients' acceptance of the new technology was likewise investigated.

Methods

Study Design and Study Sample

This study is an integrative qualitative research involving convergent, narrative, problem-centered, and discursive interviewing designed to investigate the following aspects: (1) integration of the new tool into the classical PD training framework and professionals' perspectives on its usability, (2) target patient groups who can benefit from the new technology, (3) benefits and advantages, (4) limitations and weaknesses, (5) HCPs' and patients' acceptance of the new technology, and (6) additional applications. The collected information was then used to derive a process through which VR training can be implemented in the conventional training curriculum. All nursing staff and nephrologists in the 3 NephroCare dialysis centers where this technology was piloted, who already gained practical experience with the *stay•safe* MyTraining VR, were invited by mail to the respective medical administration for voluntary participation. Overall, 6 nurses and 1 nephrologist agreed and participated in this study.

Ethical Considerations

All participants read and signed the informed consent document, including a privacy policy explaining data collection, data use, and data storage, before participating in the study. To ensure confidentiality, all participant data were anonymized before analysis. No compensation was provided to the participants for their involvement in this study. Participants were free to withdraw from the study at any time and were informed so, and participation was carried out on a voluntary basis.

This study did not undergo formal ethics review. This is justified based on the anonymity of responses and because no risk was expected to survey participants and basic ethical principles (individual autonomy, self-determination, avoidance of harm,

XSL•FO
RenderX

care, and justice) were not violated, as mentioned by the ethics commission of Bavarian universities (Gemeinsamen Ethikkommission der Hochschulen Bayerns [33]), where it is outlined that no ethics commission review is needed when no risk of damage is to be expected for participants and if basic ethical principles are not violated. As an example for such an exemption, a questioning of experts is mentioned, as this does not cause any particular risk or burden beyond what is witnessed in everyday life, as in this study.

Data Collection

To maximize information gathering, the interviews included open-ended questions and were conducted digitally as individual and group video interviews.

More specifically, the interviews included the following phases (Multimedia Appendix 1): (1) introduction and transparency of objectives, approach, and methodology; (2) introductory question for topic identification and prioritization by the interviewee; (3) narrative phase with nondirective (only encouraging) interventions by the interviewer; (4) discursive phase, in which the interviewee is questioned about their views, opinions, and evaluations and is asked to analyze and reflect in greater depth; and (5) open guiding questions according to the interview guide.

During the discursive phase, initial hypotheses were confirmed, concretized, adjusted, or rejected in an iterative process through repeated testing and feedback. To this aim, different forms of intervention were used: request for justification, explication of gaps and contradictions, validation of conclusions, target-actual comparison, and hypothesis-guided and solution-oriented questions.

All the interviews were conducted by SP and TR, took approximately 60 to 90 minutes each, and were recorded in writing.

Definitions

VR "effectiveness" was investigated considering the following aspects: (1) patient satisfaction (question 1: what is the feedback from patients? question 2: what is the range here?), (2) patient learning success (question 1: how do you rate the learning success in comparison to classic training? question 2: where is it greater, where less?), and (3) risks related to the use of the VR technology in the PD educational setting (question 1: what risks do you see in a full transition to VR?)

By contrast, "efficiency" involved the evaluation of (1) time committed by HCPs to present the technology to patients (question 1: how does your time commitment compare to traditional training?), (2) ratio of effort to benefit (question 1: is the cost-benefit ratio appropriate?), and (3) patient throughput (question 1: can you care for more patients with the new method?)

Qualitative and Statistical Analyses

A category system (Multimedia Appendix 1) for the core statements was inductively formed based on the material so that frequency distributions of topics or core statements and the associated evaluation were possible (descriptive statistics). The results were compressed, structured, and interpreted by means of a qualitative content analysis.

We computed the absolute value and relative frequency for categorical variables.

Results

HCPs' Characteristics

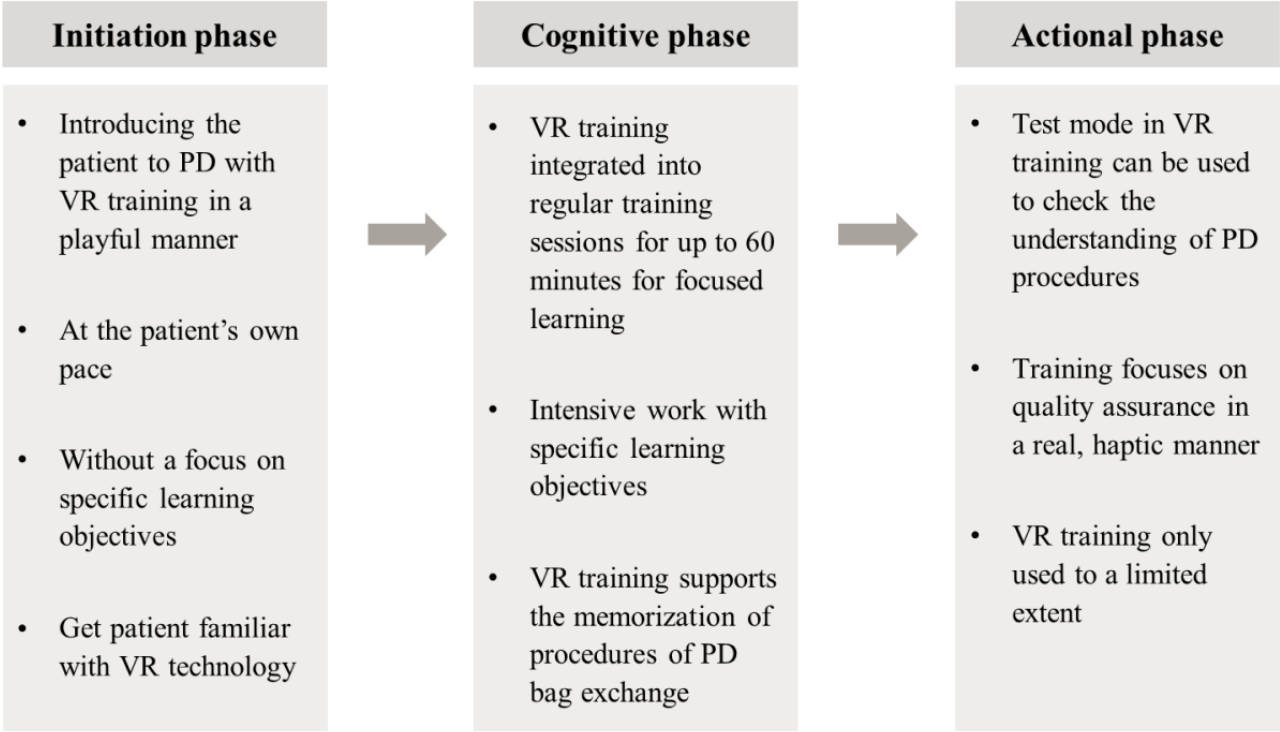
A total of 7 HCPs were interviewed, of whom 5 (71%) were nurses and 2 (29%) were nephrologists. All the respondents had >10 years of work experience with patients with CKD. HCPs had 2 to 20 years of experience in PD training, with an average of 1.5 months of experience with the *stay•safe* MyTraining VR. The number of patients trained using the VR technology ranged from 2 to 5 for each professional.

HCPs' Use of the VR Technology Within the PD Training Program

HCPs believed that the VR technology was a valuable complementary tool to the traditional PD training program. On the basis of the collected statements, the classical PD training takes approximately 2 weeks and comprises 3 different phases (Figure 1). It typically begins by providing patients with essential information needed to get a basic understanding about how PD works and how it will impact their lifestyle and habits. This process imposes a significant emotional overload on patients (motivational phase). Next, patients are instructed with all the theoretical concepts needed to successfully perform home dialysis (cognitive phase). Finally, once basic knowledge is acquired, patients start practicing the procedures using the dedicated materials and consumables. In this phase, HCPs assess patients' understanding, detect eventual risks or mistakes in performing the different tasks, and provide instructions to successfully manage any critical events (actional phase).

XSL•FO
RenderX

Figure 1. Use of the Stay•safe MyTraining virtual reality (VR) within the classical workflow of patient education for home dialysis. According to the interviewed health care professionals (HCPs), peritoneal dialysis (PD) training involves three consecutive phases: (1) motivational phase, where patients are instructed with the basic information about PD without excessive pressure due to initial emotional load; (2) cognitive phase, where patients acquire all the theoretical concepts as well as the procedures to perform home dialysis alone; and (3) actional phase, where patients practice using the various materials and consumables and HCPs check whether the different learning objectives are effectively achieved. HCPs reported a successful integration of the novel VR tool into the routine activities during the motivational and cognitive phases. Notably, the VR training provided substantial advantages and benefits, allowing the improvement of patients' learning experience. By contrast, the use of the VR tool was limited in the actional phase because the acquisition of the practical skills required a real, haptic experience with the various materials and aids.



According to the interviewees, the novel approach was easily integrated into these routine training activities. Timings and modalities of VR administration were selected by each professional based on individual patient characteristics, receptivity, and interests. More specifically, our analysis indicated that the respondents used the VR tool with different aims depending on the particular training phase. For instance, after catheter insertion, VR was frequently offered to help patients approach PD in a playful manner, without immediately focusing on specific learning objectives:

> *I think the tool is especially good for providing systematic information and orientation to strengthen compliance to treatment.*

The VR tool was particularly helpful during the cognitive phase. Indeed, the use of VR enabled HCPs to more effectively explain the relevant instructions and to make the procedures clearer, facilitating patients' internalization of both cognitive and procedural skills. Hence, most of the interviewees thought that the VR technology significantly improved patient learning experience by enabling patients to anchor learning success, process routine, procedural safety, and efficiency in home dialysis. HCPs believed that at this training stage, the VR-based approach could replace parts of the classical program:

> *VR is a focused learning medium that can be used when the patient has established orientation and acceptance [for the therapy]*

> *It's probably not only good for decision making, but maybe also for learning the basic home dialysis process in more depth.*

> *VR can serve as a support, it facilitates learning...But with VR, we have a different learning level. As a patient, you have to work it out for yourself. The training serves to deepen the knowledge and sets a different focus...With the VR glasses you can internalize the sequence, that is what it is good for, that is what this technology supports.*

By contrast, HCPs reported limited use of the VR technology during the actional phase, when patients need to perform tactile dialysis-related tasks using all the equipment or consumables in sterile conditions. In fact, in the view of the interviewees, the VR technology could not replace classical training during this final phase because safe handling requires real experience with the various materials and aids. In addition, HCPs believe that error prevention and management as well as "quality control" need to be achieved through classical one-to-one sessions:

> *VR training does not replace classical training. In any case, the patient has to learn well the handling during bag change even without VR training. VR training is a supplement.*

> *The patient should first see the practical procedures such as turning on the exercise material first. Only then can VR be used. However, the handling still has to be experienced haptically.*

XSL•FO
RenderX

> *After the virtual trial, he then has to perform it independently in real life under observation.*
>
> *Definitive coverage with VR training is not possible. However, it can be a large component. But you also need a haptic test in real life.*

Target Patient Groups for VR Training According to HCPs

In the interviewees' opinions, the VR technology was suitable for most patients, including older adults. Patients' basic motivation and willingness to engage in the VR technology were indicated as essential prerequisites for the success of VR-based training:

> *There are no clear patient groups that you can include or exclude...Mental flexibility is crucial.*
>
> *The use of VR technology is not age-dependent. The interest depends on the personal attitude.*

PD may also be used in patients with limited capabilities in some cases. This was also addressed by the respondents in the study, but overall, the use of VR training was not considered suitable for patients with dementia. Limited use with patients experiencing severe visual impairment or lacking sensomotoric competence was also mentioned:

> *VR training is basically possible for almost all patients, but the key is how the patient receives it. Only in patients with dementia VR training is not applicable.*
>
> *The glasses are not suitable for participants with severe visual impairment because they cannot see and recognize sufficiently. A younger colleague with impaired vision was also unbale to see sufficiently in the glasses here.*

Benefits and Advantages of VR Training

VR training was perceived as a tool able to improve cognitive learning effects and make the training process more efficient in terms of personnel and material resources. The advantages and strengths of the technology pointed out by HCPs are listed in Table 1.

First of all, according to responders, VR introduction into PD training facilitated patient education and improved learning experience. First, VR training allowed patients' immersion into a focused learning environment and provided an efficient shielding from external stimuli (point 1, Table 1). This maximized patients' engagement in the different educational activities. Second, VR offered learners the opportunity to repeat training activities according to their individual needs (point 2, Table 1). This enabled patients to correct errors and learn a standardized curriculum in a safe learning environment. Training session repetition allowed learners to get familiar with the use of dialysis equipment and supplies, boosting their confidence in performing the specific tasks. Third, HCPs noted that by providing multiple learning stimuli, VR-based training improved the efficiency of knowledge transmission (point 3, Table 1). In addition, the creation of a playful learning environment contributed to enhancing patients' engagement and, consequently, to improving their learning experience while avoiding overwhelming them with too much information (point 4, Table 1).

Besides facilitating patients' learning process, the use of technology conferred different benefits to HCPs (point 2, Table 1). In fact, VR helped professionals explain the operating principles of PD, allowing them to convey the information in a simplified and intuitive manner (point 5, Table 1). Moreover, because HCPs were able to perform different activities (point 7, Table 1) while the patient was practicing with the devices, VR was perceived as a significant tool to increase their productivity (point 6, Table 1). Nurses also reported that VR-based training enabled an individual operator to care for more patients at the same time due to higher training efficiency (point 7, Table 1). In the HCPs' view, another relevant advantage provided by the VR-based training was the opportunity to standardize the educational program, following the systematic content of the VR software (point 10, Table 1).

Finally, as the virtual experience does not require the use of additional PD supplies, the costs of training can be significantly reduced, as the number of dialysis fluid bags needed for training can be reduced (point 11, Table 1).

XSL•FO
RenderX

Table 1. Benefits and advantages of virtual reality (VR)–based training according to the interviewed health care professionals (HCPs).

Benefits and advantages	Representative quotes
For patients	
1. Focused learning environment	• "In VR, the patient has no distractions, but is completely concentrated in the system. Those who like to move around in this system and are focused in it learn quickly." • "In conventional training, patients are very distracted by stimuli from their immediate environment and their minds are then elsewhere. In VR training, for example, you do not see the nurse who suddenly comes into the room and disturbs you." • "VR can serve as a support, it facilitates learning...But with VR, we have a different learning level. As a patient, you have to work it out for yourself. The training serves to deepen the knowledge and sets a different focus...With the VR glasses you can internalize the sequence, that's what it's good for, that's what this technology supports."
2. Unlimited repetition of the procedures	• "The patient then repeats it as often as he wants and until it works." • "It is good for the mindless rehearsal of the procedures. The patient then repeats it as often as he wants and until it works. This is a learning effect." • "More security is created because the process is replayed over and over again."
3. Multiple learning stimuli	• "With VR, we have a different learning level. As a patient, you have to work it out for yourself. The training serves to deepen the knowledge and sets a different focus." • "The VR training can connect well to the therapy concept in the center to also bring variety into the training."
4. Playful learning environment	• "VR mainly serves to support the theory in a playful way, not doctrinal, but a different approach, not like an exam." • "The VR is playful, not schoolmasterly. It's a different approach with less pressure." • "VR takes the pressure out of the learning process."
For HCPs	
5. Simplified knowledge transfer	• "In classic training, I do a lot of words and that may then lead to mental shutdown. VR shows that in a much shorter time." • "The training replaces what I do verbally otherwise. Important aspects like turning off air conditioning, pets out of the room, etc. they see in the movie."
6. Enhanced HCPs' efficiency or productivity	• "The training replaces what I do verbally otherwise. Important aspects like turning off air conditioning, pets out of the room, etc. they see in the movie. In classic training, I do a lot of words and that may then lead to shutdown. VR shows that in a much shorter time." • "The patient had good observation skills...Therefore, I was only fully present the first time. The second time, I just put on my glasses and was gone for 15-30 minutes. I went to another patient and she stayed alone in the training room."
7. More effective time management	• "We can perform the following activities in parallel, for example, in the same room in the presence of the patient: Prepare laboratory for the next patient, evaluate laboratory tests, calculate peritonial equilibrium test (PÄT), write prescription, arrange appointments, sort and file findings, etc..." • "I can imagine that you can still do something in parallel in the same room at the desk (e.g., look at lab values, do documentation, prepare classic PD, e.g., tear open bags)." • "The patient had good observation skills...Therefore, I was only fully present the first time. The second time, I just put on my glasses and was gone for 15-30 minutes. I went to another patient and she stayed alone in the training room." • "I can do other things in parallel, there I have a discretion. There would be further efficiency if several patients are cared for in parallel in one room with VR goggles. I would try this." • "Many parallel/routine activities in the room are possible, which are not possible in classical training."
8. Induction and motivation of HCPs	• "New young non-specialized nursing staff, can get orientation via the VR. One can also see in the VR glasses a motivational approach for employees who do not yet have a connection to dialysis or home dialysis and are in training or further education. They can then realize 'oh, this is a very interesting area of work for me.'" • "The VR training is useful for the introduction of staff, including trainees and physicians."
9. HCPs' training	• "It is good for nurse training, doctor training and patient training."
For dialysis centers	
10. Standardization of training programs	• "The VR technique is not so subjective in its application. The program allows objectification, unification and a standardized approach. One forgets then nothing."
11. Cost saving	• "Material costs: The number of bag exchanges is reduced. One consumes 11-15 € per training. You can save about 5 bags."

XSL•FO
RenderX

Limitations and Weaknesses of VR Training

A list of the potential drawbacks of VR training is provided in Textbox 1.

According to HCPs, the most significant weakness of VR training resides in the inherent inability to provide learning effects in the area of tactile perception and sensorimotor fine control (point 1, Textbox 1).

Moreover, the traditional approach was still perceived as the best strategy to manage the emotional aspects related to PD treatment and address any patient-specific educational needs. In fact, according to the interviewees, the establishment of a trusting relationship between HCPs and learners is an essential prerequisite to identify emotional barriers and motivational obstacles toward PD (point 2, Textbox 1). Despite being indicated as a tool to reduce HCPs' workload, some respondents were concerned about a possible additional workload to acquire the concept of VR training both in their curriculum and for their patients during the first VR-based training sessions (point 3, Textbox 1).

Textbox 1. Drawbacks and weaknesses of virtual reality (VR)–based training according to the interviewed health care professionals (HCPs).

Drawbacks, weaknesses, and representative quotes

1. Lack of tactile education
   - "It's difficult in VR technology to get that shown virtually with the dressings. You don't get a feeling for the catheter connection in VR, for example the resistance when screwing it shut."
   - "VR cannot replace the haptic real-world experience."
   - "VR training does not replace classical training. In any case, the patient must learn well how to handle the bag change even without VR training. VR training is a supplement and VR training does not replace classical training. In any case, the patient must learn well how to handle the bag change even without VR training. VR training is a supplement."
   - "...the handling still has to be experienced haptically...VR is not a substitute for traditional training."
2. Relationship with the patient remains important
   - "The nurse cannot be replaced by the device. The events and intermediate questions must be clarified. There are quite spontaneous questions."
   - "The relationship aspect of care is indescribably important. VR does not change that."
   - "The device is a supplement, the nurse cannot be replaced by the device. The events and intermediate questions must be clarified. Questions arise spontaneously."
   - "It's all schematic in VR training. Unusual events can't be handled in the system."
   - "Everything is schematic. Unusual events cannot be processed in the system. There are typical critical events that are not mapped."
3. Initial workload for HCPs
   - "The time for preparation and implementation of VR technology is then missing in the dialogue with the patient."
   - "In the beginning you have some effort, but only later there is a time effect."

HCPs' Perspectives, Motivation, and Attitudes Toward the VR Technology

An investigation of respondents' attitudes toward VR revealed some differences among HCPs (Textbox 2). A large proportion of professionals was willing to use the new tool but felt the need to first gain self-confidence in using the technology. Some HCPs found VR exciting and were interested and enthusiastic. By contrast, others showed a nonpositive feeling and distrustful attitude toward the tool.

XSL•FO
RenderX

Textbox 2. Health care professionals' (HCPs') perspectives, motivation, and attitudes toward the virtual reality (VR) technology.

HCPs' attitudes and representative quotes

Positive

- "The VR training can connect well to the therapy concept in the center to also bring variety into the training."
- "I find the technology innovative."
- "You also have to see it as a technology of the future. I think it is good when things develop. I'm fundamentally interested and open."

Open to learn

- "I have to develop confidence first."
- "There is probably already an acceptance in the team, but for me it means a learning curve. The trust in VR has to develop first."
- "The first attempt was very confusing for me. Each subsequent one went better each time."
- "I find the VR application basically exhausting, but interesting. I find it fascinating."
- "VR is still a bit awkward to use: to grip the disinfectant, you have to use your hand, but before that you have to do something else...Other than that, I'm thrilled."
- "In the beginning you have some effort, but only later there is a time effect."

Negative

- "...there are also employees who only have a smartphone but no other IT skills. They tend to reject it."

VR Acceptance by Patients According to HCPs

In HCPs' opinions, patients' feedback on VR training was overall positive, and the acceptance was high, including that of older patients:

> *One patient [aged 82 years] has always wanted to exercise, but has not felt that well yet. VR is appropriate for him because he is really interested, because he really wants to learn. He thought it was great, is open minded.*
>
> *...we initially dealt with the topic of home dialysis rather "playfully."...She [older patient aged approximately 75 years] found the VR training enjoyable and interesting.*

Of note, HCPs were aware of the importance of adopting a patient-tailored approach to present the technology. In fact, the respondents highlighted the need to determine the patient's learning style and then to provide an initial clear and simple explanation of the use and purposes of the VR technology:

> *VR training is basically possible for almost all patients, but the key is how the patient receives it.*
>
> *You first have to check whether it makes sense, taking into account native language and level of education, acceptance of the medium, etc.*
>
> *You have to think carefully about when exactly to use the VR. The patient first has to get to know a lot of the basics and deal with the PD. Then, the patient needs a technical briefing: How does VR work? How do the individual elements such as the controller, etc. work? Then you first have to do 2-3 learning units and then a certain habit develops.*

HCPs reported a lower VR acceptance among patients not familiar with electronic or audiovisual media and devices, irrespective of their age:

> *Not many, but some find it awful if there was no computer experience. One patient did not understand it properly.*
>
> *People who are inexperienced with computers find it difficult to use them.*

Additional Applications of VR Training

HCPs believed that, besides patient education, the VR technology can be applied with other relevant goals and functions in the context of PD.

The administration of VR training before obtaining informed consent could significantly support patients in their decision-making about home dialysis. In fact, allowing an in-depth understanding of the PD procedures, VR training can improve patients' awareness and understanding about home-based treatment and, consequently, help them take more informed decisions:

> *The VR glasses can be used at the beginning in the patient consultation before the treatment decision is made...It can support the patient's decision making for home dialysis.*

At a later stage, VR can be used as a motivational tool to enhance and reinforce patients' decision about home PD:

> *If it is clear that the patient needs dialysis, then you can explain different procedures to them and then for explaining PD you can use VR. That is great.*

In addition, HCPs believed that VR could be offered to patients' family caregivers, who often play an important role in patients' decision-making as well as in their care activities:

XSL•FO
RenderX

> *VR training can also be used for training of family members.*

Another field of application of VR training could be within educational programs addressed to the HCPs themselves. Target users include both professionals already working with patients with CKD, such as nurses from cooperating ambulatory care services, and HCPs who are not directly involved in the PD process:

> *The VR training is useful for the introduction of staff, including trainees and physicians. VR training can also be used as training for family members. Retirement homes: nursing staff there, are yes very tightly staffed. The VR training can also be used in nursing homes and in hospitals with nursing staff for further training purposes.*

It is good for nurse training, doctor training and patient training.

Discussion

Principal Findings

This qualitative assessment shows that according to nephrology professionals, the *stay•safe* MyTraining VR may significantly improve PD training efficiency and efficacy. In fact, VR was perceived as a valuable complementary tool able to enhance patients' learning experience by providing the experiential learning necessary for a deeper understanding of medical information. In addition to the positive effects for patients, the *stay•safe* MyTraining VR may also provide HCPs with the opportunity to improve their productivity, both by saving their time and by facilitating the transmission of medical information to patients.

VR is an immersive experiential technology that emulates the physical world through digital simulation. Although originally intended for entertainment purposes, VR has a huge potential as a learning tool due to its unique ability to combine procedural information, interactivity, acoustic and visual information, and behavioral instructions [34]. In the medical field, VR-based education opened a new era of professionals' training on surgical and endoscopy techniques [19-25,35]. More recently, the VR technology has also been implemented for patients' education and training [26-32]. An increasing number of studies demonstrated that the use of simulation media can support different patients' learning styles by providing a mix of visual, auditory, interactive, and text elements [29,30] and improve patients' understanding and comprehension of their diseases as well as of the treatment or lifestyle interventions required to cope with them [26,29,32,36-38]. In addition, by enabling the repetition of training sessions according to individual patient needs, the technology facilitates the acquisition of competences and skills by learners. The possibility to build muscle memory through experiential training [30] and to learn from mistakes in a safe environment further improves patients' learning experience. VR is similarly emerging as an innovative strategy to foster self-management and self-care in patients with chronic conditions [28]. Finally, VR-based education could increase patient engagement and empowerment [26] while reducing anxiety and pressure related to medical procedures [26,39-41].

In nephrology, effective patient education and training have an impact on the risk of peritonitis [13-15]. In fact, training for home therapies poses unique educational challenges, as patients need to acquire not only the theoretical concepts of dialysis but also the technical skills required to manage all the procedures on their own [2,42]. Recently, Zgoura et al [43] proposed the use of VR headsets and gamification elements in support of PD training with the aim to standardize, facilitate, and accelerate patients' learning process. Here, we report HCPs' perspectives, opinions, and attitudes toward the implementation, into PD training programs, of the *stay•safe* MyTraining VR, a novel PD training program based on the use of a VR headset and a handheld controller [18]. We first explored the timings and modalities of VR integration within the classical training curriculum. Overall, the VR technology was perceived as a helpful supplementary tool, but HCPs specified some functional differences between the traditional and VR-based training programs. According to HCPs, VR-based training was particularly useful during the cognitive phase of the PD training program, where it not only enhanced patients' learning experience and facilitated information internalization but also assisted nurses in explaining the theory and procedures in a more effective and simplified manner. VR was a valuable support also during the initiation phase, during which it helped HCPs overcome potential patients' emotional barriers and lack of self-confidence. Conversely, according to the respondents, the actional phase still required a more classical approach, as a real, haptic experience was indicated as essential to learn the correct handling of the various materials as well as how to manage any critical steps or mistakes.

In HCPs' opinions, a significant added value of the VR technology resides in its ability to improve patients' learning experience and make the entire training process more effective and efficient. These relevant effects are achieved, thanks to the inherent characteristics of immersive or interactive media, which were clearly identified by the interviewees, including a focused learning environment and multiple learning stimuli. The opportunity to repeat the training modules based on patients' individual needs and cognitive skills was also highlighted. The impact of task repetition in education and training is currently well known [17,44]. In the context of health care educational interventions, the repetition of training activities resulted in faster skill acquisition and improved transfer of learning to practice [45]. The mechanisms underlying repetition benefits in cognitive learning include the induction of long-term memory [46] and internalization of procedures as unconscious habits [47]. With each repetition, the cognitive effort required for memory performance, behavior planning, and action control is reduced [30]. This allows rapid recall of habits from memory and improves learners' performance and confidence. Moreover, HCPs noted that repeated training sessions can offer patients the opportunity to correct errors and refine their skills in a safe learning environment. Therefore, patients could get instructions and learn from their mistakes. Learning by doing is one of the central features of interactive education because it allows acquiring a high level of experience and becoming accustomed

XSL•FO
RenderX

to the therapy or procedure [17,25,48]. The playful and interactive learning environment may increase motivation in PD training. As shown by Kyaw et al [20], VR interventions with more interactivity showed better results in terms of knowledge and skills outcomes than interventions with less interactivity. Therefore, the *stay•safe* My Training VR, as a highly immersive training program, may lead to higher patients' motivation and, consequently, increased learning success. These unique attributes of VR could bring substantial benefits especially for older adults, who often need more training time to learn the procedure of PD [42].

Besides the positive effects on patients' cognitive learning, HCPs identified distinctive benefits associated with the use of *stay•safe* MyTraining VR as an informative tool. In fact, VR helped HCPs to illustrate the home dialysis process to patients without providing them with too much stressful information. Similar results were obtained with the use of immersive media to prepare patients for medical procedures [32] and radiotherapy [37] and to increase patients' knowledge about their disease and its associated consequences [26,29].

Concerning patients' acceptance of the new tool, HCPs reported that most of them showed a positive attitude toward VR. These observations confirm and expand the results provided by previous studies in populations with CKD, in which the use of VR either as a distraction or an educational tool was associated with high levels of satisfaction [31,49-51]. On the basis of the HCPs' statements, there was no difference in patients' engagement in the VR-based training between older and younger people. Consistently, recent research in participants with abdominal aortic aneurism and stroke indicated that patients' age does not affect their engagement in VR-based educational programs and is not associated with cybersickness [32,52,53]. As CKD is becoming more prevalent in older individuals [54], these observations appear particularly relevant for a future implementation of the VR training tool into PD education. By contrast, HCPs noted that people who were not familiar with the use of electronic or audiovisual devices tended to build emotional barriers or showed a greater uncertainty in their use of VR. These findings are in line with the studies performed by Specht et al [52] and Huygelier et al [53], who found that patients with a negative attitude toward electronic media were more likely unwilling to use VR. Psychological research showed that acceptance toward computer media is significantly influenced by users' perceptions of the ease of use, usefulness, and playfulness of the different tools [55,56]. Of interest, studies conducted among older adults showed that users' perception that VR was useful, easy to use, and fun promoted a positive attitude toward the tool [57,58]. Moreover, interactivity was indicated as an important factor influencing the intention of continuous use of virtual practices [59]. Therefore, recommendations to boost patients' initial motivation toward VR include maximizing the positive aspects through increasing interactivity, enhancing users' perceptions of utility, reducing the difficulty associated with its use, and enhancing the playful nature of the training.

In addition to the several advantages for patients with CKD, the interviewees were able to identify different positive effects on their working activity. Indeed, in the HCPs' experience, professional assistance during VR-based training became increasingly unnecessary once patients acquired the skills to use the technology. Thus, HCPs were able to focus on numerous parallel activities and routine administrative tasks during training sessions. This allowed more effective time management among nurses [60]. Moreover, the respondents reported that by providing simple and schematic contents, VR helped convey the medical information more easily. This facilitated HCPs' teaching activity and contributed to boosting patient engagement in PD training. Consistent with these observations, professionals' acceptance toward the new tool was overall high. However, many employees highlighted that they felt the need to undergo structured training to get familiar with VR before using the technology with patients. Therefore, confidence in using the VR technology must be built up to recognize and use its full potential.

Another important improvement offered by VR-based PD training is related to cost-effectiveness. In fact, as the virtual experience does not require the use of additional dialysis fluid bags for mock training, the costs required to successfully train each patient in performing PD can be reduced [43]. Given the increasing prevalence of CKD [61], a more effective health care resource use is essential to ameliorate patient care. By providing the opportunity to save money on dialysis supplies and improving nurses' time management, VR-based training could contribute to reducing the economic burden of PD training while ensuring a high-quality educational offer.

HCPs mentioned other possible applications of VR beyond direct patient preparation for home dialysis that have not yet been tested in practice. The *stay•safe* MyTraining VR was used within the educative conversation with people who need to start dialysis with the aim to enhance the visibility of home therapies as a treatment option and to prepare patients psychologically and emotionally for PD. In this context, VR can support patients to make informed decisions about their treatment of choice. Of note, the gaming factor could have a crucial role in reassuring patients and boosting their self-confidence. Barriers for home dialysis training could thus be overcome by using VR as a complementary tool, which may in turn lead to increased uptake of home dialysis. HCPs also proposed to take advantage of the VR tool to inform patients' relatives about CAPD as a treatment option. As patients' and caregivers' low awareness about and poor understanding of home therapies are important factors underlying the low uptake of PD [6,62], the use of the VR technology as an informative motivational tool can have an important clinical impact among patients with CKD.

Some drawbacks related to the technology were also highlighted by HCPs. In the respondents' view, VR's inability to provide real, tactile education limits its use during the final phase of the PD training program. At this stage, while patients need to practice the procedural skills required to complete each task, HCPs check patients' ability to perform the procedures independently and manage eventual critical steps or mistakes. Therefore, the conventional approach was perceived as the best option for both patients' skill acquisition and professionals' quality assurance. The respondents similarly identified a potential limited use of VR training with specific patient groups,

XSL•FO
RenderX

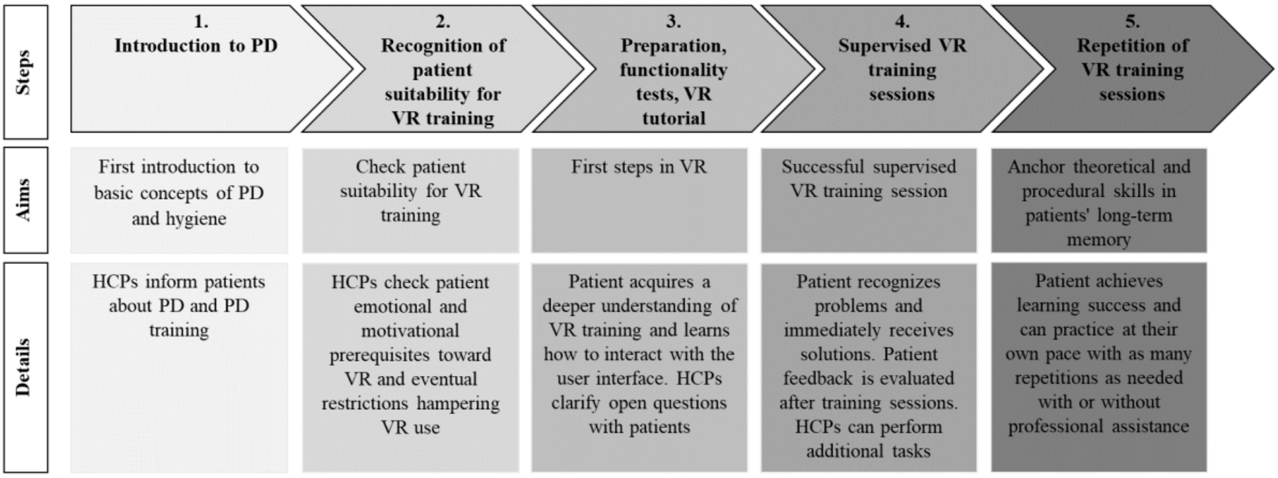
mainly among people experiencing cognitive disabilities or visual impairment.

Overall, the collected information indicates that VR implementation into clinical practice may have a profound impact in the PD training scenario. In fact, increasing patients' proficiency in self-management may lead to improved outcomes among patients with CKD [63] and to lower patient concerns related to home dialysis procedures, thereby improving their quality of life. This has a great relevance considering the high burden of anxiety in patients receiving PD [41]. Moreover, the use of VR as a motivational tool during decision-making helps patients in making informed choices and, as a consequence, can promote the uptake of home dialysis over in-center therapies. The use of VR software could similarly improve the educational or training offer because the medical contents can be given in a standardized manner in every dialysis clinic.

On the basis of these observations, we developed a 5-step process proposal that can serve as a guide to implement VR-based PD training within dialysis centers (Figure 2): (1) introduction to PD; (2) recognition of the patient's suitability for VR training; (3) preparation, functionality tests, and VR training tutorial; (4) supervised VR-based PD training sessions; and (5) repetition of VR-based PD training sessions. During the first step, HCPs evaluate patients' psychological, emotional, and physical barriers toward home dialysis to determine whether the patient meets the requirements for PD treatment (step 1, Figure 2). Before starting VR-based training, patients' attitude toward the technology as well as the presence of cognitive or physical disabilities potentially hampering the use of VR must be thoroughly evaluated (step 2, Figure 2). After checking equipment suitability, pretraining VR tutorials are administered, allowing patients to get familiar with the use of the VR tools (step 3, Figure 2). These preliminary sessions are intended to provide learners with all the instructions necessary to use the devices before starting PD training, with the ultimate goal to reduce patients' cognitive load at later steps. Next, supervised VR-based PD training sessions can be started (step 4, Figure 2). HCPs can check patients' understanding and eventually address any application-oriented questions directly. In addition, professionals evaluate patients' training needs and assess whether obstacles to the learning process are present. This approach enables HCPs to not only customize the information delivery process but also improve patient empowerment and self-confidence. The final step involves the repetition of the learning activities based on individual educational needs (step 5, Figure 2). HCPs can eventually check the acquisition of the procedural skills and provide suggestions to reinforce patient engagement in training activities.

This study has some limitations. One of the limitations is the small sample size, which may limit the generalizability of our findings to a broader population. In addition, while qualitative interviews provide valuable insights and generate working hypotheses, such an approach may introduce potential biases and subjectivity in the data collection process. Furthermore, the study may not have captured the full range of perspectives and experiences. Future research with a larger sample size could provide a more comprehensive understanding about the use of VR training for dialysis.

Figure 2. Process proposal to implement virtual reality (VR)–based peritoneal dialysis (PD) training programs. On the basis of the collected data, we propose a 5-step workflow that can serve as a guide for health care professionals (HCPs) and stakeholders to introduce the VR-based PD training within other dialysis centers: (1) HCPs' assessment of patients' suitability for home dialysis, (2) evaluation of patients' attitude and suitability toward the VR technology, (3) pretraining VR tutorials, (4) supervised VR-based PD training sessions, and (5) repetition of the various activities based on individual educational needs.



Conclusions

In conclusion, VR-based training is intended to facilitate and accelerate patients' skill acquisition required to perform a real PD bag exchange by themselves. Using this innovative technology for PD training is well accepted and feasible and provides the potential for nephrology HCPs to reach a large population and promote and facilitate home dialysis uptake. Further research is required to investigate the long-term effects of VR training on patient satisfaction, infection rates, and the longevity of PD treatment.

XSL•FO
RenderX

Data Availability

All data generated or analyzed during this study are included in this published article (Multimedia Appendix 1).

Authors' Contributions

CL contributed to writing the original draft, data visualization, reviewing, and editing. MW contributed to conceptualization, writing–review and editing, data visualization, and project administration. SP and TR contributed to investigation, formal analysis, reviewing, and editing. FK contributed to reviewing and editing. SA contributed to reviewing, editing, and data visualization. UT contributed to conceptualization, reviewing, editing, project administration, and supervision. All authors discussed the results and revised the final version of the manuscript.

Conflicts of Interest

MW, FK, SA, and UT are full-time employees at Fresenius Medical Care. CL provided medical writing services on behalf of Fresenius Medical Care. SP and TR conducted interviews on behalf of Fresenius Medical Care.

Multimedia Appendix 1
Interview guide.
[DOCX File , 28 KB - mededu_v10i1e46220_app1.docx]

References

1. François K, Bargman JM. Evaluating the benefits of home-based peritoneal dialysis. Int J Nephrol Renovasc Dis 2014;7:447-455 [FREE Full text] [doi: 10.2147/IJNRD.S50527] [Medline: 25506238]
2. Jacquet S, Trinh E. The potential burden of home dialysis on patients and caregivers: a narrative review. Can J Kidney Health Dis 2019 Dec 18;6:2054358119893335 [FREE Full text] [doi: 10.1177/2054358119893335] [Medline: 31897304]
3. Mendu ML, Divino-Filho JC, Vanholder R, Mitra S, Davies SJ, Jha V, International Home Dialysis Roundtable Steering Committee. Expanding utilization of home dialysis: an action agenda from the first international home dialysis roundtable. Kidney Med 2021 Jul;3(4):635-643 [FREE Full text] [doi: 10.1016/j.xkme.2021.04.004] [Medline: 34401729]
4. Stern LD, Waikar S. Time to expand access and utilization of home dialysis: lessons from the COVID-19 pandemic. Mayo Clin Proc 2020 Jul;95(7):1323-1324 [FREE Full text] [doi: 10.1016/j.mayocp.2020.04.038] [Medline: 32622441]
5. Annual report 2014. Fresenius Medical Care. 2014. URL: https://www.freseniusmedicalcare.com/fileadmin/data/com/pdf/Media_Center/Publications/Annual_Reports/FMC_AnnualReport_2014_en.pdf [accessed 2024-04-29]
6. Chan CT, Collins K, Ditschman EP, Koester-Wiedemann L, Saffer TL, Wallace E, et al. Overcoming barriers for uptake and continued use of home dialysis: an NKF-KDOQI conference report. Am J Kidney Dis 2020 Jun;75(6):926-934. [doi: 10.1053/j.ajkd.2019.11.007] [Medline: 32057468]
7. Chan CT, Wallace E, Golper TA, Rosner MH, Seshasai RK, Glickman JD, et al. Exploring barriers and potential solutions in home dialysis: an NKF-KDOQI conference outcomes report. Am J Kidney Dis 2019 Mar;73(3):363-371. [doi: 10.1053/j.ajkd.2018.09.015] [Medline: 30545707]
8. Mckeon K, Sibbel S, Brunelli SM, Matheson E, Lefeber N, Epps M, et al. Utilization of home dialysis and permanent vascular access at dialysis initiation following a structured CKD education program. Kidney Med 2022 Jul;4(7):100490 [FREE Full text] [doi: 10.1016/j.xkme.2022.100490] [Medline: 35801188]
9. Easom AM, Shukla AM, Rotaru D, Ounpraseuth S, Shah SV, Arthur JM, et al. Home run-results of a chronic kidney disease telemedicine patient education study. Clin Kidney J 2020 Oct;13(5):867-872 [FREE Full text] [doi: 10.1093/ckj/sfz096] [Medline: 33123362]
10. Bernardini J, Price V, Figueiredo A. Peritoneal dialysis patient training, 2006. Perit Dial Int 2020 Feb 23;26(6):625-632. [doi: 10.1177/089686080602600602]
11. Cheetham MS, Zhao J, McCullough K, Fuller DS, Cho Y, Krishnasamy R, et al. International peritoneal dialysis training practices and the risk of peritonitis. Nephrol Dial Transplant 2022 Apr 25;37(5):937-949. [doi: 10.1093/ndt/gfab298] [Medline: 34634100]
12. Bernardini J, Price V, Figueiredo A, Riemann A, Leung D. International survey of peritoneal dialysis training programs. Perit Dial Int 2006;26(6):658-663. [Medline: 17047232]
13. Figueiredo AE, Moraes TP, Bernardini J, Poli-de-Figueiredo CE, Barretti P, Olandoski M, BRAZPD Investigators. Impact of patient training patterns on peritonitis rates in a large national cohort study. Nephrol Dial Transplant 2015 Jan;30(1):137-142. [doi: 10.1093/ndt/gfu286] [Medline: 25204318]
14. Chang JH, Oh J, Park SK, Lee J, Kim SG, Kim SJ, et al. Frequent patient retraining at home reduces the risks of peritoneal dialysis-related infections: a randomised study. Sci Rep 2018 Aug 27;8(1):12919 [FREE Full text] [doi: 10.1038/s41598-018-30785-z] [Medline: 30150627]

XSL•FO
RenderX

15. Perl J, Fuller DS, Bieber BA, Boudville N, Kanjanabuch T, Ito Y, et al. Peritoneal dialysis-related infection rates and outcomes: results from the peritoneal dialysis outcomes and practice patterns study (PDOPPS). Am J Kidney Dis 2020 Jul;76(1):42-53 [FREE Full text] [doi: 10.1053/j.ajkd.2019.09.016] [Medline: 31932094]
16. Piraino B, Bernardini J, Brown E, Figueiredo A, Johnson DW, Lye WC, et al. ISPD position statement on reducing the risks of peritoneal dialysis-related infections. Perit Dial Int 2011;31(6):614-630. [doi: 10.3747/pdi.2011.00057] [Medline: 21880990]
17. Figueiredo AE, Bernardini J, Bowes E, Hiramatsu M, Price V, Su C, et al. A syllabus for teaching peritoneal dialysis to patients and caregivers. Perit Dial Int 2016 Nov 01;36(6):592-605 [FREE Full text] [doi: 10.3747/pdi.2015.00277] [Medline: 26917664]
18. Fresenius Medical Care uses virtual reality technology to train patients for home dialysis. Fresenius Medical Care AG. URL: https://www.freseniusmedicalcare.com/en/news/mytraining/ [accessed 2024-04-29]
19. Logeswaran A, Munsch C, Chong YJ, Ralph N, McCrossnan J. The role of extended reality technology in healthcare education: towards a learner-centred approach. Future Healthc J 2021 Mar 03;8(1):e79-e84 [FREE Full text] [doi: 10.7861/fhj.2020-0112] [Medline: 33791482]
20. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967 [FREE Full text] [doi: 10.2196/12967] [Medline: 31456579]
21. Xie B, Liu H, Alghofaili R, Zhang Y, Jiang Y, Lobo FD, et al. A review on virtual reality skill training applications. Front Virtual Real 2021 Apr 30;2:645153. [doi: 10.3389/frvir.2021.645153]
22. Halbig A, Babu SK, Gatter S, Latoschik ME, Brukamp K, von Mammen S. Opportunities and challenges of virtual reality in healthcare – a domain experts inquiry. Front Virtual Real 2022 Mar 23;3:837616. [doi: 10.3389/frvir.2022.837616]
23. Gasteiger N, van der Veer SN, Wilson P, Dowding D. How, for whom, and in which contexts or conditions augmented and virtual reality training works in upskilling health care workers: realist synthesis. JMIR Serious Games 2022 Feb 14;10(1):e31644 [FREE Full text] [doi: 10.2196/31644] [Medline: 35156931]
24. Barré J, Michelet D, Truchot J, Jolivet E, Recanzone T, Stiti S, et al. Virtual reality single-port sleeve gastrectomy training decreases physical and mental workload in novice surgeons: an exploratory study. Obes Surg 2019 Apr;29(4):1309-1316. [doi: 10.1007/s11695-018-03680-9] [Medline: 30689172]
25. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games 2021 Jul 08;9(3):e29080 [FREE Full text] [doi: 10.2196/29080] [Medline: 34255668]
26. van der Kruk SR, Zielinski R, MacDougall H, Hughes-Barton D, Gunn KM. Virtual reality as a patient education tool in healthcare: a scoping review. Patient Educ Couns 2022 Jul;105(7):1928-1942. [doi: 10.1016/j.pec.2022.02.005] [Medline: 35168856]
27. Gao Y, Ma L, Lin C, Zhu S, Yao L, Fan H, et al. Effects of virtual reality-based intervention on cognition, motor function, mood, and activities of daily living in patients with chronic stroke: a systematic review and meta-analysis of randomized controlled trials. Front Aging Neurosci 2021 Dec 13;13:766525 [FREE Full text] [doi: 10.3389/fnagi.2021.766525] [Medline: 34966267]
28. Reagan L, Pereira K, Jefferson V, Evans Kreider K, Totten S, D'Eramo Melkus G, et al. Diabetes Self-management training in a virtual environment. Diabetes Educ 2017 Aug;43(4):413-421. [doi: 10.1177/0145721717715632] [Medline: 28643607]
29. van der Linde-van den Bor M, Slond F, Liesdek OC, Suyker WJ, Weldam SW. The use of virtual reality in patient education related to medical somatic treatment: a scoping review. Patient Educ Couns 2022 Jul;105(7):1828-1841 [FREE Full text] [doi: 10.1016/j.pec.2021.12.015] [Medline: 35000833]
30. Maddox T, Chmielewski C, Fitzpatrick T. Virtual reality in chronic kidney disease education and training. Nephrol Nurs J 2022;49(4):329-381. [Medline: 36054805]
31. van Praet YV. The use of virtual reality in patient education: the case of chronic kidney patients. University of Twente. 2018. URL: http://essay.utwente.nl/76995/ [accessed 2024-04-29]
32. Pandrangi VC, Gaston B, Appelbaum NP, Albuquerque FC, Levy MM, Larson RA. The application of virtual reality in patient education. Ann Vasc Surg 2019 Aug;59:184-189. [doi: 10.1016/j.avsg.2019.01.015] [Medline: 31009725]
33. Häufig gestellte Fragen zum Ethikantrag. GEHBa. URL: https://www.gehba.de/fileadmin/daten/Gehba/GEHBa-FAQ_2.1.pdf [accessed 2024-04-29]
34. Burrai F, Othman S, Brioni E, Micheluzzi V, Luppi M, Apuzzo L, et al. Effects of virtual reality in patients undergoing dialysis: study protocol. Holist Nurs Pract 2019;33(6):327-337. [doi: 10.1097/HNP.0000000000000330] [Medline: 31045610]
35. Saab MM, Hegarty J, Murphy D, Landers M. Incorporating virtual reality in nurse education: a qualitative study of nursing students' perspectives. Nurse Educ Today 2021 Oct;105:105045 [FREE Full text] [doi: 10.1016/j.nedt.2021.105045] [Medline: 34245956]
36. Aardoom JJ, Hilt AD, Woudenberg T, Chavannes NH, Atsma DE. A preoperative virtual reality app for patients scheduled for cardiac catheterization: pre-post questionnaire study examining feasibility, usability, and acceptability. JMIR Cardio 2022 Feb 22;6(1):e29473 [FREE Full text] [doi: 10.2196/29473] [Medline: 35191839]

XSL•FO
RenderX

37. Wang LJ, Casto B, Luh JY, Wang SJ. Virtual reality-based education for patients undergoing radiation therapy. J Cancer Educ 2022 Jun 24;37(3):694-700 [FREE Full text] [doi: 10.1007/s13187-020-01870-7] [Medline: 32970303]
38. Cikajlo I, Peterlin Potisk K. Advantages of using 3D virtual reality based training in persons with Parkinson's disease: a parallel study. J Neuroeng Rehabil 2019 Oct 17;16(1):119 [FREE Full text] [doi: 10.1186/s12984-019-0601-1] [Medline: 31623622]
39. Găină MA, Szalontay AS, tefănescu G, Bălan GG, Ghiciuc CM, Bolo A, et al. State-of-the-art review on immersive virtual reality interventions for colonoscopy-induced anxiety and pain. J Clin Med 2022 Mar 17;11(6):1670 [FREE Full text] [doi: 10.3390/jcm11061670] [Medline: 35329993]
40. Koo C, Park J, Ryu J, Han S. The effect of virtual reality on preoperative anxiety: a meta-analysis of randomized controlled trials. J Clin Med 2020 Sep 29;9(10):3151 [FREE Full text] [doi: 10.3390/jcm9103151] [Medline: 33003411]
41. Mok MM, Liu CK, Lam MF, Kwan LP, Chan GC, Ma MK, et al. A longitudinal study on the prevalence and risk factors for depression and anxiety, quality of life, and clinical outcomes in incident peritoneal dialysis patients. Perit Dial Int 2019;39(1):74-82. [doi: 10.3747/pdi.2017.00168] [Medline: 29991560]
42. Schaepe C, Bergjan M. Educational interventions in peritoneal dialysis: a narrative review of the literature. Int J Nurs Stud 2015 Apr;52(4):882-898. [doi: 10.1016/j.ijnurstu.2014.12.009] [Medline: 25616708]
43. Zgoura P, Hettich D, Natzel J, Özcan F, Kantzow B. Virtual reality simulation in peritoneal dialysis training: the beginning of a new era. Blood Purif 2019;47(1-3):265-269 [FREE Full text] [doi: 10.1159/000494595] [Medline: 30522112]
44. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J 2019 Oct 11;6(3):181-185 [FREE Full text] [doi: 10.7861/fhj.2019-0036] [Medline: 31660522]
45. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach 2005 Jan;27(1):10-28. [doi: 10.1080/01421590500046924] [Medline: 16147767]
46. Smolen P, Zhang Y, Byrne JH. The right time to learn: mechanisms and optimization of spaced learning. Nat Rev Neurosci 2016 Feb 25;17(2):77-88 [FREE Full text] [doi: 10.1038/nrn.2015.18] [Medline: 26806627]
47. Hammar A. Automatic information processing. In: Seel NM, editor. Encyclopedia of the Sciences of Learning. Boston, MA: Springer; 2012:393-394.
48. Sirimanna P, Gladman MA. Development of a proficiency-based virtual reality simulation training curriculum for laparoscopic appendicectomy. ANZ J Surg 2017 Oct;87(10):760-766. [doi: 10.1111/ans.14135] [Medline: 28803457]
49. Omonaiye O, Smyth W, Nagle C. Impact of virtual reality interventions on haemodialysis patients: a scoping review. J Ren Care 2021 Sep;47(3):193-207. [doi: 10.1111/jorc.12362] [Medline: 33491276]
50. Burrai F, Othman S, Brioni E, Silingardi M, Micheluzzi V, Luppi M, et al. Virtual reality in dialysis: a new perspective on care. J Ren Care 2018 Dec;44(4):195-196. [doi: 10.1111/jorc.12264] [Medline: 30417581]
51. Pieterse AD, Huurman VA, Hierck BP, Reinders ME. Introducing the innovative technique of 360° virtual reality in kidney transplant education. Transpl Immunol 2018 Aug;49:5-6 [FREE Full text] [doi: 10.1016/j.trim.2018.03.001] [Medline: 29563056]
52. Specht J, Schroeder H, Krakow K, Meinhardt G, Stegmann B, Meinhardt-Injac B. Acceptance of immersive head-mounted display virtual reality in stroke patients. Comput Hum Behav Rep 2021 Aug;4:100141. [doi: 10.1016/j.chbr.2021.100141]
53. Huygelier H, Schraepen B, van Ee R, Vanden Abeele V, Gillebert CR. Acceptance of immersive head-mounted virtual reality in older adults. Sci Rep 2019 Mar 14;9(1):4519 [FREE Full text] [doi: 10.1038/s41598-019-41200-6] [Medline: 30872760]
54. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl (2011) 2022 Apr;12(1):7-11 [FREE Full text] [doi: 10.1016/j.kisu.2021.11.003] [Medline: 35529086]
55. Venkatesh V. Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model. Inf. Syst Res 2000 Dec;11(4):342-365. [doi: 10.1287/isre.11.4.342.11872]
56. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
57. Syed-Abdul S, Malwade S, Nursetyo AA, Sood M, Bhatia M, Barsasella D, et al. Virtual reality among the elderly: a usefulness and acceptance study from Taiwan. BMC Geriatr 2019 Aug 19;19(1):223 [FREE Full text] [doi: 10.1186/s12877-019-1218-8] [Medline: 31426766]
58. Roberts AR, de Schutter B, Franks K, Radina ME. Older adults' experiences with audiovisual virtual reality: perceived usefulness and other factors influencing technology acceptance. Clin Gerontol 2019;42(1):27-33. [doi: 10.1080/07317115.2018.1442380] [Medline: 29505343]
59. Huang CM, Liao JY, Lin TY, Hsu HP, Charles Lee TC, Guo JL. Effects of user experiences on continuance intention of using immersive three-dimensional virtual reality among institutionalized older adults. J Adv Nurs 2021 Sep;77(9):3784-3796. [doi: 10.1111/jan.14895] [Medline: 34051116]
60. Duffield C, Gardner G, Catling-Paull C. Nursing work and the use of nursing time. J Clin Nurs 2008 Dec 12;17(24):3269-3274. [doi: 10.1111/j.1365-2702.2008.02637.x] [Medline: 19146585]

XSL•FO
RenderX

61. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2020 Feb 29;395(10225):709-733 [FREE Full text] [doi: 10.1016/S0140-6736(20)30045-3] [Medline: 32061315]
62. Walker RC, Howard K, Morton RL, Palmer SC, Marshall MR, Tong A. Patient and caregiver values, beliefs and experiences when considering home dialysis as a treatment option: a semi-structured interview study. Nephrol Dial Transplant 2016 Jan;31(1):133-141. [doi: 10.1093/ndt/gfv330] [Medline: 26346314]
63. Lin MY, Liu MF, Hsu LF, Tsai PS. Effects of self-management on chronic kidney disease: a meta-analysis. Int J Nurs Stud 2017 Sep;74:128-137. [doi: 10.1016/j.ijnurstu.2017.06.008] [Medline: 28689160]

Abbreviations

CAPD: continuous ambulatory peritoneal dialysis
CKD: chronic kidney disease
HCP: health care professional
PD: peritoneal dialysis
VR: virtual reality

Edited by A Mavragani; submitted 02.02.23; peer-reviewed by TM Chan, M Singh, T de Azevedo Cardoso; comments to author 15.01.24; revised version received 05.03.24; accepted 11.03.24; published 06.08.24.

Please cite as:
Lonati C, Wellhausen M, Pennig S, Röhrßen T, Kircelli F, Arendt S, Tschulena U
The Use of a Novel Virtual Reality Training Tool for Peritoneal Dialysis: Qualitative Assessment Among Health Care Professionals
JMIR Med Educ 2024;10:e46220
URL: https://mededu.jmir.org/2024/1/e46220
doi:10.2196/46220
PMID:39106093

©Caterina Lonati, Marie Wellhausen, Stefan Pennig, Thomas Röhrßen, Fatih Kircelli, Svenja Arendt, Ulrich Tschulena. Originally published in JMIR Medical Education (https://mededu.jmir.org), 06.08.2024. 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.

XSL•FO
RenderX

Original Paper

Virtual Reality Simulation in Undergraduate Health Care Education Programs: Usability Study

Gry Mørk[1], MSc; Tore Bonsaksen[1,2], MSc; Ole Sønnik Larsen[1], MSc; Hans Martin Kunnikoff[1], MSc; Silje Stangeland Lie[1], PhD

[1]Department of Health, Faculty of Health Sciences, VID Specialized University, Stavanger, Norway

[2]Department of Health and Nursing Sciences, Faculty of Social and Health Sciences, Inland Norway University of Applied Sciences, Elverum, Norway

Corresponding Author:
Gry Mørk, MSc
Department of Health
Faculty of Health Sciences
VID Specialized University
Misjonsmarka 12
Stavanger, 4024
Norway
Phone: 47 47234324
Email: gry.mork@vid.no

Abstract

Background: Virtual reality (VR) is increasingly being used in higher education for clinical skills training and role-playing among health care students. Using 360° videos in VR headsets, followed by peer debrief and group discussions, may strengthen students' social and emotional learning.

Objective: This study aimed to explore student-perceived usability of VR simulation in three health care education programs in Norway.

Methods: Students from one university participated in a VR simulation program. Of these, students in social education (n=74), nursing (n=45), and occupational therapy (n=27) completed a questionnaire asking about their perceptions of the usability of the VR simulation and the related learning activities. Differences between groups of students were examined with Pearson chi-square tests and with 1-way ANOVA. Qualitative content analysis was used to analyze data from open-ended questions.

Results: The nursing students were most satisfied with the usability of the VR simulation, while the occupational therapy students were least satisfied. The nursing students had more often prior experience from using VR technology (60%), while occupational therapy students less often had prior experience (37%). Nevertheless, high mean scores indicated that the students experienced the VR simulation and the related learning activities as very useful. The results also showed that by using realistic scenarios in VR simulation, health care students can be prepared for complex clinical situations in a safe environment. Also, group debriefing sessions are a vital part of the learning process that enhance active involvement with peers.

Conclusions: VR simulation has promise and potential as a pedagogical tool in health care education, especially for training soft skills relevant for clinical practice, such as communication, decision-making, time management, and critical thinking.

(JMIR Med Educ 2024;10:e56844) doi:10.2196/56844

KEYWORDS

360° videos; health professions education; virtual reality; usability study; undergraduates; university; students; simulation

Introduction

Background

Virtual reality (VR)–based training has a generally high acceptance amongst trainees, regardless of the technology limitations, usability challenges, and cybersickness [1]. A systematic review of the effectiveness of VR-based simulation training from the past 30 years showed evidence for VR as useful for training cognitive skills, such as spatial memory, learning and remembering procedures, and psychomotor skills. VR was also found to be a good alternative where regular job training was either impossible or unsafe to implement [1].

XSL•FO
RenderX

Targeted training of health care students is paramount to ensuring the provision of patient-centered care and effective communication with patients and their families. Using 360° videos in VR headsets as a starting point for VR simulation has immense potential for the systematic design of learning experiences and for fostering social and emotional learning through collaborative interactions with students [2-4]. Social and emotional learning concerns the development of emotional intelligence skills, including self-awareness, self-management, social awareness, relationship skills, and responsible decision-making [5]. Further, VR simulation offers an active and engaging approach to learning, which can positively impact the motivation of both teachers and students while enhancing learning opportunities and creativity in the learning process [4,6]. Videos of 360° are prerecorded using an omnidirectional camera that films in every direction. This means that the viewer can look around freely as in other VR experiences, but movement and interaction are limited because the camera only records from one position at a time [4].

To ensure active learning, the vicarious VR learning experiences should be provided with an instructional component of debriefing and peer discussion [7]. According to Biggs et al [8], education is about conceptual change, not just the acquisition of information. Such conceptual change takes place when students work collaboratively and in dialogue with others, both peers and teachers. Good dialogue is an essential element of activities that shape, elaborate, and deepen understanding [8].

VR simulation has the great advantage of facilitating a rich, detailed learning environment through different scenarios that are difficult to replicate in reality (eg, frightening situations) [9]. Considering that most health professions are underpinned by a client-centered philosophy of practice, the use of VR will be more relevant when educators create content that is particularly aligned with each specific education program [10]. Consideration of design issues and student experiences of usability are central for developing VR technologies in medical education programs [11]. According to Nielsen [12], usability is defined by five quality components: (1) how easy it is to accomplish the task, (2) efficiency (when learned), (3) memorability, (4) number of user errors, and (5) satisfaction related to the design. Usability has also been extended to comprise user experience that has a more compelling emotional and motivational effect [13]. We were interested in exploring the usability of a VR simulation program designed for health care students, in particular with a view to perceived learning outcomes and possible enhancement of soft skills such as communication skills and clinical decision making.

Even though recent studies indicate that VR is increasingly being used in higher education for clinical skills training and role-playing among health care students [6,14], there is a research gap concerning the use of VR to facilitate the nontechnical skills inherent in the social and emotional competences [5], such as communication, decision-making, and ethical reflection [15-18]. There is a need for more research on VR simulation in higher education programs within health care [4,19,20]. Further research needs to be directed toward the application of immersive 360° videos and how it is experienced by students in higher education [4,19,20] and studies that explore the usability of VR simulation [2,19].

A recent review conducted on implementation of VR argued that future studies should explore viable approaches to incorporate VR into health professions education [21]. During the last 2 years, there have been published a few qualitative [22-24] as well as quantitative studies on the usability of VR in nursing education [4]. Literature searches in Google Scholar using the search words *usability AND *virtual reality AND *occupational therapy/social education, published after 2022, revealed 1 mixed methods study on the usability in physical therapy education [25]. This search did not identify any existing publications on usability of VR simulation in occupational therapy or social education programs.

The recommendations from earlier research as well as the scarcity of usability studies concerning VR in health education programs suggest that VR implementation processes and the usability of VR simulation programs should be further explored and evaluated.

Study Aim

This study aimed to explore student-perceived usability of VR simulation in 3 health care education programs in Norway. The research question was: What are the students' perceptions of the usability of the employed VR simulation and related learning activities, and are there differences in perceptions between students in different study programs?

Methods

Design and Study Context

This study is part of a larger interdisciplinary project in a Norwegian university, the Solstien 3 project, with the objective to create a VR simulation that portrays situations future health and social workers might encounter in their professional practice. The purpose was to offer students a safe and controlled environment to practice handling challenging and unexpected scenarios without the risk of compromising the well-being of clients, patients, or themselves. The VR simulation, developed and used for soft skills training, was intended to supplement teaching and field placements [26].

The project has developed a "virtual learning center" accessible on a webpage, featuring portrayals of service recipients (patients or clients) through text and images, along with 360° videos illustrating various scenarios [27]. The design of the VR simulation included scenarios and related discussion tasks based on shared learning outcomes for higher education in Norway [28]. The VR simulation in all education programs commenced with a briefing, followed by students watching the scenario using VR headsets, without any possibility to interact with the actors or the environment. The students subsequently participated in facilitated debrief discussions in groups or conducted planned assignments in groups for peer learning. The debrief was concerned with the ethical issues related to the scenario, as well as communication and interpersonal skills. The VR simulation was conducted as part of the regular curriculum for undergraduate students enrolled in the involved education programs; occupational therapy, social education,

XSL•FO
RenderX

and nursing. After graduation, those who have undergone these education programs with satisfactory results are qualified to work as health care personnel in Norway, provided authorization from The Norwegian Directorate of Health.

In the early development phase of the Solstien 3 project, both faculty members' and students' experience with the prototype 360° video was investigated. The findings showed the importance of VR being contextualized directly in educational programs to create a safe environment for learning [26]. Further, in a pilot study, the students' experience of the VR simulation was explored. Findings showed that students experienced VR simulation as valuable as a space for authentic, engaging, and reflective practice, and therefore the students felt prepared for professional practice [29]. The use of 360° videos in combination with group discussions activates social and emotional learning and appears to be promising for enhancing professional learning [30]. The process and content of the Solstien 3 project are described in detail and are accessible through the aforementioned webpage [27].

Participants

Students from 3 undergraduate health care education programs at the same university were invited to participate in this study. These students, from occupational therapy, social education, and nursing, had VR simulation scenarios that were directly associated with their professional practice (developed by the Solstien 3 project). The VR simulation was therefore carried out in different subject-specific courses in different semesters for these 3 education programs. The data were collected during the academic year of 2022-2023, specifically between October 2022 and May 2023. An overview of the VR simulation and the related learning activities in the 3 education programs are displayed in Table 1, and a description of the scenarios and the students' tasks is displayed in Multimedia Appendix 1. The social education students were in their second semester in their first year of study, while the occupational therapy students were in their fourth semester in their second year of study. The nursing students were in their fifth and sixth semesters in their third year of study.

Table 1. Overview of virtual reality (VR) simulation learning activities in 3 different education programs.

	Occupational therapy	Social education	Nursing
Semester and name of subject	Second year, fourth semester *Participation and belonging* (15 ECTS[a])	First year, second semester *Social education work processes* (10 ECTS)	Third year, fifth and sixth semesters *Home nursing care* (15 ECTS)
Students in class	38 students	95 students	280 students
Introduction	The simulation started with some initial information in plenum. The students were given a short instruction about the task and shown how to use and adjust the VR headsets	The simulation started with some initial information in plenum. The students were given a short instruction about the task and shown how to use and adjust the VR headsets	The simulation started with some initial information in plenum. The students were given a short instruction about the task and shown how to use and adjust the VR headsets
Group sizes	VR simulation: 12-15 students Discussion groups: 6-7 students	VR simulation: 14-15 students Discussion groups: 6-8 students	VR simulation: 8-9 students Discussion groups: 8-9 students
Organization of the learning activity	The class was divided into 3 VR simulation groups and 6 peer discussion groups. The students spent approximately 3 hours on the whole session.	The class was structured into 3 cohorts for teaching on different days. Each cohort was divided into 2 VR simulation groups and 5 peer discussion groups. The students spent between 2 and 4 hours on the whole session.	The class was structured into 3 cohorts for teaching on different days. The cohort was divided into approximately 11 VR simulation and peer discussion groups. The students spent 1 hour on the whole session (as a part of a full-day health care simulation program).
Peer debrief	After watching the 360° videos, students were organized into peer discussion groups of 6-7 students. Each group had 1 facilitator.	After watching the 360° videos, students were organized into peer discussion groups of 6-8 students. Only 1 facilitator circulated between the groups.	After watching the 360° videos, students stayed in the same peer groups for discussion and debrief. Each group had 1 facilitator.
Learning outcomes	The debrief is concerned with the ethical issues that arise in the situation, as well as communication and interpersonal skills and a discussion on fall prevention.	The students collaborated to discuss and create a written assignment based on their experiences in the situation. The task involved the identification of possible conflicts of values and target behaviors and a discussion on the choice of appropriate mapping tools.	The debrief is concerned with the ethical issues that arise in the situation and the relationship and communication between the nurse and the patient as well as the family.

[a]ECTS: European Credit Transfer and Accumulation System.

Measurement

To find the most suitable questionnaire for this study, research work with similar objectives was examined. We found 2 relevant studies related to the usability of VR simulation in nursing education [31,32]. Verkuyl et al [31] developed a questionnaire based on earlier usability testing in the same study context. This questionnaire of 18 items had 2 sections: *perceived easeof use* and *perceived usefulness.* The questionnaire was relevant to our study as items considered whether the students found the VR

XSL•FO
RenderX

simulation easy to use, as well as whether the learning activity would enhance soft skills such as communication skills and clinical decision-making [31]. The questionnaire of Lee et al [32] was a further refinement of the work of Verkuyl et al [31] and consisted of 17 items to be rated and 7 open-ended questions. Based on these 2 studies, a questionnaire was developed by the research group, containing 17 items to be rated and 2 open-ended questions. The questionnaire also included 6 items assessing student characteristics: study program affiliation (social education, occupational therapy, and nursing), gender (male vs female), age (years), current working status in health care or social services (yes vs no), prior experience using VR technology (yes vs no), and attitude toward VR technology in education (sceptical or indifferent vs positive).

In total, 17 statements related to the students' perceptions of the usability of VR simulation and the related learning activities were used. Of the total items, 9 concerned perceived ease of use and 8 items concerned perceived usefulness. The participants were instructed to rate their level of agreement (1=disagree, 2=disagree somewhat, 3=unsure, 4=agree somewhat, and 5=agree). Toward the end of the questionnaire, the participants were also asked to evaluate the VR simulation activity on a 1-5 scale (1=poorest evaluation and 5=best evaluation).

The questionnaire also had 2 open-ended questions: *What specifically did you learn from the learning activity (360° video, group discussions, etc.)* and *Do you have other comments about the 360° videosand the related learning activities? (Improvements? Other ideas?)*

The translation process was first undertaken by one of the research group members before the resulting translations were discussed and adjusted by the team. The students were invited to respond to a digital questionnaire in their native language (Norwegian) shortly after completing the learning activity.

Data Analysis

Differences between students enrolled in different education programs were examined with Pearson chi-square tests for categorical variables, and with 1-way ANOVA for continuous variables. To adjust for inflating type I error rates in the multiple comparisons, Tukey honestly significant difference (HSD) correction was applied to the analysis of variance. To analyze the data from the open-ended questions, we used qualitative content analysis [33]. Initially, the responses were thoroughly read to gain a broad understanding, followed by a systematic categorization into distinct codes. This process helped in identifying similarities and differences in the responses. Subsequently, these codes were grouped into 2 relevant categories, which facilitated the identification of recurring patterns focusing on students' experiences of the usability aspect of the VR simulation.

Ethical Considerations

This study is registered with the Norwegian Centre for Research Data (protocol code 423788). Research ethics were strictly adhered to throughout the study. The learning activity was conducted as a mandatory activity as part of the curriculum for all students in the relevant study programs (n=413). After having conducted the learning activity, the students were invited to complete the questionnaire, which was provided by a QR code in the classroom or a link sent by email. On the first page of the questionnaire, the students received information about the study's purpose and procedures, and that participation was voluntary. The students were also informed that they consented to participate in the study by pressing "go to questionnaire". In this way, informed consent was obtained from all participants before their involvement. The questionnaire was anonymous, and data cannot be traced back to individual informants. Moreover, the questionnaire did not ask for any sensitive information.

Results

Student Characteristics

Overall, 146 students responded to the questionnaire out of a total of 413 invited students (35.4% response rate), representing students of social education (n=74, response rate 77.9%), occupational therapy (n=27, response rate 71.1%), and nursing (n=45, response rate 16.1%). The social education students comprised 50.7% of the sample, while the nursing students comprised 30.8% and the occupational therapy students 18.5% of the sample. Most of the students were women (n=117, 80.1%), while 29 (19.9%) were men, with a larger proportion of younger students in the social education program and a larger proportion of older students in the nursing education program. The sample is described in Table 2.

At the time of the data collection, 105 (71.9%) students worked in health care or social services, and half of the students (n=73) had prior experience using VR technology. Statistically significant differences between the groups were found in the "attitude toward VR in education" and the "evaluation of the VR simulation learning activity." A larger proportion of nursing students were positive toward VR technology in their education program (n=42, 93.3%), compared with occupational therapy students (n=21, 77.8%) and social education students (n=56, 75.7%). Among the nursing students, 26 (72.2%) gave the highest possible rating (5) when evaluating the VR simulation learning activity. In contrast, this rating was given by 7 (33.3%) of the occupational therapy students and 20 (37%) of the social education students.

XSL•FO
RenderX

Table 2. Characteristics of students by study program (N=146)[a].

Characteristics	All, N	Occupational therapy, n (%)	Social education, n (%)	Nursing, n (%)	*P* value[b]
Age (years)					.009
18-20	17	2 (7.4)	15 (20.3)	0 (0)	
21-25	78	16 (59.3)	40 (54.1)	22 (48.9)	
26-30	32	7 (25.9)	11 (14.9)	14 (31.1)	
31+	19	2 (7.4)	8 (10.8)	9 (20)	
Gender					.75
Male	29	4 (14.8)	16 (21.6)	9 (20)	
Female	117	23 (85.2)	58 (78.4)	36 (80)	
Currently working in health care or social services					.07
Yes	105	17 (63)	50 (67.6)	38 (84.4)	
No	41	10 (37)	24 (32.4)	7 (15.6)	
Prior experience from using VR[c] technology					.16
yes	73	10 (37)	36 (48.6)	27 (60)	
No	73	17 (63)	38 (51.4)	18 (40)	
Attitude toward VR technology in education					.048
Positive	119	21 (77.8)	56 (75.7)	42 (93.3)	
Skeptical or indifferent	27	6 (22.2)	18 (24.3)	3 (6.7)	
Evaluation of the VR simulation learning activity[d]					.002
1-2	5	3 (14.3)	2 (3.7)	0 (0)	
3-4	53	11 (52.4)	32 (59.3)	10 (27.8)	
5	53	7 (33.3)	20 (37)	26 (72.2)	

[a]N=146 unless otherwise stated.

[b]*P*-values are based on Pearson chi-square test.

[c]VR: virtual reality.

[d]n=111, and higher ratings indicate more positive attitudes.

Perceived Usability of the VR Simulation and Related Learning Activities

In the pairwise comparisons between study programs, using the occupational therapy program as the reference group, the nursing students had the highest mean scores, while the occupational therapy students, for the most part, had the lowest mean scores.

Students' Perceptions of Ease of Use

In the pairwise comparisons, statistically significant differences were found on all items related to the perceived ease of use. The highest mean scores for all students were on the items "the pace and the narrative in the 360° video were good" (mean 4.71, SD 0.64) and "the presented situation was realistic" (mean 4.68, SD 0.64). The lowest mean score was found on the item "the audio quality of the 360° video was good" (mean 3.86, SD 1.41), which was also the lowest mean score for both the occupational therapy students (mean 2.81, SD 1.57) as well as the social education students (mean 3.72, SD 1.41) throughout the whole questionnaire.

Students' Perceptions of Usefulness

In the pairwise comparisons, statistically significant differences were found on all items related to the perceived usefulness. The highest mean scores for all students were on the item "the 360° video and the related learning activities were useful as a part of my education program" (mean 4.46, SD 0.95). The lowest mean score was found on the item "what I experienced in the 360° video improved my professional competence" (mean 3.93, SD 1.19). The difference between the study programs was rather large, with the occupational therapy students scoring this item on average 1.2 points lower than the nursing students and 0.8 points lower than the social education students. Table 3 displays the total mean scores and the pairwise comparisons of mean scores between study programs.

XSL•FO
RenderX

Table 3. Usability of VR[a] simulation and learning activities, N=140-146.

Statements	Total, mean (SD)	Occupational therapy, mean (SD)	Social education, mean (SD)	*P*[b] value	Nursing, mean (SD)	*P*[b] value
Ease of use						
The VR[a] headset and the showtime app were easy to use	4.45 (0.97)	3.67 (1.36)	4.57 (0.83)	<.001	4.73 (0.62)	<.001
I received sufficient information and instructions before using the equipment	4.65 (0.81)	4.07 (1.24)	4.77 (0.66)	<.001	4.82 (0.50)	<.001
It was easy to know what to do	4.43 (0.90)	3.89 (1.01)	4.49 (0.89)	.007	4.68 (0.71)	<.001
I did not have any technical problems	4.01 (1.38)	3.07 (1.47)	4.25 (1.31)	<.001	4.18 (1.21)	.002
The presented situation was realistic	4.68 (0.64)	3.89 (0.89)	4.86 (0.42)	<.001	4.87 (0.34)	<.001
The pace and the narrative in the 360° video were good		4.00 (1.00)	4.86 (0.39)	<.001	4.89 (0.32)	<.001
The visual quality of the 360° video was good	4.50 (0.96)	3.85 (1.57)	4.64 (0.72)	<.001	4.67 (0.67)	.001
The audio quality of the 360° video was good	3.86 (1.41)	2.81 (1.57)	3.72 (1.41)	.004	4.71 (0.59)	<.001
I would like to see more scenarios together with related learning activities	4.55 (0.97)	4.19 (1.27)	4.51 (1.00)	.28	4.82 (0.58)	.02
Usefulness						
What I experienced in the 360° video was useful for my future professional practice	4.43 (0.99)	4.00 (1.18)	4.45 (0.97)	.11	4.67 (0.82)	.02
What I experienced in the 360° video was useful in improving my communication skills	3.95 (1.34)	3.50 (1.53)	3.80 (1.29)	.57	4.45 (1.15)	.01
What I experienced in the 360° video improved my professional competence	3.93 (1.19)	3.15 (1.43)	3.94 (1.10)	.006	4.39 (0.92)	<.001
This 360° video was useful as part of my education program	4.38 (1.03)	3.70 (1.38)	4.40 (0.93)	.005	4.75 (0.69)	<.001
The learning activities were useful for my future professional practice	4.37 (1.01)	3.56 (1.34)	4.42 (0.91)	<.001	4.78 (0.60)	<.001
The learning activities were useful for improving my communication skills	4.17 (1.12)	3.62 (1.33)	4.08 (1.13)	.14	4.62 (0.75)	<.001
The learning activities improved my professional competence	4.15 (1.08)	3.13 (1.26)	4.26 (1.00)	<.001	4.53 (0.67)	<.001
The 360° video and the related learning activities were useful as a part of my education program	4.46 (0.95)	3.85 (1.29)	4.51 (0.88)	.004	4.75 (0.62)	<.001

[a]VR: virtual reality.

[b]*P*-values are based on pairwise comparisons between study programs, with 1-way analyses of variance using the Tukey HSD correction. The occupational therapy program is used as reference group.

Open-Ended Questions

Out of 146 participants, 110 answered one or both open questions. There were a total of 135 qualitative responses, which were analyzed. Two categories were created based on the responses, focusing on "perceived learning" and "difficulties and recommendations."

Perceived Learning

The participants described learning outcomes related to soft skills such as conflict management, reflection, emotion regulation, and non-verbal and verbal skills. Furthermore, they also emphasized professional conduct such as empathy, client-centered practice, and a focus on family involvement. The participants found the following debrief and group discussion particularly useful in their learning experience. One participant expressed it as follows:

> *With this type of learning, you are better prepared for clinical practice and the learning curve is steep when we reflect together. It was realistic and touched a lot of emotions. Solutions were discussed together. I think this type of learning is extremely beneficial.*

The fact that this learning took place in a safe environment was also highlighted as positive by several of the participants. Several of the participants answered the open-ended questions with few words in a general, but positive, manner. For example, the learning activity was cool, exciting, fun, realistic, and relevant for clinical practice. One participant expressed it as follows:

XSL•FO
RenderX

> *Being a fly on the wall in such a situation was very enlightening.*

Some participants also answered that this learning activity was a new type of learning and that the website was particularly useful for getting to know the material more thoroughly.

Difficulties and Recommendations

The participants' experience of difficulties related to the learning activity mostly concerned either technical or organizational problems. Particularly poor sound quality and the need for noise-reducing headsets were highlighted as technical issues. On the other hand, organizational problems revolved around too many students in the groups and that the preparations could have been more informative, especially for the students with no experience with VR simulation. Two participants reported discomfort related to a rather unpleasant scenario that they found them "in the middle of," and one of them wrote:

> *…for me, who is a little light sensitive etc., I had great difficulty having the screen so close to my face. The experience was simply quite unpleasant and triggered a headache.*

The participants recommended other types of cases and scenarios involving complex choices and interactive roles in the VR simulation, such as:

> *…options where you can "decide" the actions further.*

It was also desirable that the 360° videos would be more available for self-training and that the students could access more VR simulations at their educational program.

Discussion

Overview

This study aimed to explore student-perceived usability of VR simulation, including 360° videos in VR headsets and related learning activities, in 3 health care education programs in Norway. All in all, high mean scores indicate that the students experienced the VR simulation and the related learning activities as very useful. Nevertheless, our findings showed that the nursing students were most satisfied with the usability of the VR simulation, while the occupational therapy students were least satisfied. The qualitative data from the 2 open-ended questions confirmed the findings from the questionnaire, but also illuminated other factors regarding the usability of VR simulation and the related learning activities.

Differences Between the Study Programs

The nursing students in this study were third-year students with previous experience with other simulations, practical skill training, and clinical practice. They had considerably more experience than the first-year social education students. The second-year occupational therapy students had some prior experience with skill training and clinical practice, but not with simulation. Saab et al [22] emphasized in particular that in order to succeed, the key stakeholders, including students and educators, need to be trained in the use of VR prior to implementing VR simulation. The differences in VR experience among the students prior to this learning session may contribute to explaining why the nursing students scored the usability of the VR simulation significantly higher than students in the other 2 educational programs. They may be due to the novelty effect, where novice VR users tend to focus on managing the hand controllers and exploring the VR environment, rather than focusing on the proposed learning tasks [34].

The key findings in the systematic review of Woon et al [3] illuminate that VR increases nursing students' engagement in the learning experience when allowing them to participate in activities that are close to reality and when implementing an effective training regime consisting of short interval training (≤30 min each) for a number of sessions. In our study, the nursing students had this learning activity as part of a larger session with a multiple number of short interval trainings. The larger number of trainings for the nursing students and their embedding the VR simulation into a more comprehensive learning activity may contribute to explaining their higher scores on usability and relevance. Also, the nursing program had the smallest groups in the VR simulation (up to 9 students), while both the social education program and the occupational therapy program reported difficulties with instruction and initiation of VR simulation in somewhat larger student groups (up to 15 students). The occupational therapy program experienced technical problems, particularly with sound control, which may also be reflected in the low mean scores from the survey. Educators at the social education program experienced difficulties with one of the cohorts, due to students having little to no prior experience using VR technology. Noble et al [35] emphasized that one should not take students' knowledge and skills in using VR equipment for granted. However, their study mirrors our results, with approximately half of the students having no experience with VR even though the education program facilitates the use of VR equipment free of charge.

Group differences related to study progression, previous experience with VR simulation, as well as how the learning activity was structured, both according to group size and how many facilitators were available for the students, can partly explain why students scored the usability differently. A valuable recommendation, consistent with previous research [36], is that smaller student cohorts and having a sufficient number of staff available are important so that the students can receive sufficient help at the right time.

Practicing Clinical Skills With Realistic Scenarios in a Safe Learning Environment

Our findings in both qualitative and quantitative data showed that the students perceived the scenarios to be realistic and useful for clinical practice. Also, several students shared in the open-ended responses that they experienced the VR learning activity as a safe way to practice their clinical skills. The sober and realistic nature of the scenarios depicting real-world situations appears to be beneficial for students' skill acquisitions with relevance for clinical practice [14,37-39]. In addition, the open-ended responses indicate that students were emotionally touched and had an increasing engagement with this learning activity. This is beneficial, as increased emotions and engagement in VR simulation can enhance the learning processes [9]. In the development of the prototype for the 360°

XSL•FO
RenderX

videos, students experienced the VR simulation to be genuine and realistic due to actors using direct eye contact with the camera. Furthermore, the importance of having a safe learning environment was highlighted, especially when experiencing strong emotions [26]. Although the learning environment was described as safe by many students, a few of them pointed out that they found the VR simulation uncomfortable. Cybersickness can influence the learners' attitude toward the technology negatively and has been correlated with poorer learning outcomes [40]. For these students in particular, it will be crucial to facilitate the VR simulation to avoid discomfort. To reduce the discomfort among individuals experiencing it, one suggestion is to adapt the VR simulation to a standard desktop [36]. Nevertheless, we have reason to believe that VR simulation with realistic scenarios is an appropriate learning activity to precede clinical practice. Furthermore, this type of teaching has a particular value as students can practice skills needed in a difficult situation but still feel safe in a peer learning process.

Student Active Learning in Peer Debrief and Group Discussions

Our findings based on both qualitative and quantitative data showed that the students perceived the related learning activities useful for their future professional practice. Many students highlighted the group discussions in the debriefing sessions as particularly beneficial, which is consistent with previous research [7,17,26]. These methods are known to enhance learning by allowing students to actively process information, engage in critical thinking, and articulate their understanding or questions [41,42]. Debrief sessions, in particular, offer a reflective space where students can consolidate their learning, address misconceptions, and gain insights from peers and instructors. This discussion contributes to the larger conversation about active learning in education. Active learning, characterized by student participation and engagement, has been increasingly recognized for its effectiveness in promoting deeper understanding and retention of knowledge [43]. However, some of the students remarked that they were passive observers of the 360 video. They suggested including interactivity in the VR simulation for a better learning experience. This aligns with a Canadian study underscoring the benefits of active participation in VR simulation for occupational therapy students [10]. The learning activity in our study still required active involvement in the debrief discussion where students reflected and engaged with the teaching material together with their peers. A suitable alternative could have been active participation in the VR environment, where students would have the opportunity to interact with the virtual environment, make decisions, and influence the outcome of the scenarios. This could lead to a more engaging learning experience, probably more appropriate for training soft skills such as decision-making, time management, and critical thinking.

Our findings have important implications for educational practices. They suggest that incorporating structured group discussions and debriefing sessions can significantly enhance the VR learning experience by allowing students to engage with and process the learning material related to situations they might encounter in a real-world setting. In addition, VR simulations with the student in an active role will probably increase students' motivation and promote learning.

Study Limitations and Suggestions for Future Research

The study is based on self-report data only. The response rates among both occupational therapy and social education students were high with 71% (n=27) and 78% (n=74). In contrast, the response rate among the nursing students was only 16% (n=45), and, therefore, the findings must be interpreted with caution. Further to this, there may be a response bias, especially among the nursing students, a group possibly consisting of particularly motivated and high-achieving respondents. This may be one important reason why students in this group were more satisfied with the usability of this learning activity. Differences between the student groups may also have been caused by varying levels of VR experience and competence among the faculty facilitating the relevant sessions. While we have the impression that the faculty in the nursing program had more experience with VR compared to their counterparts in the other education programs, we did not collect data from faculty that could support or contradict this view. Thus, we are unable to assess the possible impact of faculty's VR experience and competence on the students' experience, and we suggest that future studies include this information.

Other methodological issues concern that the VR simulation and learning activity was carried out at different year levels among the 3 health care education programs and the presence of age differences between students in the different study programs. Higher age and maturity among the participating nursing students may have resulted in higher levels of satisfaction, while these results may not be representative of differences in the study population. We also note that there might be different "group cultures" related to the evaluation of teaching in different groups of students, essentially meaning that similar experiences are systematically given dissimilar ratings across groups. Thus, there is a possibility that the students' expressed levels of satisfaction differed more than their experienced levels of satisfaction.

Conclusions

In summary, the students in all 3 educational programs perceived the high usability of the VR simulation and the related learning activities. Nevertheless, the nursing students were most satisfied, while the occupational therapy students were somewhat less satisfied. By using realistic scenarios, health care students can be prepared for complex clinical situations in a safe environment. Group discussions in the debriefing sessions are a valuable part of the learning process and enhance active involvement with peers. Particular attention should be given to the organization of the VR simulation with small student groups and a sufficient number of involved educators so that students receive guidance when problems occur. These findings contribute to the overall growing evidence showing that VR simulation has promise and potential as a pedagogical tool in health care education, perhaps particularly for training soft skills such as communication, decision-making, time management, and critical thinking—skills that are unquestionably relevant for clinical practice. However, future studies need to expand

XSL•FO
RenderX

beyond the pilot phase to study if VR simulation exceeds the value of traditional teaching methods.

Acknowledgments

The authors thank the students who volunteered to take part in this study. In addition, we thank our colleagues for conducting the VR simulation and their contribution to the data collection for this study. The larger project was funded by the Norwegian Directorate for Higher Education and Skills and VID Specialized University (project AKTIV-2019/10162).

Data Availability

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

Conflicts of Interest

None declared.

Authors' Contributions

GM, TB, and SSL developed the study design. All authors contributed to collecting the data. GM, TB, and SSL conducted the data analyses, but all the authors interpreted the data in light of the learning activities. GM wrote the first manuscript draft. All authors contributed to the editing of the manuscript. Lastly, all authors read and approved the final manuscript.

Multimedia Appendix 1
Descriptions of scenarios and tasks.
[PDF File (Adobe PDF File), 159 KB - mededu_v10i1e56844_app1.pdf]

References

1. Renganayagalu SK, Mallam SC, Nazir S. Effectiveness of VR head mounted displays in professional training: a systematic review. Tech Know Learn 2021 Jan 01;26(4):999-1041. [doi: 10.1007/s10758-020-09489-9]
2. Radianti J, Majchrzak TA, Fromm J, Wohlgenannt I. A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. Comput Educ 2020 Apr;147:103778. [doi: 10.1016/j.compedu.2019.103778]
3. Woon APN, Mok WQ, Chieng YJS, Zhang HM, Ramos P, Mustadi HB, et al. Effectiveness of virtual reality training in improving knowledge among nursing students: a systematic review, meta-analysis and meta-regression. Nurse Educ Today 2021 Mar;98:104655. [doi: 10.1016/j.nedt.2020.104655] [Medline: 33303246]
4. Haugan S, Kværnø E, Sandaker J, Hustad JL, Thordarson GO. Playful learning with VR-SIMI model: the use of 360-video as a learning tool for nursing students in a psychiatric simulation setting. In: Akselbo I, Aune I, editors. How Can We Use Simulation to Improve Competencies in Nursing?. Switzerland: Springer, Cham; 2023:103-116.
5. Conley CS. SEL in higher education. In: Durlak JAD, Weissberg CE, Gullotta RP, Thomas P, editors. Handbook of Social and Emotional Learning: Research and Practice. New York, USA: Guilford Publications; 2015:197-212.
6. Choi J, Thompson CE, Choi J, Waddill CB, Choi S. Effectiveness of immersive virtual reality in nursing education: systematic review. Nurse Educ 2022;47(3):E57-E61. [doi: 10.1097/NNE.0000000000001117] [Medline: 34657101]
7. Luo H, Yang T, Kwon S, Li G, Zuo M, Choi I. Performing versus observing: investigating the effectiveness of group debriefing in a VR-based safety education program. Comput Educ 2021 Dec;175:104316. [doi: 10.1016/j.compedu.2021.104316]
8. Biggs J, Tang C, Kennedy G. Teaching for Quality Learning at University. 5th ed. Maidenhead: Open University Press; 2022.
9. Allcoat D, von Mühlenen A. Learning in virtual reality: effects on performance, emotion and engagement. Res Learn Technol 2018;26:2140. [doi: 10.25304/rlt.v26.2140]
10. Kim J, Nowrouzi-Kia B, Ho ES, Thomson H, Duncan A. Appraising occupational therapy students' perceptions of virtual reality as a pedagogical innovation. Comput Educ: X Real 2023;3:100039. [doi: 10.1016/j.cexr.2023.100039]
11. Atas G, Topalli D, Cagiltay N. A systematic review of virtual reality and user experience in medicine. 2023 Presented at: 15th International Conference on Education and New Learning Technologies; July 3-5, 2023; Palma, Spain.
12. Nielsen J. Usability 101: introduction to usability. 2012. URL: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ [accessed 2024-07-04]
13. Lewis JR. Usability: lessons learned … and yet to be learned. Int J Hum-Comput Interact 2014;30(9):663-684. [doi: 10.1080/10447318.2014.930311]
14. Beverly E, Rigot B, Love C, Love M. Perspectives of 360-degree cinematic virtual reality: interview study among health care professionals. JMIR Med Educ 2022;8(2):e32657 [FREE Full text] [doi: 10.2196/32657] [Medline: 35486427]

XSL•FO
RenderX

15. Plotzky C, Lindwedel U, Sorber M, Loessl B, König P, Kunze C, et al. Virtual reality simulations in nurse education: a systematic mapping review. Nurse Educ Today 2021;101:104868. [doi: 10.1016/j.nedt.2021.104868] [Medline: 33798987]
16. Bracq MS, Michinov E, Jannin P. Virtual reality simulation in nontechnical skills training for healthcare professionals: a systematic review. Simul Healthc 2019;14(3):188-194 [FREE Full text] [doi: 10.1097/SIH.0000000000000347] [Medline: 30601464]
17. Coyne E, Calleja P, Forster E, Lin F. A review of virtual-simulation for assessing healthcare students' clinical competency. Nurse Educ Today 2021;96:104623. [doi: 10.1016/j.nedt.2020.104623] [Medline: 33125979]
18. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students' education: scoping review. JMIR Med Educ 2022;8(1):e34860 [FREE Full text] [doi: 10.2196/34860] [Medline: 35107421]
19. Blair C, Walsh C, Best P. Immersive 360° videos in health and social care education: a scoping review. BMC Med Educ 2021;21(1):590 [FREE Full text] [doi: 10.1186/s12909-021-03013-y] [Medline: 34819063]
20. Cypress BS, Caboral-Stevens M. "Sense of presence" in immersive virtual reality environment: an evolutionary concept analysis. Dimens Crit Care Nurs 2022;41(5):235-245. [doi: 10.1097/DCC.0000000000000538] [Medline: 35905425]
21. Lie SS, Helle N, Sletteland NV, Vikman MD, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023;9:e41589 [FREE Full text] [doi: 10.2196/41589] [Medline: 36692934]
22. Saab MM, McCarthy M, O'Mahony B, Cooke E, Hegarty J, Murphy D, et al. Virtual reality simulation in nursing and midwifery education: a usability study. Comput Inform Nurs 2023;41(10):815-824 [FREE Full text] [doi: 10.1097/CIN.0000000000001010] [Medline: 36749836]
23. Andreasen EM, Høigaard R, Berg H, Steinsbekk A, Haraldstad K. Usability evaluation of the preoperative ISBAR (identification, situation, background, assessment, and recommendation) desktop virtual reality application: qualitative observational study. JMIR Hum Factors 2022;9(4):e40400 [FREE Full text] [doi: 10.2196/40400] [Medline: 36580357]
24. Mäkinen H, Haavisto E, Havola S, Koivisto JM. Graduating nursing students' user experiences of the immersive virtual reality simulation in learning - a qualitative descriptive study. Nurs Open 2023;10(5):3210-3219 [FREE Full text] [doi: 10.1002/nop2.1571] [Medline: 36598872]
25. Hartstein AJ, Verkuyl M, Zimney K, Yockey J, Berg-Poppe P. Virtual reality instructional design in orthopedic physical therapy education: a mixed-methods usability test. Simul Gaming 2022;53(2):111-134. [doi: 10.1177/10468781211073646]
26. Lie SS, Røykenes K, Sæheim A, Groven KS. Developing a virtual reality educational tool to stimulate emotions for learning: focus group study. JMIR Form Res 2023;7:e41829 [FREE Full text] [doi: 10.2196/41829] [Medline: 36939819]
27. Solstien 3. URL: https://www.solstien3.no/index_en.html [accessed 2024-09-28]
28. National regulations relating to a common curriculum for health and social care education. URL: https://www.regjeringen.no/contentassets/389bf8229a3244f0bc1c7835f842ab60/national-regulations-relating-to-a-common-curriculum-for-health-and-social-care-education.pdf [accessed 2024-09-28]
29. Lie SS, Alvestad RW, Helle N, Vikman MD, Dahl-Michelsen T. Exploring VR simulation in healthcare and social work education: students' experiences with VR simulation as preparation for professional practice. Uniped 2024;47(1):18-31.
30. Helle N, Vikman MD, Dahl-Michelsen T, Lie SS. Health care and social work students' experiences with a virtual reality simulation learning activity: qualitative study. JMIR Med Educ 2023;9:e49372 [FREE Full text] [doi: 10.2196/49372] [Medline: 37728988]
31. Verkuyl M, Romaniuk D, Mastrilli P. Virtual gaming simulation of a mental health assessment: a usability study. Nurse Educ Pract 2018;31:83-87. [doi: 10.1016/j.nepr.2018.05.007] [Medline: 29800764]
32. Lee Y, Kim SK, Eom MR. Usability of mental illness simulation involving scenarios with patients with schizophrenia via immersive virtual reality: a mixed methods study. PLoS One 2020;15(9):e0238437 [FREE Full text] [doi: 10.1371/journal.pone.0238437] [Medline: 32936813]
33. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
34. Miguel-Alonso I, Checa D, Guillen-Sanz H, Bustillo A. Evaluation of the novelty effect in immersive virtual reality learning experiences. Virtual Real 2024;28(1):27. [doi: 10.1007/s10055-023-00926-5]
35. Noble SM, Saville JD, Foster LL. VR as a choice: what drives learners' technology acceptance? Int J Educ Technol High Educ 2022;19(1):6. [doi: 10.1186/s41239-021-00310-w]
36. Saab MM, Hegarty J, Murphy D, Landers M. Incorporating virtual reality in nurse education: a qualitative study of nursing students' perspectives. Nurse Educ Today 2021;105:105045 [FREE Full text] [doi: 10.1016/j.nedt.2021.105045] [Medline: 34245956]
37. Coyne E, Frommolt V, Rands H, Kain V, Mitchell M. Simulation videos presented in a blended learning platform to improve Australian nursing students' knowledge of family assessment. Nurse Educ Today 2018;66:96-102 [FREE Full text] [doi: 10.1016/j.nedt.2018.04.012] [Medline: 29689461]
38. Wang W, Liang Z, Blazeck A, Greene B. Improving Chinese nursing students' communication skills by utilizing video-stimulated recall and role-play case scenarios to introduce them to the SBAR technique. Nurse Educ Today 2015;35(7):881-887. [doi: 10.1016/j.nedt.2015.02.010] [Medline: 25753352]

XSL•FO
RenderX

39. Massey D, Byrne J, Higgins N, Weeks B, Shuker MA, Coyne E, et al. Enhancing OSCE preparedness with video exemplars in undergraduate nursing students. a mixed method study. Nurse Educ Today 2017;54:56-61 [FREE Full text] [doi: 10.1016/j.nedt.2017.02.024] [Medline: 28477564]
40. Polcar J, Horejsi P. Knowledge acquisition and cyber sickness: a comparison of VR devices in virtual tours. MM Sci J 2015;2015(02):613-616. [doi: 10.17973/mmsj.2015_06_201516]
41. Luctkar-Flude M, Tyerman J, Verkuyl M, Goldsworthy S, Harder N, Wilson-Keates B, et al. Effectiveness of debriefing methods for virtual simulation: a systematic review. Clin Simul Nurs 2021;57:18-30. [doi: 10.1016/j.ecns.2021.04.009]
42. Gardner R. Introduction to debriefing. Semin Perinatol 2013;37(3):166-174. [doi: 10.1053/j.semperi.2013.02.008] [Medline: 23721773]
43. Børte K, Nesje K, Lillejord S. Barriers to student active learning in higher education. Teach High Educ 2020;28(3):597-615. [doi: 10.1080/13562517.2020.1839746]

Abbreviations

HSD: honestly significant difference
VR: virtual reality

Edited by B Lesselroth; submitted 28.01.24; peer-reviewed by A Bustillo, A Carrillo, E Rincon; comments to author 20.06.24; revised version received 06.08.24; accepted 24.09.24; published 19.11.24.

Please cite as:
Mørk G, Bonsaksen T, Larsen OS, Kunnikoff HM, Lie SS
Virtual Reality Simulation in Undergraduate Health Care Education Programs: Usability Study
JMIR Med Educ 2024;10:e56844
URL: https://mededu.jmir.org/2024/1/e56844
doi:10.2196/56844
PMID:

©Gry Mørk, Tore Bonsaksen, Ole Sønnik Larsen, Hans Martin Kunnikoff, Silje Stangeland Lie. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.11.2024. 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.

XSL•FO
RenderX

Objective Comparison of the First-Person–View Live Streaming Method Versus Face-to-Face Teaching Method in Improving Wound Suturing Skills for Skin Closure in Surgical Clerkship Students: Randomized Controlled Trial

Freda Halim[1], MD, PhD; Allen Widysanto[2], MD, PhD; Petra Octavian Perdana Wahjoepramono[3,*], MD, PhD; Valeska Siulinda Candrawinata[1,*], MD; Andi Setiawan Budihardja[4,*], DDS, PhD; Andry Irawan[1,*], MD; Taufik Sudirman[1,*], MD; Natalia Christina[1,*], MD; Heru Sutanto Koerniawan[1,*], MD; Jephtah Furano Lumban Tobing[5,*], MD; Veli Sungono[6,*], MS, PhD; Mona Marlina[7,*], MD, MMedEd; Eka Julianta Wahjoepramono[3,*], MD, PhD

1
2
3
4
5
6
7

*these authors contributed equally

Corresponding Author:
Freda Halim, MD, PhD

Abstract

Background: The use of digital online teaching media in improving the surgical skills of medical students is indispensable, yet it is still not widely explored objectively. The first-person–view online teaching method may be more effective as it provides more realism to surgical clerkship students in achieving basic surgical skills.

Objective: This study aims to objectively assess the effectiveness of the first-person–view live streaming (LS) method using a GoPro camera compared to the standard face-to-face (FTF) teaching method in improving simple wound suturing skills in surgical clerkship students.

Methods: A prospective, parallel, nonblinded, single-center, randomized controlled trial was performed. Between January and April 2023, clerkship students of the Department of Surgery, Pelita Harapan University, were randomly selected and recruited into either the LS or FTF teaching method for simple interrupted suturing skills. All the participants were assessed objectively before and 1 week after training, using the direct observational procedural skills (DOPS) method. DOPS results and poststudy questionnaires were analyzed.

Results: A total of 74 students were included in this study, with 37 (50%) participants in each group. Paired analysis of each participant's pre-experiment and postexperiment DOPS scores revealed that the LS method's outcome is comparable to the FTF method's outcome (LS: mean 27.5, SD 20.6 vs FTF: mean 24.4, SD 16.7; *P*=.48) in improving the students' surgical skills.

Conclusions: First-person–view LS training sessions could enhance students' ability to master simple procedural skills such as simple wound suturing and has comparable results to the current FTF teaching method. Teaching a practical skill using the LS method also gives more confidence for the participants to perform the procedure independently. Other advantages of the LS method, such as the ability to study from outside the sterile environment, are also promising. We recommend improvements in the audiovisual quality of the camera and a stable internet connection before performing the LS teaching method.

Trial Registration: ClinicalTrials.gov NCT06221917; https://clinicaltrials.gov/study/NCT06221917

(JMIR Med Educ 2024;10:e52631) doi:10.2196/52631

KEYWORDS

teaching method; live streaming; first-person view; face-to-face; simple wound suturing

XSL•FO
RenderX

Introduction

Using a combination of traditional and online teaching methods in the training of medical students is unavoidable and indispensable in the 21st century, especially in the Education 4.0 framework [1]. Although blended learning methods have been applied in many disciplines, its use in surgical clerkship training has not been thoroughly explored [2,3]. This gap was made obvious during the COVID-19 pandemic, as the training of medical students in various countries was disrupted since digital online tools were not ready to be used in the medical education field [4-6].

Compounding this problem is the discrepancy between the growth rate of new medical students compared to the training rate of certified medical school lecturers [7,8]. The Indonesian Ministry of Education stated that the ideal ratio of lecturers to medical students for effective teaching is 1:5, which is not always achievable [9]. Online teaching methods are also especially useful in the operating theater environment, as the number of personnel in the operating theater must remain as few as possible to decrease the risk of surgical infections [10,11].

A proposed solution for these problems is by teaching procedural skills using live-streamed media with strict quality assurance to ensure the quality of the graduates [12,13]. In this manner, a certified lecturer could educate a number of students simultaneously, while reducing the number of people in the operating theater. While the surgeon is doing the procedure in the operating theater, the students or participants can see and learn the procedural skill in other places simultaneously via the internet [14,15]. Although a previous study by Shikino et al [16] suggested that video training of students are generally better accepted, this may not be applicable in learning a manual dexterity skill such as suturing.

The viewpoint shown in the live stream could also affect the learners' understanding. Typically, live-streamed videos are presented in either first-person or third-person view, where a first-person view simulates the viewer being the person doing the procedure, and a third-person view shows the viewer looking at the surgeon doing the procedure from the side. In the context of surgical skills training, a first-person view could improve the students' skills acquisition, as it provides a more realistic simulation of the procedure performed, especially concerning the hand movements, instrument handling, tissue handling, knot tying, and so on [17-19]. A first-person view could also bring the students' viewpoint closer to the procedure compared to being there in person, as onlookers in the operating theater must maintain their distance due to hygiene and sterility issues [20].

An operator-mounted vlogging camera is also superior compared to fixed operating theater cameras, installed in the light fixtures or dedicated mounts, which require complicated installment, are not readily available in many theaters, and are less cumbersome compared to digital cameras with tripod settings [21-23]. Previous researchers have studied and published procedural learning methods using a minimalist and portable vlogging camera such as a GoPro, which could be easily brought into the operating theater, outpatient clinic, or classrooms [23-25]. This device is easily mountable and wearable, which also means that surgeons can easily wear it on their heads while operating, and a teaching assistant can help operate it with a simple click [26]. Head-mounted cameras are also easier to use and less intrusive to the operator compared to body mounts [23,27].

Previous studies have researched and published procedural learning methods using digital online platforms [6,13,28-32]. However, to our knowledge, there are still no studies that objectively evaluate the effectiveness of first-person–view live streaming (LS) methods in surgical training such as simple wound suturing, which is unique to this study. The aim of this study is to objectively assess whether performing simple wound sutures via LS using a first-person–view GoPro camera has the same effectiveness as traditional face-to-face (FTF) teaching.

Mastery in suturing skills for simple and clean wounds is a requirement for medical doctors. Simple wound suturing has internationally established techniques and assessment methods [33,34]. The most basic wound closure technique is the simple interrupted suture, which is a required skill for Indonesian medical doctors [35-37]. Objective assessment of this procedural skill is performed using the Objective Structured Clinical Examination (OSCE), which is routinely carried out at the Faculty of Medicine, Pelita Harapan University [32]. To improve participants' skills, the direct observational procedural skills (DOPS) method has been incorporated into the curriculum [38].

Methods

Ethical Considerations

This study was reviewed and approved by the Pelita Harapan University Faculty of Medicine Ethical Board (ethical approval 011/K-LKJ/ETIK/I/2023). This study also has been registered at ClinicalTrials.gov (registration NCT06221917). Details about the study were explained to the participants, and informed consent were obtained from all the participants. All the data were already deidentified. No compensation was given to participants.

Recruitment, Randomization, and Allocation

This study was a prospective, parallel, nonblinded, single-center, randomized controlled trial, conducted between January and April 2023. This study was not funded by any sponsor or institution. This study was conducted and reported in accordance with CONSORT (Consolidated Standards of Reporting Trials) guidelines [39] (Checklist 1).

A total of 74 surgical clerkship students of Pelita Harapan University were recruited as study participants based on a sample calculation from Lemeshow et al [40], from a previous study by Sakurai et al [41]. They were selected from a pool of 254 fifth- and sixth-year active clerkship students using simple computer randomization. They were in the final years of study in the Faculty of Medicine and had just begun their surgical rotation. These students had learned suturing in a clinical skills module during their second year of medical school but had no previous clinical experience of wound suturing in their clinical rotations, such as from a previous obstetrics and gynecology or surgical rotation. Participants who dropped out in the 1-week

XSL•FO
RenderX

period between preintervention and postintervention time points were excluded. It was made clear to the students that their participation in this study would not affect their academic results in any way.

The students were then randomized into 2 groups: of the 74 participants, the first 37 (50%) selected by simple computer randomization were allocated to the FTF group, and the next 37 (50%) were allocated into the LS group. Each recruited participant underwent a pre-experiment simple suturing DOPS assessment with a randomly assigned clinical preceptor from the Department of Surgery. These 8 clinical preceptors are active surgical specialists and subspecialists, with previous experience in DOPS assessment and tutoring medical students. The assessment rubrics used in this study have been reviewed by the Medical Education Unit of Pelita Harapan University and were routinely used in OSCEs (Multimedia Appendix 1).

The FTF group was taught how to perform simple sutures on a mannequin, and they then watched from the side as a surgeon (FH) performed the simple suturing procedure on a real patient. FH is an assistant professor at the Faculty of Medicine and an active surgeon with more than 10 years of practice. The students were allowed to interact with the operator and ask questions.

The operator simultaneously wore a head-mounted GoPro Hero 8 device, which was performing a LS function. Two assistants, HSK and VSC, helped ensure that the audiovisual quality of the demonstration was adequate. When the visual exposure was not adequate, HSK would help by adjusting the camera [42].

The LS group was taken into a different room, and they watched the live stream from the GoPro on their own devices while being monitored by HSK or VSC. All participants were instructed to use a university Wi-Fi network to ensure connectivity. LS participants were encouraged to be actively involved in the teaching process, asking questions or giving feedback directly through a speakerphone when they were not clear regarding the demonstration or explanation.

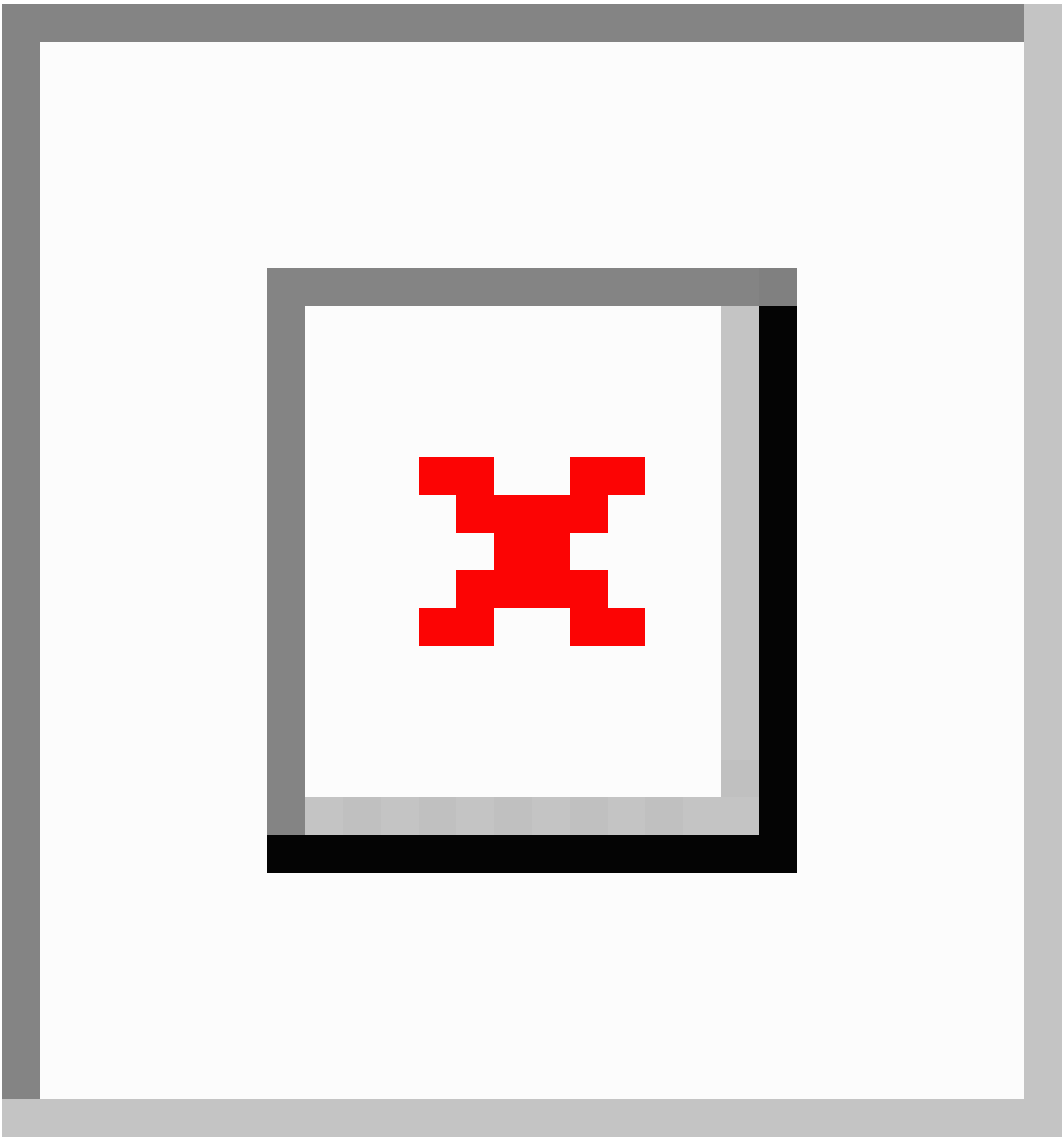
Participants in both groups were allowed to ask the instructor to stop or redo the process. If the audiovisual quality of the live stream was poor, the camera setup was immediately modified, and the instructor would repeat the unclear teaching process to make sure every participant got the same explanation before proceeding to other steps. The live-streamed session was not recorded, and students were not allowed to record it on their device under supervision from HSK or VSC.

One week after the initial training, the participants performed a postexperiment DOPS assessment with the same examiner as the pre-experiment DOPS assessment, using the same rubric to avoid interexaminer bias. Data on the grade point average (GPA) index and frequency of self-training within a 1-week period of both groups were collected.

At the end of the teaching process, we asked both groups using a Likert-scale questionnaire for their opinion regarding the quality of surgical teaching, whether the training enhanced their skill, and the confidence of the participants to do the procedure by themselves. We also asked about the audiovisual quality of the online video as well as the internet connection for the LS group, directly after the training was finished. The participant flow is shown in Figure 1.

XSL•FO
RenderX

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of participant flow. DOPS: direct observational procedural skills; FTF: face-to-face; LS: live streaming.



Statistical Analysis

Data were analyzed using SPSS (version 23.0; IBM Corp). Paired-samples 1-tailed *t* test was used to determine the difference between the preintervention and postintervention DOPS scores. Fisher exact analysis was used to analyze the subjective evaluation of FTF versus LS effectiveness to enhance participants' skills. Descriptive statistics were used to describe the audiovisual quality and internet connection quality.

The difference between DOPS scores (Δ) was defined as the numerical difference between the scores before and after the teaching process. This numerical difference was calculated from each participant's preintervention and postintervention scores (paired analysis). By calculating this Δ, we could objectively review the ability of the LS method compared with the traditional FTF method in enhancing suturing skills in this study.

Results

A total of 74 study participants were included in this study, with 37 (50%) participants each in the FTF and LS groups. The characteristics of the study participants are described in Table 1. The mean GPA index of the FTF and LS groups did not show significant differences (mean 3.26, SD 0.21 vs mean 3.20, SD 0.21; *P*=.20).

XSL•FO
RenderX

Table . Study participant characteristics.

Characteristics		Value (N=74)	*P* value
Sex, n (%)			—[a]
	Male	26 (35)	
	Female	48 (65)	
Age (years), mean (IQR)		22.4 (21-26)	—
GPA[b] index, mean (SD)			.20
	FTF[c]	3.26 (0.21)	
	LS[d]	3.20 (0.21)	
	Overall	3.23 (0.21)	

[a]Not applicable.

[b]GPA: grade point average.

[c]FTF: face-to-face.

[d]LS: live streaming.

Table 2 shows the objective evaluation of FTF versus LS effectiveness to enhance participants' skill. There was a significant increase between the preintervention and postintervention DOPS evaluation scores ($P<.001$), and this difference was more apparent in the FTF group. The LS group spent significantly more time performing self-training than the FTF group ($P=.04$).

Table 3 shows the subjective evaluation of teaching method effectiveness. Most students rated the FTF or LS method as good or very good (FTF: 36/37, 97% and LS: 35/37, 95%). Most students (28/35, 76%) in the FTF group thought that the training improved their skill, while most students (24/37, 65%) in the LS group did not find the training very useful.

Table 4 shows the student assessment of LS method quality. Most students found the first-person–view quality to be good or passable (30/37, 81%). Most students (36/37, 97%) had good or acceptable internet connection, while 1 (3%) student had frequent disconnections.

XSL•FO
RenderX

Table . Objective evaluation of FTF[a] versus LS[b] effectiveness to enhance participants' skill.

Variable		Value, mean (SD)	Value, range	*P* value[c]
Overall DOPS[d] score				*<.001*
	Preintervention	56.7 (19.5)	15-91.7	
	Postintervention	82.7 (13.9)	41.7-100	
Preintervention score				.33
	FTF	58.9 (21.8)	15-91.7	
	LS	54.5 (17.1)	20-91.7	
Postintervention score				*.02*
	FTF	86.4 (11)	58.3-100	
	LS	78.9 (15.5)	41.7-100	
FTF group score				*<.001*
	Preintervention	58.9 (21.8)	15-91.7	
	Postintervention	86.44 (11)	58.33-100	
LS group score				*<.001*
	Preintervention	54.5 (17)	20-91.7	
	Postintervention	78.9 (15.5)	41.67-100	
Difference between preintervention and postintervention scores (Δ)				.48
	FTF	27.5 (20.6)	0-76.6	
	LS	24.4 (16.7)	16.6-63.3	
Total self-training frequency in 1 week				*.048*
	LS	6.3 (3.4)	2-20	
	FTF	4.9 (2.3)	0-12	

[a]FTF: face-to-face.
[b]LS: live streaming.
[c]Mean difference by 1-tailed *t* test.
[d]DOPS: direct observational procedural skills.

XSL•FO
RenderX

Table . Subjective evaluation of FTF[a] versus LS[b] effectiveness to enhance participants' skill.

Variable	FTF method (n=37), n (%)	LS method (n=37), n (%)	*P* value
Teaching quality from instructor			*.02*
Very good	29 (78)	18 (49)	
Good	7 (19)	17 (46)	
Passable	1 (3)	0 (0)	
Poor	0 (0)	2 (5)	
Does the training improve your skill?			*<.001*
Yes, it improves my skill a lot	26 (70)	7 (19)	
Yes, it does	2 (5)	6 (16)	
Not too much	9 (24)	21 (57)	
No, it doesn't improve my skill at all	0 (0)	3 (8)	
Confidence in doing the procedure by themselves			*<.001*
Very confident	0 (0)	2 (5)	
Confident	24 (65)	34 (92)	
Not confident	13 (35)	1 (3)	

[a]FTF: face-to-face.
[b]LS: live streaming.

Table . Subjective evaluation of audiovisual quality and internet connection quality for the live streaming group.

Variable	Value (n=37), n (%)
Audiovisual quality of live streaming	
Very good	5 (14)
Good	17 (47)
Passable	13 (36)
Poor	1 (3)
Internet connection quality	
Good	25 (68)
Passable (some signal disconnections)	11 (29)
Poor (frequent signal disconnections)	1 (3)

Discussion

Principal Findings

This study aims to prove that first-person–view LS teaching has the same effectiveness compared to traditional FTF teaching in enhancing medical students' practical skills in performing simple wound suturing. As of this writing, no other study has compared these methods before.

We considered these 2 groups to have equal basic abilities prior to their training, as their GPA index and preintervention scores were similar. It is good to see that the overall DOPS scores increased significantly between the preintervention and postintervention periods (*P*<.001), suggesting that the training process generally had good results in enhancing participants' skills regardless of their training method.

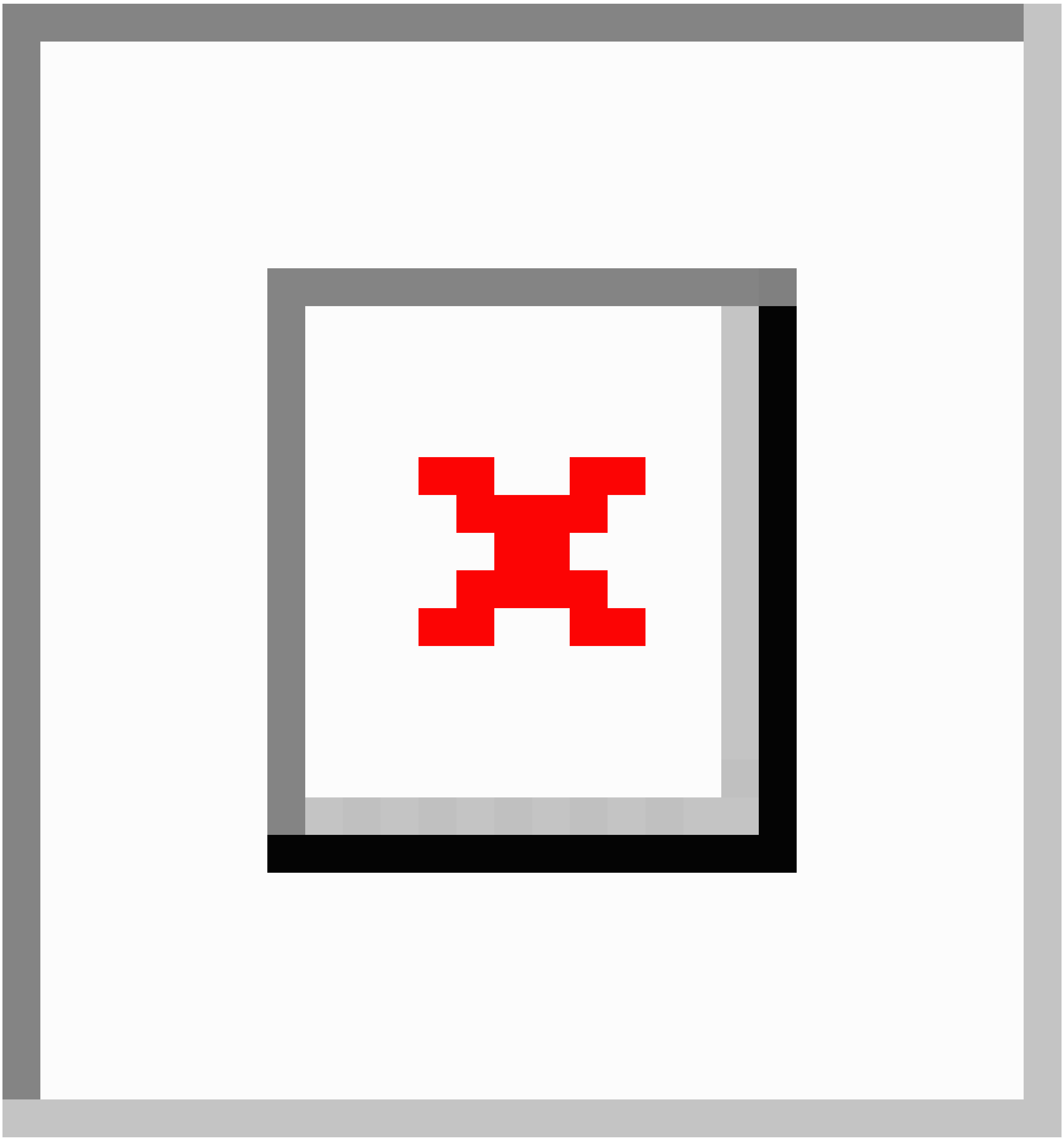
However, the posttest scores of the FTF participants were significantly better than those of the LS participants (FTF: mean 86.4, SD 1 vs LS: mean 78.9, SD 15.5; *P*=.02). As seen on the box-plot graph, the data variation in the LS group is wider than that in the FTF group (Figure 2, pink box plot). This wide range of data suggests significant variability in the results in the LS group, ranging from high to poor values (score).

We compared the ability of the LS method to enhance the participants' skills with the FTF method by performing a paired analysis of the numerical differences between each participant's preintervention and postintervention scores (Δ). Based on this analysis, we found that the score increase between the FTF and LS groups was not significantly different (FTF: mean 27.5, SD 20.6 vs LS: mean 24.4, SD 16.7; *P*=.48). Nevertheless, when we observed the data variation as depicted in box-plot graph (Figure 3), we noted that the data spread of the LS group was

XSL•FO
RenderX

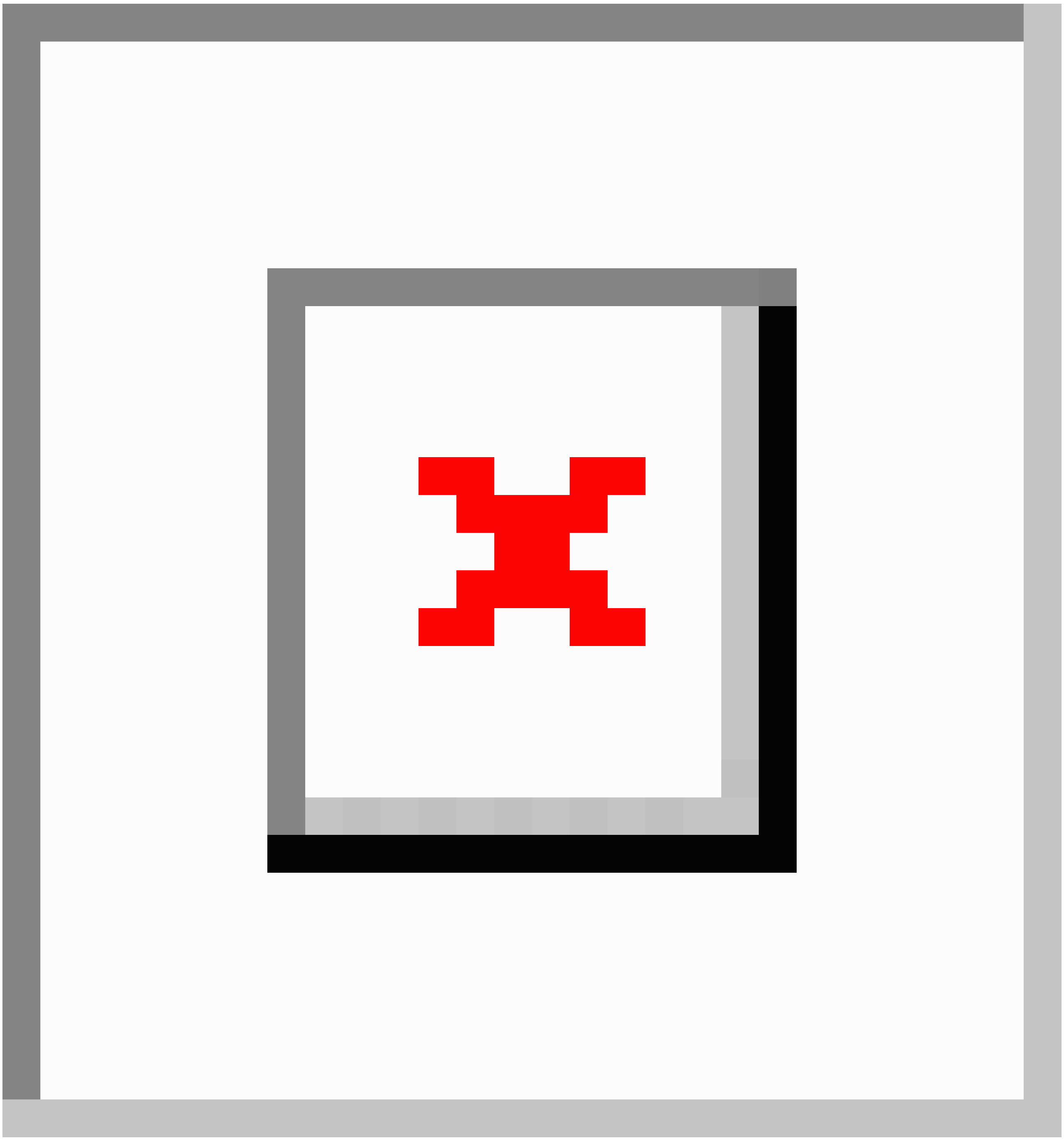
narrower in its numerical differences compared to the FTF group, which suggested more limited ability of the LS method to enhance participants' procedural skills compared to the FTF method. The mean score of the 2 groups were 27.5 (SD 20.6) for the FTF group and 24.4 (SD 16.7) for the LS group, which showed that the FTF group had higher score differences than the LS group. Therefore, we deduced that the LS method was still inferior to the FTF method in enhancing participants' ability to do simple procedural skills.

Figure 2. Box-plot graph of pretest and posttest scores of FTF versus LS group. FTF: face-to-face; LS: live streaming.



XSL•FO
RenderX

Figure 3. Box-plot graph of numerical differences of both groups' scores. FTF: face-to-face; LS: live streaming.



Procedural skills differ from cognitive matters as they need to be mastered and self-trained within some period. We encouraged the participants to train themselves as often as possible in a 1-week period. In an effort to reduce bias, we asked the participants at the end of the 1-week period about their self-training frequency during that period. This analysis showed the LS group had more self-training frequency on average (mean 6.3, SD 3.40 vs mean 4.9, SD 2.3; *P*=.048). It is debatable whether the participants in LS group performed more self-training because they felt compelled to by the LS demonstration as mentioned by Offiah et al [23] or because of something else. It is interesting to see that even though LS participants performed more self-training than FTF participants, they did not acquire the same increase in posttest DOPS scores.

The quality of the instructions given during the FTF and LS methods was also evaluated. Participants were instructed to give feedback regarding the teaching quality, asking if the instructor gave a good, clear demonstration and explanation on the technique. We found that the majority of the FTF group thought that the teaching quality was "very good" (29/37, 78%), but the LS group was dispersed in "very good" (18/37, 49%) and "good" (17/37, 46%) responses. This result may be caused by the FTF group being physically present at the room with the instructor and, therefore, feeling more at ease to ask questions in a natural manner. Although we encouraged the participants in the LS group to actively participate in training sessions, the LS group may have had questions or comments as well but did not express them simply because they felt less engaged in the LS system.

XSL•FO
RenderX

The lack of social interaction, collaborative learning, and teacher-student engagement issues are known to be barriers to online learning [43]. More specifically, the poor engagement between students and instructor in LS settings was also reported in the study of Mill et al [15]. Connectivity problems may also be an issue, as 1 participant in the LS group rated their connectivity as "poor."

Students were also subjectively asked if their method of training improved their mastery of the skill. In the LS group, most participants (21/37, 57%) said the method did not improve their skills much, while some (3/37, 8%) said it did not improve their skills at all. This contrasted sharply with the perception of the FTF group, where most participants (26/37, 70%) said the method improved their skills a lot. These results are different from the meta-analysis performed by Mao et al [44], which found that skills proficiency improvement was not significantly different between video and conventional methods. Unfortunately, we did not specifically ask which part of the teaching method that the participants were unsatisfied with.

For the LS group, we also inquired about the audiovisual quality of the LS method. Most participants answered with "good" (17/37, 47%) and "passable" (13/37, 36%), reflecting that the quality of the teaching material needed to be enhanced. In the LS method, the participants could not move their viewpoint, head, or body position to get a better picture of what is going on compared to being present in the FTF group. The GoPro itself needed to be adjusted several times during the training due to limited visual ability, causing the participants in the LS group to not see the demonstration clearly. We also thought that the visual exposure in the LS method was still lacking, even when we used the GoPro Hero 8, which came with a 4000-pixel resolution [42,45]. This experience was also noted in LS of neurosurgery cases by Jack et al [46] using the GoPro Hero 5. The LS group also mentioned of an audio delay during the live demonstration, which could be why participants' opinions of the quality of teaching and the training ability to improve procedural skills were varied. This audio delay is a common problem with the LS method and should be minimized in the future to enhance the effectiveness of LS in teaching procedural skills [47]. Future studies may also considered virtual reality for teaching technical skills, as it is a more immersive experience for the students [48]. Perhaps it is the quality of the teaching materials that needs to be improved to enhance the first-person–view LS method results.

Finally, we asked the participants about their confidence in performing simple wound suturing by themselves after the training. Interestingly, although the majority of both groups are confident, participants of the FTF group were less confident in performing the procedure compared with the LS group (13/37, 35% vs 1/37, 3%). We previously thought that participants of the gold standard FTF teaching method would be more confident in performing the procedure, as this method gives the participant direct visualization of the procedure and better proximity to the instructor to ask questions and, therefore, would impart more confidence to perform the procedure independently. This finding may be an effect of the first-person–view LS method, since this method puts the viewers directly in the instructor's field of view, as if they are doing the procedure themselves. This way, the participants felt as if they have done the procedure before and are more confident in performing it independently [19,49]. Another reason may be that the LS group could learn in a more relaxed setting, as they did not have the stress and tension of trying to learn a skill from inside the high-stress environment of an operating theater and, therefore, could enhance their confidence and willingness to practice [50,51].

Limitations

Some methods in this study could be improved. Several confounding factors could not be controlled, such as the exposure of individual students to the practice of suturing when asked to assist their preceptors in surgery during their rotation, or the enthusiasm of some students to perform self-training. As such, we limited the duration between preintervention and postntervention testing to 1 week, to reduce the effects of these factors. The retention of skills over a longer period was not explored here. We were also unable to limit contact and communication between participants from both groups during the 1-week period.

We also noted that 33% (12/37) of the LS participants had a "passable" or "poor" connection when using their own mobile devices, even though the participants were encouraged to use the university internet connection. Connectivity problems need to be more stringently monitored in the future, with all students being required to connect to university Wi-Fi.

We recommend future studies to use higher-quality recording devices to improve the quality of the teaching materials. Each participant has a different learning curve, and therefore, providing a standardized recording of the procedural skill for students would be helpful in giving them a chance to review and gain confidence before they do it independently. Using a prerecorded video to standardize the teaching material could be used, as suggested by Tackett et al [52], although using recorded media will remove the interactive quality of the live-streamed, first-person–view method. The effects of the teaching method on confidence could also be explored, to see if the first-person–view method could independently increase the participants' confidence.

Conclusions

Using first-person–view LS teaching of simple procedural skills such as simple wound suturing could provide many benefits for the educator, students, and teaching hospital. This method is comparable to standard FTF teaching for improving the students' skill in performing manual tasks. Teaching a practical skill using the LS method also gives more confidence for the participants to perform the procedure independently. Further improvement to the quality of the recording device, better internet connection, and better teaching materials could improve this method in the future.

XSL•FO
RenderX

Acknowledgments

The authors would like to express our gratitude to Professor Cucunawangsih, PhD, for her help regarding ethical review for this publication; Rhendy Wijayanto, MD; Flora Agustina Situmorang, MD; and Hendry Lie, MD, for their help in conducting the research; Ian Huang, MD, for his help in writing the manuscript; and all surgery clerkship students who were willing to contribute to this research as participants.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Assessment rubrics.
[PDF File, 193 KB - mededu_v10i1e52631_app1.pdf]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 988 KB - mededu_v10i1e52631_app2.pdf]

References

1. Chituc CM. A framework for Education 4.0 in digital education ecosystems. In: Camarinha-Matos LM, Boucher X, Afsarmanesh H, editors. Smart and Sustainable Collaborative Networks 4.0. PRO-VE 2021. IFIP Advances in Information and Communication Technology, vol 629: Springer; 2021:702-709. [doi: 10.1007/978-3-030-85969-5_66]
2. Tuma F, Malgor RD, Nassar AK. Actions to enhance interactive learning in surgery. Ann Med Surg (Lond) 2021 Mar 26;64:102256. [doi: 10.1016/j.amsu.2021.102256] [Medline: 33889404]
3. Foo CC, Cheung B, Chu KM. A comparative study regarding distance learning and the conventional face-to-face approach conducted problem-based learning tutorial during the COVID-19 pandemic. BMC Med Educ 2021 Mar 3;21(1):141. [doi: 10.1186/s12909-021-02575-1] [Medline: 33658015]
4. Doulias T, Gallo G, Rubio-Perez I, Breukink SO, Hahnloser D. Doing more with less: surgical training in the COVID-19 era. J Invest Surg 2022 Jan;35(1):171-179. [doi: 10.1080/08941939.2020.1824250] [Medline: 32959688]
5. McKechnie T, Levin M, Zhou K, Freedman B, Palter VN, Grantcharov TP. Virtual surgical training during COVID-19: operating room simulation platforms accessible from home. Ann Surg 2020 Aug;272(2):e153-e154. [doi: 10.1097/SLA.0000000000003999] [Medline: 32675522]
6. Navia A, Parada L, Urbina G, Vidal C, Morovic CG. Optimizing intraoral surgery video recording for residents' training during the COVID-19 pandemic: comparison of 3 point of views using a GoPro. J Plast Reconstr Aesthet Surg 2021 May;74(5):1101-1160. [doi: 10.1016/j.bjps.2020.10.068] [Medline: 33199220]
7. Mustika R, Nishigori H, Ronokusumo S, Scherpbier A. The odyssey of medical education in Indonesia. The Asia-Pacific Scholar 2019 Jan 2;4(1):4-8. [doi: 10.29060/TAPS.2019-4-1/GP1077]
8. Song JSA, McGuire C, Vaculik M, Morzycki A, Plourde M. Cross sectional analysis of student-led surgical societies in fostering medical student interest in Canada. BMC Med Educ 2019 Mar 8;19(1):77. [doi: 10.1186/s12909-019-1502-5] [Medline: 30849966]
9. Makarim N, Sadikin BG. Surat keputusan bersama (SKB) nomor 02/KB/2022 tentang peningkatan kuota penerimaan mahasiswa program sarjana kedokteran, program dokter spesialis dan penambahan program studi dokter spesialis melalui sistem kesehatan akademik [Article in Indonesian]. Scribd. 2022 Jul 12. URL: https://www.scribd.com/document/582964419/SKB-Kemendikbud-dan-Kemkes-tentang-Peningkatan-Kuota [accessed 2024-08-20]
10. Hagopian TM, Vitiello GA, Hart AM, Perez SD, Pettitt BJ, Sweeney JF. Do medical students in the operating room affect patient care? an analysis of one institution's experience over the past five years. J Surg Educ 2014;71(6):817-824. [doi: 10.1016/j.jsurg.2014.04.011] [Medline: 24931415]
11. Wathen C, Kshettry VR, Krishnaney A, et al. The association between operating room personnel and turnover with surgical site infection in more than 12 000 neurosurgical cases. Neurosurgery 2016 Dec;79(6):889-894. [doi: 10.1227/NEU.0000000000001357] [Medline: 27465846]
12. Awad M, Chowdhary M, Hermena S, Falaha SE, Slim N, Francis NK. Safety and effectiveness of live broadcast of surgical procedures: systematic review. Surg Endosc 2022 Aug;36(8):5571-5594. [doi: 10.1007/s00464-022-09072-6] [Medline: 35604484]
13. van Bonn SM, Grajek JS, Schneider A, Oberhoffner T, Mlynski R, Weiss NM. Interactive live-stream surgery contributes to surgical education in the context of contact restrictions. Eur Arch Otorhinolaryngol 2022 Jun;279(6):2865-2871. [doi: 10.1007/s00405-021-06994-0] [Medline: 34424381]
14. Grafton-Clarke C, Uraiby H, Abraham S, Kirtley J, Xu G, McCarthy M. Live streaming to sustain clinical learning. Clin Teach 2022 Aug;19(4):282-288. [doi: 10.1111/tct.13488] [Medline: 35365976]

XSL•FO
RenderX

15. Mill T, Parikh S, Allen A, et al. Live streaming ward rounds using wearable technology to teach medical students: a pilot study. BMJ Simul Technol Enhanc Learn 2021 May 25;7(6):494-500. [doi: 10.1136/bmjstel-2021-000864] [Medline: 35520979]
16. Shikino K, Nishizaki Y, Fukui S, et al. Development of a clinical simulation video to evaluate multiple domains of clinical competence: cross-sectional study. JMIR Med Educ 2024 Feb 29;10:e54401. [doi: 10.2196/54401] [Medline: 38421691]
17. Schmidt MW, Friedrich M, Kowalewski KF, et al. Learning from the surgeon's real perspective - first-person view versus laparoscopic view in e-learning for training of surgical skills? study protocol for a randomized controlled trial. Int J Surg Protoc 2017 Jan 23;3:7-13. [doi: 10.1016/j.isjp.2017.01.001] [Medline: 31851752]
18. Lin C, Andersen D, Popescu V, et al. A first-person mentee second-person mentor AR interface for surgical telementoring. Presented at: 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct); Oct 16-20, 2018; Munich, Germany p. 3-8. [doi: 10.1109/ISMAR-Adjunct.2018.00021]
19. Jaeger D. The impact of the use of video recording eyewear on skill acquisition: a comparison of first-person and third-person perspective video modeling [Dissertation]. : Florida Gulf Coast University; 2016 URL: https://scholarscommons.fgcu.edu/esploro/outputs/99383341626906570 [accessed 2024-08-16]
20. Croghan SM, Phillips C, Howson W. The operating theatre as a classroom: a literature review of medical student learning in the theatre environment. Int J Med Educ 2019 Apr 23;10:75-87. [doi: 10.5116/ijme.5ca7.afd1] [Medline: 31012867]
21. Madani A, Hirpara D, Chadi SA, Dhar P, Okrainec A. Leveraging videoconferencing technology to augment surgical training during a pandemic. Ann Surg Open 2021 Apr 15;2(2):e035. [doi: 10.1097/AS9.0000000000000035] [Medline: 36590033]
22. Smith CD, Skandalakis JE. Remote presence proctoring by using a wireless remote-control videoconferencing system. Surg Innov 2005 Jun;12(2):139-143. [doi: 10.1177/155335060501200212] [Medline: 16034503]
23. Offiah G, Ekpotu LP, Murphy S, et al. Evaluation of medical student retention of clinical skills following simulation training. BMC Med Educ 2019 Jul 16;19(1):263. [doi: 10.1186/s12909-019-1663-2] [Medline: 31311546]
24. Alameddine MB, Englesbe MJ, Waits SA. A video-based coaching intervention to improve surgical skill in fourth-year medical students. J Surg Educ 2018 Nov;75(6):1475-1479. [doi: 10.1016/j.jsurg.2018.04.003] [Medline: 29699931]
25. Hummel RL. Teaching with a gopro camera! simultaneously incorporate technology and learning while creating flipped classroom content. 2015 Presented at: Society for Information Technology & Teacher Education International Conference.
26. Baatjes KJ, Keiller AV, Louw AJ, van Rooyen M. Point‐of‐view technology to teach surgery. Clin Teach 2021 Apr;18(2):147-151. [doi: 10.1111/tct.13272] [Medline: 33090688]
27. Kapi E. Surgeon-manipulated live surgery video recording apparatuses: personal experience and review of literature. Aesthetic Plast Surg 2017 Jun;41(3):738-746. [doi: 10.1007/s00266-017-0826-y] [Medline: 28280896]
28. Faiz T, Marar O, Kamel MK, Vance S. Teaching operative surgery to medical students using live streaming during COVID-19 pandemic. Surg Innov 2021 Apr;28(2):253-254. [doi: 10.1177/1553350620967242] [Medline: 33040715]
29. Ganry L, Sigaux N, Ettinger KS, Salman SO, Fernandes RP. Modified GoPro Hero 6 and 7 for intraoperative surgical recording-transformation into a surgeon-perspective professional quality recording system. J Oral Maxillofac Surg 2019 Aug;77(8):1703.e1-1703.e6. [doi: 10.1016/j.joms.2019.03.026] [Medline: 31009633]
30. Koh W, Khoo D, Pan LTT, et al. Use of GoPro point-of-view camera in intubation simulation-a randomized controlled trial. PLoS One 2020 Dec 1;15(12):e0243217. [doi: 10.1371/journal.pone.0243217] [Medline: 33259536]
31. Moore MD, Abelson JS, O'Mahoney P, Bagautdinov I, Yeo H, Watkins AC. Using GoPro to give video-assisted operative feedback for surgery residents: a feasibility and utility assessment. J Surg Educ 2018;75(2):497-502. [doi: 10.1016/j.jsurg.2017.07.024] [Medline: 28838833]
32. Zulharman Z. Perancangan Objective Structured Clinical Examination (OSCE) untuk menilai kompetensi klinik [Article in Indonesian]. Jurnal Ilmu Kedokteran 2011 Mar;5(1):7-12. [doi: 10.26891/JIK.v5i1.2011.7-12]
33. Temple CLF, Ross DC. A new, validated instrument to evaluate competency in microsurgery: the University of Western Ontario Microsurgical Skills Acquisition/Assessment instrument [outcomes article]. Plast Reconstr Surg 2011 Jan;127(1):215-222. [doi: 10.1097/PRS.0b013e3181f95adb] [Medline: 21200214]
34. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997 Feb;84(2):273-278. [doi: 10.1046/j.1365-2168.1997.02502.x] [Medline: 9052454]
35. Konsil Kedokteran Indonesia. Standar Kompetensi Dokter: Konsil Kedokteran Indonesia; 2006. URL: https://simpus.mkri.id/opac/detail-opac?id=7442 [accessed 2024-08-16]
36. Moy RL, Waldman B, Hein DW. A review of sutures and suturing techniques. J Dermatol Surg Oncol 1992 Sep;18(9):785-795. [doi: 10.1111/j.1524-4725.1992.tb03036.x] [Medline: 1512311]
37. Kudur MH, Pai SB, Sripathi H, Prabhu S. Sutures and suturing techniques in skin closure. Indian J Dermatol Venereol Leprol 2009;75(4):425-434. [doi: 10.4103/0378-6323.53155] [Medline: 19584482]
38. Lörwald AC, Lahner FM, Nouns ZM, et al. The educational impact of mini-clinical evaluation exercise (Mini-CEX) and direct observation of procedural skills (DOPS) and its association with implementation: a systematic review and meta-analysis. PLoS One 2018 Jun 4;13(6):e0198009. [doi: 10.1371/journal.pone.0198009] [Medline: 29864130]
39. Schulz KF, Moher D, Altman DG. CONSORT. In: Moher D, Altman DG, Schulz KF, Simera I, Wager E, editors. Guidelines for Reporting Health Research: A User's Manual: John Wiley & Sons, Ltd; 2014:80-92. [doi: 10.1002/9781118715598.ch9]

XSL•FO
RenderX

40. Lemeshow S, Hosmer DJJ, Klar J, Lwanga SK. Adequacy of Sample Size in Health Studies: J. Wiley for the World Health Organization; 1990. URL: https://digitallibrary.un.org/record/49698?ln=en [accessed 2024-08-16]
41. Sakurai H, Kanada Y, Sugiura Y, et al. Standardization of clinical skill evaluation in physical/occupational therapist education -effects of introduction of an education system using OSCE-. J Phys Ther Sci 2013 Sep;25(9):1071-1077. [doi: 10.1589/jpts.25.1071] [Medline: 24259918]
42. GoPro Hero 8 Black User Manual 130-28314-000 REVB.: GoPro, Inc; 2019. URL: https://gopro.com/content/dam/help/hero8-black/manuals/HERO8Black_UM_ENG_REVB.pdf?srsltid=AfmBOoqq9-3dcNl_sr0i_7e1dqXCwtexCUD3fK5LH40YrIJPKpcmre7J [accessed 2024-08-30]
43. Zaki MS. Advantages and disadvantages of online learning. J Int Soc Res 2022 Sep;15(92) [FREE Full text]
44. Mao BP, Teichroeb ML, Lee T, Wong G, Pang T, Pleass H. Is online video-based education an effective method to teach basic surgical skills to students and surgical trainees? a systematic review and meta-analysis. J Surg Educ 2022;79(6):1536-1545. [doi: 10.1016/j.jsurg.2022.07.016] [Medline: 35933308]
45. Hyndman B, Papatraianou LH. The technological integration of a simulation pedagogical approachfor physical education: the GoPro PE Trial 1.0. Learn Comm Int J Learn Soc Cont 2017;21(1):6-18. [doi: 10.18793/LCJ2017.21.02]
46. Jack MM, Gattozzi DA, Camarata PJ, Shah KJ. Live-streaming surgery for medical student education - educational solutions in neurosurgery during the COVID-19 pandemic. J Surg Educ 2021;78(1):99-103. [doi: 10.1016/j.jsurg.2020.07.005] [Medline: 32747320]
47. Celebi KC, Bailey SKT, Burns MW, Bansal K. Is virtual reality streaming ready for remote medical education? measuring latency of stereoscopic VR for telementoring. Proc Hum Factors Ergon Soc Annu Meet 2021 Nov 12;65(1):757-761. [doi: 10.1177/1071181321651332]
48. Queisner M, Pogorzhelskiy M, Remde C, Pratschke J, Sauer IM. VolumetricOR: a new approach to simulate surgical interventions in virtual reality for training and education. Surg Innov 2022 Jun;29(3):406-415. [doi: 10.1177/15533506211054240] [Medline: 35137646]
49. Hatt D, Zimmerman E, Chang E, Vane J, Hollenbach KA, Shah A. First-person point-of-view instructional video on lumbar puncture procedure. Pediatr Emerg Care 2023 Dec 1;39(12):953-956. [doi: 10.1097/PEC.0000000000003084] [Medline: 38019714]
50. Miandoab NY, Behmaneshpour F, Arbabisarjou A. Stressors of clinical education in operating room students. Dr Inven Today 2019 Nov;12(11):2795-2799 [FREE Full text]
51. Norouzi N, Imani B. Clinical education stressors in operating room students: a qualitative study. Invest Educ Enferm 2021 Feb;39(1):e08. [doi: 10.17533/udea.iee.v39n1e08] [Medline: 33687812]
52. Tackett S, Green D, Dyal M, et al. Use of commercially produced medical education videos in a cardiovascular curriculum: multiple cohort study. JMIR Med Educ 2021 Oct 7;7(4):e27441. [doi: 10.2196/27441] [Medline: 34617911]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
DOPS: direct observational procedural skills
FTF: face-to-face
GPA: grade point average
LS: live streaming
OSCE: Objective Structured Clinical Examination

Edited by B Lesselroth; submitted 10.09.23; peer-reviewed by P Callas, SP Dewi; revised version received 02.06.24; accepted 02.08.24; published 30.08.24.

Please cite as:
Halim F, Widysanto A, Wahjoepramono POP, Candrawinata VS, Budihardja AS, Irawan A, Sudirman T, Christina N, Koerniawan HS, Tobing JFL, Sungono V, Marlina M, Wahjoepramono EJ
Objective Comparison of the First-Person–View Live Streaming Method Versus Face-to-Face Teaching Method in Improving Wound Suturing Skills for Skin Closure in Surgical Clerkship Students: Randomized Controlled Trial
JMIR Med Educ 2024;10:e52631
URL: https://mededu.jmir.org/2024/1/e52631
doi:10.2196/52631

© Freda Halim, Allen Widysanto, Petra Octavian Perdana Wahjoepramono, Valeska Siulinda Candrawinata, Andi Setiawan Budihardja, Andry Irawan, Taufik Sudirman, Natalia Christina, Heru Sutanto Koerniawan, Jephtah Furano Lumban Tobing, Veli Sungono, Mona Marlina, Eka Julianta Wahjoepramono. Originally published in JMIR Medical Education (https://mededu.jmir.org),

XSL•FO RenderX

30.8.2024. 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.

XSL•FO
RenderX

Original Paper

Effectiveness of Blended Versus Traditional Refresher Training for Cardiopulmonary Resuscitation: Prospective Observational Study

Cheng-Yu Chien[1,2,3,4,5], MD, PhD; Shang-Li Tsai[1,2], MD; Chien-Hsiung Huang[1,6,7], MD; Ming-Fang Wang[1], MD; Chi-Chun Lin[1,3], MD; Chen-Bin Chen[1,8], MD; Li-Heng Tsai[1], MD; Hsiao-Jung Tseng[1], MS; Yan-Bo Huang[1], MD; Chip-Jin Ng[1,2,9], MD

[1]Department of Emergency Medicine, Chang Gung Memorial Hospital Linkou Branch, Taoyuan, Taiwan
[2]Department of Emergency Medicine, Chang Gung Memorial Hospital Taipei Branch, Taipei, Taiwan
[3]Department of Emergency Medicine, Ton-Yen General Hospital, Zhubei, Taiwan
[4]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan
[5]Department of Senior Service Industry Management, Minghsin University of Science and Technology, Hsinchu, Taiwan
[6]Graduate Institute of Management, College of Management, Chang Gung University, Taoyuan, Taiwan
[7]Department of Emergency Medicine, New Taipei City Hospital, New Taipei City, Taiwan
[8]Department of Emergency Medicine, New Taipei Municipal TuCheng Hospital and Chang Gung University, New Taipei, Taiwan
[9]Department of Nursing, Chang Gung University of Science and Technology, Taoyuan, Taiwan

Corresponding Author:
Chip-Jin Ng, MD
Department of Emergency Medicine
Chang Gung Memorial Hospital Linkou Branch
No 5 Fushing Street
Gueishan District
Taoyuan, 33305
Taiwan
Phone: 886 3 3281200 ext 2505
Email: ngowl@ms3.hinet.net

Abstract

Background: Generally, cardiopulmonary resuscitation (CPR) skills decline substantially over time. By combining web-based self-regulated learning with hands-on practice, blended training can be a time- and resource-efficient approach enabling individuals to acquire or refresh CPR skills at their convenience. However, few studies have evaluated the effectiveness of blended CPR refresher training compared with that of the traditional method.

Objective: This study investigated and compared the effectiveness of traditional and blended CPR training through 6-month and 12-month refresher sessions with CPR ability indicators.

Methods: This study recruited participants aged ≥18 years from the Automated External Defibrillator Donation Project. The participants were divided into 4 groups based on the format of the CPR training and refresher training received: (1) initial traditional training (a 30-minute instructor-led, hands-on session) and 6-month traditional refresher training (Traditional6 group), (2) initial traditional training and 6-month blended refresher training (an 18-minute e-learning module; Mixed6 group), (3) initial traditional training and 12-month blended refresher training (Mixed12 group), and (4) initial blended training and 6-month blended refresher training (Blended6 group). CPR knowledge and performance were evaluated immediately after initial training. For each group, following initial training but before refresher training, a learning effectiveness assessment was conducted at 12 and 24 months. CPR knowledge was assessed using a written test with 15 multiple-choice questions, and CPR performance was assessed through an examiner-rated skill test and objectively through manikin feedback. A generalized estimating equation model was used to analyze changes in CPR ability indicators.

Results: This study recruited 1163 participants (mean age 41.82, SD 11.6 years; n=725, 62.3% female), with 332 (28.5%), 270 (23.2%), 258 (22.2%), and 303 (26.1%) participants in the Mixed6, Traditional6, Mixed12, and Blended6 groups, respectively. No significant between-group difference was observed in knowledge acquisition after initial training (*P*=.23). All groups met the

XSL•FO
RenderX

criteria for high-quality CPR skills (ie, average compression depth: 5-6 cm; average compression rate: 100-120 beats/min; chest recoil rate: >80%); however, a higher proportion (98/303, 32.3%) of participants receiving blended training initially demonstrated high-quality CPR skills. At 12 and 24 months, CPR skills had declined in all the groups, but the decline was significantly higher in the Mixed12 group, whereas the differences were not significant between the other groups. This finding indicates that frequent retraining can maintain high-quality CPR skills and that blended refresher training is as effective as traditional refresher training.

Conclusions: Our findings indicate that 6-month refresher training sessions for CPR are more effective for maintaining high-quality CPR skills, and that as refreshers, self-learning e-modules are as effective as instructor-led sessions. Although the blended learning approach is cost and resource effective, factors such as participant demographics, training environment, and level of engagement must be considered to maximize the potential of this approach.

Trial Registration: IGOGO NCT05659108; https://www.cgmh-igogo.tw

(JMIR Med Educ 2024;10:e52230) doi:10.2196/52230

KEYWORDS

cardiopulmonary resuscitation; blended method; blended; hybrid; refresher; refreshers; teaching; instruction; observational; training; professional development; continuing education; retraining; traditional method; self-directed learning; resuscitation; CPR; emergency; resuscitation; rescue; life support; cardiac; cardiopulmonary

Introduction

Sudden cardiac arrest is a severe condition, particularly when it occurs outside a medical facility, and the corresponding survival rates are very low. In Europe and North America, these survival rates range from 7% to 13%, whereas in Asia, they are even lower at 0.5% to 8.5% [1-3]. Furthermore, these survival rates vary significantly by location and demography. Some countries exhibit higher survival rates, ranging from 20% to 40%. In contrast, according to a database, the survival rate in Taiwan is 8% to 10% [3-6]. Therefore, survival after out-of-hospital cardiac arrest (OHCA) exhibits substantial variability across regions [7].

The survival status for OHCA is closely linked to the Chain of Survival of the American Heart Association (AHA), which emphasizes the early activation of emergency medical services (EMSs), early cardiopulmonary resuscitation (CPR), and early defibrillation as the first 3 critical links [8]. These 3 interventions can be administered in a prehospital setting, and achieving high-quality outcomes following these interventions is pivotal to enhancing OHCA survival rates. Owing to significant disparities in EMSs, bystander CPR rates, and public access to automated external defibrillators (AEDs) in different regions, OHCA survival rates exhibit corresponding variations [7]. However, through CPR training and dispatcher-assisted CPR, the global bystander CPR rate has improved from approximately 20% in 2001 to 40% to 55% in 2023 [9-11]. In Taiwan, the government has implemented legally mandated continuous public CPR education and training programs aimed at improving the response of bystanders to sudden cardiac arrest [12]. This effort has resulted in significant increases in bystander CPR rates and the use of public AEDs [7,13]. Over a decade, 14% and 3.8% increases have been noted in the bystander CPR rate and the use rate of public AEDs, respectively [6,9,14].

Research has demonstrated a significant decline in CPR skills over time, especially regarding chest compression depth and rate [15]. Consequently, maintaining the public's CPR skills and their motivation for learning CPR is challenging. In response to this challenge, the AHA recommended self-directed training for CPR during the COVID-19 pandemic [16]. Similarly, the European Resuscitation Council recognized blended training models as an alternative to traditional face-to-face teaching models [17,18]. Furthermore, previous studies have indicated that blended training is not inferior to traditional methods and offers advantages such as resource saving and time saving, making it an effective approach for CPR education [15]. By using blended training models, which combine web-based self-guided learning with hands-on practice, individuals can acquire or refresh their CPR skills at their own pace and convenience [15]. Such flexibility fosters increased levels of engagement and enhanced retention of CPR knowledge and thus ultimately enhances the public's preparedness for treating sudden cardiac arrests. Therefore, blended approaches are valuable both during a pandemic and when in-person training cannot be conducted, ensuring widespread CPR education for a broad audience [19].

Limited research has been conducted regarding the effective implementation of relearning stimuli to maintain CPR skills within the framework of blended training. Therefore, the primary objective of this study was to provide relearning stimuli in a blended training setting after using both traditional and blended teaching methods; this study also investigated the effectiveness and most appropriate frequency of blended training. Finally, this study compared learners' performance in 2 educational settings. We hypothesized that using the blended method with 6-month interventions would yield outcomes comparable to those achieved through the traditional method.

Methods

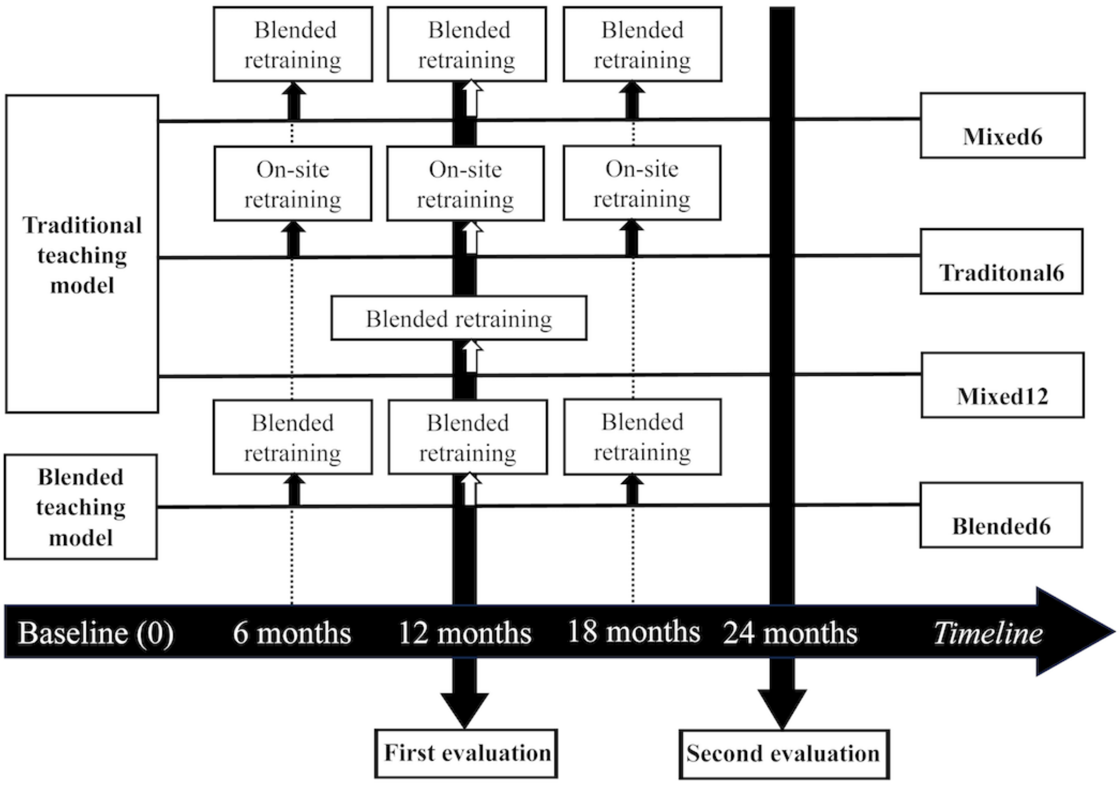
Study Design, Setting, and Participants

This study used a prospective observational design, and participants were recruited from the AED Donation Project, also called the Love GOGO program, implemented by Chang Gung Memorial Hospital, Taiwan. The Love GOGO program aims to establish an educational training system for CPR and build a comprehensive teaching database encompassing participants' attributes, learning models, and CPR parameters. Individuals from government agencies, nonprofit organizations,

XSL•FO
RenderX

schools, and organizations required by current Taiwanese regulations to have AED facilities participate in this education and training program. These include transportation hubs, large long-distance vehicles, tourist spots, schools or large assembly places, large leisure places, large shopping malls, hotels, large public bathhouses, hot springs, and public service sectors such as police stations. These organizations voluntarily participated in the Love GOGO program and proactively contacted the research assistant (YTK) of this study. For this study, participants were enrolled in the Love GOGO program from January to December 2017. Based on our previous study, both traditional and blended teaching models showed a noticeable decline in skill retention after approximately 6 months [12,15]. In this study, mandatory retraining was administered every 6 months or 1 year (Figure 1), spanning a comprehensive training regimen conducted over 2 years. In the initial training phase, the participants were assigned to either traditional teaching or blended teaching modes. Learning effectiveness assessments were conducted every 12 months, with a retraining frequency of 6 or 12 months. Before refresher courses but following initial training, each group underwent evaluation at 12 and 24 months. The results of the 12-month learning effectiveness assessment were disclosed only at 24 months. The research assistants independently allocated training methodologies and the frequencies of subsequent follow-up assessments, using unit convenience and considering the practicalities of the study context. Those responsible for the execution of course training and assessments were not involved in the allocation process.

Figure 1. Schematic diagram illustrating the arrangement of four training courses: Mixed6, Traditional6, Mixed12 and Blended6.



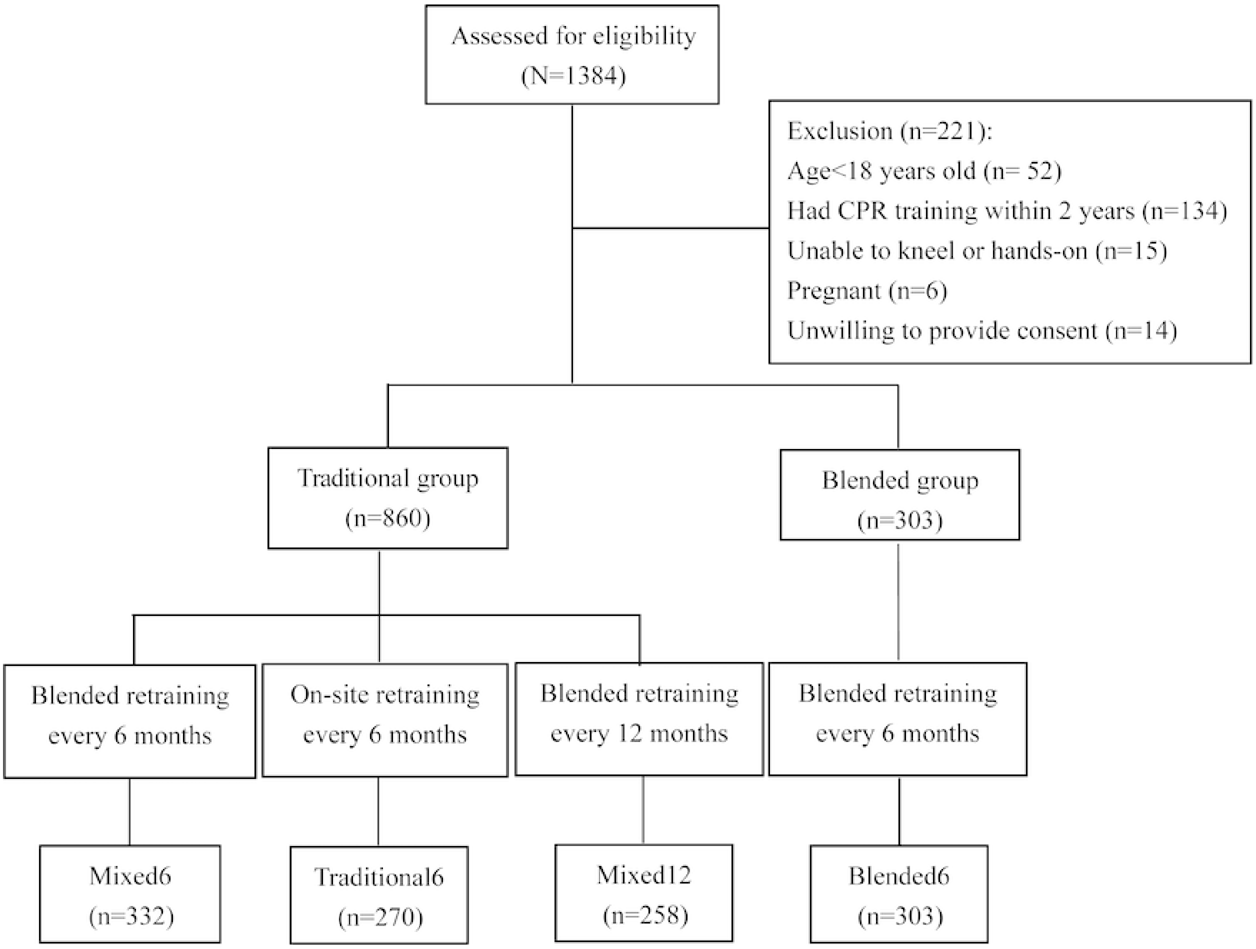
The inclusion criteria are described as follows: (1) aged at least 18 years and (2) not having undergone any CPR training within the preceding 2 years. Individuals who had physical limitations preventing them from kneeling to perform CPR, who were pregnant, or who were unwilling to sign the informed consent form were excluded from this study. Before initial training, the research assistant divided the participants into groups, and their basic characteristics—namely age, sex, educational level, exercise habits, whether they were receiving CPR training for the first time, their most recent CPR training, and their basic life support (BLS) knowledge scores—were collected through a web-based survey. The assessment of CPR learning should encompass the status of both knowledge and skills. After initial training but before refresher training, we collected data regarding BLS knowledge, skill tests, and CPR quality at the scene at 12 and 24 months. The BLS knowledge and skill tests received approval from the Chairman of the Taiwan Society of Emergency Medicine and have also been published in previous studies [12,15] (Multimedia Appendices 1 and 2).

Ethical Considerations

This study was approved by the institutional review board of the Chang Gung Memorial Foundation (approvals: 201600149B0, 201900399B0, 202200559B0, CMRPG1M0081, and CMRPG1N0081), and this study was performed in accordance with relevant guidelines and regulatory requirements. The IGOGO database is anonymized or deidentified, and no type of compensation is provided to participants. Written informed consent was obtained from all the participants (Figure 2).

XSL•FO
RenderX

Figure 2. Flow diagram of participants' inclusion and allocation. CPR: cardiopulmonary resuscitation.



Sample Size

An appropriate sample size for this study was estimated based on a pilot study, in which the expected percentage of correct compression depth was 65.4 (SD 29.5) cm for traditional training. To achieve a statistical power of 90% by using a 2-tailed *t* test with a significance level of *P*<.05, each group was required to have 225 participants. We planned to enroll at least 900 participants in total.

Interventions

The Love GOGO program offers 2 teaching models for CPR training: the traditional instructor-led, classroom-based model and the blended model. In the traditional model, participants undergo a 90-minute session, which includes a 60-minute CPR knowledge education session involving a CPR lecture and demonstration, an AED use demonstration, an introduction to relevant laws, and a 30-minute hands-on practice session focusing on compression-only CPR. The blended program, which was approved by the Chairman of the Taiwan Society of Emergency Medicine in 2016, combines an 18-minute e-learning module with a 30-minute hands-on session for compression-only CPR. The e-learning module comprises a video that covers essential knowledge related to CPR and AEDs, including knowledge related to cardiac arrest scenes, the technique of compression-only CPR, the benefits of using CPR and AEDs for OHCA treatment, CPR and AED use steps, and an introduction to relevant laws. In this study, the participants assigned to the blended program were granted access to the e-learning video 3 days before the hands-on session. After completing the e-learning module, the participants practiced their skills in a 30-minute instructor-led, hands-on session in a classroom setting. Both CPR training programs were conducted by AHA instructors who were also emergency physicians. For hands-on CPR practice, both groups used sensor-equipped manikins (Resusci Anne with QCPR, Laerdal Medical AS). The participant-to-manikin-to-instructor ratio per class was 6:3:1, involving 4 instructors and 6 examiners. The study team provided different certification learning stimuli (traditional and blended learning) at 2 frequencies: every 6 months (at 6, 12, 18, and 24 months) and every 12 months (at 12 and 24 months). To establish groups with unique frequencies, the research assistant (YTK) conducted allocation during the initial training phase. Therefore, the traditional teaching model was applied for initial training, and certification sessions for retraining occurring every 6 or 12 months were conducted using either the blended retraining model (18-minute e-learning module with self–hands-on practice for compression-only CPR) or the on-site retraining model (30-minute instructor-led, hands-on

XSL•FO
RenderX

session). These groups were called Mixed6 (initial traditional training and 6-month blended refresher training), Traditional6 (initial traditional training and 6-month traditional refresher training), and Mixed12 (initial traditional training and 12-month blended refresher training). For the Blended6 group, initial training was conducted using the blended teaching model, and for certification stimuli every 6 months, the blended retraining model was applied (Figure 1).

Outcome Measures

This study systematically assessed the participants' CPR knowledge and performance at multiple time points. Initially, the CPR knowledge and performance of the participants were assessed immediately after training. Following initial training but before refresher training, subsequent evaluations of knowledge and performance were conducted at 12 and 24 months. CPR knowledge was examined through a written test comprising 15 multiple choice questions, with a maximum total score of 100. CPR performance was assessed through 2 methods: examiner-rated assessment and manikin feedback. Individual examiners meticulously assessed the participants' ability to execute the BLS sequence, encompassing tasks from verifying scene safety to using an AED, with a maximum total score of 40. Objective assessment data regarding CPR quality—including compression depth, compression rate, and full chest recoil—were collected from manikin feedback. The assessment adhered to the 2015 AHA guidelines update for CPR and emergency cardiovascular care; high-quality CPR was characterized by the following three criteria: (1) achieving a compression depth of 5-6 cm, (2) maintaining a compression rate of 100-120 beats per minute (bpm), and (3) facilitating complete chest wall recoil of >80%. Notably, because of the focus on compression-only CPR, ventilation was excluded because it was therefore beyond the scope of the assessment in this study. The primary outcome measure was the comparison of high-quality CPR among the 4 groups. Secondary outcome measures were differences in the percentage of full chest recoil, the percentage of compressions delivered with adequate depth (5-6 cm), the percentage of compressions delivered at an adequate rate (100-120 bpm), written test scores, and examiner-rated skill test scores.

Statistical Methods

Descriptive statistics are expressed as mean (SD) for continuous variables and as counts and percentages for categorical variables. Linear regression analysis was conducted to determine any differences in the mean values of baseline characteristics among the groups, with adjustment for control variables—namely age, sex, educational level, exercise habits, whether CPR training was being received for the first time, most recent CPR training, and pretest BLS knowledge scores, which were based on the significance test result and which were proposed in previous research [12,15,20]. After allocation, differences in characteristics among groups were observed. To mitigate potential biases introduced by this allocation method, we applied multiple linear regression analyses and generalized estimating equation (GEE) to adjust for these variations when evaluating outcomes (Multimedia Appendices 3 and 4). The chi-square test was used to assess the differences in proportions among the groups, and the general linear model, such as analysis of covariance, was used to test differences among the groups. The control variables—namely age, sex, educational level, exercise habits, whether CPR training was being received for the first time, and pretest BLS knowledge scores—may have influenced skill retention and test scores. Therefore, the model was adjusted for these variables.

We conducted the assessments of the participant's skill levels and BLS knowledge scores at multiple time points. Accordingly, we used a GEE to examine changes over time in CPR ability indicators among the groups. This allows us to comprehend the changes in CPR skills among trainees under different training methods, using a GEE model to analyze the change over time in CPR ability indicators among groups. The GEE analysis was adjusted for the control variables. To ensure fairness, statistical analysis was conducted using data obtained at time points specific to each group. That is, only data from the postinitial training (baseline), 12-month, and 24-month assessments were included in the analysis.

CPR performance is displayed by line charts, bar charts, and radar charts. In particular, we generated radar charts to illustrate the relative CPR performance in each session. The scores were converted using percent ranking, and the average score was then calculated to represent the performance of each skill for each training method. Statistical analysis was conducted using SPSS Statistics (version 26.0; IBM Corp) and STATA (MP 16.0; Stata Corp LLC).

Results

Baseline Characteristics

A total of 1163 participants were recruited for this study, and they were allocated to 4 training groups. The mean age of the participants was 41.82 (SD 11.6) years, and 62.3% (n=725) of participants were female. In this study, 332 (28.5%), 270 (23.2%), 258 (22.2%), and 303 (26.1%) participants were placed in the Mixed6, Traditional6, Mixed12, and Blended6 groups, respectively. Table 1 displays the baseline characteristics of these 4 training groups. As this study was observational rather than randomized, significant differences were observed among the 4 training groups in terms of age (P<.001), sex (P=.008), educational level (P=.006), and CPR training experience (P<.001; Table 1). Notably, the Traditional6 group had the highest average age (45.30, SD 11.39 years) and consisted of 68.9% (186/270) female participants. Additionally, this group had the highest proportion of individuals receiving CPR training for the first time (92/270, 34.1%). However, no statistically significant difference was observed in the BLS pretest knowledge score (P=.11), with an overall mean score of 67.96 (SD 15.08); this finding indicated similar baseline performance across the groups before BLS training.

XSL•FO
RenderX

Table 1. Baseline characteristics of the 4 training groups.

Variables	Mixed6 (n=332)	Traditional6 (n=270)	Mixed12 (n=258)	Blended6 (n=303)	*P* value
Age (years), mean (SD)	40.78 (9.97)	45.30 (11.39)	40.72 (12.34)	40.78 (12.28)	*<.001*[a]
Sex, n (%)					*.008*
Male	117 (35.2)	84 (31.1)	104 (40.3)	133 (43.9)	
Female	215 (64.8)	186 (68.9)	154 (59.7)	170 (56.1)	
Education, n (%)					*.006*
Below high school	2 (0.6)	26 (9.6)	6 (2.3)	23 (7.6)	
High school, college education, and above	330 (99.4)	244 (90.4)	252 (97.7)	280 (92.4)	
Exercise habits, n (%)	142 (42.8)	116 (45.5)	123 (48.6)	120 (41.5)	.35
First time for CPR[b] training, n (%)	33 (9.9)	92 (34.1)	34 (13.2)	92 (30.4)	*<.00*[a]
Last CPR training, n (%)					*<.001*
Within 2-3 years	122 (36.7)	62 (23)	138 (53.5)	73 (24.1)	
Over 3 years	181 (54.5)	196 (72.6)	109 (42.3)	205 (67.7)	
Not clear	29 (8.8)	12 (4.4)	11 (4.2)	25 (8.2)	
BLS pretest knowledge score[b], mean (SD)	67.78 (13.15)	67.96 (15.08)	70.57 (15.97)	68.17 (16.12)	.11

[a]Italic formatting indicates that there is a statistically significant difference in the *P* value.

[b]CPR: cardiopulmonary resuscitation.

[c]BLS: basic life support.

Posttraining Assessment

According to the results of the objective assessment after the first training session, significant differences were found among the 4 groups in skill tests (*P*=.002), average chest compression depth (*P*<.001), and average compression rate (*P*<.001; Table 2) after adjustment for the control variables in the multivariate analysis (Multimedia Appendix 5). In the multivariate analysis, higher skill test scores were associated with younger age (*P*=.003), higher educational level (*P*<.001), more previous CPR training experience (*P*=.04), and higher BLS pretest scores (*P*=.004). Furthermore, the average compression depth was significantly associated with age (*P*=.02) and sex (*P*<.001), and the average compression rate was significantly associated with educational level (*P*=.04) and CPR training experience (*P*=.02). Although the mean chest compression depths differed among the 4 groups, the proportion of participants achieving the correct chest compression depth did not differ on average (*P*=.11). For the overall performance assessment, the proportion of participants achieving high-quality CPR ranged from 27.4% (91/332) to 32.3% (98/303). The lowest proportion was observed in the Mixed6 group, and the highest proportion was found in the Blended6 group. In the multivariate analysis, high-quality CPR was negatively correlated with the Mixed12 training method (adjusted odds ratio 0.65, 95% CI 0.45-0.93; *P*=.02; Multimedia Appendix 6).

XSL•FO
RenderX

Table 2. Postinitial training evaluation (baseline) for the 4 training groups.

Variables	Mixed6 (n=332)	Traditional6 (n=270)	Mixed12 (n=258)	Blended6 (n=303)	*P* value[a]
BLS[b] knowledge score, mean (SD)	86.05 (11.38)	84.61 (12.96)	86.76 (11.79)	84.10 (11.19)	.23
Skill test, mean (SD)	35.09 (3.26)	35.81 (2.78)	35.73 (3.76)	35.26 (4.05)	*.002*[c]
Average chest compression depth (cm), mean (SD)	5.07 (0.74)	5.01 (0.73)	5.23 (0.43)	5.33 (0.57)	*<.001*
Average chest compression rate (times per minute), mean (SD)	113.88 (13.87)	110.56 (14.34)	116.07 (11.33)	116.65 (10.28)	*<.001*
Correct compression depth, mean (SD)	70.79 (32.83)	71.24 (30.55)	74.75 (32.21)	75.88 (33.31)	.11
Correct compression rate, mean (SD)	61.14 (31.87)	66.16 (30.57)	68.61 (34.15)	61.98 (34.94)	*.01*
Correct recoil, mean (SD)	84.39 (35.29)	87.16 (30.32)	79.72 (37.57)	80.35 (35.65)	.20
High-quality CPR[d,e], n (%)	91 (27.4)	86 (31.8)	77 (29.8)	98 (32.3)	.52

[a]The *P* value was obtained from the general linear regression model adjusted for age, sex, educational level, exercise habits, whether CPR training was being received for the first time, and BLS pretest knowledge score.

[b]BLS: basic life support.

[c]Italic formatting indicates that there is a statistically significant difference in the *P* value.

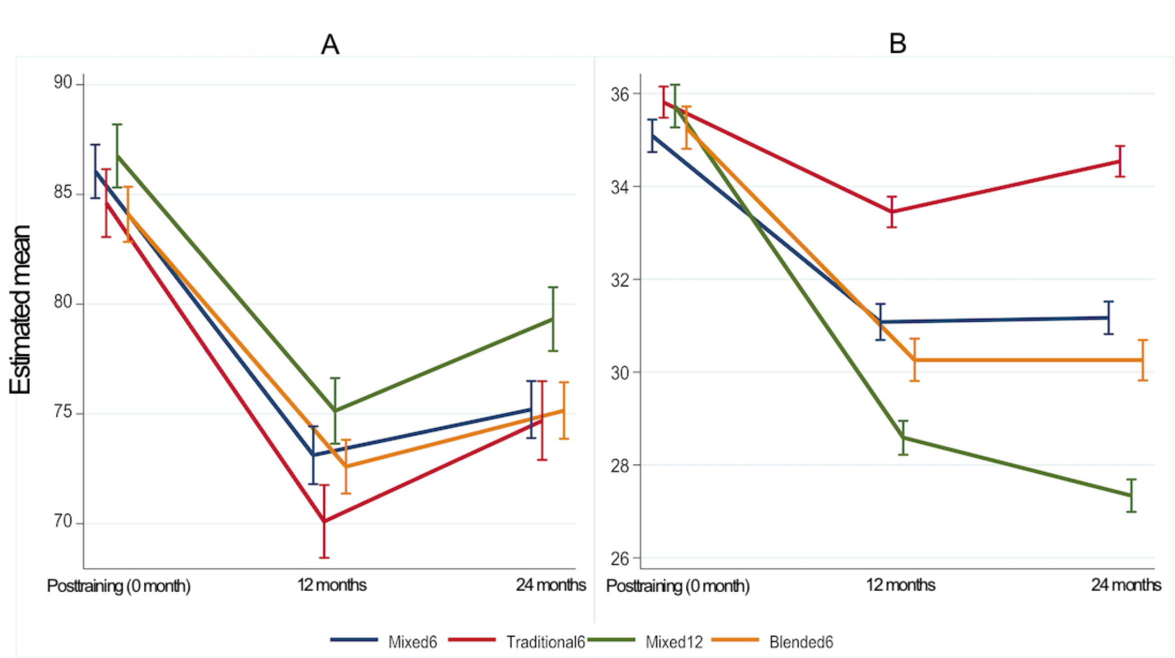
[d]CPR: cardiopulmonary resuscitation.

[e]*P* values obtained from the chi-square test. High-quality CPR was denoted by an average compression depth between 5 and 6 cm, an average compression rate of 100-120 beats per minute, and 80% chest recoil.

Posttraining Follow-Up and Maintenance

Multimedia Appendix 7 provides the descriptive statistics for the posttraining follow-up data. The results revealed that the Mixed12 group exhibited consistent BLS knowledge scores at baseline (postinitial training), with the highest average scores observed at 12 and 24 months after training. The Traditional6 group exhibited the highest average scores on the skill test at all 3 measurement time points. Figure 3 illustrates the estimated mean scores of BLS knowledge and skill tests for each group, as assessed over time using GEE models. At 12 months after initial training, the Traditional6 group had the lowest average BLS knowledge score (mean 70.10, SE 0.854), which was significantly different from that of the Mixed12 group (mean 75.14, SE 0.762; Figure 3A presents a nonoverlapping 95% CI). Subsequently, at 24 months following initial training, the Mixed12 group exhibited significantly higher scores (mean 79.32, SE 0.741) compared with the other groups.

Figure 3. Estimated mean scores with 95% CI for (A) BLS knowledge and (B) skill tests in different training courses by generalized estimating equation models. BLS: basic life support.



XSL•FO
RenderX

Furthermore, at baseline, a notable difference was observed in the average scores of the skill tests between the Mixed6 and Traditional6 groups (*P*=.003; Figure 3B shows a nonoverlapping 95% CI). Moreover, in the follow-up assessment, the Traditional6 group exhibited significantly higher scores than the other groups. Table 3 presents the proportion in each group for the achievement of high-quality CPR. At 12 and 24 months after initial training, this proportion in the Mixed12 group exhibited the most substantial decrease compared with those at 12 and 24 months after training. At baseline, no substantial differences were observed in these proportions among the 4 groups. However, no substantial differences were observed among these proportions among the Blend6, Mixed6, and Traditional6 groups at 12 or 24 months after initial training. We concurrently used multiple linear regression and GEE models to examine the performance indicators; the corresponding results are provided in Multimedia Appendices 5, 6, 8, and 9.

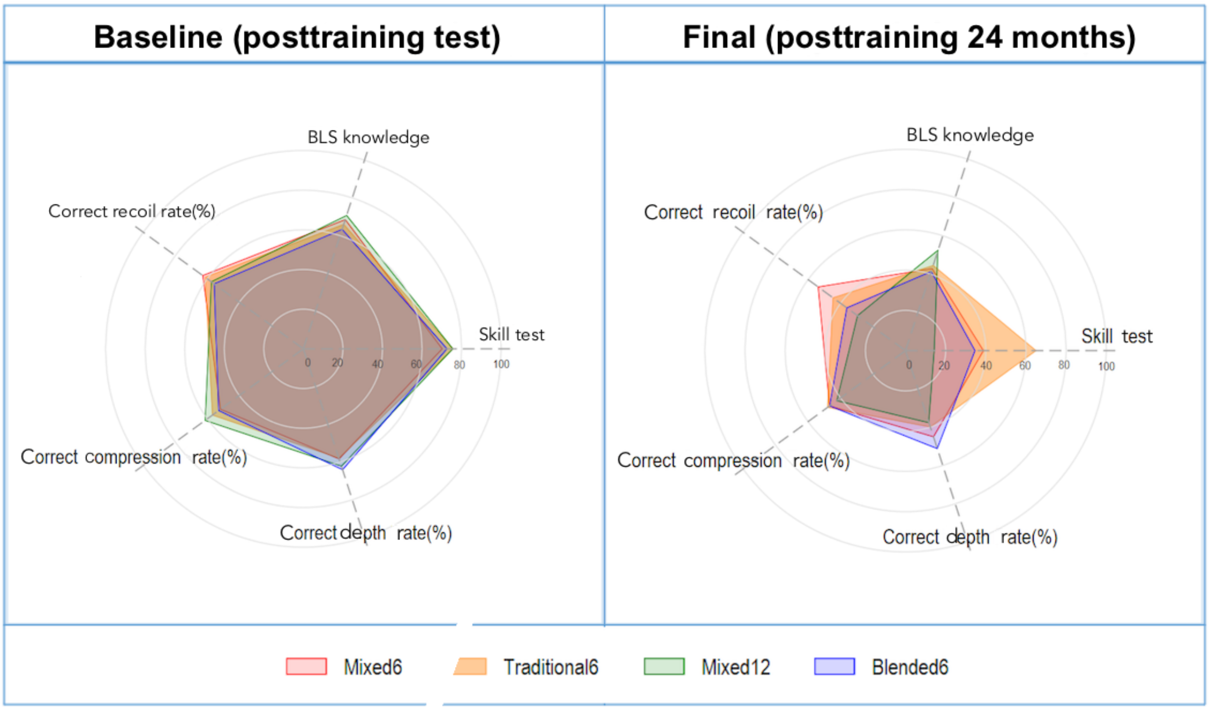
Table 3. Proportions of the achievement of high-quality CPRa at 0, 12, and 24 months after training for the different training courses.

Variables	Mixed6 (n=332), n (%)	Traditional6(n=270), n (%)	Mixed12 (n=258), n (%)	Blended6 (n=303), n (%)
Posttraining (0 month)	91 (27.4)	86 (31.9)	79 (30.6)	98 (32.3)
Posttraining (12 months)	83 (25)	61 (22.6)	2 (0.8)	63 (20.8)
Posttraining (24 months)	79 (23.8)	53 (19.6)	7 (2.7)	84 (27.7)

[a]CPR: cardiopulmonary resuscitation.

We used an alternative method to rank the 4 training methods based on objectively evaluated items. The scores were converted using percent ranking, and the average score was then calculated to represent the performance of each skill in each training method. Subsequently, we visualized the results as a radar chart (Figure 4). Overall, the 4 groups exhibited comparable average performance in the tests after the first training session. However, in the follow-up assessment (ie, 12 and 24 months after training), differences emerged among the groups (Multimedia Appendix 10). The Traditional6 group exhibited outstanding performance in the skill test and correct recoil rate. The Blended6 group demonstrated superiority in correct depth rate, whereas no significant difference was observed between the Blended6 and Traditional6 groups in terms of correct compression rate or high-quality CPR achievement. The Mixed12 group exhibited a lower correct recoil rate, compression rate, depth rate, and skill test performance compared with the other 3 groups.

Figure 4. Radar charts for posttraining evaluation at baseline and final visit (posttraining 24 months). BLS: basic life support.



Discussion

Principal Findings

This study provides 3 major findings regarding the effectiveness of traditional and blended training methods for CPR education. First, no significant difference was observed in knowledge acquisition after initial training, and all the training groups exhibited proficient CPR skills that met the requirements for high-quality CPR. However, a higher proportion of participants receiving blended training initially achieved high-quality CPR; this finding served as the basis for our comparative analysis. The second major finding highlights the importance of timely

XSL•FO
RenderX

retraining. When retraining was conducted 12 months after initial training, significant decreases were observed in the proficiency of CPR skills and the proportion of participants achieving high-quality CPR. Our third major finding suggests that more frequent retraining could maintain CPR skills more effectively. The participants who underwent retraining every 6 months exhibited slight decreases in their proficiency in CPR skills and their achievement of high-quality CPR. Additionally, we explored the potential of web-based self-directed learning as an alternative, and this learning method demonstrated effectiveness for skill retention regardless of the initial training method (traditional or blended), with no significant difference observed between the 2 methods.

Research has demonstrated that blended learning and traditional CPR methods [19,21,22] are practical and reasonably effective alternatives to traditional CPR training; however, large-scale comparisons of these methods or the integration of these instructional methods into CPR education have not been conducted. To the best of our knowledge, this study was the first study to demonstrate that blended learning and retraining stimuli are not inferior to traditional methods when it comes to CPR performance. Chien et al [15] found that blended learning for CPR training does not have inferior learning outcomes relative to traditional methods but that CPR skills at 6 months did not meet the AHA's CPR guidelines. This finding was consistent with our findings. Although traditional instruction may lead to slightly more favorable performance initially, providing self-directed blended learning stimuli every 6 months is effective for maintaining CPR skills. We found that among learners who received CPR training every 12 months, the performance of high-quality CPR decreased by 35% more than that of those retrained every 6 months. Therefore, consistent with previous research recommendations, stimulating learning every 6 months appears to be favorable to doing so every 12 months. This observation aligns with the AHA's 2020 guidelines, which suggest that for the general public, the use of convenient learning methods alongside retraining is a viable alternative to traditional face-to-face CPR training.

The blended learning method used in this study offers considerable economic benefits and is time saving for both learners and instructors. By incorporating 18 minutes of web-based learning and self-training into a course, the face-to-face instruction and relearning time were collectively shortened by approximately 72 minutes initially and by 12 minutes in subsequent training. These decreases reduced the expenditure, human resources, and time requirements for learners and instructors in CPR training courses [21]. One study investigated the cost-effectiveness of blended learning for CPR training; the results revealed that blended learning decreased training costs while achieving similar maintenance of CPR skills relative to the traditional method [23]. However, some researchers have indicated that despite the costs and time reductions offered by blended learning, such learning does not ensure that participants will acquire further professional knowledge and proficiency in a demanding training environment [22]. The maintenance of CPR skills contributes to the willingness of the public to perform CPR. When EMSs are activated, guiding individuals to identify cardiac arrest and to implement CPR with dispatcher assistance is challenging as is ensuring that members of the public are able to perform high-quality CPR [24]. Accordingly, blended teaching and retraining models, which appear to be as effective as traditional learning models, can address the challenge of instructing individuals during emergency calls. The characteristics of blended teaching models, including time saving and environmental efficiency, can be beneficial for promoting CPR education among the public and for addressing challenges in maintaining CPR skills among the public.

In this study, 95.1%% (1106/1163) of the participants were high school graduates who were approximately 40 years old and who exhibited higher learning and web-based operating abilities. This demographic advantage likely contributed to the success of blended learning in this study. Moreover, this study used a participant-to-manikin ratio of 2-3:1, leading to higher costs compared with the traditional method (1 manikin to 6 students). The increased investment in training infrastructure may affect the overall cost-effectiveness of blended learning in various settings. The study did not record the frequency of learners' usage of blended relearning stimuli; the effectiveness of self-paced web-based learning may be related to the time spent engaging with the material. Nevertheless, the primary objective of blended web-based learning is to enable individuals to learn at their convenience. In contrast to traditional face-to-face classroom learning, in blended learning, participants have the flexibility to arrange their web-based and in-class training according to their convenience and location. Accordingly, this learner-centric approach can lead to an environment that is more conducive to the maintenance of CPR skills.

In this study, favorable exercise habits and previous CPR learning experiences enhanced the effectiveness of CPR training. Even if learning had occurred more than 2 years previously, blended CPR training could effectively maintain CPR skills. Ettl et al [20] found that incorporating CPR learning into fitness exercise training increased learners' motivation and confidence in performing CPR. Therefore, establishing exercise habits helps maintain CPR skills and for fostering rescue skills.

Finally, although blended learning with a retraining frequency of 6 months demonstrated significant economic benefits and time-saving ability in this study, its cost-effectiveness depended on factors such as participant demographics, the training environment, and the level of engagement with web-based learning opportunities. Accordingly, consideration of these factors could maximize the potential of blended learning in various CPR training scenarios.

Limitations

This study had some limitations. First, in observational studies, the random allocation of samples is infeasible and could result in disparities between groups. Consequently, we used a multivariate regression model to mitigate the impact of variables; thus, we impartially assessed the differences between the groups. Moreover, this study involved tracking the training status of each group to understand the importance of the interval between retraining sessions and whether the given training method was appropriate. Second, we collected demographic data from a subset of learners, but our comprehension of these

XSL•FO
RenderX

learners' economic backgrounds and technology use was limited; consequently, whether blended learning is effective among individuals with relatively low socioeconomic status should be further explored. Third, our research cohort lacked the representation of older adults. As a result, uncertainties persist regarding the applicability of blended training for this demographic; accordingly, future studies are recommended to address this crucial gap. Finally, the absence of an analysis of the participants' willingness to perform CPR leaves a significant gap in our understanding. Accordingly, individuals' willingness to administer CPR after blended retraining should be investigated in future research.

Conclusions

Blended learning for CPR with a retraining frequency of 6 months provides higher retention of high-quality CPR skills than does retraining every 12 months. Notably, the blended method demonstrated effects similar to those of traditional relearning methods.

Acknowledgments

This manuscript was edited by Wallace Academic Editing. We are also thankful for the support of Chang Gung Memorial Hospital, Taiwan (CMRPG1M0081 and CMRPG1N0081).

Data Availability

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

Conflicts of Interest

None declared.

Multimedia Appendix 1
The written test of cardiopulmonary resuscitation knowledge.
[PDF File (Adobe PDF File), 291 KB - mededu_v10i1e52230_app1.pdf]

Multimedia Appendix 2
The skill test of cardiopulmonary resuscitation practice checklist.
[PDF File (Adobe PDF File), 169 KB - mededu_v10i1e52230_app2.pdf]

Multimedia Appendix 3
Generalized estimating equation models for the performance indicators.
[DOCX File , 26 KB - mededu_v10i1e52230_app3.docx]

Multimedia Appendix 4
Generalized estimating equation models for the performance indicators.
[DOCX File , 27 KB - mededu_v10i1e52230_app4.docx]

Multimedia Appendix 5
Multiple linear regression model for the performance indicators at baseline: basic life support knowledge, skill test, average compression depth, and rate.
[DOCX File , 27 KB - mededu_v10i1e52230_app5.docx]

Multimedia Appendix 6
Multiple analysis for the performance indicators at baseline: the proportion of correct compression depth, speed rate, and recoil.
[DOCX File , 27 KB - mededu_v10i1e52230_app6.docx]

Multimedia Appendix 7
Summary statistics for outcome assessment at baseline, post-12M, post- 24M in different training courses.
[DOCX File , 26 KB - mededu_v10i1e52230_app7.docx]

Multimedia Appendix 8
Estimated mean with 95% CI for compression depth and rate in different training courses by generalized estimating equation models.
[PDF File (Adobe PDF File), 25 KB - mededu_v10i1e52230_app8.pdf]

XSL•FO
RenderX

Multimedia Appendix 9
Estimated mean with 95% CI for correct compression depth, rate, and recoil in different training courses by generalized estimating equation models.
[PDF File (Adobe PDF File), 52 KB - mededu_v10i1e52230_app9.pdf]

Multimedia Appendix 10
The radar chart for posttraining evaluation after 12 months.
[PDF File (Adobe PDF File), 78 KB - mededu_v10i1e52230_app10.pdf]

References

1. Yan S, Gan Y, Jiang N, Wang R, Chen Y, Luo Z, et al. The global survival rate among adult out-of-hospital cardiac arrest patients who received cardiopulmonary resuscitation: a systematic review and meta-analysis. Crit Care 2020;24(1):61 [FREE Full text] [doi: 10.1186/s13054-020-2773-2] [Medline: 32087741]
2. Phattharapornjaroen P, Nimnuan W, Sanguanwit P, Atiksawedparit P, Phontabtim M, Mankong Y. Characteristics and outcomes of out-of-hospital cardiac arrest patients before and during the COVID-19 pandemic in Thailand. Int J Emerg Med 2022;15(1):46 [FREE Full text] [doi: 10.1186/s12245-022-00444-2] [Medline: 36085002]
3. Chien CY, Tsai SL, Tsai LH, Chen CB, Seak CJ, Weng YM, et al. Impact of transport time and cardiac arrest centers on the neurological outcome after out-of-hospital cardiac arrest: a retrospective cohort study. J Am Heart Assoc 2020;9(11):e015544 [FREE Full text] [doi: 10.1161/JAHA.119.015544] [Medline: 32458720]
4. Bunch TJ, White RD, Gersh BJ, Meverden RA, Hodge DO, Ballman KV, et al. Long-term outcomes of out-of-hospital cardiac arrest after successful early defibrillation. N Engl J Med 2003;348(26):2626-2633 [FREE Full text] [doi: 10.1056/NEJMoa023053] [Medline: 12826637]
5. Becker L, Gold LS, Eisenberg M, White L, Hearne T, Rea T. Ventricular fibrillation in King County, Washington: a 30-year perspective. Resuscitation 2008;79(1):22-27. [doi: 10.1016/j.resuscitation.2008.06.019] [Medline: 18687513]
6. Bækgaard JS, Viereck S, Møller TP, Ersbøll AK, Lippert F, Folke F. The effects of public access defibrillation on survival after out-of-hospital cardiac arrest: a systematic review of observational studies. Circulation 2017;136(10):954-965 [FREE Full text] [doi: 10.1161/circulationaha.117.029067]
7. Myat A, Song KJ, Rea T. Out-of-hospital cardiac arrest: current concepts. Lancet 2018;391(10124):970-979. [doi: 10.1016/S0140-6736(18)30472-0] [Medline: 29536861]
8. Nolan J, European Resuscitation Council. European resuscitation council guidelines for resuscitation 2005. section 1. introduction. Resuscitation 2005;67(Suppl 1):S3-S6. [doi: 10.1016/j.resuscitation.2005.10.002] [Medline: 16321715]
9. Huang CH, Fan HJ, Chien CY, Seak CJ, Kuo CW, Ng CJ, et al. Validation of a dispatch protocol with continuous quality control for cardiac arrest: a before-and-after study at a city fire department-based dispatch center. J Emerg Med 2017;53(5):697-707. [doi: 10.1016/j.jemermed.2017.06.028] [Medline: 28943036]
10. Wissenberg M, Lippert FK, Folke F, Weeke P, Hansen CM, Christensen EF, et al. Association of national initiatives to improve cardiac arrest management with rates of bystander intervention and patient survival after out-of-hospital cardiac arrest. JAMA 2013;310(13):1377-1384 [FREE Full text] [doi: 10.1001/jama.2013.278483] [Medline: 24084923]
11. Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023;147(8):e93-e621 [FREE Full text] [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
12. Cheng-Yu C, Yi-Ming W, Shou-Chien H, Chan-Wei K, Chung-Hsien C. Effect of population-based training programs on bystander willingness to perform cardiopulmonary resuscitation. Signa Vitae 2016;12(1):63-69 [FREE Full text] [doi: 10.22514/sv121.102016.11]
13. Chien CY, Chien WC, Tsai LH, Tsai SL, Chen CB, Seak CJ, et al. Impact of the caller's emotional state and cooperation on out-of-hospital cardiac arrest recognition and dispatcher-assisted cardiopulmonary resuscitation. Emerg Med J 2019;36(10):595-600. [doi: 10.1136/emermed-2018-208353] [Medline: 31439715]
14. Agerskov M, Nielsen AM, Hansen CM, Hansen MB, Lippert FK, Wissenberg M, et al. Public access defibrillation: great benefit and potential but infrequently used. Resuscitation 2015;96:53-58 [FREE Full text] [doi: 10.1016/j.resuscitation.2015.07.021] [Medline: 26234893]
15. Chien CY, Fang SY, Tsai LH, Tsai SL, Chen CB, Seak CJ, et al. Traditional versus blended CPR training program: a randomized controlled non-inferiority study. Sci Rep 2020;10(1):10032 [FREE Full text] [doi: 10.1038/s41598-020-67193-1] [Medline: 32572100]
16. Cheng A, Magid DJ, Auerbach M, Bhanji F, Bigham BL, Blewer AL, et al. Part 6: resuscitation education science: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 2020;142(16_suppl_2):S551-S579 [FREE Full text] [doi: 10.1161/CIR.0000000000000903] [Medline: 33081527]
17. Nolan JP, Monsieurs KG, Bossaert L, Böttiger BW, Greif R, Lott C, et al. European Resuscitation Council COVID-19 guidelines executive summary. Resuscitation 2020;153:45-55 [FREE Full text] [doi: 10.1016/j.resuscitation.2020.06.001] [Medline: 32525022]

XSL•FO
RenderX

18. Han S, Park HJ, Nah S, Lee EH, Lee HJ, Park JO, et al. Instructor-led distance learning for training students in cardiopulmonary resuscitation: a randomized controlled study. PLoS One 2021;16(5):e0251277 [FREE Full text] [doi: 10.1371/journal.pone.0251277] [Medline: 33956873]
19. Ali DM, Hisam B, Shaukat N, Baig N, Ong MEH, Epstein JL, et al. Cardiopulmonary Resuscitation (CPR) training strategies in the times of COVID-19: a systematic literature review comparing different training methodologies. Scand J Trauma Resusc Emerg Med 2021;29(1):53 [FREE Full text] [doi: 10.1186/s13049-021-00869-3] [Medline: 33781299]
20. Ettl F, Schöll A, Zupa B, Schantl C, Gramberger J, Krammel M, et al. The CPR-workout: a new training concept. Resuscitation 2016;106(Supplement 1):e54-e55. [doi: 10.1016/j.resuscitation.2016.07.131]
21. Elgohary M, Palazzo FS, Breckwoldt J, Cheng A, Pellegrino J, Schnaubelt S, et al. Blended learning for accredited life support courses—a systematic review. Resusc Plus 2022;10:100240 [FREE Full text] [doi: 10.1016/j.resplu.2022.100240] [Medline: 35592876]
22. Hsieh MJ, Bhanji F, Chiang WC, Yang CW, Chien KL, Ma MHM. Comparing the effect of self-instruction with that of traditional instruction in basic life support courses-a systematic review. Resuscitation 2016;108:8-19. [doi: 10.1016/j.resuscitation.2016.08.021] [Medline: 27581252]
23. Chong KM, Yang HW, He HC, Lien WC, Yang MF, Chi CY, et al. The effectiveness of online-only blended cardiopulmonary resuscitation training: static-group comparison study. J Med Internet Res 2023;25:e42325 [FREE Full text] [doi: 10.2196/42325] [Medline: 37018023]
24. Huang CH, Chien CY, Ng CJ, Fang SY, Wang MF, Lin CC, et al. Effects of dispatcher-assisted public-access defibrillation programs on the outcomes of out-of-hospital cardiac arrest: a before-and-after study. J Am Heart Assoc 2024;13(3):e031662 [FREE Full text] [doi: 10.1161/JAHA.123.031662] [Medline: 38240326]

Abbreviations

AED: automated external defibrillator
AHA: American Heart Association
BLS: basic life support
bpm: beats per minute
CPR: cardiopulmonary resuscitation
EMS: emergency medical service
GEE: generalized estimating equation
OHCA: out-of-hospital cardiac arrest

Edited by T de Azevedo Cardoso, AH Sapci, MD; submitted 29.08.23; peer-reviewed by A Missel, T Tangoaisarn; comments to author 28.09.23; revised version received 08.10.23; accepted 31.03.24; published 29.04.24.

Please cite as:
Chien CY, Tsai SL, Huang CH, Wang MF, Lin CC, Chen CB, Tsai LH, Tseng HJ, Huang YB, Ng CJ
Effectiveness of Blended Versus Traditional Refresher Training for Cardiopulmonary Resuscitation: Prospective Observational Study
JMIR Med Educ 2024;10:e52230
URL: https://mededu.jmir.org/2024/1/e52230
doi:10.2196/52230
PMID:38683663

©Cheng-Yu Chien, Shang-Li Tsai, Chien-Hsiung Huang, Ming-Fang Wang, Chi-Chun Lin, Chen-Bin Chen, Li-Heng Tsai, Hsiao-Jung Tseng, Yan-Bo Huang, Chip-Jin Ng. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.04.2024. 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.

XSL•FO
RenderX

A Student's Viewpoint on ChatGPT Use and Automation Bias in Medical Education

Jeanne Maria Dsouza

Kasturba Medical College, Manipal, India

Corresponding Author:
Jeanne Maria Dsouza

Related Article:

Comment in: http://mededu.jmir.org/2024/1/e50174/

(JMIR Med Educ 2024;10:e57696) doi:10.2196/57696

KEYWORDS

AI; artificial intelligence; ChatGPT; medical education

The editorial *ChatGPT in Medical Education: A Precursor for Automation Bias?* by Nguyen [1] is very timely, appropriate, and informative. Being a medical student myself, I find that it gives a balanced view on the use of ChatGPT, which is sweeping across the globe at a spectacular pace. One of the hallmarks of this tool is that it is almost universally accessible, even in parts of the world where there may be limited access to quality medical education. As authors have rightly pointed out, ChatGPT is useful for summarizing information, generating practice questions, and giving instantaneous feedback [2-4], and it could serve as an effective personalized tutor. It provides high-quality scientific text gleaned from a quick and comprehensive review of the literature and presents text in an efficient, readable, and versatile style [1]. It is no wonder that it is gaining immense popularity among students, including medical students, who 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 aptly underscored in the editorial [1].

The downside of this powerful tool has also been well portrayed. There is a very real risk of automation bias, especially among medical students in the younger generation, who are digitally savvy but often lack experience and confidence in their clinical skills. The blind dependence on ChatGPT and other artificial intelligence (AI) tools could corrode their thinking and decision-making skills and lead to erroneous medical outcomes. The clinical setting is undoubtedly the best classroom for students to develop the skills for understanding and accommodating the needs, expectations, and values of patients and their caregivers in the real-world scenario, as well as cultivate leadership qualities and work in a team. It is vital for us students to retain our originality, identity, and critical analytical skills to avoid falling into the trap of AI solutionism.

The need for AI education at this crucial juncture has been well brought out. At present, only a minority of students have received AI education [5]. Incorporating it into the medical curriculum is a challenging, multidisciplinary endeavor. Knowing how and when to use this powerful tool in a responsible manner, without clouding clinical judgment and in keeping with the tenets of medical ethics, is paramount. I agree with Nguyen's [1] view that ChatGPT should be used as a supplementary tool rather than as the default resource for medical education. There is a need to exercise vigilance in the utilization of this tool right from the formative years of medical professionals.

AI is here to stay, and ChatGPT will undoubtedly have an all-pervading influence on medical education and the practice of medicine itself. Therefore, its optimal utilization is the need of the hour. Imparting AI education would help unleash the power of ChatGPT, but appropriate pre-emptive measures to keep its disruptive potential in check are needed to pave the way for an AI-savvy generation of medical professionals with sound clinical judgment and skills.

Editorial Notice

The corresponding author of "ChatGPT in Medical Education: A Precursor for Automation Bias?" declined to respond to this letter.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Nguyen T. ChatGPT in medical education: a precursor for automation bias? JMIR Med Educ 2024 Jan 17;10:e50174. [doi: 10.2196/50174] [Medline: 38231545]
2. 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 6;9:e46885. [doi: 10.2196/46885] [Medline: 36863937]
3. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med 2023 Aug 1;98(8):867-868. [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
4. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
5. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772. [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]

Abbreviations

AI: artificial intelligence

Edited by T Leung; submitted 24.02.24; this is a non–peer-reviewed article;accepted 28.03.24; published 15.04.24.

Please cite as:
Dsouza JM
A Student's Viewpoint on ChatGPT Use and Automation Bias in Medical Education
JMIR Med Educ 2024;10:e57696
URL: https://mededu.jmir.org/2024/1/e57696
doi:10.2196/57696

© Jeanne Maria Dsouza. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.4.2024. 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.

XSL•FO
RenderX

Digital Skills to Improve Levels of Care and Renew Health Care Professions

Massimo De Martinis[1,2,3,4,*], MD; Lia Ginaldi[1,5,*], MD

1

2

3

4

5

*all authors contributed equally

Corresponding Author:
Massimo De Martinis, MD

Related Article:

Companion article: https://mededu.jmir.org/2024/1/e51112

(JMIR Med Educ 2024;10:e58743) doi:10.2196/58743

KEYWORDS

digital competence; telehealth; nursing; health care workforce; health care professionals; informatics; education; curriculum; interdisciplinary education; health care education

We read with great interest the article by Rettinger et al [1], "Telehealth education in allied health care and nursing: web-based cross-sectional survey of students' perceived knowledge, skills, attitudes, and experience," recently published in *JMIR Medical Education*.

The authors, addressing an extremely current topic, highlight the need to integrate telehealth into health care education curricula. More generally, we think that the development of digital competence is essential for all health care professionals. The digitalization of care processes requires ever-greater digital skills to ensure high-level care suited to current knowledge. Another recent investigation [2] summarizes the educational intervention methods that have been implemented to develop digital competence and the effects of these educational interventions on health care workforce; this study suggests the best method for enhancing the digital skills of nurses and allied professionals in the context of continuing professional education. This research turned attention to the active workforce, who need to adapt their knowledge to renewed working contexts where digital technology is forcefully entering. However, we must note, as emphasized by Rettinger et al [1], that our curricula often neglect the need to equip health care degree students with adequate digital skills. We observe that few of our students are keeping up with the development of technology. Digital skills can range from the simplest to the most sophisticated technological applications commonly used in a hospital environment, including the use of virtual simulators and extending to artificial intelligence, which, especially in the coming years, will become a precious tool for improving care processes [3]. Even for delivering high-quality care in digitally enabled health care environments, nursing informatics competency is a required core competency [4]. In light of this, it would be necessary to introduce programs dedicated to the acquisition of these skills into our study courses; these programs could be spread across all curricular disciplines. To achieve these objectives, it is necessary to ensure that teachers have the necessary skills in this field or have the ability to acquire them to pass them on to their students. We are well aware that the nursing profession is going through a period of crisis and that it is essential to implement all available forces and strategies to renew it, making it attractive and satisfying again [5]. There are numerous proposals for this renewal, and they must also address the active workforce; however, the updating of the study contents for degree courses in health professions must be one of the first and fundamental steps to achieve these results. The acquisition of adequate digital skills is a necessity that can no longer be postponed to train professionals capable of providing the best levels of care possible today.

Editorial Notice

The corresponding author of "Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Students' Perceived Knowledge, Skills, Attitudes, and Experience" declined to respond to this letter.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Rettinger L, Putz P, Aichinger L, et al. Telehealth education in allied health care and nursing: web-based cross-sectional survey of students' perceived knowledge, skills, attitudes, and experience. JMIR Med Educ 2024 Mar 21;10:e51112. [doi: 10.2196/51112] [Medline: 38512310]
2. Kulju E, Jarva E, Oikarinen A, Hammarén M, Kanste O, Mikkonen K. Educational interventions and their effects on healthcare professionals' digital competence development: a systematic review. Int J Med Inform 2024 May;185:105396. [doi: 10.1016/j.ijmedinf.2024.105396] [Medline: 38503251]
3. Simms RC. Work with ChatGPT, not against: 3 teaching strategies that harness the power of artificial intelligence. Nurse Educ 2024;49(3):158-161. [doi: 10.1097/NNE.0000000000001634] [Medline: 38502607]
4. O'Connor S, Cave L, Philips N. Informing nursing policy: an exploration of digital health research by nurses in England. Int J Med Inform 2024 May;185:105381. [doi: 10.1016/j.ijmedinf.2024.105381] [Medline: 38402804]
5. Ginaldi L, Di Mascio R, Sepe I, Colleluori N, De Martinis M. The necessary change of direction for the nursing profession - letter on Petrosino et al. Intensive Crit Care Nurs 2024 Jun;82:103638. [doi: 10.1016/j.iccn.2024.103638] [Medline: 38325226]

Edited by T Leung; submitted 23.03.24; this is a non–peer-reviewed article;accepted 03.04.24; published 01.05.24.

Please cite as:
De Martinis M, Ginaldi L
Digital Skills to Improve Levels of Care and Renew Health Care Professions
JMIR Med Educ 2024;10:e58743
URL: https://mededu.jmir.org/2024/1/e58743
doi:10.2196/58743

© Massimo De Martinis, Lia Ginaldi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 1.5.2024. 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.

XSL•FO
RenderX

Authors' Reply: A Use Case for Generative AI in Medical Education

Tricia Pendergrast[1], MD; Zachary Chalmers[2], PhD

1

2

Corresponding Author:
Tricia Pendergrast, MD

Related Articles:

https://mededu.jmir.org/2023/1/e48780/

https://mededu.jmir.org/2024/1/e56117/

(JMIR Med Educ 2024;10:e58370) doi:10.2196/58370

KEYWORDS

ChatGPT; undergraduate medical education; large language models

We thank the authors for their thoughtful comments on our paper titled, "Anki Tagger: A Generative AI Tool for Aligning Third-Party Resources to Preclinical Curriculum" [1,2]. The authors' discussion of the ethical issues and limitations of generative artificial intelligence is both timely and important. As the capabilities of ChatGPT and other similar tools evolve, so must our conversations about the use of generative artificial intelligence in medicine and medical education.

With respect to the production of educational materials for medical trainees, ChatGPT's ability to "hallucinate" and thereby provide misinformation should be of particular concern to educators. For example, when asked to summarize the research output of 50 scientists and cite relevant literature related to Chagas disease, ChatGPT made a major error in 86.7% of its outputs [3]. The problem of hallucination is more pronounced with smaller training data sets and may therefore disproportionately affect medical education content related to rare diseases, which are emphasized in licensing examinations. The problem of hallucination remains a substantial barrier to the widespread use of generative artificial intelligence in medical education.

We circumvented the issue of hallucination by embedding existing Anki flashcard decks in a large language model, rather than prompting ChatGPT to generate flashcards de novo from scientific literature [1]. Anki flashcard decks are among the third-party resources used by medical students to bridge perceived gaps in school curricula, especially regarding preparation for the USMLE (United States Medical Licensing Examination). Medical students report feeling overwhelmed with the number of third-party resources at their disposal and experience tension between these resources and their in-house curricula [4]. Their educators experience tension among different domains of responsibility including clinical practice, research, professional development, and education [5]. Therefore, it is beneficial to both teachers and students for medical education to be as efficient as possible. To this end, ChatGPT can organize and stratify third-party learning resources by relevance to lectures and other curricular elements [1].

While the integration of third-party resources into lesson plans for undergraduate medical education may be controversial, it is important to note that medical students are already using third-party resources instead of lectures by clinical educators [4]. Instead of viewing these learning materials as competition, our application of ChatGPT suggests the possibility of integrating third-party resources into existing medical curricula. Future studies should examine the impact of such an intervention on medical students' academic performance and satisfaction as well as medical educator burnout.

Conflicts of Interest

None declared.

References

1. Pendergrast T, Chalmers Z. Anki Tagger: a generative AI tool for aligning third-party resources to preclinical curriculum. JMIR Med Educ 2023 Sep 20;9:e48780. [doi: 10.2196/48780] [Medline: 37728965]
2. Sekhar TC, Nayak YR, Abdoler EA. A use case for generative AI in medical education. JMIR Med Educ 2024 Jun;10:e56117. [doi: 10.2196/56117]

XSL•FO
RenderX

3. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. Bibliographic research with ChatGPT may be misleading: the problem of hallucination. J Pediatr Surg 2024 Jan;59(1):158. [doi: 10.1016/j.jpedsurg.2023.08.018] [Medline: 37735041]
4. Lawrence ECN, Dine CJ, Kogan JR. Preclerkship medical students' use of third-party learning resources. JAMA Netw Open 2023 Dec 1;6(12):e2345971. [doi: 10.1001/jamanetworkopen.2023.45971] [Medline: 38048132]
5. Arvandi Z, Emami A, Zarghi N, Alavinia SM, Shirazi M, Parikh SV. Linking medical faculty stress/burnout to willingness to implement medical school curriculum change: a preliminary investigation. J Eval Clin Pract 2016 Feb;22(1):86-92. [doi: 10.1111/jep.12439] [Medline: 26563562]

Abbreviations

USMLE: United States Medical Licensing Examination

Edited by T Leung; submitted 13.03.24; this is a non–peer-reviewed article;accepted 28.03.24; published 07.06.24.

Please cite as:
Pendergrast T, Chalmers Z
Authors' Reply: A Use Case for Generative AI in Medical Education
JMIR Med Educ 2024;10:e58370
URL: https://mededu.jmir.org/2024/1/e58370
doi:10.2196/58370

© Tricia Pendergrast, Zachary Chalmers. Originally published in JMIR Medical Education (https://mededu.jmir.org), 7.6.2024. 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.

XSL•FO
RenderX

A Use Case for Generative AI in Medical Education

Tejas C Sekhar[1], BA; Yash R Nayak[2], BA; Emily A Abdoler[2], MD, MAEd

1

2

Corresponding Author:
Emily A Abdoler, MD, MAEd

Related Articles:

https://mededu.jmir.org/2023/1/e48780/

Comment in: https://mededu.jmir.org/2024/1/e58370/

(JMIR Med Educ 2024;10:e56117) doi:10.2196/56117

KEYWORDS

medical education; med ed; generative artificial intelligence; artificial intelligence; GAI; AI; Anki; flashcard; undergraduate medical education; UME

A recent study explored the novel application of generative artificial intelligence's (GAI's) capabilities with regard to Anki using a new methodology ("Anki Tagger"), leveraging OpenAI's ChatGPT-3.5 to tag and stratify flashcards from the AnKing deck, which are most aligned with a medical school's curriculum and involve a minimal cost and time expenditure [1]. To the best of our knowledge, their work represents the first publication demonstrating early proof of concept of GAI applied to Anki, a spaced repetition flashcard application designed to promote long-term retention of learned content. A major benefit of their approach is the ability to streamline and automate the otherwise time-consuming and resource-intensive process of manually comparing medical school curricula against the widely used and crowdsourced AnKing deck.

Medical students who use Anki may use decks prepared by more senior students at their medical school, the AnKing deck (a reputable and comprehensive set of >35,000 flashcards and growing daily, collaboratively maintained largely by current and graduated medical students), or a combination thereof. Research indicates that daily Anki use is associated with increased USMLE (United States Medical Licensing Examination) Step 1 scores and higher sleep quality—indicators of academic performance and personal well-being, respectively [2]. Given the prevalent usage and growing adoption of Anki among medical students, applications of GAI and large language models (LLMs) to Anki workflows may be beneficial. Even considering their present shortcomings, LLMs may provide a unique opportunity to significantly impact medical education in the intermediate term, especially given the propensity of contemporary medical students to supplement didactic learning with web-based learning resources [3].

Furthermore, LLMs with GAI capabilities, such as ChatGPT and Med-PaLM, have the potential to answer medically related questions [4] and—intriguingly for the medical education community—can pass the USMLE [5]. Such a notable feat by LLMs necessitates reevaluation of the future of medical training and practice while carefully considering the relevant ethical issues and current limitations of GAI, such as their susceptibility for generating misinformation through a process known as "hallucination." As GAI and LLMs become increasingly integrated in daily practice, similar and iteratively improved methodologies represent a way for educators and learners alike to benefit considerably by better aligning flashcards from the comprehensive AnKing deck with in-house curricula in preparation for medical licensing examinations such as USMLE Step 1. Future applications of GAI in undergraduate medical education may involve the implementation of AI-assisted features directly built into preferred educational tools and resources, allowing students increased customization with options for multimodal output beyond solely text.

Conflicts of Interest

None declared.

References

1. Pendergrast T, Chalmers Z. Anki tagger: a generative AI tool for aligning third-party resources to preclinical curriculum. JMIR Med Educ 2023 Sep 20;9:e48780. [doi: 10.2196/48780] [Medline: 37728965]

XSL•FO
RenderX

2. Wothe JK, Wanberg LJ, Hohle RD, et al. Academic and wellness outcomes associated with use of Anki spaced repetition software in medical school. J Med Educ Curric Dev 2023;10:23821205231173289. [doi: 10.1177/23821205231173289] [Medline: 37187920]
3. Wynter L, Burgess A, Kalman E, Heron JE, Bleasel J. Medical students: what educational resources are they using? BMC Med Educ 2019 Jan 25;19(1):36. [doi: 10.1186/s12909-019-1462-9] [Medline: 30683084]
4. Anastasio AT, Mills FB, Karavan MP, Adams SB. Evaluating the quality and usability of artificial intelligence–generated responses to common patient questions in foot and ankle surgery. Foot Ankle Orthop 2023 Oct;8(4):24730114231209919. [doi: 10.1177/24730114231209919] [Medline: 38027458]
5. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]

Abbreviations

GAI: generative artificial intelligence
LLM: large language model
USMLE: United States Medical Licensing Examination

Edited by T Leung; submitted 06.01.24; this is a non–peer-reviewed article;accepted 28.03.24; published 07.06.24.

Please cite as:
Sekhar TC, Nayak YR, Abdoler EA
A Use Case for Generative AI in Medical Education
JMIR Med Educ 2024;10:e56117
URL: https://mededu.jmir.org/2024/1/e56117
doi:10.2196/56117

© Tejas C Sekhar, Yash R Nayak, Emily A Abdoler. Originally published in JMIR Medical Education (https://mededu.jmir.org), 7.6.2024. 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.

XSL•FO
RenderX

Original Paper

Measuring e-Professional Behavior of Doctors of Medicine and Dental Medicine on Social Networking Sites: Indexes Construction With Formative Indicators

Marko Marelić[1], PhD; Ksenija Klasnić[2], PhD; Tea Vukušić Rukavina[1,3], MD, PhD

[1]Andrija Štampar School of Public Health, School of Medicine, University of Zagreb, Zagreb, Croatia
[2]Department of Sociology, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
[3]Biomedical Research Center Šalata, School of Medicine, University of Zagreb, Zagreb, Croatia

Corresponding Author:
Tea Vukušić Rukavina, MD, PhD
Andrija Štampar School of Public Health
School of Medicine
University of Zagreb
Rockfeller Street 4
Zagreb, 10000
Croatia
Phone: 385 14590126
Email: tvukusic@snz.hr

Abstract

Background: Previous studies have predominantly measured e-professionalism through perceptions or attitudes, yet there exists no validated measure specifically targeting the actual behaviors of health care professionals (HCPs) in this realm. This study addresses this gap by constructing a normative framework, drawing from 3 primary sources to define e-professional behavior across 6 domains. Four domains pertain to the dangers of social networking sites (SNSs), encompassing confidentiality, privacy, patient interaction, and equitable resource allocation. Meanwhile, 2 domains focus on the opportunities of SNSs, namely, the proactive dissemination of public health information and maintaining scientific integrity.

Objective: This study aims to develop and validate 2 new measures assessing the e-professional behavior of doctors of medicine (MDs) and doctors of dental medicine (DMDs), focusing on both the dangers and opportunities associated with SNSs.

Methods: The study used a purposive sample of MDs and DMDs in Croatia who were users of at least one SNS. Data collection took place in 2021 through an online survey. Validation of both indexes used a formative approach, which involved a 5-step methodology: content specification, indicators definition with instructions for item coding and index construction, indicators collinearity check using the variance inflation factor (VIF), external validity test using multiple indicators multiple causes (MIMIC) model, and external validity test by checking the relationships of the indexes with the scale of attitude toward SNSs using Pearson correlation coefficients.

Results: A total of 753 responses were included in the analysis. The first e-professionalism index, assessing the dangers associated with SNSs, comprises 14 items. During the indicators collinearity check, all indicators displayed acceptable VIF values below 2.5. The MIMIC model showed good fit (χ^2_{13}=9.4, *P*=.742; χ^2/df=0.723; root-mean-square error of approximation<.001; goodness-of-fit index=0.998; comparative fit index=1.000). The external validity of the index is supported by a statistically significant negative correlation with the scale measuring attitudes toward SNSs (r=–0.225, *P*<.001). Following the removal of 1 item, the second e-professionalism index, focusing on the opportunities associated with SNSs, comprises 5 items. During the indicators collinearity check, all indicators exhibited acceptable VIF values below 2.5. Additionally, the MIMIC model demonstrated a good fit (χ^2_4=2.5, *P*=.718; χ^2/df=0.637; root-mean-square error of approximation<0.001; goodness-of-fit index=0.999; comparative fit index=1.000). The external validity of the index is supported by a statistically significant positive correlation with the scale of attitude toward SNSs (r=0.338; *P*<.001).

Conclusions: Following the validation process, the instrument designed for gauging the e-professional behavior of MDs and DMDs consists of 19 items, which contribute to the formation of 2 distinct indexes: the e-professionalism index, focusing on the dangers associated with SNSs, comprising 14 items, and the e-professionalism index, highlighting the opportunities offered by

XSL•FO
RenderX

SNSs, consisting of 5 items. These indexes serve as valid measures of the e-professional behavior of MDs and DMDs, with the potential for further refinement to encompass emerging forms of unprofessional behavior that may arise over time.

(JMIR Med Educ 2024;10:e50156) doi:10.2196/50156

KEYWORDS

e-professionalism; social media; formative index; social networking; doctors; medical; dental medicine

Introduction

Background

The development of social networking sites (SNSs) as a new form of media and communication channel has brought many changes to the health care system [1]. The widespread use of SNSs affects what we perceive as the professional behavior of health care professionals (HCPs) [2].

The rise in SNS users has sparked a growing interest in comprehending e-professionalism, particularly concerning SNSs. This specific facet of e-professionalism is becoming increasingly important. Over the past few years, numerous studies on the e-professionalism of HCPs have emerged [3,4], indicating a sustained momentum in generating scientific insights into e-professionalism.

Defining and Measuring e-Professionalism

The American Board of Internal Medicine (ABIM) guidelines on medical professionalism define 3 fundamental principles and a set of 10 professional responsibilities (or commitments). Fundamental principles are the importance of patient welfare, the principle of patient autonomy, and the principle of social justice. Professional responsibilities are commitments to professional competence, honesty with patients, patient confidentiality, maintaining appropriate relations with patients, improving the quality of care, improving access to care, a just distribution of finite resources, scientific knowledge, maintaining trust by managing conflicts of interest, and commitment to professional responsibilities [5].

E-professionalism is a specific type of professionalism. Cain and Romanelli [6] defined e-professionalism as the attitudes and behaviors (some of which may occur in private settings) reflecting traditional professionalism paradigms that are manifested through digital media.

A large number of previous research around e-professionalism measured the perception of e-professionalism [7-11] and attitude toward e-professionalism [12-18]. Through cross-validation, Kelley et al [19] created an instrument for measuring professional behaviors in pharmacy students, and even though there are some thematic overlaps, it is not suitable for measuring online behavior.

E-professionalism is often defined as a value which justifies the operationalization that directs the measurement of professionalism toward the measure of attitude. Nevertheless, from the perspective of the professions themselves, although professionalism is taught and transferred through socialization into the profession as a value, for assessing the level of e-professionalism of doctors of medicine (MDs) and doctors of dental medicine (DMDs) the behavioral component is of greater interest. Professional behavior, rather than just attitude, constitutes a visible aspect of professionalism. It is through professional behavior that not only patients and colleagues perceive a doctor's professionalism, but also it is subject to internal control according to Freidsonian principles [20], enabling the profession to enforce sanctions on the professional. Professionalism is a behavior rather than an attitude because it should not be a hypothetical or idealized concept, as Evans [21] writes, but should be perceived as a reality—an actual entity. However, it is a real entity only if it is operational. To be real, professionalism must be something that people—professionals—actually "do," not just something that the government or any other agency wants them to do, or wrongly imagines them to be doing [21]. The disconnection between behavior and attitude is termed "cognitive dissonance" [22], a phenomenon already acknowledged as a threat to the e-professionalism of HCPs on SNSs [4].

The research focused on the medical and dental professions as the target populations. These 2 fields were chosen due to their fundamental similarities, enabling comparisons, as well as their differences, suggesting potential variations in e-professionalism. Both medical and dental professions are sociologically recognized as professions [20] and share the commonality of providing health services. This entails a significant patient-practitioner relationship in both disciplines. Comparing various health professions is a valuable approach, and existing literature has already established overlaps in core competencies [23].

The primary distinction driving the selection of these 2 professions is the orientation of MDs, particularly in the Croatian context, toward the public sector, whereas DMDs are oriented toward the private sector.

This paper seeks to develop a reliable and valid instrument for assessing the e-professional behavior of both MDs and DMDs.

Normative Framework for Defining e-Professional Behavior

Overview

To define and measure e-professional behavior effectively, it is crucial to differentiate between professional and unprofessional behaviors. In our case, the primary objective of the normative framework is to delineate the content specifications, specifically the domains of instruments used to measure e-professional behavior.

The normative framework for assessing e-professionalism among MDs and DMDs draws upon 3 primary sources. While none of these sources alone is adequate for defining a comprehensive normative framework, each provides essential information crucial for its development. Some aspects of these

XSL•FO
RenderX

sources overlap conceptually, while others offer unique insights necessary for crafting the framework.

The first source comprises the e-professional conduct guidelines established by the ABIM [5]. These guidelines, among the earliest to be published, were developed through an international collaboration involving multiple institutions. They address the fundamental principles of professionalism and outline the professional responsibilities expected of MDs.

The second source consists of guidelines aimed at fostering e-professional behavior among medical and dental students [24]. While initially targeted at this specific demographic, a significant portion of the recommendations is applicable to the e-professionalism of MDs and DMDs. Consequently, these guidelines serve as a valuable resource for “reconstructing” the components of a normative framework for e-professionalism. They aid in delineating acceptable and unacceptable behaviors on SNSs within the context of medical and dental professions.

The third source is Julie Skrabal’s research [9], where she used the grounded theory method to develop a theoretical framework for e-professionalism. Her study empirically demonstrated which behaviors on SNSs are perceived as unprofessional. While the research focused on nursing students, the identification of key domains and indicators comprising professional behavior on SNSs holds significant value and applicability to MDs, DMDs, and all HCPs.

Based on the analysis of these 3 sources, e-professionalism, or e-professional behavior, can be categorized into 6 domains. Four of these domains pertain to the dangers associated with SNSs: confidentiality, privacy, contact with patients, and fair distribution of resources. The remaining 2 domains concern the opportunities afforded by SNSs: proactive dissemination of information relevant to public health and maintaining scientific objectivity. Each of these 6 domains is elaborated upon below.

Confidentiality

Confidentiality encompasses behaviors that primarily contravene the Health Insurance Portability and Accountability Act (HIPAA) of 1996. It entails safeguarding patient confidentiality to ensure that information regarding the patient is not disclosed, even to the patient’s relatives, without the patient’s explicit consent.

Concerning behavior on SNSs, HIPAA violations predominantly involve the unauthorized publication of photos or confidential patient information [9]. Additionally, adopting fake names (pseudonyms) to share posts containing medical or dental information constitutes another unprofessional behavior [24].

Privacy

This domain pertains to profile privacy settings and the management of post visibility. Barlow et al [25] established a correlation between privacy settings and unprofessional behavior, particularly among medical students. Consequently, they recommended the adoption of “private visibility settings” to mitigate such behaviors. Monitoring privacy settings [24], controlling post visibility [9,24], and seeking permission before tagging colleagues in posts to safeguard their privacy [24] are advocated practices. Furthermore, it is advisable to refrain from publishing professionally inappropriate content on SNSs, including posts containing curses, vulgar expressions, inappropriate attire, or behavior [9,24].

Contact With Patients

This domain encompasses direct contact with patients via SNSs. Inappropriate expressions, political incorrectness, or derogatory remarks toward patients or any individual or group can severely tarnish the public’s perception of doctors’ professional conduct [24]. Additionally, using unofficial channels, such as SNSs, to communicate sensitive professional information is considered unprofessional behavior within this domain [9].

Fair Distribution of Resources

Fair distribution of resources, as acknowledged in the ABIM guidelines, is considered an essential aspect of professional responsibility. While the ABIM guidelines emphasize the avoidance of unnecessary interventions and examinations, resource distribution also extends to SNSs. Time, a valuable resource allocated by MDs and DMDs to their patients, is particularly relevant in this context. Derived from the fundamental principle of professionalism known as the “Principle of Social Justice,” striving for a fair distribution of health care resources is imperative [5]. Communication with patients via SNSs typically requires the doctor’s time, often during their free time since it is an informal communication channel. According to the principle of fairness, it would be considered unprofessional behavior if a doctor selectively chooses which patients they are willing to communicate with on SNS and which they are not.

Proactive Publication of Information of Public Health Interest

The dimension of proactive publication of professional information of public health interest is one of the recognized aspects of e-professionalism that highlights the opportunity aspect of using SNSs. These behaviors are not deemed unprofessional when avoided; however, they can significantly contribute to e-professionalism when practiced by MDs and DMDs. While Skrabal [9] emphasizes creating positive postings as the absence of criticism and negative comments, proactive posting as a deliberate action toward e-professionalism is acknowledged in another research [26].

Scientific Objectivity

Sharing knowledge on SNSs is indeed desirable and constitutes professional behavior. However, it is essential to clearly differentiate between personal or subjective medical opinions and scientifically based facts [24].

Formative Approach in Measuring e-Professionalism

Most latent variables used in the social sciences are measured using reflective (effect) indicators [27,28]. According to a prevailing convention, indicators are seen as functions of the latent variable, whereby changes in the latent variable are reflected in changes in the observable indicators [27]. This is often true regarding constructs such as personality or attitude [28]. For example, attitude about SNSs affects respondents’ responses to the items posed to them. If someone has a negative attitude about SNSs, that attitude “guides” their responses.

XSL•FO
RenderX

However, in the case where the direction of "influence" is reversed, and where the indicators are "causing" the latent variable instead of "being caused by it," then we can talk about formative measures [28].

Index construction focuses on explaining the abstract (unobserved) variance, considers multicollinearity among indicators, and emphasizes the role of an indicator as a "predictor" (latent variable) rather than "a predicted variable" [27].

The choice of approach (reflexive vs formative) stems from the concept, that is, from the relationship between variables and constructs [29]. Jarvis et al [30] stated 4 conditions that can help discern whether a reflective or a formative model is appropriate: (1) the direction of causality between the construct and the indicator, (2) the interchangeability of the indicators, (3) covariance between indicators, and (4) the nomological network of construct indicators.

The first argument presented by Jarvis et al [30] is valid for our research because, unlike attitude, e-professional behavior stems from specific actions and decisions on SNSs. If someone refrains from posting pictures of patients, seeks permission from a colleague before mentioning them on SNS, actively controls the visibility of their posts, and takes similar actions, then these decisions contribute to their e-professional behavior.

For the second argument, e-professional behavior indicators are not interchangeable, even though they all measure e-professionalism. Posting a picture of a patient on an SNS is considered unprofessional behavior, but so is posting pictures from parties at work. Both behaviors are unprofessional, although they are not interchangeable in measurement (someone may frequently post photos of patients but rarely post workplace-related images).

The third argument states that covariance among indicators is unnecessary [30]. It is neither expected nor needed here because recognized behaviors within the normative framework can be entirely unrelated but still measure e-professional behavior (eg, sending a friend request to a patient and asking a colleague to mention them in a post).

The fourth argument suggests that the nomological network in the formative model can have different antecedents and consequences [30]. Indicators of e-professional behavior do not need to share the same antecedents because they can be driven by different motivations. A doctor may post pictures of patients because they believe it raises awareness about a particular illness (even though this act is unprofessional), while the motivation for unprofessional behavior, such as posting pictures from workplace parties, does not stem from the same motivation.

Based on these arguments, the behavioral component of e-professionalism measured in this paper conceptually corresponds to the formative approach.

We presume that other research in this area has not applied a formative approach in measuring e-professionalism because they have yet to define e-professionalism as a behavior.

Diamantopoulos and Winklhofer [28] proposed 4 key steps for validating indexes with formative indicators. The first step, content specification, refers to specifying the scope of the latent variable; in the second step, it is necessary to define the indicators; the third step refers to checking the collinearity of the indicators using the variance inflation factor (VIF) [28]. The fourth step is to assess the external validity of the index. Verification of the external validity of formative indices is often carried out by checking the relationship of the index with other measures and variables (as cited in [28]).

Although these 4 steps are sufficient for constructing and validating the index, it is possible to make an additional check of the external validity proposed by Diamantopoulos and Winklhofer [28]. This requires creating a model in which some reflective indicators are included (Diamantopoulos and Winklhofer [28] use 2) in the same model as the formative indicators. This model is called the multiple indicators multiple causes (MIMIC) model [28]. Acceptable overall model fit suggests retention of items in the formative model. If the exclusion of some items can significantly increase the model fit under the very strict condition that not a single exclusion would violate the content validity of the formative model, only then can the items be excluded.

In this paper, we have followed these 4 key steps for validating indexes with formative indicators. An additional step (the MIMIC model) was conducted before assessing the external validity of the index.

Methods

Sample

Quantitative survey data were collected using an online survey questionnaire. The Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [31] is available in Multimedia Appendix 1. The required sample size was defined according to a conservative estimate often used for multivariate analyses, corresponding to a 10:1 ratio between the number of observations and the number of variables used in the questionnaire's largest instrument [32]. In our case, that is a sample size of 280 (140 MDs and 140 DMDs). The type of sample was a nonprobabilistic purposive sample.

The study was a part of a long-term research project funded by the Croatian Science Foundation, UIP-05-2017 "Dangers and Benefits of Social Networks: E-Professionalism of Health Care Professionals – SMePROF" [33].

The mailing lists used to distribute the survey were the official full membership emailing lists of the Croatian Medical Chamber (CMC) and Croatian Chamber of Dental Medicine (CCDM). At the time of the survey, the CMC's emailing list contained 15,562 email addresses of MDs, and the CCDM's email list contained 7616 email addresses of DMDs. The email included a brief text about the study's objective, the expected time to complete the survey, and the person and university responsible for conducting the study.

Participation in the survey was voluntary; there was no form of incentive to complete the survey. To ensure anonymity, no identification data were collected. Data were collected from February to July 2021, with 2 reminders sent in that period.

XSL•FO
RenderX

Ethics Approval

Both the study and the questionnaire were approved by the ethical boards of the University of Zagreb School of Medicine (641-01/18-02/01) and the University of Zagreb School of Dental Medicine (05-PA-24-2/2018). In addition, formal approval was obtained from the governing bodies of both the CMC and CCDM for the use of the complete mailing lists of MDs and DMDs who are members of the CMC (900-06/20-01/11) and CCDM (900-01/21-01/02).

Measures

The instrument for measuring the e-professional behavior of MDs and DMDs, presented in this study, is part of a more extensive questionnaire called SMePROF Project Survey Questionnaire on Social Media Usage, Attitudes, Ethical Values and E-professional Behaviour of Doctors of Medicine and Doctors of Dental Medicine, available at Viskić et al [34]. Although the questionnaire contained multiple instruments partially derived from previous studies [10,34,35], the instrument for measuring the e-professional behavior of MDs and DMDs is a novel instrument created by the authors. The instrument contains 20 items measured using the self-reporting approach, used to create 2 e-professionalism indexes, and the process is explained in the following parts of this paper. In validating indexes, an MIMIC model was used, which required 4 reflexive variables (y_1-y_4) measuring attitude toward e-professionalism. These items were taken from a validated instrument for measuring attitudes toward e-professionalism [35]. Descriptives of these 4 reflexive variables are shown in Multimedia Appendix 2. The MIMIC model was exclusively used as a method for validating the external validity of the indexes, and not for theory development.

The associations of indexes with theoretically related constructs were tested to assess the external validity. For this purpose, we used a validated instrument for measuring attitudes toward SNSs [36]. The instrument was translated into the Croatian language, and after additional reliability checks, 1 item was removed from the scale ("Potential and/or existing employers may use the information found on SNS to make decisions about prospective and/or existing employees"). The final instrument used had 12 items and Cronbach α=.70.

Analytical Methods

A descriptive analysis of frequencies and percentage of responses was carried out, and distribution measures such as mean, range, SDs, and α_3 measure of asymmetry were determined depending on appropriateness. Correlations between quantitative variables were tested with the Pearson correlation coefficient and phi coefficients of associations. The multicollinearity of the instruments was tested with the VIF. The MIMIC model was used to check the external validity of instruments with formative indicators. Data analysis was performed using IBM SPSS Statistics 26. IBM SPSS Amos 22 was used to test the MIMIC model.

Results

Survey Responses

A total of 1013 responses were collected. The response rate was 4.37% (1013/23,178). The final realized sample of the entire research contained the answers of 999 respondents, of which 75.4% (753/999) use at least one SNS, 67.3% (507/753) of the respondents were MDs and 32.7% (246/753) were DMDs. The sample was predominantly female (558/753, 74.1%) with an average age of 38 (SD 10.99) years. Most respondents worked in a public health institution (412/753, 54.7%), and the second most frequent type of workplace was a private institution with a contract with the Croatian Health Insurance Fund (CHIF; 148/753, 19.7%).

Previous research on the same sample [34] showed a significant difference in age, where MDs were older than DMDs with an average age of 39.26 years as opposed to 36.58 years, respectively, and in the type of employment, with more than two-thirds of DMDs (168/246, 68.2%) being employed in the private sector compared with only 20.5% (104/507) of MDs. All specialization status levels are included in the sample (Multimedia Appendix 3).

The Construction of the e-Professionalism Index—The Danger Aspect of SNSs

Following the first step in creating the index, according to Diamantopoulos and Winklhofer [28], the content for the latent variable is specified below. In the second step, e-professional behaviors described in the normative framework were operationalized into an instrument for measuring the aspect of e-professionalism related to the dangers of SNSs (Table 1). The identified indicators are grouped into 4 domains: confidentiality, privacy, contact with patients, and fair distribution of resources. Items were evaluated on a frequency rating scale: 0=Never, 1=Rarely, 2=Occasionally, 3=Often; and the option "I have never been in a situation where this could happen" was added. It was essential to distinguish behaviors that could have happened but did not from those for which the respondent was not even in a situation to practice them. Depending on the direction and content of the items, the difference between the opportunity to behave in a certain way and the frequency of that behavior can mean the difference between professional and unprofessional behavior. In the case of items formulated in a positive direction (marked +), a higher frequency measures a higher level of e-professionalism. In the case of items formulated in a negative direction (marked –), higher frequency measures a lower level of e-professionalism.

XSL•FO
RenderX

Table 1. Domains, indicators, and items for the instrument of e-professionalism—the danger aspect of SNSs[a].

Domain and indicator	Item	Direction[b]
Confidentiality		
Disclosure of patient information.	I published some information about my patient.	–
Publication of photographs of the patient without their consent.	I posted a photo of my patient without their knowledge.	–
Hiding behind false names when posting online or anonymously posting medical information.	I shared medical/dental advice on SNS without my name being visible.	–
Confidentiality of communication also applies to SNS.	I shared some information about the patient I received through SNS with others.	–
Privacy of MDs and DMDs profiles		
Active management of the visibility of posts depending on their content.	Depending on the appropriateness of the content of my posts, I determine to whom they will be visible.	+
Controlling the visibility of other people's posts that include you, depending on their content.	If I notice that someone else has published something about me (eg, my picture, location, or similar), I control who will see it.	+
Seeking prior approval from colleagues to publish information about them.	I asked a colleague's permission to mention them in the post.	+
Appropriate behavior on published content from a professional context.	I have posted content that shows informal situations at my workplace (eg, drinks with colleagues or parties at work).	–
The use of profanity and other vulgar expressions in posts.	A curse word or some different vulgar expression occasionally slips out in my posts.	–
Contact with patients		
Inappropriate expression in posts.	In my posts, I am cautious that my expression is entirely professional.	+
Separation of professional and private communication.	I communicate with patients regarding medical/dental problems and treatment from a private profile.	–
Inclusion of patient data obtained at SNS in the medical documentation without the patient's consent.	I included information about the patient I found through SNS in the medical documentation without their knowledge.	–
Sending a friend request to a patient or a patient's family member.	Have you ever sent a "friend request" to a patient or a member of the patient's family from a private profile on an SNS?	–
Fair distribution of resources		
Communication with patients via SNS and outside working hours is selective (the doctor chooses whom they respond to; patients without SNS cannot reach them)	On SNS, I choose which patients I will make contact with and which I will not.	–

[a]SNS: social networking site.

[b]For items formulated in a positive direction (marked +), a higher frequency measures a higher level of e-professionalism. In the case of items formulated in a negative direction (marked –), a higher frequency measures a lower level of e-professionalism.

The indicator "Sending a friend request to a patient or a member of the patient's family" was not measured as frequency. Instead, the 4 offered answers were as follows: Yes, to the patient; Yes, to a family member; Yes, both; and No. The negative response is considered professional, while all other responses indicate unprofessional behavior.

The descriptive results for the items that measure the aspect of e-professionalism related to the dangers of SNSs are shown in Table 2. The items that measure e-professional behavior are marked with a "b." All other items measure e-unprofessional behavior.

XSL•FO
RenderX

Table 2. E-professionalism (the dangers aspect of SNSs[a]) descriptives (N=753).

Danger aspects	Never, n (%)	Rarely, n (%)	Occasionally, n (%)	Often, n (%)	I have never been in a situation where this could happen, n (%)
1. I asked a colleague's permission to mention them in the post.[b]	170 (22.6)	117 (15.5)	71 (9.4)	50 (6.6)[c]	345 (45.8)[c]
2. I shared some information about the patient that I received through SNS with other people.	368 (48.9)[c]	61 (8.1)	26 (3.5)	3 (0.4)	295 (39.2)[c]
3. I posted a photo of my patient without their knowledge.	492 (65.3)[c]	14 (1.9)	7 (0.9)	2 (0.3)	238 (31.6)[c]
4. I included information about the patient I found through SNS in the medical documentation without their knowledge.	484 (64.3)[c]	3 (0.4)	2 (0.3)	0 (0.0)	264 (35.1)[c]
5. I shared medical/dental advice on SNS without my name being visible.	503 (66.8)[c]	39 (5.2)	6 (0.8)	4 (0.5)	201 (26.7)[c]
6. Depending on the appropriateness of the content of my posts, I determine to whom they will be visible.[b]	295 (39.2)	99 (13.1)	93 (12.4)	71 (9.4)[c]	195 (25.9)[c]
7. If I notice that someone else has published something about me (eg, my picture, location, or similar), I control who will see it.[b]	209 (27.8)	106 (14.1)	104 (13.8)	177 (23.5)[c]	157 (20.8)[c]
8. I have published content that shows informal situations at my workplace (eg, drinks with colleagues or parties at work).	354 (47.0)[c]	181 (24.0)	84 (11.2)	17 (2.3)	114 (15.1)[c]
9. I published some information about my patient.	579 (76.9)[c]	22 (2.9)	5 (0.7)	2 (0.3)	145 (19.3)[c]
10. I communicate with patients regarding medical/dental problems and treatment from a private profile.	423 (56.2)[c]	133 (17.7)	64 (8.5)	14 (1.9)	119 (15.8)[c]
11. On SNS, I choose which patients I will make contact with and which I will not.	293 (38.9)[c]	74 (9.8)	65 (8.6)	76 (10.1)	245 (32.5)[c]
12. In my posts, I am cautious that my expression is entirely professional.[b]	51 (6.8)	61 (8.1)	111 (14.7)	366 (48.6)[c]	164 (21.8)[c]
13. A curse word or some other vulgar expression occasionally slips out in my posts.	494 (65.6)[c]	86 (11.4)	23 (3.1)	2 (0.3)	148 (19.7)[c]
14. Have you ever sent a "friend request" to a patient or a member of the patient's family from a private profile on an SNS?[d]	699 (92.8)[c]	33 (4.4)	3 (0.4)	18 (2.4)	N/A[e]

[a]SNS: social networking site.

[b]Item represents professional behavior on SNS.

[c]Response represents professional behavior on SNS.

[d]The options were "no," "yes, to a patient," "yes, to a family member," and "yes, both," respectively.

[e]N/A: not applicable.

The answer "I have never been in a situation where this could happen" is not a missing value, but it carries a conceptual meaning that must be distinguished from the answer "Never." The assessment of whether that answer is professional or unprofessional depends on the content and direction of the item. Respondents who have never engaged in unprofessional behavior are professional, but so are those who never had an opportunity to act unprofessionally. Respondents who often practice behaviors on items marked with "b" are professional, and so are those who have never been in a situation to practice these behaviors because they have not been in a situation to behave unprofessionally.

For example, in the case of positive items (those representing professional behavior), such as "I asked a colleague's permission to mention him/her in the post," professional behavior is defined as a situation where the individual has never violated this rule because they have never mentioned colleagues in their posts or seek permission each time they mention them. Any other frequency level implies that, at some point, the person has posted about colleagues without their consent, which constitutes unprofessional behavior on SNS.

XSL•FO
RenderX

It is crucial here to differentiate between the absence of behavior of interest (requesting permission from colleagues when mentioning them in posts) in situations where it should have been sought (if mentioning them in posts) from the situations where it should not have been sought (because they never mention colleagues).

By contrast, for negative items (those representing unprofessional behavior), such as "I shared some information about the patient that I received through SNS with other people," professional behavior is defined as situations where the individual has never engaged in such behavior or has not even been in a situation where they could engage in such behavior (eg, they do not communicate with patients via SNS, so they cannot receive patient information through this channel).

Therefore, the context of the absence of specific behaviors plays a pivotal role in distinguishing between professional and unprofessional behaviors. It is essential to combine the response "I have never been in a situation where this could happen" with the level of behavior frequency.

To construct the index, the frequency of behavior on each indicator was not graded but only considered as a binary value (professional vs unprofessional).

For items that measure unprofessional behavior, any degree of frequency other than "never" was considered unprofessional behavior. For items that measure professional behavior (eg, asking a colleague's permission to mention them in a post), all those who did this never, rarely, or occasionally were considered unprofessional on that indicator, because this is the behavior they are expected to do always (or often in our scale).

The Validation Process of the e-Professionalism Index—The Danger Aspect of SNSs

After specifying the scope and defining the indicators, the third step, according to Diamantopoulos and Winklhofer [28], refers to checking the collinearity of the indicators. Intercorrelations of the items in the e-professionalism instrument—the danger aspect of SNSs are shown in Multimedia Appendix 4. Given that these are binary variables, phi coefficients of associations were used. The correlation between the variables "On SNS, I choose with which patients I will make contact with and which I will not." and "From a private profile, I communicate with patients regarding medical/dental problems and treatment." (r=0.568) represents a moderate correlation and evokes the need to investigate potential multicollinearity. This suggests that those who communicated with patients via SNSs also chose with whom (patients) they would establish communication. As a formative approach is used, special care is needed before excluding indicators to preserve the instrument's validity. Therefore, the VIF and MIMIC model were calculated. Multicollinearity was tested using a VIF with an additive index of e-professionalism, an aspect of the danger of SNS that was constructed as the sum of values on binary indicators. According to the conservative threshold [37], VIF values on all indicators were below the value of 2.5, which suggests that multicollinearity is not an issue.

The MIMIC model was implemented to check the external validity of the instrument. The path diagram of the MIMIC model is shown in Multimedia Appendix 5. Variables x_1-x_{14} correspond to the items from Table 2. Items y_1 (Communication with a patient through social media can be achieved without compromising doctor-patient confidentiality) and y_2 (Social media have the potential to improve communication between a doctor and a patient) were chosen as reflective indicators.

The model showed good fit (χ^2_{13}=9.4, P=.742; χ^2/df =.723; root-mean-square error of approximation<0.001; goodness-of-fit index=0.998; comparative fit index=1.000). However, 7 of the 14 items (x_1, x_2, x_3, x_6, x_7, x_8, and x_{13}) did not have significant regression coefficients (γ) that can also be interpreted as validity coefficients [28]. The probable reason is that the measured reflective indicators did not measure the same domains as e-professional behavior; instead, they measured an attitude toward e-professionalism. Both items 11 (P<.001) and 12 (P=.02), which were investigated as potential problems of multicollinearity, have significant validity coefficients. Considering that, as well as an acceptable VIF, they were retained in the index to preserve the content validity to which formative models are particularly sensitive.

A higher value on the index means a higher degree of e-professionalism, that is, a lower incidence of unprofessional behavior on SNSs. The index results ranged from 0 to 14, and the average value in our sample was 10.60 (SD 2.173). The distribution of the index was skewed toward higher values (α_3=–.44, P=.09), that is, toward the professional behavior of our respondents on SNSs.

The external validity of the index is supported by the correlation with other measured constructs. There was a statistically significant negative correlation between the index of e-professionalism (aspects of the danger of SNSs) and the scale of attitude toward SNSs (r=–0.225, P<.001).

The Construction of the e-Professionalism Index—The Opportunity Aspect of SNSs

The construction of the e-professionalism index—the opportunity aspect of SNSs follows the same validation steps as the aspect of the dangers of SNSs [28].

E-professional behaviors described in the normative framework were operationalized into an instrument for measuring e-professionalism through the opportunity aspect of SNSs. The instrument contains 2 domains, measured by 6 items. All items are formulated in the same direction so a higher frequency measures a higher level of e-professionalism (Table 3).

XSL•FO
RenderX

Table 3. Domains, indicators, and items for the instrument of e-professionalism—opportunity aspect of SNSs[a].

Domain and indicator	Item	Direction[b]
Proactive posting of expert information of public health interest		
Sharing posts that contain general medical advice	I share posts on social media that contain general medical/dental advice.	+
Sharing new scientific knowledge in the field of medicine	I use my profile to share information about new scientific knowledge in the field of medicine/dental medicine.	+
Debunking medical myths and misinformation	I debunk medical/dental myths and misinformation by posting on SNS.	+
Calling for public health actions	I use SNS to raise public awareness of public health actions.	+
Encouraging responsible behavior	I create posts on SNS that call for responsible health behavior.	+
Scientific objectivity		
Emphasis on distinguishing personal medical opinions from facts	In the posts, I clearly separate my personal opinion on a medical/dental issue from scientifically confirmed facts.	+

[a]SNS: social networking site.

[b]All items are formulated in the same direction so a higher frequency measures a higher level of e-professionalism.

The descriptive results for the items that measure the opportunity aspect of SNSs are shown in Table 4. While measuring the danger aspect of SNSs focused on occurrence, not on the frequency of occurrence, the frequency of each behavior is relevant with this instrument. All behaviors in this instrument have the characteristic of being desirable, but the absence of such behaviors is not unprofessional. If an MD or DMD practices these behaviors, they use opportunities of SNSs and contribute to their professionalism. However, if they do not practice any of these behaviors, or have never been in a situation where they can behave like that, it is not unprofessional, but misses the opportunity to use the advantages of SNSs.

Table 4. E-professionalism—opportunity aspect of SNSs[a]—descriptives (N=753).

	Never, n (%)	Rarely, n (%)	Occasionally, n (%)	Often, n (%)	I have never been in a situation where this could happen, n (%)
1. I debunk medical/dental myths and misinformation by posting on SNS.	355 (47.1)	130 (17.3)	87 (11.6)	16 (2.1)	165 (21.9)
2. I share posts on social media that contain general medical/dental advice.	312 (41.4)	167 (22.2)	128 (17.0)	28 (3.7)	118 (15.7)
3. I use SNS to raise public awareness of public health actions.	185 (24.6)	191 (25.4)	224 (29.7)	84 (11.2)	69 (9.2)
4. I use my profile to share information about new scientific knowledge in the field of medicine/dental medicine.	248 (32.9)	183 (24.3)	184 (24.4)	54 (7.2)	84 (11.2)
5. I create posts on SNS that call for responsible health behavior.	186 (24.7)	196 (26.0)	212 (28.2)	81 (10.8)	78 (10.4)
6. In the posts, I clearly separate my personal opinion on a medical/dental issue from scientifically confirmed facts.	170 (22.6)	49 (6.5)	77 (10.2)	149 (19.8)	371 (49.3)

[a]SNS: social networking site.

To construct an index reflecting the degree of e-professionalism in utilizing social networking opportunities, responses marked as “Never” or “I have never been in a situation where this could happen” are not considered contributions to e-professionalism and are coded as 0. Conversely, responses categorized as “Rarely,” “Occasionally,” and “Often” contribute to e-professionalism, representing 3 levels of engagement with the benefits of social networks and are coded as 1, 2, and 3 respectively.

The Validation Process of the e-Professionalism Index—Opportunity Aspect of SNSs

The correlations between the items that constitute this index have higher values than those in the aspect of dangers of the SNS index (Multimedia Appendix 6). The item “I create posts on SNS that call for responsible health behavior” moderately correlates with several items (from r=0.418 to 0.714). To check

XSL•FO
RenderX

if multicollinearity is present in this instrument, paying attention to the VIF is necessary.

VIF was calculated with an additive index of e-professionalism—opportunity aspect of SNSs. VIF values on all indicators are below the value of 2.5, which suggests no risk of multicollinearity, even according to a conservative interpretation.

Before excluding the item "I create posts on SNS that call for responsible health behavior," an MIMIC model was created with all the items included, and a second model without that item was created to check for any changes in the model fit. The diagram of the MIMIC model is shown in Multimedia Appendix 5. Variables x_1-x_6 correspond to the items from Table 4. Items y_3 (As MD/DMD, it is my duty to keep abreast of current trends in the use of SNS) and y_4 (Guiding patients to online information is a new responsibility of MDs/DMDs in the digital age) were chosen as 2 reflective indicators.

The MIMIC model with all 6 items showed good fit characteristics (χ^2_5=2.880, *P*=.718; χ^2/df = 0.576; root-mean-square error of approximation<.001; goodness-of-fit index=0.999; comparative fit index=1.000). However, 3 items (x_1, x_3, and x_5) did not have significant regression coefficients (γ; *P*=.14, *P*=.44, and *P*=.19, respectively).

Considering the high correlations with other items, the VIF value that exceeds the limit of 2.5, and the regression coefficient γ that is not statistically significant (*P*=.19), item x_5 was excluded from the e-professionalism index—opportunity aspect of SNSs. After excluding item x_5, the fit of the MIMIC model did not change significantly ($\Delta\chi^2_1$=0.336, *P*=.56) and the fit of the model was χ^2_4=2.544, *P*=.718; χ^2/df = 0.637; root-mean-square error of approximation<.001; goodness-of-fit index=0.999; comparative fit index=1.000.

The index of e-professionalism—opportunity aspect of SNSs was created as the sum of the values of the remaining 5 recoded variables. A higher value on the e-professionalism index means a higher degree of e-professionalism. The index results ranged from 0 to 15 (mean 4.13, SD 3.712). The distribution of the index was skewed toward lower values (α_3=0.67, *P*=.09), showing that 24% (181/753) of respondents do not take advantage of SNSs at all.

The external validity of the index of e-professionalism—opportunity aspect of SNSs is supported by the correlation with other measured constructs. There was a statistically significant positive correlation between the index and the scale of attitude toward SNSs (*r*=0.338, *P*<.001).

Discussion

Principal Findings

As far as the authors are aware, this is the first measure constructed to measure the e-professional behavior of MDs and DMDs, with the created indexes of opportunity and the danger aspects of SNSs being the first attempt at using a formative approach in the research of professionalism in general and in e-professionalism. The final instrument for measuring the e-professional behavior of MDs and DMDs consists of 19 items that form 2 indexes. Index of e-professionalism—the danger aspect of SNS, which is formed by 14 items, and the index of e-professionalism—opportunity aspect of SNS, which is formed by 5 items.

These novel indexes can be used to measure the level of e-professional behavior among MDs and DMDs, which can have potential real-world applications. The main implications can be utilized in education for young medical and dental professionals and the development of guidelines for improving e-professionalism. If the instrument were applied on a representative sample, it could yield valuable data to enable the implementation of data-based policies with specific behaviors of interest. Investigation of the external validity of both e-professionalisms showed acceptable results. There was a statistically significant negative correlation between the index of e-professionalism—the danger aspect of SNSs and the scale of attitude toward SNSs (*r*=–0.225, *P*<.001). This is the theoretically expected direction of the correlation because the more positive attitude the respondents have about SNSs, the more inclined they are to use them when working with patients, which according to the normative framework, represents unprofessional behavior. The statistically significant positive correlation between the index of e-professionalism—opportunity aspect of SNSs and the scale of attitude toward SNSs (*r*=0.338, *P*<.001) is also theoretically expected because the more positive attitude toward SNSs doctors have, the more likely they will take advantage of the benefits of SNSs.

In the index of e-professionalism—the danger aspect of SNSs, all initially operationalized indicators were retained. In the index of e-professionalism—the opportunity aspect of SNSs, item x_5 (I create posts on SNS that call for responsible health behavior) measuring the indicator "Encouraging responsible behavior" was excluded. The formative approach suggests cautious consideration of managing the content validity of the model. It seems that respondents understood item x_5 very similarly to item x_3 (I use SNS to raise public awareness of public health actions.). After testing the indicators in the MIMIC model, the authors concurred that the content validity is not threatened by excluding this item, and multicollinearity would pose a more significant problem than losing a very subtle difference in the contents of these items.

Comparison With Prior Work

Conceptual domains recognized in this study only partially overlap with domains in the instrument of (offline) professional behavior [19] and the instrument for measuring attitudes toward e-professionalism [35]. Kelley et al [19] recognized a domain called "Upholding principles of integrity and respect," which corresponds to the domain "Confidentiality" in this study, as well as "Citizenship and professional engagement" [19], which corresponds to "Proactive posting of expert information of public health interest." In an instrument for measuring attitudes toward e-professionalism, Marelić et al [35] recognized the domain "Ethical aspects" that theoretically includes HIPAA violations and therefore corresponds to the domain "Confidentiality" in this study, and the domain "Physicians in

XSL•FO
RenderX

the digital age" that corresponds to "Contact with patients". However, the instrument of (offline) professional behavior contains domains that are not comparable to e-professional behavior, and the instrument for measuring attitudes toward e-professionalism contains domains that are not applicable for behavior measurement, and because of potential cognitive dissonance, measuring attitude is not a replacement for behavior measurement.

Limitations

The first limitation of this study is the low response rate (1013/23,178, 4.37%). Previous research has indicated that these professions have low survey response rates, especially in e-mailing surveys using web-based formats [38-42]. Time, confidentiality concerns, and topic relevance are some of the main reasons for their low survey participation [40]. Previous research has indicated that declining response rates among HCPs may be attributed to various factors, including heightened requests to participate in surveys and increased workloads. This increase in workload encompasses both the rising number of patients and administrative responsibilities [38,39].

One factor likely contributing to the low response rate in this study is the demanding schedule of MDs and DMDs. The estimated time required to complete our survey was lengthy, ranging from 10 to 15 minutes, due to the inclusion of a complex and comprehensive questionnaire containing 40 questions. Moreover, the survey was conducted during the COVID-19 pandemic (February to July 2021), a period marked by heightened strain on the health care system. MDs, especially those in Croatia, were confronted with extreme workloads and specific working conditions during this time. Additionally, MDs received numerous invitations to participate in web-based surveys, particularly regarding the impact of the COVID-19 pandemic on their physical or mental health. Given these circumstances, our study's focus on e-professionalism may have been perceived as of lower interest, potentially further reducing doctors' willingness to participate in research.

However, our objective in creating and validating new indexes did not prioritize achieving representativeness in our sample or generalizing our findings to the entire population of MDs and DMDs in Croatia. Instead, our focus was on assessing the suitability of the developed measurement instruments across various medical professions, using a nonprobabilistic purposive sample. Our final sample comprised responses obtained from the population of interest for this study, specifically MDs and DMDs who use at least one SNS. It is worth noting that the number of responses received in our survey (507 MDs and 246 DMDs) exceeded the initially planned sample size (140 MDs and 140 DMDs) by a considerable margin.

The second limitation concerns a relatively large proportion of respondents (ranging from 69/753, 9.2%, to 371/753, 49.3%) who selected the option "I have never been in a situation where this could happen" for certain items. It remains unclear why they did not simply respond with "Never." The reasons behind this choice are ambiguous. It is possible that some respondents are passive users of SNSs, thus not engaging in any content publication and consequently unable to exhibit unprofessional behavior. Alternatively, it could be that these respondents do not work directly with patients, rendering items related to violations of the HIPAA irrelevant to them. Another possibility is that they perceive their standards of professionalism to be exceptionally high, leading them to believe they would never engage in such behavior. While this issue does not affect the measurement of the occurrence of e-(un)professional behavior, it does impede a detailed understanding of the frequency of e-unprofessional behavior. Addressing this limitation could be a focus of future research and modifications to the measurement instrument, but this should be preceded by gaining new insights into the e-professional behaviors of MDs and DMDs.

The third limitation involves the potential for bias associated with using a self-reporting approach to measurement. Similar to other self-report measures in medicine, 2 key biases often arise: recall bias and social desirability bias [43]. Recall bias in our study could be attributed to the lack of a specified timeframe, such as "during the last year." We chose this approach because it represents the initial assessment of such behaviors, and we faced a scarcity of existing data on this subject. Introducing a specific timeframe in future research could aid in mitigating potential recall bias. The potential for social desirability bias stems from 2 sources. First, the nature of the measurement itself requires HCP respondents to self-report potentially unprofessional behaviors, including some that may constitute violations of HIPAA. The other factor to consider is that respondents were contacted to participate in our research through the same institutions responsible for granting and revoking licenses to practice medicine/dental medicine. Despite our assurance of anonymity in the study, respondents may have felt compelled to provide socially desirable answers on certain items. One method to mitigate or control social desirability bias is to include positive items, such as those measuring professional behaviors, alongside other items. An additional approach to address both biases, which could serve as a recommendation for future research, involves further refinement and validation of the instrument. This could be achieved by comparing self-reported data with information obtained through web scraping of respondents' SNS profiles, particularly focusing on visible behaviors.

The fourth limitation arises from the potential mismatch between the use of reflective indicators y_1-y_4 in the MIMIC model and the nature of the created indexes, which are intended to measure e-professionalism as behavior. However, the reflective variables used in the model measure attitude. While this approach was necessary for creating the MIMIC model in this study, there is a possibility that cognitive dissonance [4,21] may compromise the fit of the model.

The fifth limitation to note is that the sources used to establish a normative framework were relevant to the time and location of this research. However, their applicability to other countries and populations of HCPs, or their accuracy over time, may be limited. For example, the ABIM e-professional conduct guidelines [5] are relatively dated, and while they represent fundamental values of professionalism, they may not fully encompass changes in societal values that have occurred since the emergence of SNSs. Specific behaviors measured in these indexes may require revision or supplementation in the future.

XSL•FO
RenderX

Moreover, additional studies conducted after the development of this index may offer new insights into creating a normative framework for defining e-professional behaviors [44].

Future Directions

In considering avenues for enhancing both the instruments used in this study and future research directions, it becomes apparent that there are opportunities for improvement and deeper exploration. One potential extension of this study, which could lead to a more thorough understanding of the topic, involves testing the indexes on specific subsamples, particularly within specialties such as dermatology and reconstructive and cosmetic surgery. These specialties may involve visual representations of procedures, such as “before and after” images [34], which could pose potential threats to e-professionalism.

Improving the quality of external validity assessment can be achieved by incorporating self-evaluation of e-professionalism into the MIMIC model. This addition would enhance the content validity of the model by supplementing existing reflective indicators used in the research. Furthermore, self-evaluation of e-professionalism would serve as a valuable tool for evaluating the nomological network of the instrument. It would provide insights into the direction and strength of correlation among individual indicators of e-professionalism, the e-professionalism indices themselves, and potential predictors for model creation.

Future attempts aimed at measuring e-professionalism could focus on investigating the underlying reasons behind responses such as “I have never been in a situation where this could happen." It is plausible that a more precise definition of items or the inclusion of specific examples could serve as mechanisms to help respondents differentiate between behaviors they never engage in and those they may never encounter. By refining the clarity and specificity of survey items, researchers can facilitate a more accurate assessment of respondents’ experiences and perceptions related to e-professional behavior. This approach could lead to a deeper understanding of the nuances involved in professional conduct within the context of SNSs.

Conclusions

In this paper, an instrument for measuring the e-professional behavior of MDs and DMDs was developed and validated using the formative approach. Following the validation process, the instrument comprises 19 items, which contribute to the formation of 2 indexes. The first index, focusing on the danger aspect of SNSs, is composed of 14 items that were dichotomized before index construction. The second index, which examines the opportunity aspect of SNSs, is composed of 5 items that were recoded as 4-point items before index construction.

These innovative indexes offer a means to gauge the level of e-professional behavior among MDs and DMDs. This marks the first measure specifically designed to assess the e-professional behavior of MDs and DMDs. The paper demonstrates the feasibility of investigating e-professional behavior using a formative approach, representing an advancement over existing measuring instruments. This approach provides a means to mitigate the impact of cognitive dissonance between attitudes and the actual behavior of MDs and DMDs.

The validation process confirmed that these indexes serve as a robust measure of e-professional behavior. Nevertheless, the instrument has been scrutinized for potential areas of enhancement, and suggestions for improvements have been proposed for future iterations of the instrument.

Acknowledgments

This study was funded by the Croatian Science Foundation under project UIP-05-2017 “Dangers and Benefits of Social Networks: E-Professionalism of Health Care Professionals – SMePROF”. Generative AI was not used in any portion of the manuscript writing.

Data Availability

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

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist for Reporting Results of Internet E-Surveys (CHERRIES).
[DOCX File , 23 KB - mededu_v10i1e50156_app1.docx]

Multimedia Appendix 2
Descriptive characteristics of reflective indicators for the MIMIC models of e-professionalism (N=753).
[DOC File , 34 KB - mededu_v10i1e50156_app2.doc]

Multimedia Appendix 3
Type of workplace and specialization status of the respondents.

XSL•FO
RenderX

[DOCX File , 14 KB - mededu_v10i1e50156_app3.docx]

Multimedia Appendix 4
Intercorrelations of items in the e-professionalism instrument - danger aspect of SNSs (N=753).
[DOCX File , 19 KB - mededu_v10i1e50156_app4.docx]

Multimedia Appendix 5
MIMIC e-Professionalism models.
[DOCX File , 219 KB - mededu_v10i1e50156_app5.docx]

Multimedia Appendix 6
Intercorrelations of items in the e-professionalism instrument - opportunity aspect of SNSs (N=753).
[DOC File , 36 KB - mededu_v10i1e50156_app6.doc]

References

1. Hazzam J, Lahrech A. Health care professionals' social media behavior and the underlying factors of social media adoption and use: quantitative study. J Med Internet Res 2018 Nov 07;20(11):e12035 [FREE Full text] [doi: 10.2196/12035] [Medline: 30404773]
2. Gholami-Kordkheili F, Wild V, Strech D. The impact of social media on medical professionalism: a systematic qualitative review of challenges and opportunities. J Med Internet Res 2013 Aug 28;15(8):e184 [FREE Full text] [doi: 10.2196/jmir.2708] [Medline: 23985172]
3. Chretien KC, Tuck MG. Online professionalism: a synthetic review. Int Rev Psychiatry 2015 Apr;27(2):106-117. [doi: 10.3109/09540261.2015.1004305] [Medline: 25804627]
4. Vukušić Rukavina T, Viskić J, Machala Poplašen L, Relić D, Marelić M, Jokic D, et al. Dangers and benefits of social media on e-professionalism of health care professionals: scoping review. J Med Internet Res 2021 Nov 17;23(11):e25770 [FREE Full text] [doi: 10.2196/25770] [Medline: 34662284]
5. ABIM Foundation, ACP-ASIM Foundation, European Federation of Internal Medicine. Medical professionalism in the new millennium: a physician charter. Ann Intern Med 2002 Feb 05;136(3):243-246 [FREE Full text] [doi: 10.7326/0003-4819-136-3-200202050-00012] [Medline: 11827500]
6. Cain J, Romanelli F. E-professionalism: a new paradigm for a digital age. Currents in Pharmacy Teaching and Learning 2009 Dec;1(2):66-70. [doi: 10.1016/j.cptl.2009.10.001]
7. Kitsis EA, Milan FB, Cohen HW, Myers D, Herron P, McEvoy M, et al. Who's misbehaving? Perceptions of unprofessional social media use by medical students and faculty. BMC Med Educ 2016 Feb 18;16:67 [FREE Full text] [doi: 10.1186/s12909-016-0572-x] [Medline: 26887561]
8. Pronk SA, Gorter SL, van Luijk SJ, Barnhoorn PC, Binkhorst B, van Mook WNKA. Perception of social media behaviour among medical students, residents and medical specialists. Perspect Med Educ 2021 Aug;10(4):215-221 [FREE Full text] [doi: 10.1007/s40037-021-00660-1] [Medline: 33826108]
9. Skrabal J. Factors and processes that influence e-professionalism among pre-licensure baccalaureate nursing students when utilizing social media (EdD Dissertation). ERIC. Omaha, NE: College of Saint Mary; 2017. URL: https://eric.ed.gov/?id=ED578156 [accessed 2024-02-12]
10. Viskić J, Jokić D, Marelić M, Machala Poplašen L, Relić D, Sedak K, et al. Social media use habits, and attitudes toward e-professionalism among medicine and dental medicine students: a quantitative cross-sectional study. Croat Med J 2021 Dec 31;62(6):569-579. [doi: 10.3325/cmj.2021.62.569] [Medline: 34981689]
11. White J, Kirwan P, Lai K, Walton J, Ross S. 'Have you seen what is on Facebook?' The use of social networking software by healthcare professions students. BMJ Open 2013;3(7):e003013 [FREE Full text] [doi: 10.1136/bmjopen-2013-003013] [Medline: 23883886]
12. Adilman R, Rajmohan Y, Brooks E, Urgoiti GR, Chung C, Hammad N, et al. Social media use among physicians and trainees: results of a national medical oncology physician survey. J Oncol Pract 2016 Jan;12(1):79-80, e52. [doi: 10.1200/JOP.2015.006429] [Medline: 26443837]
13. Avcı K, Çelikden SG, Eren S, Aydenizöz D. Assessment of medical students' attitudes on social media use in medicine: a cross-sectional study. BMC Med Educ 2015 Feb 15;15:18 [FREE Full text] [doi: 10.1186/s12909-015-0300-y] [Medline: 25890252]
14. Cain J, Scott DR, Akers P. Pharmacy students' Facebook activity and opinions regarding accountability and e-professionalism. Am J Pharm Educ 2009 Oct 01;73(6):104 [FREE Full text] [doi: 10.5688/aj7306104] [Medline: 19885073]
15. Fuoco M, Leveridge MJ. Early adopters or laggards? Attitudes toward and use of social media among urologists. BJU Int 2015 Mar;115(3):491-497. [doi: 10.1111/bju.12855] [Medline: 24981237]
16. Kenny P, Johnson IG. Social media use, attitudes, behaviours and perceptions of online professionalism amongst dental students. Br Dent J 2016 Nov 18;221(10):651-655. [doi: 10.1038/sj.bdj.2016.864] [Medline: 27857111]

XSL•FO
RenderX

17. Ness GL, Sheehan AH, Snyder ME. Graduating student pharmacists' perspectives on e-professionalism and social media: qualitative findings. J Am Pharm Assoc (2003) 2014;54(2):138-143. [doi: 10.1331/JAPhA.2014.13188] [Medline: 24632929]
18. Rocha PN, de Castro NAA. Opinions of students from a Brazilian medical school regarding online professionalism. J Gen Intern Med 2014 May;29(5):758-764 [FREE Full text] [doi: 10.1007/s11606-013-2748-y] [Medline: 24395103]
19. Kelley KA, Stanke LD, Rabi SM, Kuba SE, Janke KK. Cross-validation of an instrument for measuring professionalism behaviors. Am J Pharm Educ 2011 Nov 10;75(9):179 [FREE Full text] [doi: 10.5688/ajpe759179] [Medline: 22171107]
20. Freidson E. Profession of Medicine: A Study of the Sociology of Applied Knowledge. Chicago, IL: University of Chicago Press; 1970.
21. Evans L. Professionalism, professionality and the development of education professionals. British Journal of Educational Studies 2008 Mar;56(1):20-38. [doi: 10.1111/j.1467-8527.2007.00392.x]
22. George DR, Navarro AM, Stazyk KK, Clark MA, Green MJ. Ethical quandaries and Facebook use: how do medical students think they (and their peers) should (and would) act? AJOB Empirical Bioethics 2014 Apr 14;5(2):68-79. [doi: 10.1080/23294515.2013.864344]
23. Spielman AI, Fulmer T, Eisenberg ES, Alfano MC. Dentistry, nursing, and medicine: a comparison of core competencies. J Dent Educ 2005 Nov;69(11):1257-1271. [Medline: 16275689]
24. Viskić J, Jokić D, Marelić M, Machala Poplašen L, Relić D, Sedak K, et al. Social media use habits, and attitudes toward e-professionalism among medicine and dental medicine students: a quantitative cross-sectional study. Croat Med J 2021 Dec 31;62(6):569-579 [FREE Full text] [doi: 10.3325/cmj.2021.62.569] [Medline: 34981689]
25. Barlow CJ, Morrison S, Stephens HO, Jenkins E, Bailey MJ, Pilcher D. Unprofessional behaviour on social media by medical students. Med J Aust 2015 Dec 14;203(11):439. [doi: 10.5694/mja15.00272] [Medline: 26654611]
26. Gomes AW, Butera G, Chretien KC, Kind T. The development and impact of a social media and professionalism course for medical students. Teach Learn Med 2017;29(3):296-303. [doi: 10.1080/10401334.2016.1275971] [Medline: 28272900]
27. Diamantopoulos A, Siguaw JA. Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. Br J Manage 2006 Jun 16;17(4):263-282. [doi: 10.1111/j.1467-8551.2006.00500.x]
28. Diamantopoulos A, Winklhofer HM. Index construction with formative indicators: an alternative to scale development. Journal of Marketing Research 2018 Oct 10;38(2):269-277. [doi: 10.1509/jmkr.38.2.269.18845]
29. Khan EA, Dewan MNA, Chowdhury MMH. Reflective or formative measurement model of sustainability factor? A three industry comparison. Corporate Ownership and Control Journal 2016;13(2):83-92 [FREE Full text] [doi: 10.22495/cocv13i2p9]
30. Jarvis C, MacKenzie S, Podsakoff P. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. J Consum Res 2003 Sep;30(2):199-218. [doi: 10.1086/376806]
31. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
32. Knapp TR, Campbell-Heider N. Numbers of observations and variables in multivariate analyses. West J Nurs Res 1989 Oct;11(5):634-641. [doi: 10.1177/019394598901100517] [Medline: 2815734]
33. Dangers and Benefits of Social Networks: E-Professionalism of Health Care Professionals – SMePROF. University of Zagreb, School of Medicine. URL: https://mef.unizg.hr/en/znanost-2/istrazivanje/web-stranice-projekata/projekt-hrzz-smeprof/ [accessed 2024-02-18]
34. Viskić J, Marelić M, Machala Poplašen L, Vukušić Rukavina T. Differences between doctors of medicine and dental medicine in the perception of professionalism on social networking sites: the development of the e-professionalism assessment compatibility index (ePACI). BMC Med Ethics 2022 Dec 06;23(1):129 [FREE Full text] [doi: 10.1186/s12910-022-00870-0] [Medline: 36474221]
35. Marelić M, Viskić J, Poplašen LM, Relić D, Jokić D, Rukavina TV. Development and validation of scale for measuring attitudes towards e-professionalism among medical and dental students: SMePROF-S scale. BMC Med Educ 2021 Aug 23;21(1):445 [FREE Full text] [doi: 10.1186/s12909-021-02879-2] [Medline: 34425792]
36. Gerlich RN, Browning L, Westermann L. The Social Media Affinity Scale: implications For education. CIER 2010 Nov 15;3(11):35 [FREE Full text] [doi: 10.19030/cier.v3i11.245]
37. Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Qual Quant 2018;52(4):1957-1976 [FREE Full text] [doi: 10.1007/s11135-017-0584-6] [Medline: 29937587]
38. Funkhouser E, Vellala K, Baltuck C, Cacciato R, Durand E, McEdward D, National Dental PBRN Collaborative Group. Survey methods to optimize response rate in the national dental practice-based research network. Eval Health Prof 2017 Sep;40(3):332-358 [FREE Full text] [doi: 10.1177/0163278715625738] [Medline: 26755526]
39. Klabunde CN, Willis GB, McLeod CC, Dillman DA, Johnson TP, Greene SM, et al. Improving the quality of surveys of physicians and medical groups: a research agenda. Eval Health Prof 2012 Dec;35(4):477-506. [doi: 10.1177/0163278712458283] [Medline: 22947596]
40. Barnhart BJ, Reddy SG, Arnold GK. Remind me again: physician response to web surveys: the effect of email reminders across 11 opinion survey efforts at the American Board of Internal Medicine from 2017 to 2019. Eval Health Prof 2021 Sep;44(3):245-259. [doi: 10.1177/01632787211019445] [Medline: 34008437]

XSL•FO
RenderX

41. Cunningham CT, Quan H, Hemmelgarn B, Noseworthy T, Beck CA, Dixon E, et al. Exploring physician specialist response rates to web-based surveys. BMC Med Res Methodol 2015 Apr 09;15:32 [FREE Full text] [doi: 10.1186/s12874-015-0016-z] [Medline: 25888346]
42. Aitken C, Power R, Dwyer R. A very low response rate in an on-line survey of medical practitioners. Aust N Z J Public Health 2008 Jun;32(3):288-289 [FREE Full text] [doi: 10.1111/j.1753-6405.2008.00232.x] [Medline: 18578832]
43. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016;9:211-217 [FREE Full text] [doi: 10.2147/JMDH.S104807] [Medline: 27217764]
44. Vukušić Rukavina T, Machala Poplašen L, Majer M, Relić D, Viskić J, Marelić M. Defining Potentially Unprofessional Behavior on Social Media for Health Care Professionals: Mixed Methods Study. JMIR Med Educ 2022 Aug 09;8(3):e35585 [FREE Full text] [doi: 10.2196/35585] [Medline: 35758605]

Abbreviations

ABIM: The American Board of Internal Medicine
CCDM: Croatian Chamber of Dental Medicine
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
CHIF: Croatian Health Insurance Fund
CMC: Croatian Medical Chamber
DMD: doctor of dental medicine
HCP: health care professional
HIPAA: Health Insurance Portability and Accountability Act
MD: doctor of medicine
MIMIC: multiple indicators multiple causes
SNS: social networking site
VIF: variance inflation factor

Edited by T Leung, T de Azevedo Cardoso; submitted 21.06.23; peer-reviewed by C Asaad, A Jeroncic; comments to author 25.08.23; revised version received 29.09.23; accepted 28.12.23; published 27.02.24.

Please cite as:
Marelić M, Klasnić K, Vukušić Rukavina T
Measuring e-Professional Behavior of Doctors of Medicine and Dental Medicine on Social Networking Sites: Indexes Construction With Formative Indicators
JMIR Med Educ 2024;10:e50156
URL: https://mededu.jmir.org/2024/1/e50156
doi:10.2196/50156
PMID:38412021

©Marko Marelić, Ksenija Klasnić, Tea Vukušić Rukavina. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Exploring Anesthesia Provider Preferences for Precision Feedback: Preference Elicitation Study

Zach Landis-Lewis[1], MLIS, PhD; Chris A Andrews[2], PhD; Colin A Gross[3], BSc; Charles P Friedman[1], PhD; Nirav J Shah[4], MD

1
2
3
4

Corresponding Author:
Zach Landis-Lewis, MLIS, PhD

Abstract

Background: Health care professionals must learn continuously as a core part of their work. As the rate of knowledge production in biomedicine increases, better support for health care professionals' continuous learning is needed. In health systems, feedback is pervasive and is widely considered to be essential for learning that drives improvement. Clinical quality dashboards are one widely deployed approach to delivering feedback, but engagement with these systems is commonly low, reflecting a limited understanding of how to improve the effectiveness of feedback about health care. When coaches and facilitators deliver feedback for improving performance, they aim to be responsive to the recipient's motivations, information needs, and preferences. However, such functionality is largely missing from dashboards and feedback reports. Precision feedback is the delivery of high-value, motivating performance information that is prioritized based on its motivational potential for a specific recipient, including their needs and preferences. Anesthesia care offers a clinical domain with high-quality performance data and an abundance of evidence-based quality metrics.

Objective: The objective of this study is to explore anesthesia provider preferences for precision feedback.

Methods: We developed a test set of precision feedback messages with balanced characteristics across 4 performance scenarios. We created an experimental design to expose participants to contrasting message versions. We recruited anesthesia providers and elicited their preferences through analysis of the content of preferred messages. Participants additionally rated their perceived benefit of preferred messages to clinical practice on a 5-point Likert scale.

Results: We elicited preferences and feedback message benefit ratings from 35 participants. Preferences were diverse across participants but largely consistent within participants. Participants' preferences were consistent for message temporality (α=.85) and display format (α=.80). Ratings of participants' perceived benefit to clinical practice of preferred messages were high (mean rating 4.27, SD 0.77).

Conclusions: Health care professionals exhibited diverse yet internally consistent preferences for precision feedback across a set of performance scenarios, while also giving messages high ratings of perceived benefit. A "one-size-fits-most approach" to performance feedback delivery would not appear to satisfy these preferences. Precision feedback systems may hold potential to improve support for health care professionals' continuous learning by accommodating feedback preferences.

(JMIR Med Educ 2024;10:e54071) doi:10.2196/54071

KEYWORDS

audit and feedback; dashboard; motivation; visualization; anesthesia care; anesthesia; feedback; engagement; effectiveness; precision feedback; experimental design; design; clinical practice; motivational; performance; performance data

Introduction

Health care professionals must learn continuously as a core part of their work. As the rate of knowledge production in biomedicine increases, better support for continuous learning is needed [1]. Feedback about care quality and outcomes is pervasive in health systems and widely considered to be essential for learning that drives improvement. Clinical performance feedback is one form of feedback that is commonly delivered to health care professionals in clinical quality dashboards and reports. However, engagement with these resources is generally low, and their impact has been less than optimal [2-5], resulting in missed opportunities to improve the quality and safety of care. A large proportion of randomized controlled trials of feedback interventions (also known as *audit and feedback*) show limited influence on clinical practice [5]. Moreover, what is

XSL•FO
RenderX

considered as best practice for feedback interventions has not changed meaningfully for decades, even after hundreds of trials and repeated calls for new approaches to feedback interventions [6-8].

To our knowledge, most clinical performance feedback interventions use a "one-size-fits-most" approach to both the prioritization of performance information and its visual display as feedback, with the same metrics and visualizations being sent to all recipients. One-size-fits-most feedback may not be effective due to a host of characteristics such as individuals' knowledge, skills, and motivational orientation to their work [2,3,9-11]. Methods used by coaches, educators, and quality improvement facilitators to deliver feedback suggest that these factors are important [2,12,13]. Furthermore, in the context of routine feedback interventions (eg, with monthly or quarterly measurement cycles), the value of performance information [14-16] may be reduced when performance is stable, but feedback interventions are not commonly prioritized accordingly. Given the increasing use and digitization of performance measures and clinical quality dashboards [17,18], health care systems need to understand how to better accommodate health care professionals' feedback preferences and the corresponding value of performance information.

Precision feedback is feedback that has been prioritized based on its motivational potential for a specific recipient [19-23]. Using this approach, high-value feedback messages can be selected to enhance reports and emails, such as "You reached the top performer benchmark" and "Your performance dropped below the peer average." The potential impact of precision feedback increases with greater variability and differences in individuals' knowledge, skills, and motivational orientation, but these differences and their interactions are not well understood, as studies of health care professionals' feedback preferences appear to be scarce. Qualitative studies have explored feedback preferences by asking participants to discuss their experiences with prior feedback; for example, they can be prompted by a published feedback report [24] or a performance report belonging to the participant or their organization [25]. Quantitative preference elicitation methods have been used extensively in health decision-making [26,27], but uncertainty about the measurement properties of preferences contributes to controversy around their use [28]. To our knowledge, no instruments of health care professional feedback preferences with validity evidence have been developed. To begin to explore and understand these differences, we designed a preference elicitation study for motivating performance information and its display format.

We conducted this study in the context of anesthesia care quality improvement. In this context, data generated about care processes are produced primarily by anesthesia machines that report the administration of anesthetics and the patient's corresponding state with relatively high accuracy and reliability. Attribution of performance to individual anesthesia providers is feasible due to their authenticated use of an anesthesia machine for each operative case. A national-scale quality improvement consortium, the Multicenter Perioperative Outcomes Group (MPOG) [29,30], has developed approximately 70 performance measures for anesthesia care quality and outcomes. Feedback is delivered through its infrastructure via monthly emails and a clinical quality dashboard to more than 8000 health care professionals in more than 20 US states. Thus, a relatively large set of measures are routinely assessed using high-quality clinical data, representing performance information that health care professionals have limited natural sources for across their patient populations.

Multiple types of motivation are recognized as mechanisms through which feedback influences performance [2,10,11,31-33]. These various types of motivation can be understood as a consequence of the cognitive processing of performance information. We use the term *motivating performance information* to mean performance information that has the potential to motivate a feedback recipient through a known mechanism of action (Table 1). A key type of motivating performance information is a *comparison* that represents a discrepancy between the performance level of a feedback recipient and some *comparator* [22]. There are multiple types of comparators, including *benchmarks* having a performance level that is determined by a population-based analysis. Benchmarks are commonly calculated as a summary statistic of top performers, such as choosing the performance level for a population that occurs at the 90th percentile, or the achievable benchmark of care (ABC) method [34]. Another type of comparator is an *explicit target*, including goals or standards that set expectations for attaining a specific performance level that is not necessarily dependent upon peers or another reference group's performance [35]. The choice of comparators can result in the use of alternate mechanisms of motivation, such as motivation related to social norms versus personal goal-setting. Another key type of motivating information is *trends* that represent change in performance (getting better or worse) [22]. Comparisons and trends may co-occur in performance data to represent an *achievement*, such as reaching a goal, or a *loss*, such as losing top-performer status [22].

XSL•FO
RenderX

Table . Glossary.

Term	Description	Source
Performance information	Information about measures, levels, time intervals, comparators, and feedback recipient	[20,22]
Feedback	Information about performance that can guide future action	[36]
Feedback recipient	A person, team, or organization to whom a feedback intervention is directed	[22]
Precision feedback	Feedback that is prioritized according to its motivational potential for a specific recipient	[23]
Motivating performance information	Performance information that holds motivational potential	[23]
Comparison	Motivating performance information that is about a discrepancy between the performance levels of a feedback recipient and a comparator	[22]
Trend	Motivating performance information that is about a change in performance	[22]
Achievement	Motivating performance information that is about a change from a negative comparison to a positive comparison	[22]
Loss	Motivating performance information that is about a change from a positive comparison to a negative comparison	[22]
Comparator	Information that is used to identify a discrepancy with the performance level of a feedback recipient	[22]
Benchmark	A comparator with a performance level that is calculated from the performance of other health professionals or peers	[22,35]
Explicit target	A comparator with a performance level that is explicitly expected	[22,35]
Time point information	Performance information that is about a single time interval	—[a]
Time series information	Performance information that is about multiple time intervals	—[a]
Causal pathway model	A specification of influential elements in a causal process, including preconditions, mechanisms, moderators, and outcomes	[37]

[a]Not available.

Comparisons and trends are represented using a wide range of visualizations in clinical quality dashboards and feedback reports [20]. These visualizations vary both in their content, such as the use of measures, comparators, and duration of time intervals, as well as the display format, such as bar charts, line charts, and tables to represent performance data. A review of published displays from feedback reports and dashboards identified 6 unique combinations of visualized performance information content [20]. For example, feedback displays vary in the number of performance measures, time intervals, and comparators that they visualize.

The display of feedback is theorized as one of many factors affecting the success of clinical performance feedback in Clinical Performance Feedback Intervention Theory (CP-FIT) [38], a leading theory of audit and feedback. Motivating performance information in clinical performance data concerns configurations of types of feedback display, but is also closely related to CP-FIT's *goal* construct, which concerns the importance and relevance of feedback to health care professionals. Precision feedback may contribute to additional CP-FIT constructs, including health professional characteristics (knowledge and skills in quality improvement), feedback delivery (function), and implementation process (adaptability and ownership).

To understand anesthesia provider preferences for motivating performance information and feedback display format, we investigated the following four research questions:

1. To what extent do anesthesia providers' selected messages reveal an overall preference for
    a. messages containing time series versus time point information (temporality)?

XSL•FO
RenderX

b. messages relative to benchmarks versus explicit performance targets (basis of comparison)?
c. messages formatted as bar charts versus line charts and text only (display format)?

2. How consistent are individual anesthesia provider preferences?
3. To what extent do anesthesia provider preferences depend on performance level, trend, and their professional background?
4. To what extent are preferred feedback messages perceived to hold potential to improve future clinical practice?

Methods

Overview

To address these questions, we developed a test set of feedback messages that a software application could generate. We formatted these as brief email messages, but designed them as "least common denominator" content that could also be delivered via other channels for feedback, such as clinical quality dashboards.

In the absence of instruments with validity evidence for assessing health care professional feedback preferences, we created an experimental design to elicit preferences that would expose participants, who were anesthesia providers, to contrasting message versions. To enable measurement validity assessment, we developed performance scenarios in which the same motivating performance information and display characteristics could be repeated in contrasting messages.

Ethical Considerations

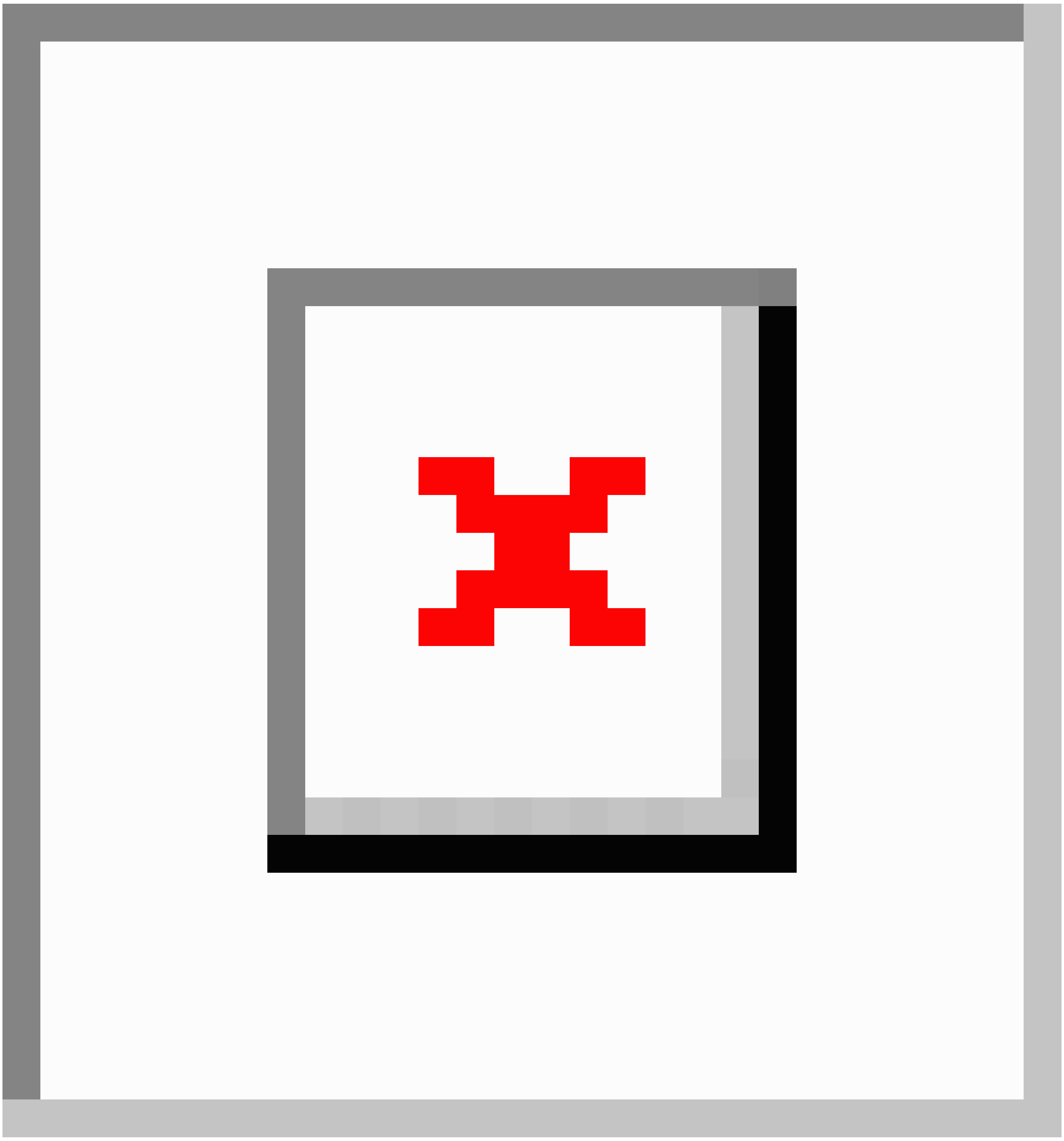
This study was approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board (IRB-HSBS HUM00167426). All participants provided consent to participate and were informed about the ability to opt out of the study. No participant identifiers were collected with the research data for this study, preventing the linking of participants' responses with their identities. No incentives for participation were provided. We offered participants an opportunity to receive a copy of the study results upon completion.

Email Test Set Development

We developed the email message test set iteratively in three phases: (1) knowledge modeling, (2) display format development, and (3) message set development (Figure 1).

XSL•FO
RenderX

Figure 1. Development of a precision feedback email test set.



Phase 1: Knowledge Modeling

In the first phase we modeled knowledge about the elements of performance information, types of motivating information, and the influence of motivating performance information (Figure 1). We iteratively refined a model of the elements of performance information through an analysis of published feedback reports [20], resulting in the identification of 5 key elements: measures, recipients, comparators, performance levels, and time intervals. We developed a model of motivating information that combined the 5 elements of performance information into types of motivating information, including comparisons, trends, achievement, and loss. Each type of motivating information was defined using the elements of performance information. For example, a comparison (a kind of motivating performance information) was defined as a discrepancy between the performance levels of a feedback recipient and a comparator.

Through modeling types of motivating performance information, we recognized that the choice of comparator could affect which type of motivation was used to influence a recipient. For example, choosing a 90th-percentile peer benchmark as a comparator does not necessarily leverage motivation from goal-setting when recipients do not form an intention to reach the benchmark as their personal goal. By inviting anesthesia providers to set goals, feedback that shows performance improving toward a goal may leverage motivation arising from a desire for growth and achievement, rather than a desire for

XSL•FO
RenderX

safety and avoidance of harm. These sources of motivation can differentially interact with the feedback sign (ie, valence) to have counterintuitive effects, such as goal abandonment, relaxation, or the delivery of low-value feedback [2,10].

To understand how different types of motivating performance information might relate to theoretical mechanisms of influence, we created causal pathway models [37] for each type of motivating information with benchmark and explicit target comparators (Multimedia Appendix 1). For example, in one causal pathway we modeled the expected influence of a feedback intervention that combines three elements of a recipient's performance: (1) performance below a comparator (low performance level), (2) a benchmark (such as a peer average), and (3) performance getting better (improving trend). This pathway could represent the influence of precision feedback emails that show performance approaching a peer average, which could indicate to recipients that efforts to improve performance appear to be succeeding. Based on the theoretical construct of *positive velocity* [31] (ie, showing performance improvement), this causal pathway (which we named *social approach* due to the recipient reducing a performance gap with a peer benchmark) uses motivation as a mechanism of action, through which a feedback recipient may decide to increase or sustain effort to improve performance.

We drafted and refined example messages for each type of motivating information. For the causal pathway *social approach*, an example message is "Your performance is approaching the benchmark." We implemented the causal pathway models in computer-interpretable form in a knowledge base to enable automation of the processing of performance information to identify motivating information in a precision feedback system.

Phase 2: Display Format Development

In the second phase we developed display formats for motivating information in the body of an email message. We selected visualizations (ie, bar charts and line charts) common in health care organizations so that a familiar format would convey the minimal amount of information necessary for each causal pathway. We developed software to generate visualizations within an email message using R (version 4.3.3; R Foundation for Statistical Computing). We included the absence of a visualization (ie, text only) to accommodate recipient preferences for concise, text-based communication (Figure 1).

Phase 3: Message Set Development

In the third and final phase we created a test set of email messages with balanced characteristics of motivating information and display formats. We began by creating four performance scenarios with alternate performance levels (high vs low) and trends (improvement vs worsening vs stable): (1) improvement to a high level, (2) worsening to a low level, (3) consistently high (stable) performance, and (4) consistently low (stable) performance (Table 2). In all scenarios, the recipient's performance could be compared with either the peer average (benchmark comparator) or an organizational goal (explicit target comparator). We set the recipient's performance level to have the same relationship with each comparator (better or worse), enabling either comparator to be displayed while maintaining balance with other elements.

XSL•FO
RenderX

Table . Precision feedback email message test set specification.

Performance data scenario			Motivating information characteristics		Key message	Display format	
Level	Trend	Performance description	Temporality	Comparator		Shown to group A	Shown to group B
High	Improving	Performance level moves above comparators	Time series	Benchmark	You have become a top performer	Line chart	Bar chart
				Explicit target	You reached the goal	Bar chart	Line chart
			Time point	Benchmark	You are a top performer	Bar chart	Text only
				Explicit target	Congratulations on your high performance	Text only	Bar chart
Low	Worsening	Performance level moves below comparators	Time series	Benchmark	You are no longer a top performer	Bar chart	Line chart
				Explicit target	Your performance dropped below the goal	Line chart	Bar chart
			Time point	Benchmark	You are not a top performer	Text only	Bar chart
				Explicit target	You may have an opportunity to improve	Bar chart	Text only
High	No change	Performance level is consistently above comparators	Time series	Benchmark	You are a consistent top performer	Bar chart	Line chart
				Explicit target	Your performance is consistently high	Line chart	Bar chart
			Time point	Benchmark	You are a top performer	Text only	Bar chart
				Explicit target	Congratulations on your high performance	Bar chart	Text only
Low	No change	Performance level is consistently below comparators	Time series	Benchmark	Your performance has remained low	Line chart	Bar chart
				Explicit target	Your performance has not improved	Bar chart	Line chart
			Time point	Benchmark	You are not a top performer	Bar chart	Text only
				Explicit target	You may have an opportunity to improve	Text only	Bar chart

We selected types of motivating information and their example messages across three characteristics: (1) performance temporality (time series vs time point), (2) performance comparison basis (benchmark vs explicit target), and (3) performance display format (bar chart or other). We selected the bar chart format as a key display format because of its common use in health care organizations. We further divided the *other* display format into *line chart* and *text only*. We composed emails with example messages from each type, based on a single quality measure (Avoiding postoperative nausea and vomiting [PONV-03]) for anesthesia providers. The resulting emails contained information from the same performance scenarios, but not all information from each scenario was provided in each message. For example, of the 4 emails that each participant read in each scenario, 2 messages contained a

XSL•FO
RenderX

goal comparator (explicit target), while the other 2 messages showed a peer benchmark comparator instead.

Study Design

We designed a within-subjects, repeated measures study of anesthesia provider preferences for precision feedback using a test set of prototype email messages printed on paper. We created 2 versions of the test set with alternate display formats for each message (group A vs group B) to enable randomization of the pairing of display format with motivating information. We created a document containing all of the email messages in the test set (Multimedia Appendix 2). We printed paper copies of the messages and organized them into packets in varying order for a paper card selection task. Based on our experience, we estimated that a sample of more than 30 participants would provide adequate power to detect meaningful differences in summary statistics and internal consistency of preferences.

Population and Setting

We recruited anesthesia providers from a single academic medical center in the midwestern United States. Anesthesiologists (physicians) and certified registered nurse anesthetists (CRNAs) were eligible to participate. A member of the study team recruited anesthesia provider participants by email. All participants received monthly anesthesia provider feedback emails from MPOG.

Data Collection

Upon enrollment, we scheduled a 15-minute proctored video call with each participant and sent them a paper packet with email prototypes before the call. Participants were randomized to receive a paper packet of messages from either group A or group B of the message test set, each of which contained 16 email messages grouped in 4 packets of 4. Each packet of 4 messages contained alternate message formats for 1 of 4 performance scenarios, with balanced message formats and performance information across the 4 scenarios. We created a questionnaire to collect data from participants about their preferred emails using Qualtrics (Qualtrics, Inc). We created 2 versions of the questionnaire (A and B), 1 for each message group to be used based on the participant's random assignment at the time of enrollment. At the time of enrollment, we also instructed participants to have a desk space or table available for placing printed email messages in front of them, and to wait to open the packets until asked to do so during the video call.

Preference Elicitation and Message Usability Assessment

At the start of the proctored video call, a research team member introduced the study, confirmed the participant's preparation, and provided a link to the questionnaire. During the completion of the questionnaire, the participant repeated a preferred email message selection task 4 times, following the instructions in their packet, once for each performance data scenario. The questionnaire software randomized the scenario presentation order. We described the scenarios as hypothetical performances that the participant could imagine as being their own. At the start of each scenario, participants were asked to find the corresponding set of emails, identified with a cover sheet. Participants were then asked to lay out all 4 of the printed email messages for that scenario in front of them. Next, participants read each message and selected their preferred message. After selecting a preferred message, participants responded to the following statement: "I gained information from this email that would benefit my practice." We adapted this question from an instrument with good validity evidence for assessing the usability of feedback displays [39]. Responses were collected on a 5-point Likert scale, ranging from strongly disagree to strongly agree. The survey questions did not ask directly about preference for information content or display format. Instead, participants' preferences were inferred through the types of content and display format that the selected message contained. After participants completed the questionnaire, we conducted brief interviews and collected qualitative data that were analyzed separately and will be reported elsewhere.

Analysis

To identify preferences, we analyzed 2 characteristics of the selected messages: motivating information (including temporality type and comparator type) and display format. We summed the selected messages with each type of motivating information and display format and calculated descriptive statistics for these sums (Q1). To investigate the consistency of participants' preferences, we calculated the Cronbach α for each preference characteristic in participants' selected messages across the 4 performance scenarios (Q2). We used descriptive statistics to assess relationships between participants' preferences and the characteristics of the 4 performance scenarios, including performance level (high vs low) and trend presence (present vs absent). Similarly, we considered relationships between participants' preferences and their professional background using descriptive statistics (Q3).

To understand participants' perceptions of the potential benefit of precision feedback to their clinical practice, we analyzed ratings of perceived benefit for selected messages using descriptive statistics (Q4). We conducted analyses using R and Google Sheets (Google LLC).

Results

We recruited 35 anesthesia providers, including 18 anesthesiologists and 17 CRNAs (Table 3). All participants completed all message selection tasks, resulting in the selection of 140 preferred precision feedback messages.

XSL•FO
RenderX

Table . Study participant characteristics (N=35).

Characteristics		Participants, n (%)
Professional role		
	Anesthesiologist	18 (51)
	Certified registered nurse anesthetist	17 (49)
Race/ethnicity		
	African-American	0 (0)
	Asian	2 (6)
	Hispanic	0 (0)
	White	31 (89)
	Other	2 (6)
Gender		
	Female	19 (54)
	Male	16 (46)
	Nonbinary/other	0 (0)

To What Extent Do Anesthesia Providers' Selected Messages Reveal an Overall Preference for Temporality (Q1a), Basis of Comparison (Q1b), and Display Format (Q1c)?

An overall preference for multiple time intervals (ie, time series) was apparent, with 110 of 140 (79%) messages being selected over those with a single time interval (ie, time point) (Q1a). Preferences for display format were highly varied, with selected messages being equally distributed between bar charts versus other formats (Table 4) (Q1c). Preferred messages were also highly varied in their comparators, with 74 of 140 (53%) preferred cards containing explicit target comparators (ie, organizational goals not dependent on population performance) (Q1b), but our assessment of the consistency suggests that the comparator result was not reliable as a preference characteristic (see Q2 below).

Table . Characteristics of preferred precision feedback messages.

Message characteristic and subtype		Preferred messages (n=140), n (%)	Message characteristic preference (n=4), mean (SD)	α
Temporality				.85
	Time series	110 (79)	3.14 (1.38)	
	Time point	30 (21)	0.86 (1.38)	
Comparators				–.40
	Benchmark	66 (47)	1.89 (0.87)	
	Explicit target	74 (53)	2.11 (0.87)	
Display format				.80
	Bar chart	70 (50)	2.00 (1.61)	
	Other display	70 (50)	2.00 (1.61)	

How Consistent Are Individual Anesthesia Provider Preferences (Q2)?

Participants' preferences were consistent for temporality (α=.85) and display format (α=.80). For performance comparators, participants' selected messages were negatively correlated (α=−.40), indicating an absence of consistency, perhaps from an incorrect measurement model [40]. We consider this result to be an artifact of the study design, given that our message test set balanced several characteristics and created opportunities to select them in combination. We anticipate that comparators were not salient for participants, relative to the visual display and temporality characteristics; therefore, we are unable to draw conclusions about preferences for comparators.

To What Extent Do Anesthesia Provider Preferences Depend on Performance Level and Trend and Their Professional Background (Q3)?

Participant preferences for temporality and display format did not appear to depend on the messages' performance level, with relatively similar means for the selection of each type of message content. Similarly, these preferences did not appear to vary with the presence or absence of performance trends (Table 5).

XSL•FO
RenderX

Table . Precision feedback preferences by performance scenario characteristics.

	Temporality preference, mean (SD)		Comparator preference, mean (SD)		Display format preference, mean (SD)			
	Time series	Time point	Benchmark	Explicit target	Bar chart	Other display	Other display: line chart	Other display: text only
Level: high	1.60 (0.69)	0.40 (0.69)	1.09 (0.56)	0.91 (0.56)	0.94 (0.87)	1.06 (0.87)	0.77 (0.88)	0.29 (0.62)
Level: low	1.54 (0.78)	0.46 (0.78)	0.80 (0.53)	1.20 (0.53)	1.06 (0.87)	0.94 (0.87)	0.66 (0.84)	0.29 (0.62)
Trend present	1.66 (0.68)	0.34 (0.68	1.06 (0.54)	0.95 (0.54)	1.09 (0.89)	0.91 (0.89)	0.69 (0.83)	0.23 (0.60)
Trend absent	1.49 (0.78)	0.51 (0.78)	0.83 (0.51)	1.17 (0.51)	0.91 (0.85)	1.09 (0.85)	0.74 (0.85)	0.34 (0.64)

Preferences for temporality and display format varied with participants' professional background (Table 6). Some professional role-based differences in means were apparent, such as a higher preference for time point messages among CRNAs than anesthesiologists (mean message characteristics preference 1.18, SD 1.59 vs mean message characteristic preference 0.56, SD 1.10). However, a majority of CRNAs preferred time series messages, and all message characteristics were repeatedly observed in selections by participants from both professional background–based groups.

Table . Precision feedback preferences by professional background.

	Temporality preference, mean (SD)		Comparator preference, mean (SD)		Display format preference, mean (SD)			
	Time series	Time point	Benchmark	Explicit target	Bar chart	Other display	Other display: line chart	Other display: text only
Anesthesiologist	3.44 (1.10)	0.56 (1.10)	1.83 (0.79)	2.17 (0.79)	2.28 (1.60)	1.72 (1.60)	1.33 (1.68)	0.39 (0.78)
Certified registered nurse anesthetist	2.82 (1.59)	1.18 (1.59)	1.94 (0.97)	2.06 (0.97)	1.71 (1.61)	2.29 (1.61)	1.53 (1.55)	0.76 (1.48)

To What Extent Are Preferred Feedback Messages Perceived to Hold Potential to Improve Future Clinical Practice (Q4)?

Participants' ratings of perceived benefit from all precision feedback messages were positive, with a mean rating of 4.27 (SD 0.77). Although positive overall, the anesthesiologists' ratings were lower than the CRNAs' ratings (mean rating 4.08, SD 0.85 vs mean rating 4.47, SD 0.61). Ratings for messages did not appear to vary across performance levels or with trends (Table 7). Average ratings of perceived benefit were similar across message content characteristics. One exception to this was for explicit target comparators, which appeared to receive slightly higher ratings (mean rating 4.38, SD 0.7) over benchmark comparators (mean rating 4.15, SD 0.83).

XSL•FO
RenderX

Table . Perceived benefit of selected messages.

Characteristics			Mean rating (SD)
Participant professional background			
	Anesthesiologist		4.08 (0.85)
	Certified registered nurse anesthetist		4.47 (0.61)
Performance scenario			
	Performance level		
		High performance	4.23 (0.76)
		Low performance	4.31 (0.77)
	Performance trend		
		Trend present	4.34 (0.72)
		Trend absent	4.2 (0.81)
Message content			
	Temporality		
		Time series	4.27 (0.81)
		Time point	4.27 (0.58)
	Comparator		
		Benchmark	4.15 (0.83)
		Explicit target	4.38 (0.7)
	Display format		
		Bar chart	4.27 (0.76)
		Other display	4.27 (0.78)
		Other display: Line chart	4.28 (0.83)
		Other display: text only	4.25 (0.64)

Discussion

Principal Results

In this study, we found that anesthesia provider preferences for motivating information and display format varied, which suggests that individual difference characteristics may represent a barrier to improving the effectiveness of feedback interventions. Across a set of 4 diverse performance scenarios, we observed preference variability that precision feedback could better address than one-size-fits-most feedback in this anesthesia provider population.

We observed consistency in participant preferences for the temporality of motivating information and for display format. Even though a large majority of participants preferred messages with time-series information, the participants who preferred time-point messages reliably selected them. The consistency of preferences for display format was similar, and also more varied, with exactly half of participants choosing bar charts over other visual displays. We also did not observe differences in preferences associated with performance scenario characteristics or professional background that could be used to design one-size-fits-most feedback interventions.

While participants exhibited diverse preferences, their ratings of the benefit of the messages were consistently high across performance scenarios. These findings suggest that anesthesia providers would welcome the enhancement of feedback interventions with precision feedback that prioritizes motivating information. These findings are important because they point to a possible approach for improving audit and feedback that can leverage both high and low performance, as well as increasing or decreasing trends, to prioritize performance feedback.

To our knowledge, this is the first quantitative study of preferences for clinical performance feedback. As an exploratory study, the findings primarily demonstrate the existence of differences in preferences for feedback, rather than speaking to the significance of their role in the success of clinical performance feedback. Our findings are related to CP-FIT, which recognizes that health professional knowledge and skills for engaging with feedback can be important factors for the success of feedback [38]. Differences in feedback preferences could be driven by differences in health care professionals' knowledge and skills related to the interpretation of performance data. For example, participants' variable and consistent selection of messages could be related to their graph literacy skills [41,42]. Precision feedback could be used to accommodate these and other individual differences by enabling health professionals to configure their feedback delivery and display, which further holds potential to increase feelings of ownership of feedback.

XSL•FO
RenderX

By prioritizing motivating information according to recipients' preferences, precision feedback could be a strategy for reducing the cognitive load required by health professionals to recognize and assess the priority of learning opportunities. Precision feedback has also potential to improve feedback cycle completion by delivering information that is more likely to be perceived and accepted, resulting in increased formation of intentions to sustain or improve performance. In terms of CP-FIT, precision feedback can be understood as an approach for prioritization of feedback messages that are more likely to result in successful completion of the feedback cycle.

Our findings are aligned with the idea that positive feedback can be effective for learning and improvement [13], as well as sustainment of high performance. It is noteworthy that participants rated precision feedback messages as beneficial even when performance was high, such as the messages "you are a top performer" or "you reached the goal." This finding points to the possibility that a key function of feedback may be to motivate recipients through appreciation of accomplishments [43], including recognition of high performance, in addition to motivating recipients to learn to improve.

Limitations

As an exploratory study for a novel type of feedback intervention, there are several important limitations for this study. The poor consistency of preferences demonstrated for performance comparators suggests that participants did not meaningfully differentiate between peer-based benchmarks and explicit targets, as presented in the message test set. This may be a function of the labels used for the comparators message test set, and during the study we discovered that some of the printed messages contained the abbreviation "ave" instead of "avg" for the peer average comparator. Competing explanations are that (1) anesthesia providers equated the value of both comparator types or did not perceive them as fundamentally different, and (2) that this characteristic was less salient than the others, such that its significance was negligible.

Using performance scenarios based on synthetic performance data may have introduced bias in participants' responses. However, the consistency of participant preferences for temporality of motivating information and display format suggests that this bias was not significant. Nevertheless, our study design assessed preferences within types of motivating information (eg, high and improving performance or low and worsening performance) that were presented with unambiguous motivating information, such as trends showing marked improvement or worsening. As such, our results do not address the appropriateness of using performance scenarios to elicit the strength of anesthesia provider preferences directly; rather, they primarily demonstrate the existence of individual differences as an exploration of factors that may moderate the influence of feedback on health care professional learning and improvement.

We asked participants to rate the perceived benefit of messages that they had already selected as their preferred message, which may have resulted in positively biased ratings. Furthermore, we used a single performance measure for all messages (avoiding postoperative nausea and vomiting) that may not be representative of other performance measures, both in terms of perceived benefit and preferences for motivating information. We did not evaluate feedback about clinical outcome measures, which may have resulted in a different preference profile across this population. We also did not evaluate participants' skills or knowledge to engage effectively in feedback, which is a recognized factor [38] that may have resulted in further insight into participant preferences.

Additional limitations include the context and nature of the preference elicitation task, which was done in a video call with paper prototypes and thus differed from the context of email use in health care organizations. When designing this study, we chose to use email messages printed on paper because we could not identify a remote, video call–proctored approach that would allow participants to consider 4 different messages types in the same field of view on their personal or work computer without a risk of technical complications from participants' particular computer monitor and device configurations.

Our model of preferences in this study was linear and static and assumed that available information was complete, but anesthesia provider preferences may be nonlinear, dynamic, and depend on missing information that we did not consider. When designing the test set of messages, we paired the text-only display format consistently with time-point information, and line charts with the time-series format. As such, preferences for line charts and text-only display formats were not independent from temporality. We recruited anesthesia providers from a single academic institution whose population is not necessarily representative of other anesthesia provider populations. We did not recruit any anesthesia providers who identified as Black or Hispanic, increasing the likelihood that our results are racially and ethnically biased toward the perspectives of anesthesia providers who identify as White and non-Hispanic. In spite of all of these limitations, we note that the variability that we observed demonstrates that preferences were nonuniform in this small population, which suggests that a one-size-fits-all solution may be inadequate for feedback reporting to anesthesia providers more generally.

Future Studies

We anticipate that preference clusters may exist and may be identifiable in studies that are better powered to detect such differences. Such clusters could be used to develop profiles for precision feedback, such as profiles for anesthesia providers who prefer text-only messages about low performance or those who prefer visualization of performance changes (ie, trends) using time-series displays in line charts. Future studies may be able to detect preference clusters to better understand the diversity of preferences for performance feedback across a larger anesthesia provider population that is more racially, ethnically, and geographically diverse. Furthermore, we would welcome studies that aim to better understand the diversity of anesthesia provider preferences in association with additional anesthesia provider characteristics, such as duration of professional experience, clinical setting, and organization type.

Conclusions

Clinical performance feedback to health care professionals has potential to support continuous learning and influence practice,

XSL•FO
RenderX

but this potential is frequently not achieved. By prioritizing motivating performance information based on the preferences and needs identified for a health care professional population, precision feedback may increase the effectiveness of clinical performance feedback for health care professionals' continuous learning and resulting quality improvement. Among a sample of anesthesia providers, preferences for precision feedback were varied, yet consistent within participants. Furthermore, participants' perceived benefits of precision feedback messages were observed to be high across a diverse set of performance scenarios. Based on these findings, it appears that precision feedback holds potential to improve support for health care professionals' continuous learning.

Acknowledgments

The authors would like to thank Ms Astrid Fishstrom for contributing to the investigation. We are grateful to Dr Anne Sales, Dr John Rincón-Hekking, Ms Dahee Lee, Mr Cooper Stansbury, and Ms Veena Panicker for their contributions to foundational work for this study. We thank the NIH (National Institutes of Health) National Library of Medicine for funding this research. This study was funded by a grant from the NIH National Library of Medicine (K01LM012528; principal investigator ZLL) paid to the University of Michigan.

Data Availability

Data collected without identifiers are included as a supplementary file (Multimedia Appendix 3).

Authors' Contributions

ZLL and CPF conceptualized this study. ZLL and NJS contributed to the investigation. Formal analysis of the study was conducted by ZLL, CPF, and CAA. ZLL, NJS, CAG, and CPF contributed to the methodology. CAG developed the software for this study. ZLL and CAG developed the visualizations. ZLL wrote the original draft, and ZLL, CAA, CAG, CPF, and NJS contributed to the review and editing. No generative artificial intelligence was used in any portion of the manuscript writing.

Conflicts of Interest

ZLL has received research support paid to the University of Michigan and related to this work from the National Library of Medicine (K01 LM012528; R01 LM013894). NJS has received research support paid to University of Michigan and unrelated to this work from Merck & Co. NJS received support paid to the University of Michigan for his role as program director of the Anesthesiology Performance Improvement and Reporting Exchange (ASPIRE) Collaborative Quality Initiative, and has received research support from Edwards Lifesciences, Apple Inc, and the National Institute on Aging (R01 AG059607) paid to the University of Michigan and unrelated to this work.

Multimedia Appendix 1
Causal pathway models for precision feedback interventions.
[PDF File, 51 KB - mededu_v10i1e54071_app1.pdf]

Multimedia Appendix 2
Test set of precision feedback emails.
[PDF File, 1400 KB - mededu_v10i1e54071_app2.pdf]

Multimedia Appendix 3
Study data.
[XLSX File, 61 KB - mededu_v10i1e54071_app3.xlsx]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.
[PDF File, 104 KB - mededu_v10i1e54071_app4.pdf]

References

1. Stead WW, Searle JR, Fessler HE, Smith JW, Shortliffe EH. Biomedical informatics: changing what physicians need to know and how they learn. Acad Med 2011 Apr;86(4):429-434. [doi: 10.1097/ACM.0b013e3181f41e8c] [Medline: 20711055]
2. Kluger AN, Van Dijk D. Feedback, the various tasks of the doctor, and the feedforward alternative. Med Educ 2010 Dec;44(12):1166-1174. [doi: 10.1111/j.1365-2923.2010.03849.x] [Medline: 21091758]
3. Kluger AN, DeNisi A. Feedback interventions: toward the understanding of a double-edged sword. Curr Dir Psychol Sci 1998 Dec;7(3):67-72. [doi: 10.1111/1467-8721.ep10772989]

XSL•FO
RenderX

4. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012 Jun 13(6):CD000259. [doi: 10.1002/14651858.CD000259.pub3] [Medline: 22696318]
5. Ivers NM, Grimshaw JM, Jamtvedt G, et al. Growing literature, stagnant science? Systematic review, meta-regression and cumulative analysis of audit and feedback interventions in health care. J Gen Intern Med 2014 Nov;29(11):1534-1541. [doi: 10.1007/s11606-014-2913-y] [Medline: 24965281]
6. Grimshaw JM, Ivers N, Linklater S, et al. Reinvigorating stagnant science: implementation laboratories and a meta-laboratory to efficiently advance the science of audit and feedback. BMJ Qual Saf 2019 May;28(5):416-423. [doi: 10.1136/bmjqs-2018-008355] [Medline: 30852557]
7. Ivers NM, Sales A, Colquhoun H, et al. No more "business as usual" with audit and feedback interventions: towards an agenda for a reinvigorated intervention. Implement Sci 2014 Jan 17;9:14. [doi: 10.1186/1748-5908-9-14] [Medline: 24438584]
8. Foy R, Skrypak M, Alderson S, et al. Revitalising audit and feedback to improve patient care. BMJ 2020 Feb 27;368:m213. [doi: 10.1136/bmj.m213] [Medline: 32107249]
9. Van Dijk D, Kluger AN. Task type as a moderator of positive/negative feedback effects on motivation and performance: a regulatory focus perspective. J Organ Behavior 2011 Nov;32(8):1084-1105. [doi: 10.1002/job.725]
10. Higgins ET. Value from regulatory fit. Curr Dir Psychol Sci 2005 Aug;14(4):209-213. [doi: 10.1111/j.0963-7214.2005.00366.x]
11. Ilgen DR, Fisher CD, Taylor MS. Consequences of individual feedback on behavior in organizations. J Appl Psychol 1979;64(4):349-371. [doi: 10.1037//0021-9010.64.4.349]
12. Sargeant J, Lockyer J, Mann K, et al. Facilitated reflective performance feedback: developing an evidence- and theory-based model that builds relationship, explores reactions and content, and coaches for performance change (R2C2). Acad Med 2015 Dec;90(12):1698-1706. [doi: 10.1097/ACM.0000000000000809] [Medline: 26200584]
13. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007 Mar;77(1):81-112. [doi: 10.3102/003465430298487]
14. Coiera E. Assessing technology success and failure using information value chain theory. Stud Health Technol Inform 2019 Jul 30;263:35-48. [doi: 10.3233/SHTI190109] [Medline: 31411151]
15. Gude WT, van der Veer SN, de Keizer NF, Coiera E, Peek N. Optimizing digital health informatics interventions through unobtrusive quantitative process evaluations. Stud Health Technol Inform 2016;228:594-598. [Medline: 27577453]
16. Coiera E. A new informatics geography. Yearb Med Inform 2016 Nov 10(1):251-255. [doi: 10.15265/IY-2016-018] [Medline: 27830259]
17. Dowding D, Randell R, Gardner P, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform 2015 Feb;84(2):87-100. [doi: 10.1016/j.ijmedinf.2014.10.001] [Medline: 25453274]
18. Tuti T, Nzinga J, Njoroge M, et al. A systematic review of electronic audit and feedback: intervention effectiveness and use of behaviour change theory. Implement Sci 2017 May 12;12(1):61. [doi: 10.1186/s13012-017-0590-z] [Medline: 28494799]
19. Landis-Lewis Z, Flynn A, Janda A, Shah N. A scalable service to improve health care quality through precision audit and feedback: proposal for a randomized controlled trial. JMIR Res Protoc 2022 May 10;11(5):e34990. [doi: 10.2196/34990] [Medline: 35536637]
20. Lee D, Panicker V, Gross C, Zhang J, Landis-Lewis Z. What was visualized? A method for describing content of performance summary displays in feedback interventions. BMC Med Res Methodol 2020 Apr 23;20(1):90. [doi: 10.1186/s12874-020-00951-x] [Medline: 32326895]
21. Landis-Lewis Z, Brehaut JC, Hochheiser H, Douglas GP, Jacobson RS. Computer-supported feedback message tailoring: theory-informed adaptation of clinical audit and feedback for learning and behavior change. Implement Sci 2015 Jan 21;10:12. [doi: 10.1186/s13012-014-0203-z] [Medline: 25603806]
22. Landis-Lewis Z, Stansbury C, Rincon J, Gross C. Performance summary display ontology: feedback intervention content, delivery, and interpreted information. Presented at: International Conference on Biomedical Ontology; Sep 25 to 28, 2022; Ann Arbor, MI URL: https://icbo-conference.github.io/icbo2022/papers/ICBO-2022_paper_2172.pdf [accessed 2024-05-31]
23. Landis-Lewis Z, Janda AM, Chung H, Galante P, Cao Y, Krumm AE. Precision feedback: a conceptual model. Learn Health Syst 2024:e10419. [doi: 10.1002/lrh2.10419]
24. van Overveld LFJ, Takes RP, Vijn TW, et al. Feedback preferences of patients, professionals and health insurers in integrated head and neck cancer care. Health Expect 2017 Dec;20(6):1275-1288. [doi: 10.1111/hex.12567] [Medline: 28618147]
25. Ross JS, Williams L, Damush TM, Matthias M. Physician and other healthcare personnel responses to hospital stroke quality of care performance feedback: a qualitative study. BMJ Qual Saf 2016 Jun;25(6):441-447. [doi: 10.1136/bmjqs-2015-004197] [Medline: 26253122]
26. Weernink MGM, Janus SIM, van Til JA, Raisch DW, van Manen JG, IJzerman MJ. A systematic review to identify the use of preference elicitation methods in healthcare decision making. Pharm Med 2014 Aug;28(4):175-185. [doi: 10.1007/s40290-014-0059-1]
27. Weernink MGM, van Til JA, Witteman HO, Fraenkel L, IJzerman MJ. Individual value clarification methods based on conjoint analysis: a systematic review of common practice in task design, statistical analysis, and presentation of results. Med Decis Making 2018 Aug;38(6):746-755. [doi: 10.1177/0272989X18765185] [Medline: 29592585]

XSL•FO
RenderX

28. Llewellyn-Thomas HA, Crump RT. Decision support for patients: values clarification and preference elicitation. Med Care Res Rev 2013 Feb;70(1 Suppl):50S-79S. [doi: 10.1177/1077558712461182] [Medline: 23124615]
29. Colquhoun DA, Shanks AM, Kapeles SR, et al. Considerations for integration of perioperative electronic health records across institutions for research and quality improvement: the approach taken by the Multicenter Perioperative Outcomes Group. Anesth Analg 2020 May;130(5):1133-1146. [doi: 10.1213/ANE.0000000000004489] [Medline: 32287121]
30. Kheterpal S. Clinical research using an information system: the Multicenter Perioperative Outcomes Group. Anesthesiol Clin 2011 Sep;29(3):377-388. [doi: 10.1016/j.anclin.2011.06.002] [Medline: 21871400]
31. Kluger AN, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 1996;119(2):254-284. [doi: 10.1037//0033-2909.119.2.254]
32. Carver CS, Scheier MF. Control theory: a useful conceptual framework for personality -- social, clinical, and health psychology. Psychol Bull 1982 Jul;92(1):111-135. [doi: 10.1037/0033-2909.92.1.111] [Medline: 7134324]
33. Locke EA, Latham GP. Building a practically useful theory of goal setting and task motivation. A 35-year odyssey. Am Psychol 2002 Sep;57(9):705-717. [doi: 10.1037//0003-066x.57.9.705] [Medline: 12237980]
34. Weissman NW, Allison JJ, Kiefe CI, et al. Achievable benchmarks of care: the ABCs of benchmarking. J Eval Clin Pract 1999 Aug;5(3):269-281. [doi: 10.1046/j.1365-2753.1999.00203.x] [Medline: 10461579]
35. Gude WT, Brown B, van der Veer SN, et al. Clinical performance comparators in audit and feedback: a review of theory and evidence. Implement Sci 2019 Apr 24;14(1):39. [doi: 10.1186/s13012-019-0887-1] [Medline: 31014352]
36. Ambrose SA, Bridges MW, DiPietro M, Lovett MC, Norman MK, Mayer RE. How Learning Works: Seven Research-Based Principles for Smart Teaching, 1st edition: Jossey-Bass; 2010.
37. Lewis CC, Klasnja P, Powell BJ, et al. From classification to causality: advancing understanding of mechanisms of change in implementation science. Front Public Health 2018 May;6:136. [doi: 10.3389/fpubh.2018.00136] [Medline: 29868544]
38. Brown B, Gude WT, Blakeman T, et al. Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. Implement Sci 2019 Apr 26;14(1):40. [doi: 10.1186/s13012-019-0883-5] [Medline: 31027495]
39. Karlin B, Ford R. The Usability Perception Scale (UPscale): a measure for evaluating feedback displays. In: Marcus A, editor. Design, User Experience, and Usability. Design Philosophy, Methods, and Tools. DUXU 2013. Lecture Notes in Computer Science, vol 8012: Springer; 2013:312-321. [doi: 10.1007/978-3-642-39229-0_34]
40. Thompson B. Score Reliability: SAGE; 2003. [doi: 10.4135/9781412985789]
41. Hegarty M. Advances in cognitive science and information visualization. In: Zapata-Rivera D, editor. Score Reporting Research and Applications: Routledge; 2018. [doi: 10.4324/9781351136501]
42. Galesic M, Garcia-Retamero R. Graph literacy: a cross-cultural comparison. Med Decis Making 2011 May;31(3):444-457. [doi: 10.1177/0272989X10373805] [Medline: 20671213]
43. Stone D, Heen S. Thanks for the Feedback: The Science and Art of Receiving Feedback Well: Viking; 2014.

Abbreviations

ABC: achievable benchmark of care
CP-FIT: Clinical Performance Feedback Intervention Theory
CRNA: certified registered nurse anesthetist
MPOG : Multicenter Perioperative Outcomes Group

Edited by TDA Cardoso; submitted 28.10.23; peer-reviewed by AA Bhurane, É Dufour; revised version received 05.03.24; accepted 26.04.24; published 11.06.24.

Please cite as:
Landis-Lewis Z, Andrews CA, Gross CA, Friedman CP, Shah NJ
Exploring Anesthesia Provider Preferences for Precision Feedback: Preference Elicitation Study
JMIR Med Educ 2024;10:e54071
URL: https://mededu.jmir.org/2024/1/e54071
doi:10.2196/54071

© Zach Landis-Lewis, Chris A Andrews, Colin A Gross, Charles P Friedman, Nirav J Shah. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.6.2024. 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.

XSL•FO
RenderX

Original Paper

Enhancing Digital Health Awareness and mHealth Competencies in Medical Education: Proof-of-Concept Study and Summative Process Evaluation of a Quality Improvement Project

Fatma Sahan[1*], BSc; Lisa Guthardt[1*], MA; Karin Panitz[1], MSc; Anna Siegel-Kianer[1]; Isabel Eichhof[2], Dr Rer Nat; Björn D Schmitt[2], Dr Rer Nat; Jennifer Apolinario-Hagen[1], Dr Rer Medic, Dipl-Psych

[1]Institute of Occupational, Social and Environmental Medicine, Centre for Health and Society, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Startup4MED, Dean's Office of the Medical Faculty, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[*]these authors contributed equally

Corresponding Author:
Jennifer Apolinario-Hagen, Dr Rer Medic, Dipl-Psych
Institute of Occupational, Social and Environmental Medicine, Centre for Health and Society
Medical Faculty and University Hospital Düsseldorf
Heinrich Heine University Düsseldorf
Moorenstraße 5
Düsseldorf, 40225
Germany
Phone: 49 211 8106477
Email: jemac100@hhu.de

Abstract

Background: Currently, there is a need to optimize knowledge on digital transformation in mental health care, including digital therapeutics (eg, prescription apps), in medical education. However, in Germany, digital health has not yet been systematically integrated into medical curricula and is taught in a relatively small number of electives. Challenges for lecturers include the dynamic field as well as lacking guidance on how to efficiently apply innovative teaching formats for these new digital competencies. Quality improvement projects provide options to pilot-test novel educational offerings, as little is known about the acceptability of participatory approaches in conventional medical education.

Objective: This quality improvement project addressed the gap in medical school electives on digital health literacy by introducing and evaluating an elective scoping study on the systematic development of different health app concepts designed by students to cultivate essential skills for future health care professionals (ie, mobile health [mHealth] competencies).

Methods: This proof-of-concept study describes the development, optimization, implementation, and evaluation of a web-based elective on digital (mental) health competencies in medical education. Implemented as part of a quality improvement project, the elective aimed to guide medical students in developing app concepts applying a design thinking approach at a German medical school from January 2021 to January 2024. Topics included defining digital (mental) health, quality criteria for health apps, user perspective, persuasive design, and critical reflection on digitization in medical practice. The elective was offered 6 times within 36 months, with continuous evaluation and iterative optimization using both process and outcome measures, such as web-based questionnaires. We present examples of app concepts designed by students and summarize the quantitative and qualitative evaluation results.

Results: In total, 60 students completed the elective and developed 25 health app concepts, most commonly targeting stress management and depression. In addition, disease management and prevention apps were designed for various somatic conditions such as diabetes and chronic pain. The results indicated high overall satisfaction across the 6 courses according to the evaluation questionnaire, with lower scores indicating higher satisfaction on a scale ranging from 1 to 6 (mean 1.70, SD 0.68). Students particularly valued the content, flexibility, support, and structure. While improvements in group work, submissions, and information transfer were suggested, the results underscore the usefulness of the web-based elective.

Conclusions: This quality improvement project provides insights into relevant features for the successful user-centered and creative integration of mHealth competencies into medical education. Key factors for the satisfaction of students involved the

XSL•FO
RenderX

participatory mindset, focus on competencies, discussions with app providers, and flexibility. Future efforts should define important learning objectives for digital health literacy and provide recommendations for integration rather than debating the need for digital health integration.

(JMIR Med Educ 2024;10:e59454) doi:10.2196/59454

KEYWORDS

medical students; digital health; design thinking; digital health literacy; medical education; digital health competencies; mobile phone

Introduction

Background

Initiated by the slow digital transformation of the German health care system, the German Federal Parliament passed the Digital Healthcare Act in December 2019, which made it possible to prescribe certain medical apps [1]. Since then, topics related to medical informatics, digital health, and telemedicine have started to appear more and more in the curriculum of some German medical schools, although mostly through few electives rather than compulsory subjects [2,3]. As digital health is a highly complex, dynamic field that constantly changes and advances and that has rarely been implemented into medical curricula [4,5], it seems necessary to compile and establish designated novel teaching formats that focus on the different subtopics, such as digital mental health interventions (DMHIs; eg, mobile health apps for dealing with depressive symptoms or managing study-related stress).

Digital mental health literacy, as well as competencies, becomes more and more important for (future) health professionals and medical education. It can be defined as “the degree to which individuals obtain, process, and understand basic mental health information and services needed to aid their recognition, management, or prevention of mental health issues” [6]. Even though younger generations are often supposed to be familiar with digitization and the corresponding competencies, research has shown that medical students do not feel adequately prepared for digitization in their course of study [2,3,7,8].

Previous studies have found that medical students perceive insufficient digital health literacy [9] and know little about available DMHIs [10,11]. At the same time, mental disorders have relatively high prevalence rates among the general population (ie, approximately 27.8% of the general German population have a mental disorder [12], and approximately 20% of the German population have depressive symptoms [13], a general decline in mental health in the last year [14]), and digital interventions could offer additional treatment and prevention options [15]. A systematic review and meta-analysis demonstrated that pooled depression prevalence among medical students worldwide was 37.9% [16]. DMHIs are especially relevant for medical students because, on the one hand, students are less likely to seek psychological help (due to barriers such as fear of stigmatization or lack of awareness) [17]. On the other hand, they themselves will eventually treat patients with (mental) health issues and may prescribe and use telemedicine, including digital health applications (in German: *Digitale Gesundheitsanwendungen* [DiGAs]) as future physicians [18]. Since October 2020, physicians and psychotherapists can prescribe DiGAs that are listed in the DiGA register by the German Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte), including different DMHIs for mental disorders, on the expense of statutory health insurance companies [1]. DiGAs are certified medical products mainly based on digital technologies that can be used to detect, surveil, treat, or mitigate diseases, injuries, or disabilities [19]. The goals of DiGAs include the monitoring and improvement of current treatments in patient care. In the face of aging societies and the rise of chronic diseases, high-income industrial nations are confronted with rising health care costs. Therefore, digital health care and digital self-care practices are linked to efforts to better prevent disease, calculate disease risks and life expectancy through algorithm-based personalized medicine, and at the same time delegate clinical treatment responsibilities to the affected individuals themselves [20]. The potential of digital therapeutics or DiGAs in particular to improve the uptake of health care services has not been fully exploited yet, and the uptake is relatively low compared to prevalence rates [21]. In preventive medicine and disease monitoring, digital interventions could improve patients’ health and personal motivation [22]. In addition, digital data collection promises to optimize processes and increase the efficiency of the health care system at an institutional level [20].

Studies have shown that physicians are open to the idea of DiGAs [21], but the current prescription rates of DiGAs are low, with approximately 203,000 DiGAs prescribed or granted by health insurances in Germany [23]. Physicians in a mixed methods study described that they were skeptical (eg, due to technical insecurities) and said that they lacked adequate information sources on how to prescribe DiGAs and how to guide and advise patients concerning their use [19]. From a patient perspective, there is a clear interest in digital health as well. A recent German study on health app acceptance found that 76% of participants, including those without previous app experience, expressed willingness to use DiGAs [18]. Information measures can effectively increase acceptance of quality-assured digital health services among health care providers and patients [24]. To address the knowledge gap and enhance digital health competencies, practicing physicians are considering continuing education opportunities [21].

However, digital health literacy and digital competencies, or the acquisition of knowledge on DMHIs, including DiGAs for the treatment of mental disorders and the management of chronic conditions, need to be part of the curriculum, which could be piloted in elective subjects in medical schools. To the best of our knowledge, there are only a few studies concerning the

XSL•FO
RenderX

teaching of digital health and digital competencies in German medical schools. Thus, little is known about strategies to implement such new teaching offerings in medical education. In one study, a total of 16 universities in Germany were identified that had included digital skills in their curricula (17 elective and 8 compulsory courses) [2]. For example, a study investigated the impact of an interdisciplinary and cross-faculty course concerning digital medicine with the help of a web-based questionnaire before and after the course [3]. Aulenkamp et al [2] found a positive impact of such courses on the students' digital competencies and concluded that more efforts to integrate them into the curriculum would be necessary. One comparative study examined the implementation of a module on digital health among undergraduate medical students at a German university, the knowledge gain of students, and their attitudes toward digital health and suggested a firm implementation of digital competencies in medical education [25]. Another German study examined interdisciplinary teaching with the help of teaching teams of medical informatics professionals and physicians. In different academic years, new seminars on digital competencies were designed and implemented, and the usefulness of interdisciplinary teaching teams was demonstrated [26]. Further studies covered the integration of an elective on digital health in diabetes for pharmacy students [27] or the design, implementation, and evaluation of a course with a focus on telemedical components [28].

Overall, it is essential to permanently integrate digital health education into the curricula of medical schools [29,30]. Practical, competence-oriented didactic concepts offer promising strategies to enhance knowledge transfer and enable students to proficiently handle DMHIs and health apps.

Goals of the Quality Improvement Project and the Case Study

The overarching objective of the quality improvement project was to develop and iteratively optimize an innovative learning and teaching offering for medical students based on their preferences and needs in a German medical school. The goals were targeted via an elective subject.

The elective subject in this proof-of-concept study aimed to provide students with basic knowledge and practical skills and promote a comprehensive understanding of designing concepts for digital health interventions in prevention and therapy with the potential user in mind. The focus of the elective was on digital health competencies in the field of mental health; the use of DMHIs in occupational and social medicine, especially in the area of primary prevention (eg, stress management); and DiGAs for somatic and mental diseases based on the students' choices.

Our proof-of-concept study on quality improvement in medical education is meant to provide insights into the piloting of an innovative digital elective subject concerning the development of theoretical prototypes for DMHIs. We aimed to evaluate the implementation and realization of the elective subject using a design thinking process and analyze students' feedback, ideas, and preferences concerning the competence-based education on mental health apps ("learning by doing" and cocreating).

Methods

Setting and Background of Teaching Innovation

The focus of the innovative teaching offering at a German medical school was on digital mental health in areas of application relevant to prevention and health care settings; quality criteria of mobile health apps; legal framework, including structural conditions for telemedicine (eg, the so-called Digital Healthcare Act in Germany and prescription of DiGAs); and user-oriented app design (eg, persuasive design [31]). The elective also aimed to promote knowledge of app development, self-competence, collaborative learning (codevelopment of an app concept in small student groups according to the prominent design thinking approach by the Hasso Plattner Institute [32]) as well as critical reflection on the opportunities and risks of digitalization for health professionals. The main target group were medical students, but due to the interdisciplinary nature of health app development, we allowed a small number of students from other disciplines to also participate.

Quality Improvement Project

The quality improvement project was divided into 2 parts: funding (24 months) and the implementation of the web-based elective into the curriculum after the pilot project (12 months). The quality improvement project took place for 24 months from January 1, 2021, to December 31, 2022, at the Medical Faculty of Heinrich Heine University Düsseldorf (HHU) in Germany. We offered the digital elective 6 times within 3 years. The concept was based on a preceding elective subject involving a co-design workshop on digital mental health literacy in medical studies, which we conducted on campus with 26 medical students in March 2020, as described in sufficient detail by Dederichs et al [8]. In this subject, medical students developed theoretical prototypes of health apps in small groups using appropriate background knowledge and the established 5-step design thinking principle [33]. The students gave us feedback and indicated that they would have liked more flexibility, digital content in ILIAS (Integriertes Lern-, Informations- und Arbeitskooperations-System [German for "Integrated Learning, Information and Work Cooperation System"]), and additional guest lectures, which is why we developed the idea of a digital or hybrid implementation of the seminar (blended learning approach) [8]. The corresponding author conceived the project idea and acquired funding for an innovative educational project by the Commission for Quality Improvement in Teaching and Studies of the Dean's Office of the Medical Faculty in November 2020. In January 2021, we prepared the digital elective "Fit for digitalization and 'apps on prescription'?—Understanding digital health applications and developing digital health offers such as health apps" for the summer semester of 2021. Medical students were able to take the elective subject with 2 semester hours per week. Over time, we also extended the elective to students from other fields if there were enough available openings. This was achieved either in the framework of "studium universale" or as a psychology minor, including credit points for active participation. Finally, the following 12 months (January 2023-January 2024) were used to implement the project into medical education practice without additional staff, which required some adaptations,

XSL•FO
RenderX

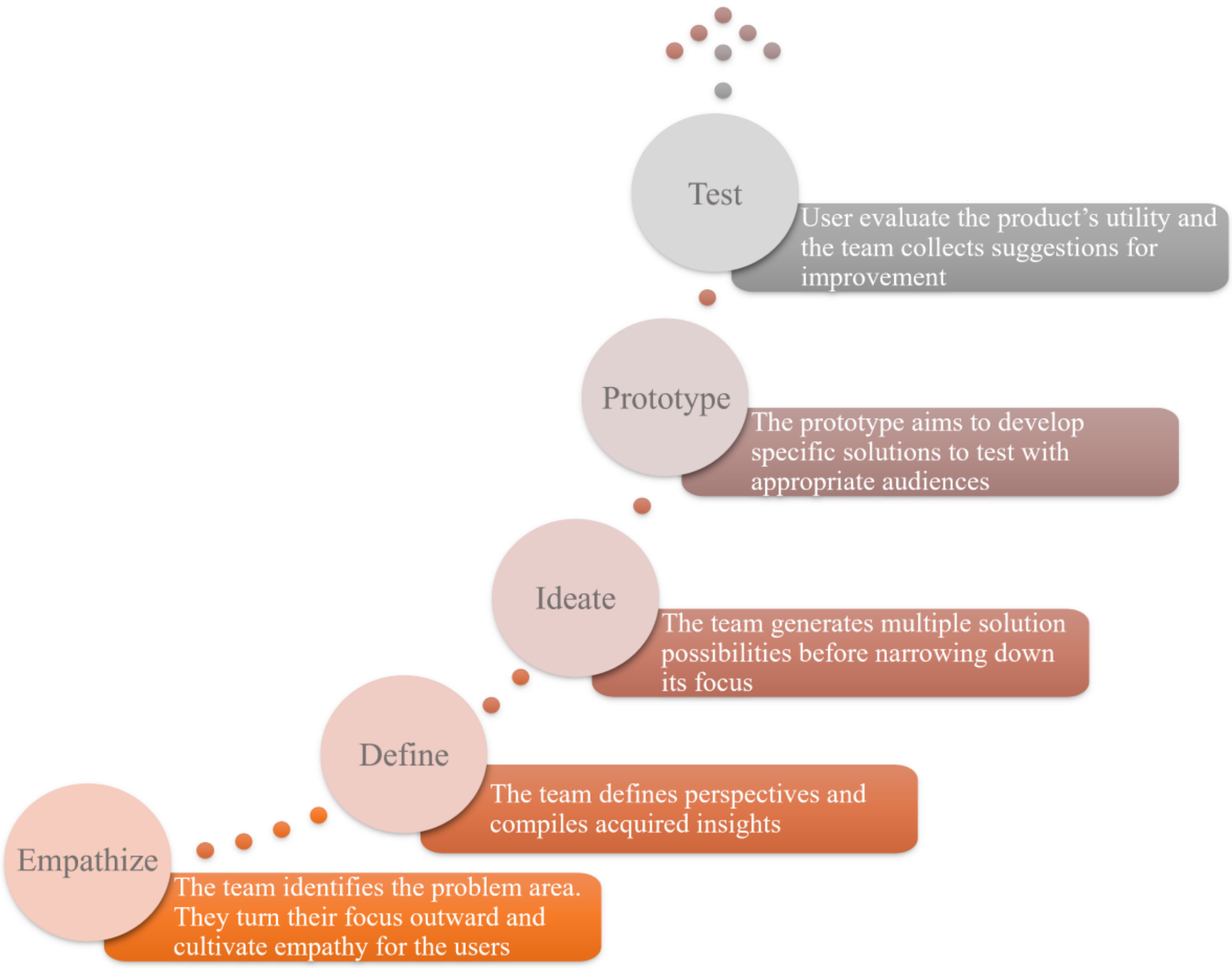
including providing a screen cast (ie, video tutorial) on the technical creation of app mock-ups using the collaborative design tool Figma (Figma, Inc) instead of personal assistance for each group.

Structure and Composition of the Elective

We varied the elective either as an intensive block course (5 consecutive days, full time during the lecture time, and provided once) or as weekly seminars (7 times, part time during the lecture-free time, and provided 5 times), equating to a time effort of 2 semester hours for both options. For more intensive support, we increased the proportion of synchronous live lectures after the first run in 2021. In addition, students were able to access the content (eg, presentations) and additional material via the web-based learning management system ILIAS (ILIAS open source e-Learning e.V). Students could request feedback on their group work via ILIAS. We refined the ILIAS environment steadily to ensure its immediacy and to meet students' preferences.

The focus was on competence-oriented learning and participatory design approaches. The acquired knowledge about health apps was directly transferred into a concrete app concept in the group work. In creating the group assignments, we were guided by the concept of design thinking and its phases for developing an app. Design thinking is an iterative, user-centered approach to innovation and problem-solving. This framework emphasizes the comprehensive analysis and understanding of human needs to generate original approaches to solving complex problems [31-33]. The process, which typically involves multiple stages, includes understanding the problem, carefully eliciting user-specific requirements, generating ideas, creating prototypes, and testing solutions (Figure 1 [31]).

Figure 1. A simplified explanation of the 5 stages of design thinking (based on the work by Meinel et al [31]).



The elective included an introductory session at the beginning of the semester, which introduced the procedure, contents, and objectives and brought students together in small groups with different app topics (eg, insomnia). Initial content-related issues were also covered (eg, definitions and fields of application). In lessons 2 to 5, the students were given a group task aligned with the content of the lessons and a design thinking process, which systematically guided them in the development of a hypothetical app.

A detailed description of the structure of the elective can be found in Table 1.

XSL•FO
RenderX

Table 1. Lessons and main contents of the elective course[a].

Session number	Main topic	Contents
1	Organizational matters and introduction	• Overview and examination • Presentation of group results (app concepts) from previous semesters • Introduction—working definition of digital mental health • Areas of application • Support with digital health interventions • Evidence of e–mental health • Telemedicine and digital health in occupational medicine • Division into small groups for the development of own app concepts and introducing tools for group work • First introduction to design thinking • Option of test access to DiGAs[b] (GAIA AG)
2	Quality criteria of health apps	• Description of quality seals for health apps • Description of quality principles • Presentation of the DiGA register • Quality measurement using the MARS[c] [34] (German version [35]) • Presentation of the search for quality-assessed health apps and e–mental health interventions • Presentation of an app (eg, blood pressure app Manoa) • Group work
3	DiGAs—“apps on prescription”	• Regulation of the apps listed in the DiGA directory • Acceptance and use of DiGAs • Various guest lectures on mental health DiGAs (eg, Somnio, Deprexis, Elona Therapy Depression, and Velibra DiGAs) • Guest lecture ReHappy (former stroke DiGA), including screencast (ie, video files of recorded lectures) • Group work
4	Design thinking and persuasive design (app development)	• Description of the phases of the design thinking process • Presentation of concept mapping • Teaching persuasive design and persuasive design in mental health apps • Discussion on the “ideate” phase in the design thinking cycle and idea generation • Incentives for managing successful apps • Group work
5	Strategies to promote acceptance and adherence	• Measures to promote user adherence and motivation by communicating acceptance models, such as the UTAUT[d] and UTAUT2[e] adapted to digital health [36,37] • Presentation of usability in health apps (eg, via the SUS[f] [38]) • Description of gamification approaches for health apps • Presentation on promoting acceptance through information • Description of the distinction between health and medical apps as well as trusted health apps • Group work
6	Finish for the final presentation	• Explaining peer feedback • Guest lecture from Startup4MED, HHU[g] (each semester) • Tips for the final presentations and peer feedback • Group work
7	Final presentation of the group work	• Lecture—summary plus current developments and perspectives • Presentation of the developed app concepts in small groups • Feedback and evaluation of the elective

[a]Detailed description of the elective structure (sessions 1 to 7). The contents partly varied across the semesters based on expert availability (digital health app providers) and adaptations made following the evaluation results of the former semester.

[b]DiGA: digital health application.

[c]MARS: Mobile App Rating Scale.

[d]UTAUT: Unified Theory of Acceptance and Use of Technology.

[e]UTAUT2: extension of the UTAUT to the consumer context.

[f]SUS: System Usability Scale.

XSL•FO
RenderX

[g]HHU: Heinrich Heine University Düsseldorf (Germany).

Practical Transfer Through the Integration of Guest Lectures and Test Access

We provided students with selected guest lectures by experts who created, provided, or presented DiGAs to the professional community (eg, Somnio, Velibra, and Elona Therapy Depression). In addition, at the end of the course, we hosted lectures from staff members of the medical-specific start-up support unit "Startup4MED." Startup4MED is the internal start-up support unit of the University Medicine Düsseldorf and identifies, promotes, and supports the commercial exploitation of innovative medical projects from the Medical Faculty of HHU and the University Hospital of Düsseldorf. This allowed participants to connect with the start-up support team for extra guidance on their ideas and on how to implement their app concepts in practice.

Each semester, students had the opportunity to attend between 1 and 3 guest lectures, some of which were recorded as screencasts (ie, videos in terms of recorded lectures) and uploaded online. In addition, students were given the chance to try free demonstration versions of various DiGAs through a trial program sponsored by a German company (GAIA AG) in 2021 and 2022.

Student Support and Competence-Oriented Performance Recording

During their elective, students received comprehensive assistance through multiple media and communication channels that aligned with their individual preferences. These channels comprised Rocket.Chat, an open-source team chat platform, as well as Microsoft Teams. To reinforce their learning process, the participants were assigned concise, structured tasks following every session, and the completion of these tasks was monitored by the project team. We also offered personalized feedback and guidance on demand beyond the group feedback each week upon completion of the tasks uploaded using ILIAS. As part of the overall design thinking process, the students were gradually and systematically introduced to the app concept. This was done by providing weekly synchronous web-based lectures in addition to the educational material provided via ILIAS. This approach aimed to promote collaboration, reduce inhibitions to reach out, and provide students with more flexibility. In the last session of the elective, student groups presented their developed app concepts to their instructors and peers and received feedback. Optionally, the assessment could be completed as an exam. Finally, the students were instructed to evaluate the elective through a web-based survey provided via "evasys" (evasys GmbH).

Evaluation

In addition to the oral feedback in the last session, we used a web-based evaluation questionnaire created using templates provided by the Dean's Office (Department of Evaluation, Medical Faculty). The questionnaire, implemented using "evasys," included standardized questions on digital teaching offerings and was completed by students online after finishing the course. Participation was voluntary, and thus, we tried to increase it by sending reminders. In the first part of the questionnaire, we asked students to provide information about their gender and their field of study. In the second part of the questionnaire, we asked students to rate the module through 19 questions. The first 13 were answerable on a 6-point Likert scale ranging from 1 ("completely agree") to 6 ("do not agree at all"). These questions concerned the content of the elective as well as the visualization and access to ILIAS, such as "The learning module was well structured." Finally, we assessed the perceived difficulty level of the learning material, overall satisfaction with the elective (both in general and regarding digital implementation), and the students' own estimated learning gains after completing the elective.

A total of 3 additional questions dealt with the difficulty of the learning content, the scope of the learning and reading material, and the assessment of particularly helpful elements in the course (5 response options, eg, text units and screencasts). The last 3 questions served as feedback and were presented in the form of open questions (eg, what students liked most about the course, which competence area they benefited most from, and which suggestions for improvement they had). The evaluation was centralized and anonymized by the Dean's Office. Results of the evaluation were provided in an aggregated format by the Dean's Office if a minimum of 5 completed surveys per elective were available. Following the completion of the elective in the summer semester of 2021, the questionnaire was revised and used in the subsequent courses. We further analyzed quantitative data from the elective descriptively (eg, means and proportions) and summarized qualitative data (comment fields) using Microsoft Excel and SPSS (version 27.0; IBM Corp). Answers to the open-ended questions and comments were analyzed using MAXQDA 2020 (VERBI GmbH). We then formed categories both deductively based on our questionnaire and inductively from the material. The preliminary code system was discussed, revised, and agreed upon.

Ethical Considerations

This proof-of-concept study received an ethics approval for retrospective analyses by the IRB of the Medical Faculty at the Heinrich Heine University Düsseldorf (ref. no. 2024-3033). We obtained informed consent for the publication of mockup figures designed by the student groups. The evaluation in this quality improvement project was conducted in a regular teaching context. The participation in the web-based survey on the subjective evaluation of the elective was voluntarily and conducted anonymously by the central evaluation of the study dean office upon completion of the elective, so that the authors had no access to personal data linking individual survey responses to specific participants. We also had no access to information on individual data such as gender, age, study subject or semester based on the aggregated results of the web-based survey. The web-based survey did not address sensitive topics and the respondents were not considered as a vulnerable group according to the nature of the survey as we only ask for them to indicate their views on the quality of the attended elective and optionally provide suggestions for improvements.

XSL•FO
RenderX

Results

Sample Characteristics

A total of 75 students (women: n=45, 60%) from HHU registered for the elective in 6 seminars over 3 years (2021-2024; ie, 5 semesters [no elective in summer 2023 due to parental leave of the project lead]). Of these students, 72% (54/75) were medical students (median 5, range 3-11 semesters), 12% (9/75) were psychology students (bachelor's program; median 7, range 5-21 semesters), and 4% (3/75) were economics students (bachelor's program; median 9, range 7-16 semesters). In addition, 1% (1/75) studied business administration (bachelor's program; semester 1); 5% (4/75) were bachelor's biology students (mean 9.50, SD 2.96 semesters; range 5-13); 1% (1/75) studied art history (semester 17), philosophy (semester 9), or medical physics (semester 7; bachelor's program each case) each; and 1% (1/75) were master's biology students (semester 1). Overall, 20% (15/75) of the students dropped out of the elective, mainly before its start, as they did not attend the first session. The remaining 80% (60/75) of the students successfully finished the elective (including course achievement). Of these 60 students (women: n=36, 60%), 77% (46/60) were medical students (median 5, range 3-11 semesters), 15% (9/60) were psychology students (bachelor's program; as mentioned above), and 8% (5/60) studied one of the aforementioned subjects (2/5, 40% studied economics; median 12.5, range 9-16 semesters; and 1/5, 20% studied a bachelor's biology program; semester 13; art history, and a master's biology program).

Insights Into Common Themes and App Development From the Sessions and Group Discussions

According to individual preferences, students focused on both mental health promotion and dealing with mental and somatic diseases in the development of the hypothetical app, but for this case study, we only reported some examples of health apps. After the first round of the elective, the scope was expanded beyond DMHIs in the course description for upcoming electives to better meet the preferences of more medical students. Overall, 25 projects were finished and presented, 10 (40%) of which covered somatic conditions and 15 (60%) of which covered mental health conditions or indications. The students were asked to create a name for their app concept and to check whether this app name exists already. The task was to create a suitable, recognizable name they can explain with respect to the hypothetical product and target group. However, we did not control each app name in terms of brands or specific products worldwide as this was an elective for educational non-commercial purposes. Table 2 shows an overview of all app concepts developed by the students in the different semesters.

The following main topics emerged in the prototypes of the apps: exam anxiety in students, stress management, resilience, insomnia and sleep disturbances in children and adolescents, and depression in youths and young adults as themes for mental health apps, with some frequently chosen topics (especially stress management). Furthermore, various disease management or prevention apps for somatic conditions were conceptualized (breast cancer, diabetes, stroke and arterial hypertonia, reflux disease, musculoskeletal diseases, glioblastoma, skin diseases, and premenstrual syndrome). Recurring app topics among different electives were stress, exam anxiety, and stress-inducing learning behavior regarding procrastination, as well as depression and anxiety.

In the context of app development, the students systematically applied the knowledge they had learned in the elective subject considering various aspects, guided each week on different aspects by the elective's team (5 steps of design thinking [31,33]). The topics that students emphasized most often in their presentations of their app concepts included target group-specific information (eg, prevalence rates and relevance of the app for health care); usability; and features of persuasive design, including gamification, accessibility, and the promotion of adherence (eg, through reward systems, gamification, provision of human support, and cost reimbursement), which we will illustrate with suitable case examples in the following sections. The featured functionalities varied depending on the selected app concept. For apps related to disease management, students often incorporated symptom diaries, reminders about medications or physician's appointments, and educational resources about the corresponding illness. In contrast, apps aimed at stress management during learning focused mainly on structuring daily routines, relaxation, and avoiding procrastinating behaviors. An example of this would be the ability to lock smartphones for a set period. Most app prototypes included considerations for integrating professional guidance (eg, via chat functions). These chat functions either established contact with professionals or facilitated communication among user groups.

XSL•FO
RenderX

Table 2. Overview of the different app concepts in each semester partially translated into English.

Number	Topic or indication of the app	Target group or population	Name of the app concept	Semester
1	Chronic conditions, diabetes	Patients (adults)	My Diabetes Pass	SS[a] 2021
2	Learning support, self-management, stress management	Pupils and students	Studytime	SS 2021
3	Resilience promotion	Young people in education	Mental Power	SS 2021
4	Stress reduction	Students	MeTime	WS[b] 2021-2022
5	Musculoskeletal diseases, pain	Patients (primarily adults)	legLos (“getStarted”)	WS 2021-2022
6	Sleep disturbances, insomnia	Patients (children aged 3-12 years, supported by their parents)	Morpheus goes to sleep	WS 2021-2022
7	Depression, depressive symptoms	Patients (aged 13-25 years)	Dinotherapy	WS 2021-2022
8	Gastroesophageal reflux disease	Patients (aged >40 years)	StopGERD	WS 2021-2022
9	Breast cancer	Patients (adult female individuals)	BRUHNO	WS 2021-2022 intensive block
10	Stress reduction, exam anxiety	Students and trainees (aged 15-30 years)	Companion	WS 2021-2022 intensive block
11	Glioblastoma	Patients (aged 50-70 years)	GlioblAPP	WS 2021-2022 intensive block
12	Stress reduction, stress prevention	Company employees (ie, finance sector)	Stress Cutter	WS 2021-2022 intensive block
13	Back problems, pain	Patients (aged 18-60 years)	Backfit	SS 2022
14	Eating disorders	Patients (ie, anorexia nervosa)	Provida	SS 2022
15	Hypertension, blood pressure problems	Patients (mainly adults aged >50 years)	Eutonia	SS 2022
16	Stress management after rehabilitation	Patients (adults)	iGrow	WS 2022-2023
17	Blood pressure problems, arterial hypertension	Patients and risk groups (adults)	Tonus	WS 2022-2023
18	Sleep disturbance in depression, burnout	Patients (adults aged 35-45 years)	Happy Sleeper	WS 2022-2023
19	Exam anxiety	Students (aged 18-25 years)	Exam Anxiety	WS 2022-2023
20	Skin diseases	Patients (children and adults)	DermaDiary	WS 2023-2024
21	PMS[c], PMDD[d], self-management	Patients (mainly female individuals aged 18-30 years)	ZenCycle	WS 2023-2024
22	Self-management, (emotional) self-regulation	Primary school children (their parents or legal guardians)	MaxiKids	WS 2023-2024
23	Stress management	University students	UniRelax	WS 2023-2024
24	Resilience promotion	Health care staff, nurses (aged 16-64 years)	Pflege-Care (ie, care for nurses)	WS 2023-2024
25	Depression, depressive symptoms	Patients (adults)	Livetta	WS 2023-2024

[a]SS: summer semester.

[b]WS: winter semester.

[c]PMS: premenstrual syndrome.

[d]PMDD: premenstrual dysphoric disorder.

Case Examples: App Development and Concepts

In this section, several app prototypes that were developed by student groups are presented. All prototypes chosen for this paper serve as comprehensive and visually clear examples of app concepts. All design samples, images, and the creative theoretical work belong to the students and cannot be used without their permission.

Morpheus geht schlafen (Morpheus goes to sleep) is an app for sleep disturbance in children. A group of 2 students (n=2, 100% female) created this app concept to improve sleep hygiene and facilitate falling asleep for children. In Textbox 1, the app and its design and implementation are described in detail. Figure 2 visualizes the app concept.

XSL•FO
RenderX

Figure 2. Mock-ups of the Morpheus geht schlafen (Morpheus goes to sleep) app. Top left: welcome screen. Top right and bottom left: choice of the customizable companion (animal or human). Bottom right: assessment of individual well-being.

XSL•FO
RenderX

Textbox 1. App concept for Morpheus geht schlafen.

Morpheus geht schlafen (Morpheus goes to sleep) app concept

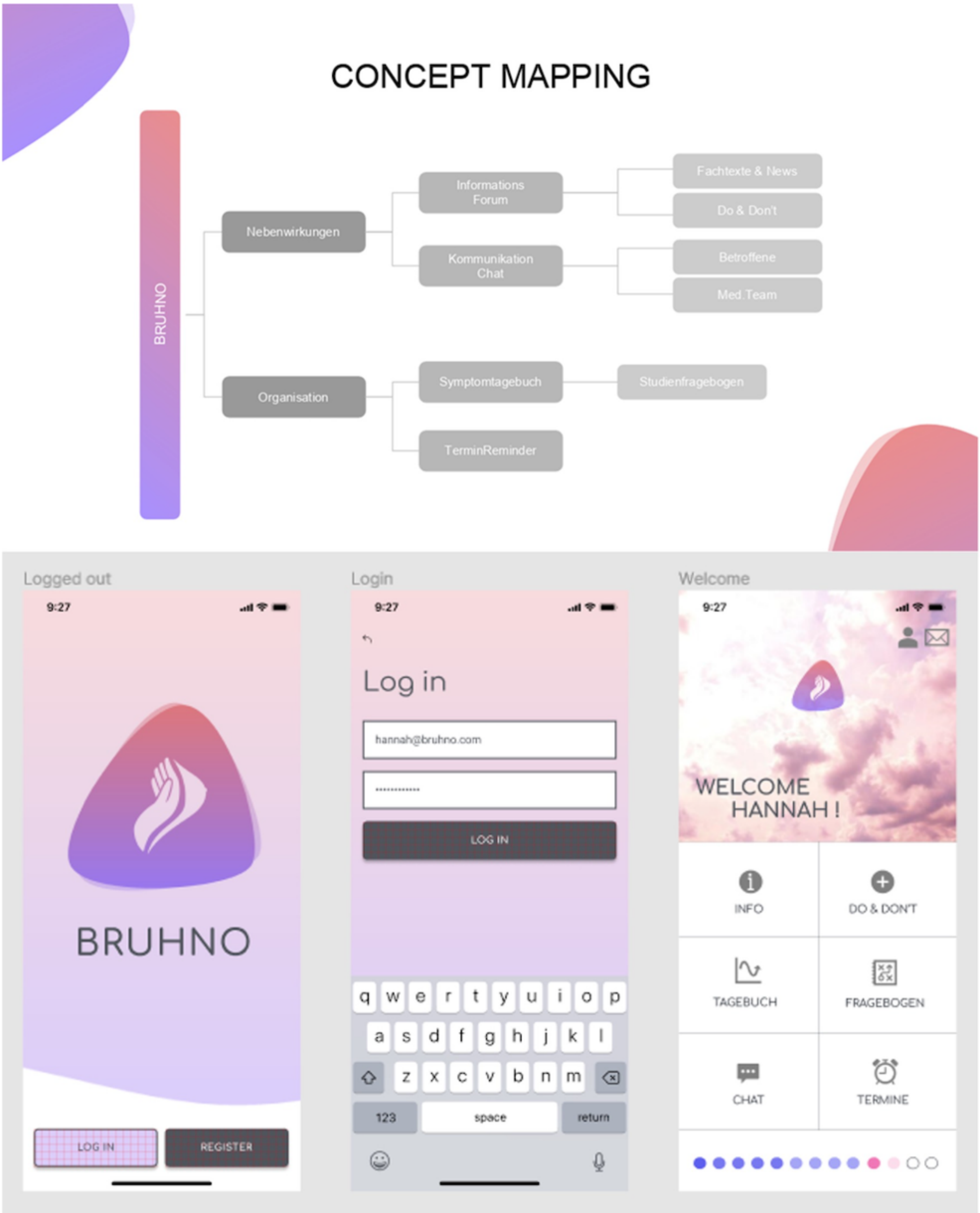
- Students developed the app for (nonorganic) sleep disturbances in children.
- The selected target group was children aged between 3 and 12 years.
- Concerning design and aesthetics, students chose dark background colors, little movement on screen, a background that remains the same, and a dark color scheme with little white. They also added a customizable companion (human or animal) with rounder body shapes to make it more appealing for children and to present a positive body scheme.
- According to the students, the app can be used to improve sleep hygiene and facilitate falling asleep and could be adapted in relation to the child's level of activity. Methods included exercises concerning movement and relaxation, fantasy journeys with progressive muscle relaxation, bedtime stories with integrated breathing instructions, "sound forests," and lullabies.
- Additional functions covered a link to the app alarm (default setting for when to get up, individualization by parents, and calculation of sleep phases and the ideal sleep time), a sleep tracker, and, optionally, a companion toy with speaker function.
- Students said that parents should be involved via either a separate app or a button for parents. Parents could be supported and informed about sleep hygiene in children (examples concerning a child-oriented evening routine, information on sleep disturbances in children, and a section with frequently asked questions).
- Concerning adherence promotion, students thought of the individualization of the companion, fun facts when brushing teeth, interesting activities, and the collection of stars for every use day. Nonadherence should also be detected (notifications in case of nonactivity).
- Data security was another very important aspect for the students. They explained that the app should include a data privacy statement. In addition, the use of the app is only possible after an active confirmation of informed consent. Data are stored on the device, and data transfer needs are separately authorized (or data can be downloaded as a PDF file).

Bruhno—Brustkrebs Helfer für Nebenwirkungen und Organisation (Breast cancer helper for side effects and organization) is a companion app for female patients with breast cancer. A group of 2 students (n=2, 100% female) created this app concept to support women in systemic therapy and aftercare.

In Textbox 2, the app and its design and implementation are described in detail. Figure 3 visualizes the app concept (for additional prototypes, see Figures S1 and S2 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 3. Mock-ups of the Bruhno app. Above: concept mapping for the Bruhno app mock-up as developed by the students. Below: mock-up of the Bruhno app-logo and log-out, log-in, and welcome screens.



XSL•FO
RenderX

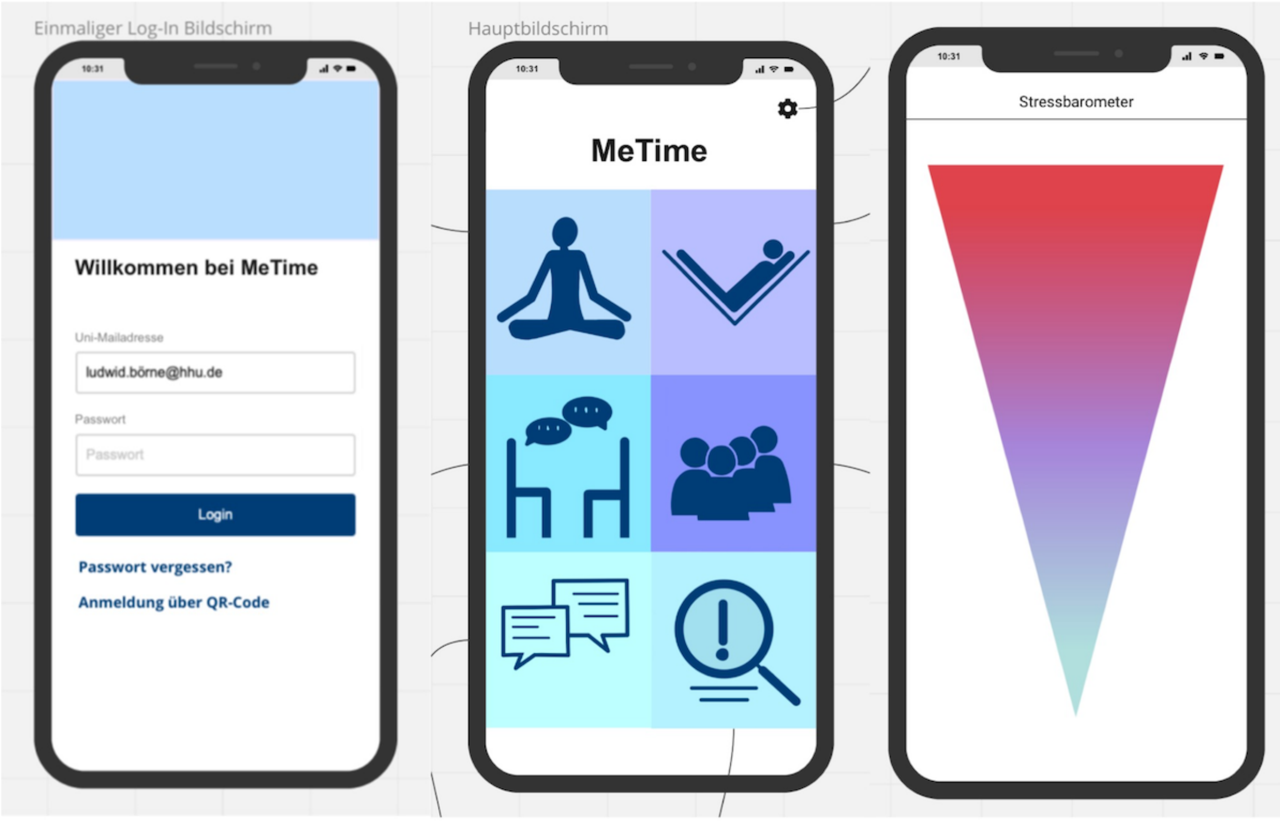
Textbox 2. App concept for Bruhno.

Bruhno app concept

- Students developed the app as a companion for female patients with breast cancer (support in systemic therapy and aftercare).
- The selected target group was predefined (participants of a clinical study with a diagnosis of breast cancer). Female patients aged ≥18 years (mean age 40-65 years) with sufficient technical abilities and equipment (mobile phone and internet access) should use the app.
- The following goals and functions were intended by the student group: reduction of fear (chat with physicians or other persons affected and short explanatory texts or videos), enhancement of well-being (eg, tips to manage side effects), data collection and monitoring (query of side effects, symptom diary, possible evaluation by physician, and questionnaires on life quality), appointment reminders (push notifications, and appointments could be entered by the clinic team or the patient), and an optional intake tracker in case of oral medication.
- Included methods were information, education, data collection, monitoring, tracking, and reminders.
- In the concept of the app, factors to promote adherence were tracking progress using a timeline and maintaining personal motivation through social exchange with other app users via chat.
- The student group also considered technical aspects related to the use of the app. Internet access is vital, and additional web applications would be useful to type longer posts or contributions. In addition, aspects concerning certification, advertisement, and financing were considered (categorization as a medical product and sale to study organizers or sponsors).

MeTime is an app for stress reduction in students. A group of 3 students (n=3, 67% female) created this app concept to promote relaxation in students. In Textbox 3, the app and its design and implementation are described in detail. Figure 4 visualizes the app concept (for additional prototypes, see Figures S3 and S4 in Multimedia Appendix 1).

Figure 4. Mock-ups of the MeTime app. Left: one-time log-in screen. Middle: welcome screen. Right: visualization of a stress barometer.



XSL•FO
RenderX

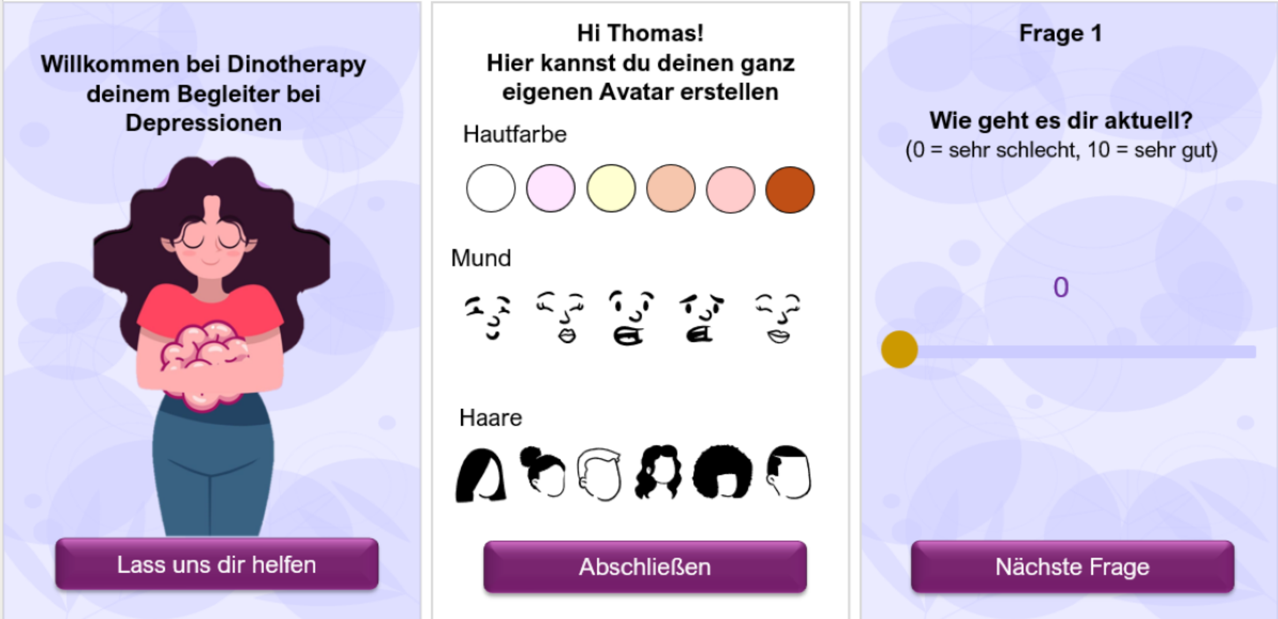
Textbox 3. App concept for MeTime.

MeTime app concept

- The app MeTime was developed by a group of students to reduce stress and promote relaxation for students.
- The relevant target group were students from Heinrich Heine University Düsseldorf in acute phases of stress (eg, exam periods).
- The concept of the app was to provide relaxation exercises and possibilities for networking and social exchange.
- Factors related to the design were also considered by the students. The app should be intuitive and minimalistic. Appealing colors, little text and many symbols, and a stress barometer should also be implemented.
- Students also added a companion, Heinrich, to the app (referring to the first name of a German writer and poet and eponym of the Heinrich Heine University Düsseldorf). This companion supports users, serves as a search engine for the lexicon, reads out instructions and texts, and wears different clothes depending on the context or situation.
- Use options and options to personalize the app were considered. Stress reduction, networking, and education were seen as important. In addition, personal exercises can be created and selected, favorite exercises can be chosen, and an anonymous chat can be used. Color scheme and text size should be customizable, and the user's personal calendar can be accessed. Individual stress levels should be assessed by the user before and after an exercise.
- Concerning app costs, students suggested including the costs in the semester fee. The chosen category for the mHealth app in app stores was lifestyle, health, and fitness.

Dinotherapy is a companion app for depressive episodes in youths and adolescents. A group of 2 students (n=2, 100% male) created this app concept to bridge the time until psychotherapy starts or as a companion during psychotherapy. In Textbox 4, the app and its design and implementation are described in detail. Figure 5 visualizes the app concept (for additional prototypes, see Figure S5 in Multimedia Appendix 1).

Figure 5. Mock-ups of the Dinotherapy app. Left: welcome screen (vivid user interface to create a positive user experience). Middle: options for the individual, customizable avatar or companion. Right: questionnaire to assess the user's mood.



XSL•FO
RenderX

Textbox 4. App concept for Dinotherapy.

Dinotherapy app concept

- Students developed the app as a companion for youths and adolescents who are diagnosed with depression.
- The selected target group were youths aged between 13 and 25 years.
- The app combines interactive provision of psychoeducational content (using gamification elements) with practical tasks and uses dinosaurs as protagonists, identification figures, and "virtual pets."
- According to the students, the app has different goals: (1) development of knowledge and understanding of the disease—"depression as a disease," (2) continuous tracking of the course of the depression, (3) establishing activities and fixed rituals that will have a beneficial effect on the course of the depression, and (4) social exchange with other persons affected via the app.
- Students said that a link between the accounts of the patient and physician or psychotherapist responsible can be possible. This way, the attending person could set priorities; assign patients to each other as peers; and specifically be informed if depression values fall below a critical threshold in case, for example, there is a suicide risk.
- To promote adherence, students thought of self-monitoring (tracking using short, daily questions), a visual presentation of the course of depression, tailoring and personalization of tasks and contents, rehearsals or reminders, appraisal and points for each completed task that can be used for the dinosaur (the "virtual pet"), and peer interaction.
- The design of the dinosaurs is also meant to promote adherence and should be appealing for the selected target group.
- The following functions were included by the students: (1) surveys (*TRACK*), (2) comics (*EDUCATE*), (3) tasks (*ACT*), (4) interaction (*EXCHANGE*, and (5) *INTERVENT*.

Companion is an app to reduce stress and exam anxiety and enhance support for students and trainees. A group of 2 students (n=2, 100% female) created this app concept to accompany the target group in the course of the study and in vocational training. In Textbox 5, the app and its design and implementation are described in detail. Figure 6 visualizes the app concept (for additional prototypes, see Figures S6 and S7 in Multimedia Appendix 1).

Figure 6. Mock-ups of the Companion app. Left: welcome screen, current level, functions, and the customizable pet. Right: selection of tasks and activities for breaks.

XSL•FO
RenderX

Textbox 5. App concept for Companion.

Companion app concept

- Students developed the app as a companion for students and trainees. The selected target group were users aged between 15 and 30 years.
- Students introduced an individual pet that accompanies users and explains the handling of the app. They also implemented daily individual and evidence-based hints to reduce stress, exam anxiety, or procrastination.
- In the app, learning plans, to-do lists, and morning and evening routines can be created. In addition, motivating reminders are used. Chatting with the companion (artificial intelligence) is possible. Further functions are structuring, a timer, time management, and the blocking of other apps to work productively.
- To promote adherence, the students implemented progress confirmation, the collection of points for leveling up, and rewards for new levels (eg, unlocking new items for the companion pet) in the app concept. They also paid attention to personalization and adaptation regarding user needs (tailoring).
- To enhance credibility, the students suggested the following: universities should recommend the app, physicians and therapists should test and also recommend it, scientific researchers should be involved in the app development, and the app should receive a seal of approval.
- No data are meant to be given to third parties. If an account is deleted, all individual data are also deleted, and there is anonymous feedback.
- Unique selling points developed by the students were rewards for breaks and not only for work phases (to promote a healthy balance between resting and working), the combination of various functions in 1 app, and evidence-based psychoeducation for users' daily lives.
- The app is meant to be a free app in the app stores or an app with a monthly subscription and a free test version (eg, for 14 days). University licenses and licenses from training facilities should enable use free of charge.

The theoretical prototypes developed as part of the elective course were designed with a strong practical orientation and regard to applicable national requirements for medical devices. This opens up the possibility of using the products designed in the course in the inpatient or outpatient medical sector in the future.

The Bruhno and Dinotherapy student groups approached Startup4MED with their ideas concerning the app concepts. Together with the start-up support unit, the students were given the opportunity to plan and work on the transfer of their theoretical prototypes into practice. For this purpose, next steps, analyses of marketability and customer needs, and financial requirements were examined in individual consultation sessions. To address the financial requirements for further development of the concepts and their implementation into initial prototypes, suitable funding was identified, and the application process was supported. Providing contact with founders of successful medical start-ups, additional team members, and mentors from the Startup4MED network further aimed to achieve translation of theory into practice.

The app concepts developed during the course showed valuable potential for the students to found their own start-ups and generate future-oriented innovations in the field of mental health. In addition, combining education and training in the field of digital health and the practical application of the designed products can make an important contribution to sustainable benefits in the field of innovation.

Evaluation and Suggestions for Improvement

Quantitative Results: Questionnaires

Students were asked to complete a questionnaire that had been designed by the research team to assess the elective and receive suggestions for further improvement. To ensure comparability between the individual semesters, a basic average for each semester from all questionnaires submitted per elective was calculated (a total of 6 seminars within 5 semesters). Overall, 73% (44/60) of the students completed the evaluation survey.

Table 3 shows the results of the different items of the questionnaires for each semester. Figures 7 and 8 display the perceived learning gain and the evaluation of the elective overall (for additional results on the satisfaction with the module as well as the evaluation of the digitalization, see Figures S8 and S9 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 3. Items from the evaluation questionnaire, including means and SDs per course[a].

Item from the evaluation questionnaire	Course					
	SS[b] 2021 (n=5[c]), mean (SD)	WS[d] 2021-2022 (n=11), mean (SD)	WS 2021-2022 block seminar (n=9), mean (SD)	SS 2022 (n=5), mean (SD)	WS 2022-2023 (n=5), mean (SD)	WS 2023-2024 (n=9), mean (SD)
Overall assessment of the elective	1.50 (0.50)	1.70 (0.80)	2.40 (1.10)	1.30 (0.40)	1.50 (0.50)	1.80 (0.80)
Digital implementation	1.70 (0.80)	1.70 (1.10)	1.80 (0.70)	1.20 (0.30)	1.40 (0.60)	1.60 (0.70)
The learning module was well structured.	1.60 (0.50)	1.50 (0.70)	1.60 (0.70)	1.00 (0.00)	1.00 (0.00)	1.90 (0.90)
The content was made easy to understand.	1.20 (0.40)	1.60 (0.90)	1.80 (0.80)	1.00 (0.00)	1.40 (0.50)	1.30 (0.50)
The distinction between crucial information and insignificant particulars became evident.	2.00 (1.00)	2.10 (1.00)	3.40 (1.60)	1.50 (0.90)	2.00 (0.70)	2.10 (1.10)
The significance of the material instructed for the medical field became evident.	1.00 (1.00)	1.50 (1.00)	2.70 (1.10)	1.40 (0.50)	1.40 (0.50)	1.70 (0.70)
The visuals (such as interactive images and video tutorials) made it easier to grasp the educational material.	1.60 (0.90)	1.40 (0.50)	1.80 (0.80)	1.60 (0.50)	1.20 (0.40)	1.80 (0.80)
The integrated tests (eg, quizzes) were aligned with the learning content.	2.40 (0.80)	—[e]	—	—	—	—
The integrated tests (eg, quizzes) helped me check my understanding of the learning content.	2.40 (1.30)	—	—	—	—	—
Concrete tips were provided for following up on the learning material.	1.20 (0.40)	1.50 (1.20)	1.60 (0.70)	1.20 (0.40)	1.50 (0.50)	1.60 (0.70)
I found it easy to stay motivated throughout the module.	2.20 (1.30)	2.10 (1.50)	2.90 (1.20)	1.00 (0.00)	1.60 (0.90)	1.80 (1.20)
Access to the learning module was successful without any issues.	1.20 (0.40)	1.80 (1.50)	1.30 (0.50)	1.00 (0.00)	1.00 (1.00)	1.20 (0.40)
The learning module was easy to use.	1.20 (0.40)	1.50 (0.70)	1.30 (0.50)	1.00 (0.00)	1.40 (0.90)	1.40 (0.50)
All participants were engaged in the course.	—	1.70 (0.80)	1.80 (1.00)	1.60 (1.30)	1.20 (0.40)	1.80 (1.30)
There were various opportunities available for inquiries and exchanges (eg, through Rocket.Chat, Webex, ILIAS, or email).	—	1.20 (0.60)	1.40 (0.50)	1.40 (0.90)	1.00 (1.00)	1.20 (0.40)

[a]With regard to the first 2 items listed in the table, students were requested to provide an assessment of the elective subject in terms of the German school grades. A 6-point Likert scale ranging from 1 (“completely agree”) to 6 (“do not agree at all”) was used for the subsequent items.

[b]SS: summer semester.

[c]Questionnaires submitted per course.

[d]WS: winter semester.

[e]These items were subsequently removed from the questionnaire, consequently they were only surveyed for SS 2021.

XSL•FO
RenderX

Figure 7. Assessment of students' perceived learning gain from the elective conducted at the end of each semester. SS: summer semester; WS: winter semester.

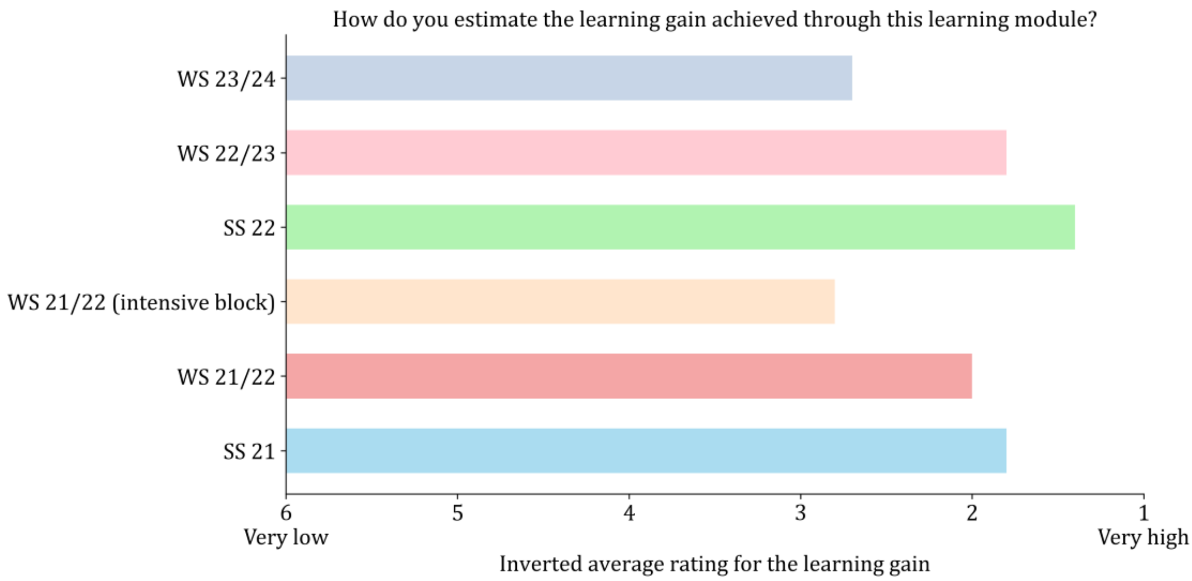
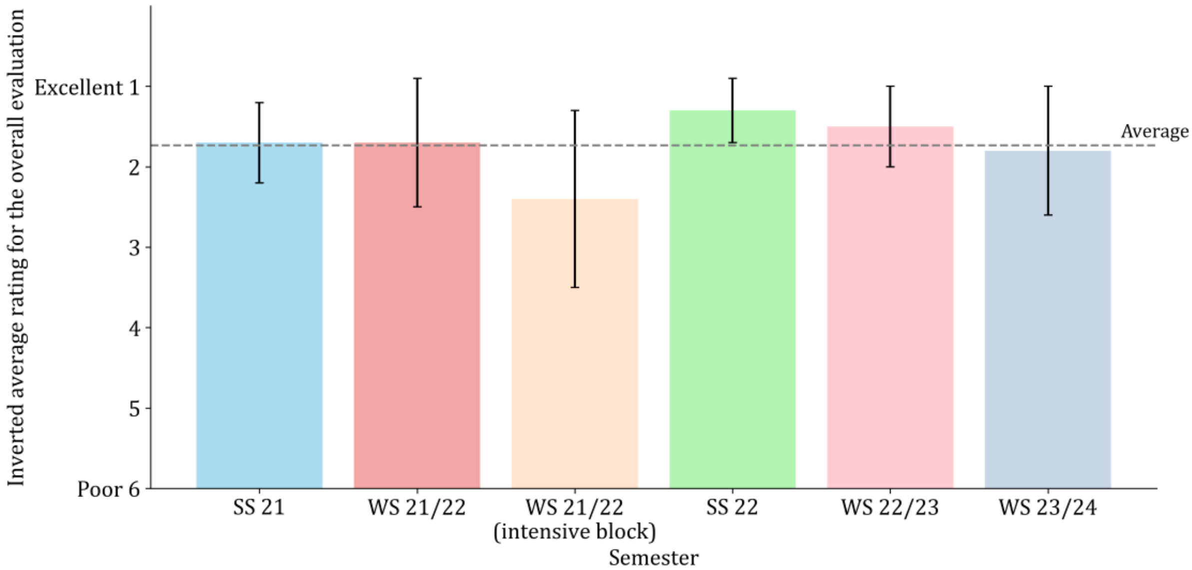


Figure 8. Overall evaluation of the elective subjects over 6 courses within 5 semesters. The ratings are based on student ratings in accordance with the German grading system (1=completely agree; 6=do not agree at all), with lower values indicating better ratings. The inversion is intended to improve readability, with higher bars now corresponding to higher satisfaction ratings. The dashed line indicates the average rating across all semesters. SS: summer semester; WS: winter semester.



The general feedback in the different semesters showed that the elective was well received by the students. The relevance of the various contents for their future work life was noticeable for many students, and there were enough opportunities to ask questions and discuss issues. It was also positively mentioned that the contents were explained in a clear and comprehensible way, that the learning module could be accessed easily, and that the elective was well structured. Room for improvement became apparent (eg, with regard to self-motivation, the perceived helpfulness of the integrated tests and quizzes, and the differentiation between important contents and less important detailed knowledge).

In the winter semester of 2021 to 2022, when the elective was presented as an intensive block seminar on 5 consecutive days, the overall rating and the rating concerning digital implementation were the lowest (Table 3) compared to those in the other semesters. However, the mean score was good in terms of German school grades even in the worst-rated intensive course. Concerning the intensive block seminar, students were less satisfied with the block course but still rated it as good.

XSL•FO
RenderX

They reported having difficulties motivating themselves to do the tasks and complete the module.

Qualitative Results: Open Questions

In the open questions, students could give additional feedback (eg, suggestions for improving the elective) in the form of free text. Overall, many of them seemed to appreciate contents concerning digital health and challenges in implementing digital health offerings and perceived a substantial knowledge gain after taking part in the elective. Students liked the autonomy, flexibility, support from teaching staff, contents, and structure best:

> *Autonomy and free time management. Interesting content that is otherwise not covered in your studies! I learned a lot. Immediate feedback from the teaching staff, absolutely great!* [Summer semester 2021]
>
> *The entire elective offered us the opportunity to freely shape our ideas, but it was all within a professional framework.* [Winter semester 2021-2022]
>
> *Great support!* [Summer semester 2022]
>
> *Freedom of choice for the app concept.* [Winter semester 2023-2024]

Additional factors that were perceived as positive were creativity, the fact that they could always ask questions, the guest lectures, fixed submission dates, and social exchange:

> *The freedom to design and be creative yourself.* [Winter semester 2021-2022]
>
> *Plenty of room to ask questions.* [Winter semester 2021-2022 intensive block]
>
> *Permanent availability of the teaching staff.* [Summer semester 2021]
>
> *The guest lectures were very interesting.* [Winter semester 2023-2024]

When being asked which competence area students benefited most from, several participants mentioned knowledge gain, self-sufficiency, and soft skills. It was also important to assess perceived deficits of the elective and aspects that could be improved. Students mentioned that submissions and arrangements within the work group seemed to be improvable (eg, submission deadlines could have been communicated more clearly right at the beginning, and communication between group members should be enhanced):

> *Addressing clearer deadlines at the beginning: For me, it was a bit difficult to understand the structure with the respective modules etc. right away.* [Summer semester 2021]

Some participants perceived group work as exhausting because of the structure of the groups (ie, group members were determined based on thematic interests and it did not always seem to be clear enough who else was part of the group) and different work attitudes. It was also mentioned that some more face-to-face sessions would have been helpful (eg, to address problems and issues and to get in contact with the other group members). More information on how to realize and implement app ideas and concepts was also desired:

> *I would have liked more guest lectures to learn more about their experiences in app development.* [Winter semester 2022-2023]
>
> *How do you actually turn a concept into an app? Who else do you need for this; how do you get the app into the app store when it's nearly finished?* [Winter semester 2021-2022]

Some students from the block seminar in February 2022 (winter semester 2021-2022) would have liked to have more time for the contents dealt with as it seemed to be a lot of input in a short period for them:

> *It's a lot of input for one week. Maybe you can try to skip or cut some parts if possible.* [Winter semester 2022-2023]

Furthermore, it was said that the elective would have been preferred as an in-class or face-to-face event to facilitate exchange and discussions:

> *For me, it would be easier to follow the content of the course in person. Unfortunately, I drifted off a lot of times.* [Winter semester 2022-2023]

Concerning missing contents or topics, the following suggestions were made: more economic contents (eg, financial aspects, approaching sponsors, and approaching programmers), more information on how to turn a concept into an actual app or how to publish the app in an app store, and more guest lectures.

In the winter semester of 2023 to 2024, the issue of artificial intelligence (AI) was also brought up:

> *The topic, or presentation on AI in the medical setting was something new, I would have liked more of that. In general, more variety would be good.* [Winter semester 2023-2024]

Discussion

This study describes findings in the piloting and iterative optimization of a new, digitally mediated elective subject in the field of digital health at a German medical school. A central objective was to involve students actively and continuously in the optimization and design of the elective from the beginning onward and successfully convey relevant content.

Principal Findings

Through multiple iterations, a participatory, optimized, innovative teaching offering was developed tailored specifically to the needs and preferences of medical students. The overall implementation, digital execution, knowledge increase, content relevance, and supervision received (very) positive ratings by the students. The overarching goal of acquiring competencies and knowledge in the field of digital health (digital skills) was clearly achieved according to the surveyed participants.

Students considered several important aspects related to digital health (especially health apps, including DiGAs) and implemented them accordingly in their app prototypes. As previous studies have shown, gamification [39], adherence promotion, and an appealing design are important elements for creating a positive user experience [40,41]. These features were

XSL•FO
RenderX

also considered by many student groups in our elective subject. However, it is important to emphasize that one should not only rely on elements of gamification to develop a useful, appealing, and persuasive app [41]. This was also considered by most of the students in their app concepts as they often included background information, useful functions depending on the purpose of the app, and information on data security. As data security concerns are one of the main barriers to the acceptance and broad implementation of mobile health apps, addressing these issues in the process of the app's development is essential [42].

According to the participants, the expansion of their knowledge in the area of mobile health apps was successfully realized as they reported having more information and an improved understanding of digital health after completing the seminar. In particular, the opportunities for independent and creative work seemed to convince the students of our approach to teaching digital health and digital competencies. Examining the mean scores for digital delivery revealed a slight positive trend. The comprehensive ratings of 4 courses ranged from "very good" to "good plus" in terms of overall implementation and digital delivery. Interestingly, the block course (on 5 consecutive days) in the winter semester of 2021 to 2022 received substantially lower ratings than the other courses during the semester despite higher investment in the supervision. Potentially, the intensive course involved a workload perceived as too high with too little time between the sessions to develop the app concepts in a creative process that may benefit from a longer time to discuss the ideas within the group. Feedback regarding improvement suggestions indicated that the content and development of an app concept required more time for processing, which negatively affected student satisfaction with the elective in the form of a block seminar.

We were able to implement the proposed improvements, which proved successful as these suggestions were not mentioned again in subsequent course evaluations. However, it must be noted that there might be other possible reasons for this (eg, students with other needs in the following courses or students with generally lower expectations or previous knowledge). Some students emphasized in their feedback that the topics covered filled a gap in their previous education. This perceived gap aligns with the results of elective subjects in the field of digital health at other universities. Before participating in the digital health elective, a survey was conducted at another German medical school, the Charité–Berlin University Medicine, where >85% of participating students stated that digital medicine was not sufficiently integrated into the current curricula [25]. A further study examining medical students' perspectives indicated their desire for a more robust integration of digital health into the curriculum [43]. While certain German medical faculties have initiated the provision of digital health electives [2], the integration of digital health topics into medical education in Germany and Europe is still not firmly established [9,26,44]. During our last elective course, the topic of AI was suggested for the first time. We had already anticipated this and included it as an excursus, which was well received by the students during the seminar. Current research indicates that most physicians and medical students have a favorable attitude toward the integration of AI into medical education. Many are already studying AI or intend to do so. The introduction of AI into the curriculum must be carefully planned to ensure that students' education remains up-to-date. The digitalization of the health care system, the use of digital health apps, and AI in medicine are interconnected. Therefore, medical curricula should be adapted to the digital age as soon as possible [45].

However, contrary to expectations, no substantial increase in the number of participants, reflecting the demand for digital health education, was observed. In all semesters, the total number of participants in the elective course remained below the maximum capacity (5 to 17 students who successfully completed the elective). One exception was the winter semester of 2021 to 2022, in which the course was offered both weekly and as a block course during the lecture-free period with 30 course places, which was the upper limit of participants. This could be attributed to the wide range of parallel electives offered at the Medical Faculty at HHU (approximately 150 different electives each semester [46]) as well as the lack of integration of digital health into the National Competence-Based Learning Objectives for Undergraduate Medical Education. Higher registration numbers for other electives indicate that students' interest and focus rather seem to lie on clinical diseases or imaging procedures, possibly due to a supposedly clearer and more tangible practical relevance. Feedback of the students in the final discussions after their app concept presentations also revealed that the content of the elective was considered very important but that there were other decisive factors in course selection, such as the integrability into the rest of their schedule and the expected workload. The number of participants should be further improved in possible subsequent courses after the intended integration of digital competencies into the medical curriculum. It might be beneficial to organize more large-scale courses and interdisciplinary sessions, maybe also in collaboration with other departments in the medical school and beyond (eg, medical informatics) to increase the range and achieve higher participation numbers. In doing so, the relevance of digital health for students' future daily work as physicians could also be highlighted. It is also worth discussing to what extent students' needs regarding organization, amount of work, and implementation could be combined (eg, preferences regarding block seminars in the semester break, face-to-face sessions, or blended learning formats).

The provision of digital health education represents an important step in supporting future practicing physicians who need to be able to use and prescribe digital health interventions. By integrating the elective into medical curricula, the willingness of physicians and psychotherapists to prescribe DiGAs could be increased. In a previous study, 63% of surveyed general practitioners indicated a relatively low willingness to prescribe DiGAs [47]. This could be attributed to concerns about safety, reliability, additional workload [48], personal uncertainty [21], lack of knowledge, lack of reliable information sources [49], and a lack of evidence of effectiveness [19]. The inclusion of digital health education in medical studies could address these concerns and positively influence the already established infrastructure of DiGAs through well-designed and trusted information based on academic training.

XSL•FO
RenderX

Limitations

The limitations of this case study are the small sample sizes and the conduction at only 1 medical school in Germany, which may restrict the generalization of the findings of the evaluation to a broader or more diverse population of medical students (eg, students with different preferences and experiences and students in more advanced semesters). We did also open the elective to further study subjects beyond medicine to broaden the perspectives and capture the interdisciplinary nature of digital health. The integration of a certain proportion of other subjects was also beneficial because there are approximately 150-200 electives per semester at our medical school that the students can freely choose from and we could only conduct the elective with at least 5 students. Nonetheless, we adhered to the requirement that the vast majority of the web-based seminars had to consist of medical students (usually at least two-thirds of the participants).

Furthermore, we did not statistically compare the evaluation outcomes (mean scores) across the semesters as the sample sizes were small and unequal. Instead, initial individual insights can serve as starting points for a broader implementation of digital health in the German curricula for medical students. In addition, there could be a potential bias due to participant self-selection—students chose the elective course themselves. Therefore, it might be possible that they already had a greater interest in the topic or were more familiar with digital health in general, which might have led to more positive evaluations.

Furthermore, the prescription of DiGAs, which was a key topic of the elective, is still unique to the German health care system, making several of the contents of the lectures and ILIAS modules for self-guided learning hardly generalizable to medical education in other countries. However, other aspects are generally important for education on digital health and app development (eg, gamification). In addition, the design thinking approach to develop app concepts is universally adaptable to electives that aim to foster digital competencies in medical education outside the context of Germany.

Another limitation is the fact that the input of lecturers and experts might have had a strong influence on the students' focus on the development of app concepts. This could have led to a potential bias on the part of the students in terms of topics and priorities in the implementation process. However, it was essential to provide sufficient knowledge beforehand, and this enabled students to put special emphasis on factors that were relevant for them (such as gamification, adherence promotion, or data security). Furthermore, the evaluation was based primarily on students' feedback. Thus, the possible gain of knowledge and usefulness for the students cannot really be assessed. In future electives, it could be helpful to assess pre- and postknowledge on digital health topics addressed in the elective in a standardized manner. In this project, we only gathered this information via verbal feedback in the first and last session as well as via 1 item in the anonymized survey (perceived knowledge gains). However, students' presentations and feedback indicated that they seemed to have learned relevant aspects concerning digital health. Finally, there was no feedback from students who started the elective but chose to terminate participation or from students who did not choose the elective in the first place. These insights could have been useful to further improve the course and find out more about the generally rather low participation rate (eg, whether this was related to the course content or to organizational and time aspects).

Implications

For Medical Schools

Currently, there is a discrepancy between the demand for and the benefits of DMHIs and their integration into medical education. Although medical students acknowledge the importance of digital health in enhancing core skills [50], participation in our elective was rather low. Existing research literature highlights that students and health care providers have a discernible knowledge gap in the domain of digital health. To address this, providing clear information about the benefits of such electives is crucial. A general lack of awareness, compounded by insufficient information from academic faculties, may hinder progress. In Germany, digital health courses are limited, mainly offered as electives [2]. Even though more and more medical schools participate in surveys in which they are asked whether they offer digital health courses, indicating an interest in the topic, the number of courses they actually provide appears to stagnate [51]. This discrepancy highlights the importance of defining standards and providing guidance for digital health education across medical schools. Integrating digital health into curricula has the potential to enhance future physicians' capabilities, but specialized training is vital for navigating the digital age. Introducing digital health competencies equips aspiring clinicians with essential skills. These include fostering positive patient-physician relationships and explaining the risks and benefits. Foundational knowledge can be integrated into standard medical curricula, with electives focusing on specific in-depth knowledge and specializations [30].

For Teaching Staff

To successfully implement and execute courses that focus on digital health and app development, it is useful to provide enough time for the students to take in relevant aspects and knowledge (eg, regular weekly courses instead of block seminars). Regular exchange among students and between students and teaching staff also seems to be important. The use of interactive web-based learning platforms and chat programs can also improve the course (eg, ILIAS and Rocket.Chat). Involving different occupational groups and experts (eg, start-up developers and economists) facilitates networking and allows for the exposure to different perspectives and in-depth knowledge for the students. The concept itself is scalable at a larger level even though it needs staff deployment, and the focus or topics of the elective are interchangeable. Finally, it could be beneficial to collaborate with other institutes or departments to increase the range of courses related to digital health.

For Researchers

Further quantitative and qualitative research methods could be useful to gain more insights into the preferences and needs of medical students with regard to digital health. It might also be beneficial to conduct research concerning outcomes of learning

XSL•FO
RenderX

success that goes beyond individual feedback in the form of questionnaires and open questions. Research among the general population of medical students in Germany is needed to deepen existing knowledge, further adapt elective courses, and prepare the integration of digital health into the medical curricula.

Conclusions and Outlook

A current challenge in health care and prevention is a lack of knowledge and competencies in the field of digital health among health care providers. To increase the prescription readiness and establish digital health in the everyday lives of patients and physicians, it needs to be implemented in medical education. Further research is needed to define specific learning objectives for digital health competencies and develop detailed recommendations for their integration into medical curricula.

Acknowledgments

First, the authors cordially thank the Dean and the Commission for Quality Improvement (QI) in Teaching and Studies of the Medical Faculty of Heinrich Heine University Düsseldorf (HHU), including the Dean’s Office, for supporting their QI project and their helpful advice (e.g., consultation and webinars on ILIAS [Integriertes Lern-, Informations- und Arbeitskooperations-System (German for “Integrated Learning, Information and Work Cooperation System”)]). The QI project was funded by decentralized budget means from the Dean’s Office of the Medical Faculty of HHU based on the endorsement by the Commission for QI for the first 2 years (QI grant no. 39/21). They would like to express their sincere gratitude to Stefan Stehl (research associate until April 2023) for his valuable support in programming and optimizing additional e-learning material, his excellent contribution to the supervision of the creation of digital mock-ups, and his support in various operative tasks (organization and coordination). His contribution was crucial to the successful completion of this QI project. The authors also thank Dr Thomas Muth for his support in administrative tasks. Furthermore, they thank the Evaluation Department of the Medical Faculty for their support with the web-based evaluation using evasys (evasys GmbH) and for providing them with aggregated evaluation results. The authors thank all students who participated in the elective. In particular, they would like to thank Sophie Smutny, Roxana Hulpoi, Dang Quoc-Dat, Annik Roßberg, Viola Neumann, Jasmin Mirheli, Grit Klöker, and Nils Gerisch, who generously shared their results and provided their prototype pictures, contributing significantly to the depth and quality of this proof-of-concept study. Without their combined efforts and contributions, this work would not have been possible. The authors also thank Professor Dr Nils Hansson, Institute of History, Theory and Ethics in Medicine (HHU), for collaborating with them and sharing ideas on the topic of digital health in medical education. In addition, they thank the DiGA providers for their guest lectures (GAIA AG, Mementor GmbH, and Elona Health GmbH). They also would like to express their gratitude to Büsra Köprücü for her swift assistance with the adapted visualization of the figure of the Dinotherapy app mockups for publication purposes. Finally, the authors thank the Heinrich Heine University Düsseldorf for covering the publication fees (University Library, Open-Access-Funds of the Heinrich Heine University Düsseldorf).

Authors' Contributions

FS contributed significantly to the methodology, conceptualization, formal analysis, software development, visualization, data curation, investigation, and the writing of the original draft. LG was involved in the methodology, conceptualization, formal analysis, data curation, visualization, and contributed to the writing of the original draft. KP participated in the conceptualization, investigation, and visualization, and was responsible for reviewing and editing the manuscript. AS-K was responsible for the software development, investigation, and visualization, and contributed to the review and editing of the manuscript. IE was involved in the conceptualization and contributed to the review and editing of the manuscript. BDS contributed to the conceptualization and was responsible for reviewing and editing the manuscript. JA-H played a key role in conceptualization, funding acquisition, project administration, and methodology. JA-H also contributed to the investigation, data curation, supervision, and was responsible for reviewing and editing the manuscript.

Conflicts of Interest

None declared. DiGA providers held their lectures for free, and the authors did not receive any kind of compensation from them.

Multimedia Appendix 1
Additional prototype visualizations.
[PDF File (Adobe PDF File), 481 KB - mededu_v10i1e59454_app1.pdf]

Multimedia Appendix 2
Additional results.
[PDF File (Adobe PDF File), 206 KB - mededu_v10i1e59454_app2.pdf]

References

XSL•FO
RenderX

1. Lauer W, Löbker W, Höfgen B. [Digital health applications (DiGA): assessment of reimbursability by means of the "DiGA Fast Track" procedure at the Federal Institute for Drugs and Medical Devices (BfArM)]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Oct;64(10):1232-1240 [FREE Full text] [doi: 10.1007/s00103-021-03409-7] [Medline: 34529095]
2. Aulenkamp J, Mikuteit M, Löffler T, Schmidt J. Overview of digital health teaching courses in medical education in Germany in 2020. GMS J Med Educ 2021;38(4):Doc80 [FREE Full text] [doi: 10.3205/zma001476] [Medline: 34056069]
3. Nitsche J, Busse TS, Ehlers JP. Teaching digital medicine in a virtual classroom: impacts on student mindset and competencies. Int J Environ Res Public Health 2023 Jan 22;20(3):1-17 [FREE Full text] [doi: 10.3390/ijerph20032029] [Medline: 36767393]
4. Waibel AM, Bischoff M. Digitale Kompetenzen im Medizinstudium: Ergebnisse einer interdisziplinären Lehrveranstaltung. HNO 2024 Mar;72(3):161-165 [FREE Full text] [doi: 10.1007/s00106-023-01411-w] [Medline: 38265753]
5. Särchen F, Springborn S, Mortsiefer A, Ehlers J. Digital learning about patients: an online survey of German medical students investigating learning strategies for family medical video consultations. Digit Health 2024;10:20552076241230070 [FREE Full text] [doi: 10.1177/20552076241230070] [Medline: 38323240]
6. Root E, Caskie G. eMental health literacy and the relationship to barriers to mental health care. Innov Aging 2020;4(Supplement_1):305. [doi: 10.1093/geroni/igaa057.978]
7. Digital health in the medical curriculum: addressing the needs of the future health workforce. European Medical Students' Association. URL: https://emsa-europe.eu/wp-content/uploads/2021/06/Policy-2019-04-Digital-Health-in-the-Medical-Curriculum_-Addressing-the-Needs-of-the-Future-Health-Workforce.pdf [accessed 2024-04-29]
8. Dederichs M, Nitsch FJ, Apolinário-Hagen J. Piloting an innovative concept of e-mental health and mHealth workshops with medical students using a participatory co-design approach and app prototyping: case study. JMIR Med Educ 2022 Jan 10;8(1):e32017 [FREE Full text] [doi: 10.2196/32017] [Medline: 35006085]
9. Poncette AS, Glauert DL, Mosch L, Braune K, Balzer F, Back DA. Undergraduate medical competencies in digital health and curricular module development: mixed methods study. J Med Internet Res 2020 Oct 29;22(10):e22161 [FREE Full text] [doi: 10.2196/22161] [Medline: 33118935]
10. Mayer G, Gronewold N, Alvarez S, Bruns B, Hilbel T, Schultz JH. Acceptance and expectations of medical experts, students, and patients toward electronic mental health apps: cross-sectional quantitative and qualitative survey study. JMIR Ment Health 2019 Nov 25;6(11):e14018 [FREE Full text] [doi: 10.2196/14018] [Medline: 31763990]
11. Dederichs M, Weber J, Pischke CR, Angerer P, Apolinário-Hagen J. Exploring medical students' views on digital mental health interventions: a qualitative study. Internet Interv 2021 Sep;25:100398 [FREE Full text] [doi: 10.1016/j.invent.2021.100398] [Medline: 34026567]
12. Zahlen und Fakten. Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde (DGPPN). URL: https://www.dgppn.de/schwerpunkte/zahlenundfakten.html [accessed 2023-12-20]
13. Robert-Koch-Institut. Hochfrequente Mental Health Surveillance (Depressive Symptome). GitHub. URL: https://tinyurl.com/52cp4emu [accessed 2024-08-30]
14. Mauz E, Walther L, Junker S, Kersjes C, Damerow S, Eicher S, et al. Time trends in mental health indicators in Germany's adult population before and during the COVID-19 pandemic. Front Public Health 2023;11:1065938 [FREE Full text] [doi: 10.3389/fpubh.2023.1065938] [Medline: 36908429]
15. Phillips EA, Himmler SF, Schreyögg J. Preferences for e-Mental health interventions in Germany: a discrete choice experiment. Value Health 2021 Mar;24(3):421-430 [FREE Full text] [doi: 10.1016/j.jval.2020.09.018] [Medline: 33641777]
16. Jia Q, Qu Y, Sun H, Huo H, Yin H, You D. Mental health among medical students during COVID-19: a systematic review and meta-analysis. Front Psychol 2022;13:846789 [FREE Full text] [doi: 10.3389/fpsyg.2022.846789] [Medline: 35619776]
17. Wada M, Suto MJ, Lee M, Sanders D, Sun C, Le TN, et al. University students' perspectives on mental illness stigma. Ment Health Prev 2019 Jun;14:200159. [doi: 10.1016/j.mph.2019.200159]
18. Uncovska M, Freitag B, Meister S, Fehring L. Patient acceptance of prescribed and fully reimbursed mHealth apps in Germany: an UTAUT2-based online survey study. J Med Syst 2023 Jan 27;47(1):14 [FREE Full text] [doi: 10.1007/s10916-023-01910-x] [Medline: 36705853]
19. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. Physicians' attitudes toward prescribable mHealth apps and implications for adoption in Germany: mixed methods study. JMIR Mhealth Uhealth 2021 Nov 23;9(11):e33012 [FREE Full text] [doi: 10.2196/33012] [Medline: 34817385]
20. Hoch P, Arets J. Die Videosprechstunde als Modell der Akzeptanz digitaler Leistungen im Gesundheitswesen. B&G Bewegungstherapie und Gesundheitssport 2021 Aug 11;37(04):151-156. [doi: 10.1055/a-1528-3793]
21. Wangler J, Jansky M. [What potential and added value do DiGA offer for primary care?-results of a survey of general practitioners in Germany]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2022 Dec;65(12):1334-1343 [FREE Full text] [doi: 10.1007/s00103-022-03608-w] [Medline: 36269336]
22. Heidel A, Hagist C. Potential benefits and risks resulting from the introduction of health apps and wearables into the German statutory health care system: scoping review. JMIR Mhealth Uhealth 2020 Sep 23;8(9):e16444 [FREE Full text] [doi: 10.2196/16444] [Medline: 32965231]

XSL•FO
RenderX

23. Bericht des GKV-Spitzenverbandes über die Inanspruchnahme und Entwicklung der Versorgung mit digitalen Gesundheitsanwendungen (DiGA-Bericht). GKV-Spitzenverband. URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2022_DiGA_Bericht_BMG.pdf [accessed 2024-02-25]
24. Byambasuren O, Beller E, Hoffmann T, Glasziou P. mHealth app prescription in Australian general practice: pre-post study. JMIR Mhealth Uhealth 2020 Jun 01;8(6):e16497 [FREE Full text] [doi: 10.2196/16497] [Medline: 32478660]
25. Seemann RJ, Mielke AM, Glauert DL, Gehlen T, Poncette AS, Mosch LK, et al. Implementation of a digital health module for undergraduate medical students: a comparative study on knowledge and attitudes. Technol Health Care 2023;31(1):157-164 [FREE Full text] [doi: 10.3233/THC-220138] [Medline: 35754241]
26. Behrends M, Paulmann V, Koop C, Foadi N, Mikuteit M, Steffens S. Interdisciplinary teaching of digital competencies for undergraduate medical students - experiences of a teaching project by medical informatics and medicine. Stud Health Technol Inform 2021 May 27;281:891-895. [doi: 10.3233/SHTI210307] [Medline: 34042802]
27. Emina O, Bushra AS, Armin D, Sabina S, Maira AD, Stephanie L. Introducing m-Health and digital diabetes apps in clinical pharmacy education in Germany. J Diabetes Clin Res 2022;4(1):17-19. [doi: 10.33696/diabetes.4.051]
28. Vogt L, Schmidt M, Follmann A, Lenes A, Klasen M, Sopka S. Telemedicine in medical education: an example of a digital preparatory course for the clinical traineeship - a pre-post comparison. GMS J Med Educ 2022;39(4):Doc46 [FREE Full text] [doi: 10.3205/zma001567] [Medline: 36310883]
29. Khurana MP, Raaschou-Pedersen DE, Kurtzhals J, Bardram JE, Ostrowski SR, Bundgaard JS. Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Med Educ 2022 Feb 26;22(1):129 [FREE Full text] [doi: 10.1186/s12909-022-03163-7] [Medline: 35216611]
30. Zainal H, Xin X, Thumboo J, Fong KY. Medical school curriculum in the digital age: perspectives of clinical educators and teachers. BMC Med Educ 2022 Jun 03;22(1):428 [FREE Full text] [doi: 10.1186/s12909-022-03454-z] [Medline: 35659212]
31. Meinel C, Leifer L, Plattner H. Design Thinking: Understand – Improve – Apply. Cham, Switzerland: Springer; 2011.
32. HPI School of Design Thinking. Hasso-Plattner-Institut. URL: https://hpi.de/school-of-design-thinking.html [accessed 2024-02-02]
33. Sandars J, Goh PS. Design thinking in medical education: the key features and practical application. J Med Educ Curric Dev 2020;7:2382120520926518 [FREE Full text] [doi: 10.1177/2382120520926518] [Medline: 32548307]
34. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
35. Terhorst Y, Philippi P, Sander LB, Schultchen D, Paganini S, Bardus M, et al. Validation of the Mobile Application Rating Scale (MARS). PLoS One 2020;15(11):e0241480 [FREE Full text] [doi: 10.1371/journal.pone.0241480] [Medline: 33137123]
36. Philippi P, Baumeister H, Apolinário-Hagen J, Ebert DD, Hennemann S, Kott L, et al. Acceptance towards digital health interventions - model validation and further development of the unified theory of acceptance and use of technology. Internet Interv 2021 Dec;26:100459 [FREE Full text] [doi: 10.1016/j.invent.2021.100459] [Medline: 34603973]
37. Azad-Khaneghah P, Neubauer N, Miguel Cruz A, Liu L. Mobile health app usability and quality rating scales: a systematic review. Disabil Rehabil Assist Technol 2021 Oct;16(7):712-721. [doi: 10.1080/17483107.2019.1701103] [Medline: 31910687]
38. Inal Y, Wake JD, Guribye F, Nordgreen T. Usability evaluations of mobile mental health technologies: systematic review. J Med Internet Res 2020 Jan 06;22(1):e15337 [FREE Full text] [doi: 10.2196/15337] [Medline: 31904579]
39. Wei PS, Lee SY, Lu HP, Tzou JC, Weng CI. Why do people abandon mobile social games? Using candy crush saga as an example. Int J Ind Manuf Eng 2015;9(1):13-18. [doi: 10.1108/intr-04-2013-0082]
40. Al-Shamaileh O, Sutcliffe A. Why people choose Apps: an evaluation of the ecology and user experience of mobile applications. Int J Hum Comput Stud 2023 Feb;170:102965. [doi: 10.1016/j.ijhcs.2022.102965]
41. Jessen S, Mirkovic J, Ruland CM. Creating gameful design in mHealth: a participatory co-design approach. JMIR Mhealth Uhealth 2018 Dec 14;6(12):e11579 [FREE Full text] [doi: 10.2196/11579] [Medline: 30552080]
42. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
43. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
44. Tudor Car L, Kyaw BM, Nannan Panday RS, van der Kleij R, Chavannes N, Majeed A, et al. Digital health training programs for medical students: scoping review. JMIR Med Educ 2021 Jul 21;7(3):e28275 [FREE Full text] [doi: 10.2196/28275] [Medline: 34287206]
45. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res 2023;15(7):4820-4828 [FREE Full text] [Medline: 37560249]
46. Wahlcurriculum. Heinrich-Heine-Universität Düsseldorf. URL: https://www.medizinstudium.hhu.de/duesseldorfer-curriculum-medizin/wahlcurriculum [accessed 2024-02-02]

XSL•FO
RenderX

47. Radić M, Donner I, Waack M, Brinkmann C, Stein L, Radić D. Digitale Gesundheitsanwendungen: Die Akzeptanz steigern. Dtsch Arztebl 2021;118(6):A286 [FREE Full text]
48. Wangler J, Jansky M. Gesundheits-Apps als Instrumente der Prävention? – Eine Interviewstudie zu Potenzialen für das hausärztliche Setting. Präv Gesundheitsf 2020 Mar 31;15(4):340-346. [doi: 10.1007/s11553-020-00769-x]
49. Byambasuren O, Beller E, Hoffmann T, Glasziou P. Barriers to and facilitators of the prescription of mHealth apps in Australian general practice: qualitative study. JMIR Mhealth Uhealth 2020 Jul 30;8(7):e17447 [FREE Full text] [doi: 10.2196/17447] [Medline: 32729839]
50. Boyers LN, Schultz A, Baceviciene R, Blaney S, Marvi N, Dellavalle RP, et al. Teledermatology as an educational tool for teaching dermatology to residents and medical students. Telemed J E Health 2015 Apr;21(4):312-314 [FREE Full text] [doi: 10.1089/tmj.2014.0101] [Medline: 25635528]
51. Pourmand A, Ghassemi M, Sumon K, Amini SB, Hood C, Sikka N. Lack of telemedicine training in academic medicine: are we preparing the next generation? Telemed J E Health 2021 Jan;27(1):62-67. [doi: 10.1089/tmj.2019.0287] [Medline: 32294025]

Abbreviations

AI: artificial intelligence
ILIAS: Integriertes Lern-, Informations- und Arbeitskooperations-System (German for “Integrated Learning, Information and Work Cooperation System”)
DiGAs: Digitale Gesundheitsanwendungen (German for “digital health applications”)
DMHI: digital mental health intervention
HHU: Heinrich Heine University Düsseldorf

Edited by L Tudor Car; submitted 12.04.24; peer-reviewed by L Fehring, D Abdel-Hady; comments to author 01.07.24; revised version received 09.07.24; accepted 11.07.24; published 20.09.24.

Please cite as:
Sahan F, Guthardt L, Panitz K, Siegel-Kianer A, Eichhof I, Schmitt BD, Apolinario-Hagen J
Enhancing Digital Health Awareness and mHealth Competencies in Medical Education: Proof-of-Concept Study and Summative Process Evaluation of a Quality Improvement Project
JMIR Med Educ 2024;10:e59454
URL: https://mededu.jmir.org/2024/1/e59454
doi:10.2196/59454
PMID:

©Fatma Sahan, Lisa Guthardt, Karin Panitz, Anna Siegel-Kianer, Isabel Eichhof, Björn D Schmitt, Jennifer Apolinario-Hagen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.09.2024. 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.

XSL•FO
RenderX

Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study

Marcos Rojas[1,*], MD; Marcelo Rojas[2,*], MD; Valentina Burgess[2,*], MD; Javier Toro-Pérez[2,*]; Shima Salehi[1,*], PhD

1

2

*all authors contributed equally

Corresponding Author:
Marcos Rojas, MD

Abstract

Background: The deployment of OpenAI's ChatGPT-3.5 and its subsequent versions, ChatGPT-4 and ChatGPT-4 With Vision (4V; also known as "GPT-4 Turbo With Vision"), has notably influenced the medical field. Having demonstrated remarkable performance in medical examinations globally, these models show potential for educational applications. However, their effectiveness in non-English contexts, particularly in Chile's medical licensing examinations—a critical step for medical practitioners in Chile—is less explored. This gap highlights the need to evaluate ChatGPT's adaptability to diverse linguistic and cultural contexts.

Objective: This study aims to evaluate the performance of ChatGPT versions 3.5, 4, and 4V in the EUNACOM (Examen Único Nacional de Conocimientos de Medicina), a major medical examination in Chile.

Methods: Three official practice drills (540 questions) from the University of Chile, mirroring the EUNACOM's structure and difficulty, were used to test ChatGPT versions 3.5, 4, and 4V. The 3 ChatGPT versions were provided 3 attempts for each drill. Responses to questions during each attempt were systematically categorized and analyzed to assess their accuracy rate.

Results: All versions of ChatGPT passed the EUNACOM drills. Specifically, versions 4 and 4V outperformed version 3.5, achieving average accuracy rates of 79.32% and 78.83%, respectively, compared to 57.53% for version 3.5 ($P<.001$). Version 4V, however, did not outperform version 4 ($P=.73$), despite the additional visual capabilities. We also evaluated ChatGPT's performance in different medical areas of the EUNACOM and found that versions 4 and 4V consistently outperformed version 3.5. Across the different medical areas, version 3.5 displayed the highest accuracy in psychiatry (69.84%), while versions 4 and 4V achieved the highest accuracy in surgery (90.00% and 86.11%, respectively). Versions 3.5 and 4 had the lowest performance in internal medicine (52.74% and 75.62%, respectively), while version 4V had the lowest performance in public health (74.07%).

Conclusions: This study reveals ChatGPT's ability to pass the EUNACOM, with distinct proficiencies across versions 3.5, 4, and 4V. Notably, advancements in artificial intelligence (AI) have not significantly led to enhancements in performance on image-based questions. The variations in proficiency across medical fields suggest the need for more nuanced AI training. Additionally, the study underscores the importance of exploring innovative approaches to using AI to augment human cognition and enhance the learning process. Such advancements have the potential to significantly influence medical education, fostering not only knowledge acquisition but also the development of critical thinking and problem-solving skills among health care professionals.

(JMIR Med Educ 2024;10:e55048) doi:10.2196/55048

KEYWORDS

artificial intelligence; AI; generative artificial intelligence; medical education; ChatGPT; EUNACOM; medical licensure; medical license; medical licensing exam

Introduction

The launch of OpenAI's ChatGPT-3.5 in November 2022 has impacted various fields, including medical education [1]. On September 25, 2023, OpenAI announced the release of a highly anticipated new functionality, ChatGPT-4 With Vision (4V; also known as "GPT-4 Turbo With Vision"), to support multimodal interaction and further exploration [2].

ChatGPT has shown promise, or some would argue that it is a threat, for medical education with its outstanding performance in several medical examinations. For example, in the Médicos Internos Residentes examination in Spain [3], ChatGPT

XSL•FO
RenderX

answered 51.4% of the questions correctly [3]. In the United States, different studies have reported an accuracy of 80%-90% on the United States Medical Licensing Examination [4]. These results highlight ChatGPT's potential to impact the future of medical education. However, there is a limited understanding of ChatGPT's performance in non-English examinations in Latin America, such as Chile's EUNACOM (Examen Único Nacional de Conocimientos de Medicina).

The EUNACOM comprises 180 multiple-choice questions from various medical areas such as internal medicine, pediatrics, obstetrics and gynecology, surgery (general surgery and anesthesia, traumatology, and urology), psychiatry, specialties (including dermatology, ophthalmology, and otorhinolaryngology), and public health. The examination assesses topics such as diagnosis, treatment, and follow-up care. Passing the EUNACOM is vital for foreign-trained doctors to practice in Chile and for Chilean medical students to complete their studies and transition to medical practice [5]. This examination, central to Chilean medical education, can potentially pose linguistic, cultural, and contextual challenges to ChatGPT. This study aimed to evaluate the performance of ChatGPT versions 3.5, 4, and the recently released 4V on EUNACOM practice drills, with the intent to guide future improvements—specifically, the integration and use of artificial intelligence (AI) in medical education—across various cultural and linguistic contexts, thereby contributing to the ongoing debate on the role and efficacy of AI as an educational tool in the global medical community.

Methods

Study Design

We adopted a quantitative, descriptive, cross-sectional approach to evaluate ChatGPT's performance in the EUNACOM practice drills. We gathered a data set of EUNACOM practice questions, categorized them, and analyzed the responses of ChatGPT versions 3.5, 4, and 4V.

EUNACOM Data Set

It is challenging to obtain an authentic and representative set of questions from the EUNACOM, as the examination is not publicly accessible for integrity and security reasons. Therefore, we used 3 official practice drills designed by the University of Chile as preparatory material for its students. These drills are not included in the data used to train ChatGPT due to their limited public availability. Each drill consists of 180 multiple-choice questions with 5 options, where only 1 is correct. The number of questions across medical areas in each drill reflects the specifications of the EUNACOM's administrative office (ie, internal medicine, n=67; pediatrics, n=29; obstetrics and gynecology, n=29; surgery, n=20; psychiatry, n=14; specialties, n=12; and public health, n=9).

Classification of Questions

The categorization of EUNACOM's questions in this study is in line with that of Carrasco et al [3] in 2023 on the Médicos Internos Residentes examination in Spain. Two of our research team members classified the questions as follows:

1. Medical area according to the EUNACOM: internal medicine, pediatrics, obstetrics and gynecology, surgery, psychiatry, specialties, and public health.
2. Category of questions: "clinical case" if they presented a clinical case in the stem of the question, or "medical knowledge" if the question asked for the retrieval of knowledge of medical content.
3. Type of question in clinical case questions: diagnosis, treatment, or follow-up.

Prompting and Application of ChatGPT

We used ChatGPT versions 3.5, 4, and 4V, trained up to January 2022, to respond to the 3 EUNACOM drills in October 2023. Each drill was conducted 3 times with each version of ChatGPT using the prompt, "Which is the correct answer to the following questions?" We excluded "EUNACOM" from the prompt to guarantee ChatGPT's responsiveness to the questions, since, according to OpenAI's policies, the model abstains from taking official assessments. When using version 4V, we prompted questions with images (eg, x-ray) individually, attaching the image to its corresponding question.

The 3 attempts at providing responses in each drill allowed us to address the variability in ChatGPT's answers, attributable to its probabilistic nature, by estimating an average performance.

Data Analysis

Data analysis was conducted using Stata (version 16.0; StataCorp). We computed the percentage of correct responses for each drill and set the passing score at >51% in accordance with the EUNACOM standard [6]. We used a 2-sample test of proportions to test for differences in performance among different versions of ChatGPT [7].

Ethical Considerations

The Human Research Ethics Committee of the Faculty of Medicine at the University of Chile determined that this study presented no ethical concerns that warranted institutional review board oversight. We used EUNACOM drills authorized by the University of Chile's School of Medicine because access to the actual examination is restricted.

Results

The three versions of ChatGPT successfully passed EUNACOM drills on average. Notably, version 4 exhibited superior performance to that of version 3.5 across all drills and attempts, while version 4V did not show a statistically significant advantage over version 4. The only instance of not passing the EUNACOM was observed with version 3.5, specifically during its third attempt at drill 2. Detailed performance metrics for each drill and attempt are provided in Table 1. To assess the robustness of our results, we also compared the performance of ChatGPT by each attempt and by each drill. The results are qualitatively similar.

XSL•FO
RenderX

Table . Correct answers of ChatGPT versions 3.5, 4, and 4 With Vision on each of the EUNACOM[a] drills (each with 180 multiple-choice questions) per attempt.

EUNACOM drill and attempt		Correct answers provided by each version of ChatGPT, n (%)		
		3.5[b]	4[c]	4 With Vision[d]
Drill 1				
	1	105 (58.33)	143 (79.44)	147 (81.67)
	2	109 (60.56)	148 (82.22)	149 (82.78)
	3	103 (57.22)	146 (81.11)	145 (80.56)
Drill 2				
	1	93 (51.67)	138 (76.67)	133 (73.89)
	2	94 (52.22)	134 (74.44)	139 (77.22)
	3	86 (47.78)[e]	132 (73.33)	137 (76.11)
Drill 3				
	1	112 (62.22)	143 (79.44)	142 (78.89)
	2	114 (63.33)	150 (83.33)	139 (77.22)
	3	116 (64.44)	151 (83.89)	146 (81.11)

[a]EUNACOM: Examen Único Nacional de Conocimientos de Medicina.

[b]Mean accuracy rate 57.53% (95% CI 55.12%-59.94%).

[c]Mean accuracy rate 79.32% (95% CI 77.35%-81.29%); $z_{3.5 \text{ vs } 4}$=–13.34, P<.001 (2-sample test of proportions).

[d]Mean accuracy rate 78.83% (95% CI 76.84%-80.82%); $z_{4 \text{ vs } 4V}$=0.35, P=.73 (2-sample test of proportions).

[e]This is the only instance of not passing the EUNACOM practice drill.

Across all attempts and the 3 practice drills, we observed a variation in average accuracy rates by both medical area and clinical case question type. In an evaluation across various medical areas, all 3 ChatGPT versions demonstrated distinct high and low performances. For version 3.5, the highest accuracy was observed in psychiatry (average 69.84%), while the lowest accuracy rate was observed in internal medicine (average 52.74%). Version 4 excelled in surgery with a 90.00% average accuracy rate, whereas its weakest performance was observed in internal medicine (average 75.62%). Similarly, version 4V's performance was strongest in surgery (average 86.11%) and weakest in public health (average 74.07%). When analyzing performance across different medical areas, ChatGPT-4 consistently outperformed ChatGPT-3.5. However, ChatGPT-4V did not significantly outperform ChatGPT-4.

The 3 drills included a total of 501 clinical case questions and 39 medical knowledge questions. In answering clinical case questions, the average accuracy rate of ChatGPT across the 3 attempts was as follows: 57.22% for version 3.5, 80.11% for version 4, and 79.71% for version 4V. In answering medical knowledge questions, the average accuracy rate of ChatGPT was as follows: 61.54% for version 3.5, 74.36% for version 4, and 67.52% for version 4V.

Among the clinical case questions, ChatGPT performed best in follow-up questions, with version 4 scoring 88.89%, while the lowest performance was observed in treatment questions, with version 3.5 scoring 48.50%. On analyzing performance over different types of clinical case questions, ChatGPT-4 regularly outperformed ChatGPT-3.5. Nonetheless, ChatGPT-4V showed no significant difference in performance compared to ChatGPT-4. Comprehensive data on average performances across all medical areas and types of clinical case questions are included in Multimedia Appendix 1.

The 3 drills had a total of 50 questions with images; therein, ChatGPT-4 had an average accuracy rate of 70.67% and version 4V had an average accuracy rate of 70.00% across the 3 attempts.

Discussion

Principal Findings

This study shows that ChatGPT successfully passed the EUNACOM, with version 4 showing a superior performance to that of version to 3.5. However, interestingly, version 4V did not significantly outperform version 4 in this examination. All versions demonstrated proficiency in various medical specialties, with version 3.5 excelling in psychiatry and versions 4 and 4V in surgery. However, unexpectedly, version 4V did not outperform the other 2 versions in questions including images. The differences in performance among versions are likely due to continuous enhancements in training and knowledge with each update, which improve the models' grasp of complex medical subjects. Nevertheless, varying success rates in specific medical fields could stem from the complexities of those specialties, unique terminologies, or the specific structure of the questions in those areas, which may align differently with the data the models were trained on.

In particular, when analyzing the question categories, all versions presented a lower accuracy rate in medical knowledge questions than in clinical case questions, indicating a possible

XSL•FO
RenderX

gap in the models' data regarding specific content knowledge. In clinical case questions, versions 4 and 4V consistently outperformed version 3.5, possibly due to the AI's advancement in pattern recognition. Interestingly, each version performed differently across various types of questions in the clinical case category: version 3.5 showed a lower performance on treatment and follow-up questions, whereas versions 4 and 4V performed better on follow-up questions, suggesting an enhanced ability to handle dynamic, evolving clinical scenarios in later versions.

The modest enhancements in visual data interpretation from ChatGPT-4 to ChatGPT-4V indicate that improvements in later versions focused more on specific refinements rather than on broad upgrades to support image processing. This trend is evident in image-based questions, where version 4V did not outperform version 4 in questions including images. For example, while ChatGPT showed improved accuracy in interpreting electrocardiograms, its performance was less consistent with dermatological images. A striking instance was its misdiagnosis of a *Staphylococcus aureus* skin infection in a toddler, where ChatGPT incorrectly identified the condition as Molluscum contagiosum, erroneously attributing significance to an area of the image that was, in fact, the patient's belly button. These variations underscore the intricate challenges AI faces in processing multimodal medical information and suggest that while ChatGPT's textual understanding has advanced, its image processing requires further contextual depth and fine-tuning.

ChatGPT's strong performance on medical licensing examinations from different parts of the world and in different languages demonstrates its adaptability and potential in medical education despite not being specifically designed for such specialized content [3,4,8-10]. However, its varying responses highlight the model's limitations in handling the depth and variability of real-life medical expertise.

This study is one of the first to evaluate ChatGPT-4, including its vision-enhanced iteration, in medical licensing examinations, notably being the first to evaluate its performance in Chile's EUNACOM. The multiple attempts per practice drill approach in our methodology is a significant strength of our study, facilitating a thorough examination of ChatGPT's response consistency. Despite these strengths, the study has some limitations. The reliance on practice drills from the University of Chile may not encompass the full breadth of the EUNACOM's questions, potentially narrowing the scope of our findings. The focus on specific versions of ChatGPT could also limit the generalizability of our results to other iterations of the model. Inherent biases in the AI's training data pose another challenge, potentially affecting the accuracy of responses.

Future studies should expand AI evaluations in medical training by including diverse medical examinations and question types, assessing adaptability to various contexts. Exploring newer AI models and their performance in practical medical scenarios will also be crucial. This research will enhance the understanding of AI's role in medicine, guiding its effective integration into health care education and practice.

The rise of generative AI in medicine, highlighted by tools such as ChatGPT and upcoming models such as Med-PaLM [11], signals a need to evolve medical education. While these tools provide extensive resources, the essence of medical practice extends beyond simple access to data, necessitating reflective and critical application of this knowledge. Therefore, medical curricula must prioritize critical thinking, enabling future practitioners to discern the quality and relevance of AI-generated information. Similarly, reflective practices are crucial, promoting continuous self-assessment and adaptation in a rapidly advancing technological landscape. As AI becomes integral, especially in diagnostics, professionals must merge AI insights with human-centric care, underscoring that medical expertise is not only about accessing information but also involves deep understanding and evaluation of that information, empathy, and ethical judgment.

Conclusions

In conclusion, this study shows the performance of ChatGPT versions 3.5, 4, and 4V in successfully passing the EUNACOM, underscoring the evolving role of AI in the field of medicine and its potential in medical education. Future studies should encompass a wider array of AI models and diverse question types, contributing to a deeper understanding of how AI can enhance medical education. Moreover, it is imperative to explore innovative directions in the application of AI, such as leveraging AI to augment human cognition and optimize the learning process. Embracing these possibilities can lead to a more profound impact on medical education, fostering not only knowledge acquisition but also critical thinking and problem-solving skills among future health care practitioners.

Acknowledgments

We thank the School of Medicine, University of Chile, for providing the EUNACOM (Examen Único Nacional de Conocimientos de Medicina) drills essential for this research. We would like to express our gratitude to Mridul Joshi for his invaluable assistance with the statistical analysis.

Disclaimer

This manuscript was prepared without the assistance of ChatGPT or similar artificial intelligence tools for writing, editing, or proofreading.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Average accuracy rate per medical area and clinical case question type.
[DOCX File, 21 KB - mededu_v10i1e55048_app1.docx]

References

1. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 1;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
2. GPT-4V(ision) system card. OpenAI. 2023. URL: https://openai.com/research/gpt-4v-system-card [accessed 2024-04-19]
3. Carrasco JP, García E, Sánchez DA, et al. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España [Article in Spanish]. Rev Esp Edu Med 2023;4(1). [doi: 10.6018/edumed.556511]
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
5. Fechas EUNACOM-ST de julio y cierre de inscripciones. EUNACOM. URL: https://www.eunacom.cl/home.html [accessed 2024-04-19]
6. Reglamento que establece los criterios generales y disposiciones sobre exigencia, aplicación, evaluación y puntuación mínima para el diseño y aplicación del examen único nacional de conocimientos de medicina [Article in Spanish]. MINSAL Chile. URL: https://www.eunacom.cl/reglamentacion/ReglamentoLey20261.pdf [accessed 2024-04-19]
7. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions, 3rd edition: Wiley; 2013.
8. Aljindan FK, Al Qurashi AA, Albalawi IAS, et al. ChatGPT conquers the Saudi Medical Licensing Exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus 2023 Sep;15(9):e45043. [doi: 10.7759/cureus.45043] [Medline: 37829968]
9. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. J Fr Ophtalmol 2023 Sep;46(7):706-711. [doi: 10.1016/j.jfo.2023.05.006] [Medline: 37537126]
10. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 4;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
11. Med-PaLM. Google Research. URL: https://sites.research.google/med-palm/ [accessed 2024-04-19]

Abbreviations

4V: ChatGPT-4 With Vision
AI: artificial intelligence
EUNACOM: Examen Único Nacional de Conocimientos de Medicina

Edited by G Eysenbach, TDA Cardoso; submitted 30.11.23; peer-reviewed by I Albalawi, S Thirunavukkarasu, U Hin Lai; revised version received 06.02.24; accepted 22.03.24; published 29.04.24.

Please cite as:
Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S
Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study
JMIR Med Educ 2024;10:e55048
URL: https://mededu.jmir.org/2024/1/e55048
doi:10.2196/55048

© Marcos Rojas, Marcelo Rojas, Valentina Burgess, Javier Toro-Pérez, Shima Salehi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.4.2024. 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

XSL•FO
RenderX

information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Integration of ChatGPT Into a Course for Medical Students: Explorative Study on Teaching Scenarios, Students' Perception, and Applications

Anita V Thomae[1], PhD; Claudia M Witt[1,2], Prof Dr; Jürgen Barth[1], PhD

1

2

Corresponding Author:
Jürgen Barth, PhD

Abstract

Background: Text-generating artificial intelligence (AI) such as ChatGPT offers many opportunities and challenges in medical education. Acquiring practical skills necessary for using AI in a clinical context is crucial, especially for medical education.

Objective: This explorative study aimed to investigate the feasibility of integrating ChatGPT into teaching units and to evaluate the course and the importance of AI-related competencies for medical students. Since a possible application of ChatGPT in the medical field could be the generation of information for patients, we further investigated how such information is perceived by students in terms of persuasiveness and quality.

Methods: ChatGPT was integrated into 3 different teaching units of a blended learning course for medical students. Using a mixed methods approach, quantitative and qualitative data were collected. As baseline data, we assessed students' characteristics, including their openness to digital innovation. The students evaluated the integration of ChatGPT into the course and shared their thoughts regarding the future of text-generating AI in medical education. The course was evaluated based on the Kirkpatrick Model, with satisfaction, learning progress, and applicable knowledge considered as key assessment levels. In ChatGPT-integrating teaching units, students evaluated videos featuring information for patients regarding their persuasiveness on treatment expectations in a self-experience experiment and critically reviewed information for patients written using ChatGPT 3.5 based on different prompts.

Results: A total of 52 medical students participated in the study. The comprehensive evaluation of the course revealed elevated levels of satisfaction, learning progress, and applicability specifically in relation to the ChatGPT-integrating teaching units. Furthermore, all evaluation levels demonstrated an association with each other. Higher openness to digital innovation was associated with higher satisfaction and, to a lesser extent, with higher applicability. AI-related competencies in other courses of the medical curriculum were perceived as highly important by medical students. Qualitative analysis highlighted potential use cases of ChatGPT in teaching and learning. In ChatGPT-integrating teaching units, students rated information for patients generated using a basic ChatGPT prompt as "moderate" in terms of comprehensibility, patient safety, and the correct application of communication rules taught during the course. The students' ratings were considerably improved using an extended prompt. The same text, however, showed the smallest increase in treatment expectations when compared with information provided by humans (patient, clinician, and expert) via videos.

Conclusions: This study offers valuable insights into integrating the development of AI competencies into a blended learning course. Integration of ChatGPT enhanced learning experiences for medical students.

(JMIR Med Educ 2024;10:e50545) doi:10.2196/50545

KEYWORDS

medical education; ChatGPT; artificial intelligence; information for patients; critical appraisal; evaluation; blended learning; AI; digital skills; teaching

Introduction

Since its public launch in November 2022, ChatGPT (OpenAI), as a text-generating artificial intelligence (AI), has garnered significant attention in academic education overall and particularly in the field of medical education. Besides endeavors such as exams in the field of medicine [1,2], there are many other opportunities to implement ChatGPT in medical education [3,4]. However, these opportunities also have certain challenges such as overreliance, plagiarism, and privacy concerns [5]. Previous research has suggested the need for the advancement of knowledge, interpretation, and application of AI in the context of medical education [6], thereby underscoring the importance of acquiring practical skills essential for using AI in one's future

XSL•FO
RenderX

professional career. The lack of integrating AI into medical education has been described [7,8]. Up to now, there are, however, few specific proposals on how to implement text-generating AI into existing courses. Recent studies exemplify the integration of ChatGPT primarily as a tool for training communication skills among medical students [9,10] or as a supporting tool in problem-based learning scenarios [11].

In our elective course, "Placebo and Nocebo," which was offered to medical students at the University of Zurich, we integrated content generated with ChatGPT into various teaching units by using different learning scenarios. The overall aim of this course was to teach medical students concepts related to the topic of placebo and nocebo.

Within the course, the importance of expectations regarding medical treatment, as raised by specific information and the corresponding impacts on treatment outcomes, were presented. Methods for phrasing information for patients concerning medical treatment, including benefits and potential side effects, was another key topic of this course. One possible application of ChatGPT in the clinical context could be to support the writing of information for patients in order to educate or prepare patients for upcoming treatments [12,13]. Such information must be clear and safe. Clarity includes readability and understandability of the presented information [14]. With regard to safety, information should present concerning potential side effects in a correct but layperson-friendly way, and the positive framing of side effects is encouraged [15,16].

In this explorative study, we wanted to investigate how medical students evaluate the integration of ChatGPT teaching units into the course and their perceived importance of text-generating AI–related competencies during their studies. Furthermore, we wanted to explore how personal characteristics such as sex and openness to digital innovation are related to these outcomes. By using information for patients written using ChatGPT, we wanted to further explore how medical students in this course assess the use of ChatGPT-created content as a source of information for patients and its respective persuasiveness.

Methods

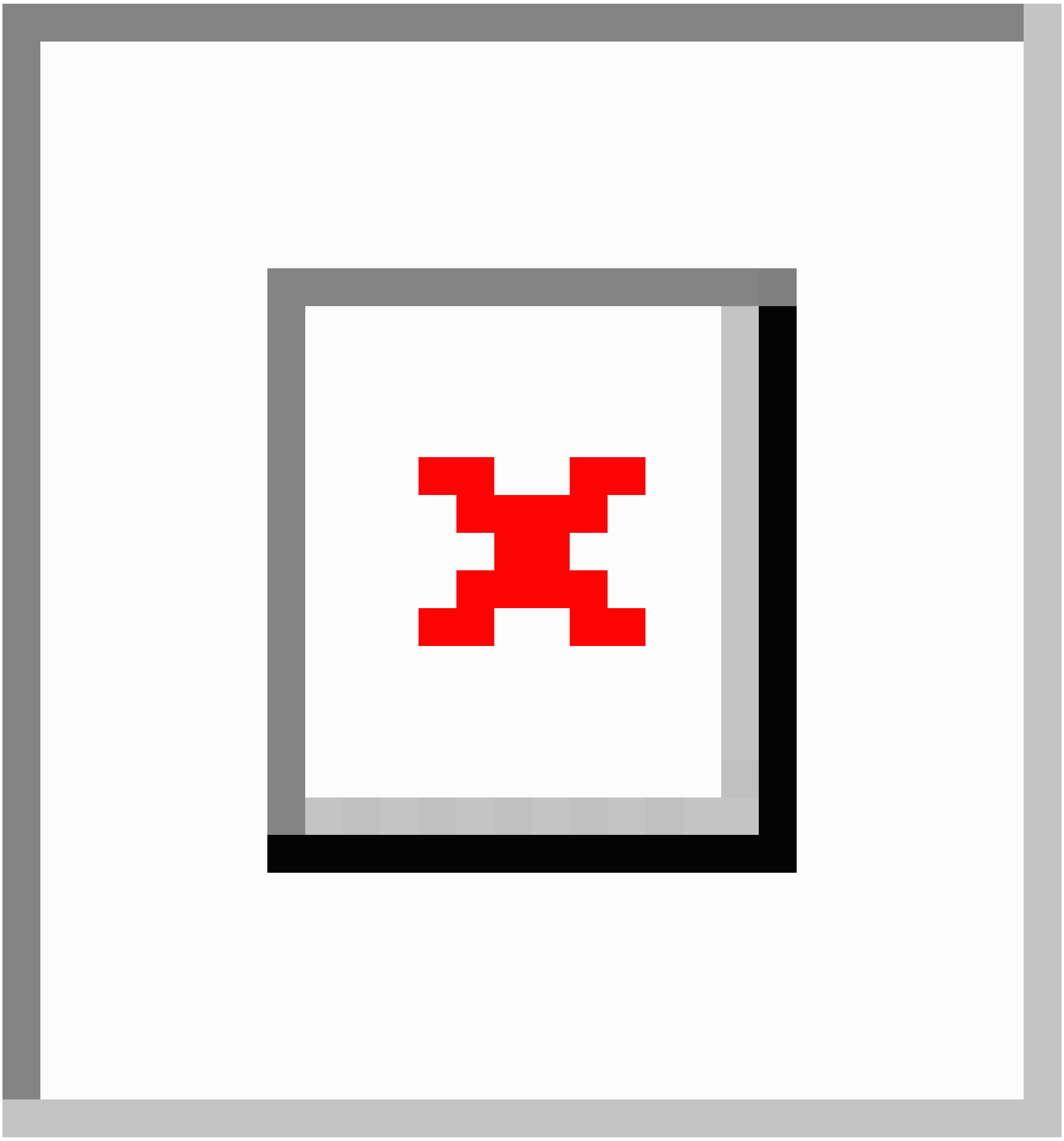
Procedure

Medical students (third bachelor and first master) were invited to enrol themselves in the elective course "Placebo and Nocebo" at the University of Zurich in spring 2023. The course comprised 28 teaching units (45 min each), 3 of which integrated ChatGPT. The course was set up as a blended learning course combining 13 teaching units delivered as e-learning and 15 teaching units delivered as in-person lectures. The course description was available to students before enrollment and indicated that the course would include a scientific evaluation and that the results would be published. All interaction with ChatGPT was performed on a personalized teacher's account (AT) rather than by students due to data privacy issues.

The teaching units of the course and the corresponding insights gained into ChatGPT application as well as the data gathered are summarized in Figure 1.

XSL•FO
RenderX

Figure 1. Integration of ChatGPT into different teaching units during the "Placebo and Nocebo" course taught at the University of Zurich.



Baseline and additional data, along with the overall course evaluation data, were collected via web both before and after the course.

In the teaching unit "Placebo control condition," the students were divided into working groups. They developed a placebo control condition for a clinical trial using a checklist in a problem-based learning scenario. Subsequently, the students presented their results in the plenary. These results were compared with suggestions from ChatGPT and discussed in the plenary, emphasizing a critical appraisal of the ChatGPT-generated solution.

In the teaching unit "Expectations," the students participated in an experiment. Through web-based questionnaires, they were asked about their individual outcome expectations regarding an acupuncture treatment for headaches. This specific symptom and treatment were selected due to the likelihood that students may have familiarity with both. We created 4 different videos that were 60 to 80 seconds in length, which were presented sequentially over 4 weeks (1 video per week). The videos were animated voice-overs. Each video delivered supportive information regarding the treatment of acute headaches using acupuncture, but different protagonists and different information were used (expert opinion, clinician opinion, patient experience, and general information from the internet). For example, the expert opinion focused on the evidence, whereas the patient reported on her own experience. The video transcripts in English are presented in Multimedia Appendix 1. Directly before and

XSL•FO
RenderX

after each video, the medical students answered 5 questions regarding their own expectations concerning acupuncture within the e-learning of the course. This pre-post-measurement provided an opportunity to assess the extent to which each video impacted the expectations of the students. The persuasiveness of text written using ChatGPT was compared with text provided by human beings. The information written using ChatGPT 3.5 was labeled as "general information from the internet." The same text was used in the subsequent teaching unit "Information for patients" using the extended prompt from that unit.

Before the teaching unit "Information for patients," students learned and practiced the rules for developing content intended for patients, along with the criteria used to review such information [15,16]. During the lecture, students were divided into 9 groups containing 4 to 6 students each. They applied these criteria within their groups to review 2 different pieces of information for patients written with ChatGPT 3.5 using digital whiteboards. Aiming to provide insight into the importance of prompt quality, the first information for patients was written with a rather basic prompt ("Can you please write a patient information about the effectiveness of acupuncture for tension headaches. The text should be no longer than 180 words."), whereas the second extended prompt included the review criteria ("Can you please change this text so that it is generally understandable for patients? Please formulate statements about effects and side effects in the sense of positive framing. Also make sure that patient safety is guaranteed. The text should be no longer than 180 words"). The students also marked content in each of the descriptions which they found "problematic" within the information for patients (eg, terminology). The review decisions and the rationale behind each group's choices were discussed in the plenary. The prompts and corresponding answers written with ChatGPT 3.5 as well as the content labeled as "problematic" by the groups are shown in Multimedia Appendix 1.

Outcomes

Treatment Expectations

During the experiment in the teaching unit "Expectation," students' expectations concerning acupuncture treatment for headaches were assessed using the Expectation for Treatment Scale [17] before and after the students watched the information provided in the video. The Expectation for Treatment Scale contains 5 items assessing the individual expectations regarding the effectiveness of a specific treatment. Total scores could range from 5 to 20, with higher values indicating higher expectations concerning acupuncture treatment.

Quality of Information for Patients

During the teaching unit "patient information," in 9 subgroups, students reviewed the quality of information for patients written with ChatGPT in terms of three criteria: (1) comprehensibility (lay language), (2) patient safety, and (3) correct application of communication rules taught in the course (positive framing). The categories for these judgments were "Fully met," "Partly met," and "Not met." Furthermore, the medical students were instructed to mark unclear, ambiguous, and critical content in the information for patients.

Evaluation

The integration of ChatGPT into different teaching units of the course was evaluated based on the Kirkpatrick Model [18,19], encompassing 3 levels: satisfaction, learning progress, and application. Ratings were assigned on a scale ranging from 1 (no agreement) to 6 (full agreement). We built aggregated scores for satisfaction (Cronbach α=.90), learning progress (Cronbach α=.80), and application (Cronbach α=.72) with good to excellent internal consistency based on 4 items per level. The results regarding single items are reported in Multimedia Appendix 1.

Additional Data

Baseline information of the students (age, sex, and previous experience with text-generating AI) was collected anonymously via the web. Openness to digital innovation was assessed using 4 items that have a similar phrasing to that used in the NeoFFI (NEO Five Factors Inventory) for assessing openness to experience [20]. Each item was assessed on a scale ranging from 1 (no agreement) to 6 (full agreement). The aggregated score ranged from 1 to 6, and the four items had excellent internal validity (Cronbach α=.95). Furthermore, medical students were asked about the potential of AI in medical education and medicine in general. The students evaluated the importance of 5 AI competencies in medical education on a scale ranging from 1 (no agreement) to 6 (full agreement). Competencies were selected based on the proposal of Caliskan et al [21]. Additionally, students shared their thoughts concerning the potential use of text-generating AI in medical education in response to an open question.

Data Analysis

Baseline data are summarized as median and IQR. Data from the teaching unit "Expectations," were analyzed using paired *t* tests to explore pre-post differences in expectations. The magnitude of the effect is expressed as mean difference with the respective 95% CI and as effect size. Data drawn from the teaching unit "Information for patients," are described as counts on group level. Data drawn from the overall evaluation are presented as median and IQR for the 3 evaluation levels due to the skewed data distribution. Spearman correlations between openness and the 3 evaluation levels were calculated for the total group and stratified by sex. The quantitative results concerning competencies are reported in a descriptive way. Quantitative analyses were conducted using IBM SPSS Statistics (Version 29). All analyses must be considered as exploratory.

We conducted a thematic analysis of the qualitative data using MAXQDA Software (release 20.4.2; Verbi) [22]. A team of 6 researchers created themes based on the answers provided by the students in 3 subgroups and coded the answers accordingly. The themes and codings associated with the subgroups were then harmonized through discussion, rearrangement, and intersubjective validation within the whole group.

Ethical Considerations

We submitted the study synopsis to the Ethics Committee of Zurich, Switzerland, and, after review, they stated that the study did not fall under the regulations of the Human Research Act of Switzerland (BASEC-Nr. Req-2023 - 00400).

XSL•FO
RenderX

Results

Study Participants

In total, 52 medical students (19 male and 33 female) participated in the "Placebo and Nocebo" course. The median age of the participants was 23 (IQR 22 - 25) years. Of the 40 students who completed the overall evaluation, 43% (n=17) of participants reported having never or rarely used text-generating AI before. A third (14/40, 35%) of participants had used text-generating AI occasionally before, and 23% (n=9) participants reported frequent or very frequent previous use of text-generating AI.

Overall Evaluation

Overall, the integration of ChatGPT into different teaching units of the course was evaluated very positively with high satisfaction scores (median 5.12, IQR 4.31 - 5.75), high perceived learning progress (median 4.37, IQR 3.75 - 5.25) due to the course and high applicability of the knowledge (median 4.75, IQR 4.25 - 5.25). The results regarding single items are reported in Multimedia Appendix 1.

The 3 levels of satisfaction, progress, and applicability were correlated by approximately 0.50 to 0.60, thus indicating rather strong associations among all 3 learning levels (Table 1). More interestingly, the associations between the overall evaluation outcomes and participants' sex and their openness to digital innovation are presented in the first 2 lines of Table 1. Sex was not associated with the evaluation outcomes of the course, but openness was strongly associated with satisfaction and to a lesser extent with applicability. No association between openness and progress was found.

Since the male sex was associated with higher openness, the associations for females and males were analyzed separately. There were no major differences between sex strata in the association of openness with satisfaction (female 0.500; male 0.652). However, the association of openness with progress (female: 0.150; male: 0.419) and applicability (female: 0.317; male: 0.783) differed between sexes. Male students with high openness also indicated higher learning progress and a higher applicability of the course content than female students.

Table . Associations of openness, sex, and the evaluation of the course (satisfaction, progress, and application).

		Sex	Openness	Satisfaction	Progress	Application
Sex[a]						
	Spearman correlation	1	0.462	0.037	−0.121	0.073
	P value	—[b]	.003	.82	.46	.65
Openness						
	Spearman correlation	0.462	1	0.483	0.119	0.399
	P value	.003	—	.002	.47	.01
Satisfaction						
	Spearman correlation	0.037	0.483	1	0.585	0.584
	P value	82	.002	—	<.001	<.001
Progress						
	Spearman correlation	−0.121	0.119	0.585	1	0.622
	P value	.46	.47	<.001	—	<.001
Application						
	Spearman correlation	0.073	0.399	0.584	0.622	1
	P value	.65	.01	<.001	<.001	—

[a]Female sex was coded as 1 and male sex as 2.

[b]Not applicable.

Potential of AI in Medical Education and Medicine in General

As shown in Table 2, the perceived importance of AI-related competencies in other courses of the medical curriculum is high for the students. Among the suggested competencies, the assessment of the opportunities and limitations of text-generating AI received the highest rating, while competencies for basic understanding of how AI functions received the lowest rating but are still regarded as important.

Qualitative analysis of the open-ended question revealed areas in teaching and learning for which students see potential uses

XSL•FO
RenderX

of ChatGPT. The themes identified and example quotations are shown in Table 3. Furthermore, students identified other potential uses of ChatGPT, namely, supporting clinical practice (eg, the use of ChatGPT in the context of documentation/patient reports, administrative work and support as a second opinion) and serving as a general information source. In addition, students made several statements regarding their opinions and values, such as the perceived lack of empathy in ChatGPT and the necessity for human supervision.

Table . Students' perceived importance[a] of artificial intelligence (AI)–related competencies in other courses.

AI-related competencies	Median (IQR)
Competencies for assessing the opportunities and limitations of text-generating AI	5.0 (5.00 - 6.00)
Competencies for combining text-generating AI with professional knowledge	5.0 (4.00 - 5.75)
Competencies for assessing the value of text-generating AI in teaching, care and research	5.0 (4.00 - 5.75)
Competencies for the efficient and effective use of text-generating AI in patient care	5.0 (4.00 - 6.00)
Competencies pertaining to a basic understanding of how text-generating AI functions	4.0 (4.00 - 5.00)

[a]Each item was evaluated on a scale ranging from 1 (no agreement) to 6 (full agreement).

Table . Students' ideas concerning the potential use of ChatGPT to support teaching and learning: Themes and quotations.

Themes	Quotations[a] (representative examples)
General support for teaching and learning	• *To generate good summaries of the learning material* (student 15) • *Instead of emailing lecturers to clarify ambiguities regarding lecture content, ask ChatGPT* (student 37) • *As exam preparation* (student 21)
ChatGPT as a form of writing support	• *Maybe ChatGPT could be used to optimize one's own texts* (student 13) • *As a form of support, workload relief and the acceleration of processes* (student 38)
ChatGPT for patient case simulation	• *ChatGPT could be used to simulate patients and thus practice what has been learned* (student 13)
Learning how to use ChatGPT	• *We could learn in the context of these courses how we could use ChatGPT optimally for learning, for research* (student 13) • *Assess the possibilities and limitations of such tools in a medical context* (student 22)

[a]These quotations were originally in German and were translated into English by the authors.

Results Regarding Specific Teaching Units

Change in Treatment Expectations by Different Information

Expectations regarding treatment showed an increase after each of the 4 information videos was presented (Table 4). The strongest increase was observed for the video that shared patient experiences, whereas the video containing the ChatGPT content did not change expectations substantially. Information presented by a clinician or the expert opinion changed expectations moderately.

Table . Changes in treatment expectations by information presented as 4 videos (expert, clinician, patient, and ChatGPT).

	Expectation, mean difference (95% CI)	*t* test (*df*)	*P* value	Effect size
Expert (n=50)	0.649 (0.102 - 1.178)	2.391 (49)	.021	0.338
Clinician (n=51)	0.941 (0.324 - 1.558)	3.063 (50)	.004	0.429
Patient (n=51)	1.431 (0.746 - 2.116)	4.198 (50)	<.001	0.588
ChatGPT (n=49)	0.449 (0.002 - 0.896)	2.021 (48)	.049	0.289

XSL•FO
RenderX

Quality of Information for Patients

The students reviewed the information for patients written using ChatGPT using a basic prompt and judged the text most often as partially comprehensive, safe, and appropriate in terms of communication rules (Table 5). The students identified many terms and phrases that they deemed problematic for use in information for patients. Reasons mentioned by students in group discussion included, for instance, the use of too specific terminology with low readability and poor understandability (Multimedia Appendix 1). The information for patients written using ChatGPT using an extended prompt was reviewed very positively. Only a minority of the student groups indicated after the review that the criteria of comprehensibility, safety, and communication rules were only partly met or not met. The number of problematic terms identified by the students was much lower than the number of such terms in the first text.

Table . Students' judgments (group decision counts) regarding the information for patients generated using ChatGPT using a basic or an extended prompt.

Review criteria	Basic prompt[a]			Extended prompt[b]		
	Fully met[c]	Partly met[c]	Not met[c]	Fully met[c]	Partly met[c]	Not met[c]
Comprehensibility	0	5	3	8	0	1
Safety	1	6	2	8	1	0
Communication rules	1	5	2	7	1	1

[a]Basic prompt: "Can you please write a patient information about the effectiveness of acupuncture for tension headaches? The text should be no longer than 180 words."

[b]Extended prompt: "Can you please change this text so that it is generally understandable for patients? Please formulate statements about effects and side effects in the sense of positive framing. Also make sure that patient safety is guaranteed. The text should be no longer than 180 words."

[c]Fully met: Rules are applied throughout the whole text; partly met: rules are sometimes applied, but not consistently throughout the whole text; not met: rules were not applied within the text.

Discussion

Principal Findings

Our study showed that integrating ChatGPT into medical courses is feasible, although the majority of the students had no or only limited experience using ChatGPT. The ChatGPT-enriched teaching units were highly appreciated by medical students, and this approach can be used as a stimulating teaching tool. Text generated using ChatGPT in a persuasion experiment (ie, information for patients to change treatment outcome expectations), a practical review exercise focusing on information for patients, and a problem-based learning scenario were suitable formats for our teaching units. All 3 formats used in this course were closely related to possible scenarios that may be relevant for medical students in their later professional careers. Medical students consider the acquisition of competencies related to text-generating AI to be highly important during their studies.

To support constructive learning for the students, the ChatGPT-enriched scenarios addressed different teaching strategies, namely, problem-solving, self-experience, and evaluation. The respective teaching units were embedded into the framework of the course rather than merely added, as also suggested by McCoy et al [23]. This goal was achieved by using ChatGPT directly to revisit or deepen teaching content drawn from other teaching units of the course. In general, the integration of ChatGPT in education demonstrated superiority in both evaluation results and knowledge outcomes. The findings from a medical communication course incorporating ChatGPT revealed positive student evaluations [24]. Direct comparisons of dental students' knowledge after a ChatGPT-integrated teaching scenario with an AI-free scenario showed better learning progress [11]. Our course evaluation was based on the Kirkpatrick Model considering satisfaction, learning progress, and applicable knowledge as key assessment levels [18,19,25]. Based on the subjective assessment of the students, the evaluation results show that our approach can facilitate students' understanding of the course content and allow them to explore the possibilities and limitations of text-generating AI.

Although nearly half of the students had no or only limited experience using text-generating AI, the evaluation was very positive. Hence, students' previous experience with or interest in innovative technologies does not seem to be a necessary prerequisite for the introduction of such technologies into medical teaching. This confirms the findings of Weidener and Fischer from a survey of medical students across Germany, Austria, and Switzerland. In this study, less than half of the students had prior experience with ChatGPT or other AI-based chat applications but indicated a need for AI in medical education [7].

For the successful integration of AI into teaching modules, facilitating and impeding factors among students should be investigated. Openness to digital innovations might be an asset to facilitate learning with AI tools. It has been demonstrated in adults in the United States that users' trust impacts both the intention to use and the actual use of ChatGPT [26]. Here, we showed that students with lower openness to digital innovation reported less satisfaction and lower applicability in our evaluations, which may be a result of their lower motivation to engage in teaching units with ChatGPT content. Consequently, less open students may also lack knowledge regarding the limitations of ChatGPT since they may avoid this technology in general. Sex is another factor that could lead to different receptions of AI-enriched courses. Higher use of AI tools in

XSL•FO
RenderX

male persons has also been shown in students from different fields in Germany [27], possibly reflecting a higher openness to digital innovations. In our study, we also found similar associations of openness to digital innovations, progress, and applicability for male students. However, for female students, other factors beyond openness might affect progress and applicability.

Our evaluation results showed that the perceived importance of AI-related competencies for students is rated very high in general and covers a wide range of different competencies. The chosen competencies are similar to the categories of knowledge, interpretation, and application of AI that were revealed by teaching experts [6]. Some of these competencies were addressed in our teaching units. For example, the teaching unit "Information for patients," illustrated the need to use a meaningful prompt and the importance of choosing relevant criteria when using ChatGPT to create information for patients. During the course, the quality of this information written using ChatGPT has been improved by incorporating important criteria for the text into the prompt. Students found information written using the extended prompt to be of higher quality, providing insights into the importance of prompt quality for the generated text. Several studies have investigated the application of ChatGPT and other large language models as tools for providing patient material, yielding promising results. According to Ayers et al [12], AI-generated text messages on health-related patient questions in a social media forum were superior to physician responses as rated by health care professionals. Tangadulrat et al [28] showed that both medical students and graduated doctors positively perceived using ChatGPT for creating patient educational materials. Patient material readability scores were considerably improved by the large language model as demonstrated by Rouhi et al [29]. Interestingly, the ChatGPT generated text using the extended prompt, was found to be the least persuasive within the expectation experiment. It did not change students' expectations regarding a specific treatment substantially compared with information provided by humans (especially when compared with a patient statement). As the information written using ChatGPT was read by an artificial voice, while the other information was read by humans, the lower persuasiveness might be due to a lower acceptance of the artificial voice. A preference for human voices has been shown in other research [30].

Open-ended questions revealed misleading concepts, such as the use of text-generating AI to support patient documentation, a potential concern due to data protection issues as discussed by Eggmann et al [31]. Particularly, to circumvent problematic applications of text-generating AI in physicians' later professional careers, the systematic integration of AI-related competencies into medical curricula is critical.

Limitations

Our results lack generalizability with respect to the use of ChatGPT in other learning environments (eg, larger groups). Furthermore, the results cannot be generalized to the use of other generative AI (such as image-generating AI).

The evaluation items were based on the Kirkpatrick Model. However, all items were self-reported. Ideally, learning progress and application would be assessed with objective indicators, eg, based on progress tests and performance evaluations. Learning effects, especially at the level of application, would be larger if students used ChatGPT on their own, entering their own prompts rather than using answers written using ChatGPT based on teachers' prompts. However, this scenario would cause problems with students' data privacy and would be a course on its own.

Given the predefined structure and learning objectives of the course, it was unfortunately not possible to explore further the use of AI in generating information for patients and its respective change in expectation. Additionally, it would have been of advantage to reflect these questions not only with medical students but with patients as actual target groups of such information.

Conclusions

According to the evaluation of medical students, integration of ChatGPT into an existing course is highly appreciated and enhances the learning experience. The development of AI-related competencies, including the phrasing of meaningful prompts during medical education, was perceived as very important by these medical students. The ability to critically appraise AI-generated information is also an important competency for medical students.

Acknowledgments

We would like to acknowledge Yuqian Yan, Institute for Complementary and Integrative Medicine, University Hospital Zurich, Switzerland, for her help in the data analysis of the expected results. We would like to acknowledge Manuela Oehler, Institute for Complementary and Integrative Medicine, University Hospital Zurich, Switzerland, for her work in setting up the expectation experiment. We would like to express our special thanks to Claudia Canella for her support during the qualitative analyses. Finally, we are grateful to Casey Murphy for conducting literature update searches and editing the manuscript.

Conflicts of Interest

CMW has active research grants to the University for digital health projects from the German Health Care Innovation Fund, Newsense Lab GmbH, Krebsforschung Schweiz, and DIZH (Digitalization Initiative of the Zurich Higher Education Institutions). JB has an active grant from Krebsforschung Schweiz for a digital health project and has received honoraria for workshops on digital health.

XSL•FO
RenderX

AVT has received honoraria from the Universitätsklinikum Heidelberg for conduction of e-learning workshops by the Universitätsklinikum Heidelberg. She is a member of the Advisory Board of the Zentrum für Kompetenzentwicklung Krebs-Selbsthilfe, University of Freiburg.

Multimedia Appendix 1

Detailed evaluation results and insight into teaching scenarios.

[PDF File, 1317 KB - mededu_v10i1e50545_app1.pdf]

References

1. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
3. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
4. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? on opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr;103:102274. [doi: 10.1016/j.lindif.2023.102274]
5. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
6. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. JMIR Med Educ 2023 Apr 24;9:e46428. [doi: 10.2196/46428] [Medline: 36946094]
7. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. JMIR Med Educ 2024 Jan 5;10:e51247. [doi: 10.2196/51247] [Medline: 38180787]
8. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
9. Gray M, Baird A, Sawyer T, et al. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of ChatGPT: exploratory observational study. JMIR Med Educ 2024 Feb 1;10:e50705. [doi: 10.2196/50705] [Medline: 38300696]
10. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024 Jan 16;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
11. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2024 Jan 31;10:e51344. [doi: 10.2196/51344] [Medline: 38111256]
12. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
13. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr 2023 Mar 1;7(2):pkad010. [doi: 10.1093/jncics/pkad010] [Medline: 36808255]
14. Jairoun AA, Al-Hemyari SS, Jairoun M, El-Dahiyat F. Readability, accuracy and comprehensibility of patient information leaflets: the missing pieces to the puzzle of problem-solving related to safety, efficacy and quality of medication use. Res Social Adm Pharm 2022 Apr;18(4):2557-2558. [doi: 10.1016/j.sapharm.2021.10.005] [Medline: 34711520]
15. Barnes K, Faasse K, Geers AL, et al. Can positive framing reduce nocebo side effects? current evidence and recommendation for future research. Front Pharmacol 2019;10:167. [doi: 10.3389/fphar.2019.00167] [Medline: 30894815]
16. Wilhelm M, Rief W, Doering BK. Decreasing the burden of side effects through positive message framing: an experimental proof-of-concept study. Int J Behav Med 2018 Aug;25(4):381-389. [doi: 10.1007/s12529-018-9726-z] [Medline: 29785686]
17. Barth J, Kern A, Lüthi S, Witt CM. Assessment of patients' expectations: development and validation of the expectation for treatment scale (ETS). BMJ Open 2019 Jun 17;9(6):e026712. [doi: 10.1136/bmjopen-2018-026712] [Medline: 31213446]
18. Kirkpatrick DL, Kirpatrick JD. Evaluation Training Programs: The Four Levels, 3rd edition: Berret-Koehler Publishers; 2006. URL: https://www.scirp.org/reference/referencespapers?referenceid=2702697 [accessed 2024-08-14]
19. Smidt A, Balandin S, Sigafoos J, Reed VA. The Kirkpatrick model: a useful tool for evaluating training outcomes. J Intellect Dev Disabil 2009 Sep;34(3):266-274. [doi: 10.1080/13668250903093125] [Medline: 19681007]
20. Costa PT, McCrae RR, Psychological Assessment Resources, Inc. NEO PI/FFI Manual Supplement for Use with the NEO Personality Inventory and the NEO Five-Factor Inventory: Psychological Assessment Resources; 1989. URL: https://sjdm.org/dmidi/NEO-FFI.html [accessed 2024-08-14]

XSL•FO
RenderX

21. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-delphi study for competencies. PLoS One 2022;17(7):e0271872. [doi: 10.1371/journal.pone.0271872] [Medline: 35862401]
22. Bourgeault I, Dingwall R, De Vries R. The SAGE Handbook of Qualitative Methods in Health Research: SAGE Publications Ltd; 2010.
23. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
24. Park J. Medical students' patterns of using ChatGPT as a feedback tool and perceptions of ChatGPT in a leadership and communication course in Korea: a cross-sectional study. J Educ Eval Health Prof 2023;20:29. [doi: 10.3352/jeehp.2023.20.29] [Medline: 38096895]
25. Frye AW, Hemmer PA. Program evaluation models and related theories: AMEE guide no. 67. Med Teach 2012;34(5):e288-e299. [doi: 10.3109/0142159X.2012.668637] [Medline: 22515309]
26. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023 Jun 14;25:e47184. [doi: 10.2196/47184] [Medline: 37314848]
27. von Garrel J, Mayer J. Artificial intelligence in studies—use of ChatGPT and AI-based tools among students in Germany. Humanit Soc Sci Commun 2023;10(1):799. [doi: 10.1057/s41599-023-02304-7]
28. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658. [doi: 10.2196/50658] [Medline: 38133908]
29. Rouhi AD, Ghanem YK, Yolchieva L, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? a pilot study. Cardiol Ther 2024 Mar;13(1):137-147. [doi: 10.1007/s40119-023-00347-0] [Medline: 38194058]
30. Kühne K, Fischer MH, Zhou Y. The human takes it all: humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. Front Neurorobot 2020;14:593732. [doi: 10.3389/fnbot.2020.593732] [Medline: 33390923]
31. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent 2023 Oct;35(7):1098-1102. [doi: 10.1111/jerd.13046] [Medline: 37017291]

Abbreviations

AI: artificial intelligence
NeoFFI: NEO Five Factors Inventory

Edited by K Venkatesh; submitted 24.07.23; peer-reviewed by A Arbabisarjou, C Klein, J Hastings, J Arango-Ibañez, L Tong, L Zhu, N Owens; revised version received 27.02.24; accepted 04.06.24; published 22.08.24.

Please cite as:
Thomae AV, Witt CM, Barth J
Integration of ChatGPT Into a Course for Medical Students: Explorative Study on Teaching Scenarios, Students' Perception, and Applications
JMIR Med Educ 2024;10:e50545
URL: https://mededu.jmir.org/2024/1/e50545
doi:10.2196/50545

© Anita V Thomae, Claudia M Witt, Jürgen Barth. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.8.2024. 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.

XSL•FO
RenderX

Viewpoint

Patients, Doctors, and Chatbots

Thomas C Erren[1], 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
Berlin-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.

Concerning the interaction between patients, doctors, and chatbots, I describe what lies ahead when using chatbots and identify many questions for the daily work of doctors. I conclude with a gloomy outlook, expectations for urgently needed ethical discourse [5,6], and a hope in relation to humans and machines [3,7].

Weighing ChatGPT's Quote

How ChatGPT describes its role [1]—"I am not a substitute for medical advice"—should be a fact. Doctors, as the only authoritative providers of professional MA, must always be in the driver's seat. Chatbots have the potential to help with the task of contributing general information to an information chain. Importantly, doctors need to review and question all AI output and see if and how it contributes to a patient's understanding and fits within MA. Depending on the expectations and hopes that ChatGPT raises in patients, this task could become an unprecedented challenge.

With their up-to-date knowledge and medical experience and expertise, doctors need to integrate personal, specific, and general information into their comprehensive MA to the patients. Chatbots are limited to general information stored in databases. Concerningly, ChatGPT invents facts, called a hallucination in

XSL•FO
RenderX

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.

XSL•FO
RenderX

References

1. 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]
2. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023 Apr 25;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
3. Shaw D, Morfeld P, Erren T. The (mis)use of ChatGPT in science and education: Turing, Djerassi, "athletics" & ethics. EMBO Rep 2023 Jul 05;24(7):e57501. [doi: 10.15252/embr.202357501] [Medline: 37259767]
4. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: ethical issues with using chatbots in mental health. Digit Health 2023;9:20552076231183542 [FREE Full text] [doi: 10.1177/20552076231183542] [Medline: 37377565]
5. Akerson M, Andazola M, Moore A, DeCamp M. More than just a pretty face? Nudging and bias in chatbots. Ann Intern Med 2023 Jul;176(7):997-998. [doi: 10.7326/M23-0877] [Medline: 37276595]
6. Erren TC, Lewis P, Shaw DM. Brave (in a) new world: an ethical perspective on chatbots for medical advice. Front Public Health 2023;11:1254334 [FREE Full text] [doi: 10.3389/fpubh.2023.1254334] [Medline: 37663854]
7. Turing AM. I.—Computing machinery and intelligence. Mind 1950 Oct;LIX(236):433-460. [doi: 10.1093/mind/LIX.236.433]
8. Frankfurt HG. On Bullshit. Princeton, NJ: Princeton University Press; 2005.
9. Metz C. ‘The Godfather of A.I.’ leaves google and warns of danger ahead. The New York Times. 2023 May 02. URL: https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html [accessed 2023-10-18]
10. Fung B. Mr. ChatGPT goes to Washington: OpenAI CEO Sam Altman testifies before Congress on AI risks. CNN. 2023 May 16. URL: https://edition.cnn.com/2023/05/16/tech/sam-altman-openai-congress/index.html [accessed 2023-10-18]
11. Hyman I. The risks of consulting Dr. Google. Psychology Today. 2020 Apr 29. URL: https://www.psychologytoday.com/us/blog/mental-mishaps/202004/the-risks-consulting-dr-google [accessed 2023-10-19]
12. Freckelton I. Internet disruptions in the doctor-patient relationship. Med Law Rev 2020 Aug 01;28(3):502-525. [doi: 10.1093/medlaw/fwaa008] [Medline: 32417891]
13. Zhu Y, Wang R, Pu C. "I am chatbot, your virtual mental health adviser." What drives citizens' satisfaction and continuance intention toward mental health chatbots during the COVID-19 pandemic? An empirical study in China. Digit Health 2022;8:20552076221090031 [FREE Full text] [doi: 10.1177/20552076221090031] [Medline: 35381977]
14. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "Black Box" medicine? Ann Intern Med 2020 Jan 07;172(1):59-60. [doi: 10.7326/M19-2548] [Medline: 31842204]
15. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019 Nov 12;322(18):1765-1766. [doi: 10.1001/jama.2019.15064] [Medline: 31584609]
16. GPT-4 technical report. OpenAI. 2023 Mar 27. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2023-10-18]
17. Currie G, Robbie S, Tually P. ChatGPT and patient information in nuclear medicine: GPT-3.5 versus GPT-4. J Nucl Med Technol 2023 Dec 05;51(4):307-313. [doi: 10.2967/jnmt.123.266151] [Medline: 37699647]
18. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. Vaccines (Basel) 2023 Jul 07;11(7):1217 [FREE Full text] [doi: 10.3390/vaccines11071217] [Medline: 37515033]
19. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023 Oct 01;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
20. Manke K. AI lectures at Berkeley to explore possibilities, implications of ChatGPT. Berkeley News. 2023 Mar 10. URL: https://news.berkeley.edu/2023/03/10/ai-lectures-at-berkeley-to-explore-possibilities-implications-of-chatgpt/ [accessed 2023-10-19]
21. The Lancet. AI in medicine: creating a safe and equitable future. Lancet 2023 Aug 12;402(10401):503. [doi: 10.1016/S0140-6736(23)01668-9] [Medline: 37573071]
22. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (Turing) test: survey study. JMIR Med Educ 2023 Jul 10;9:e46939 [FREE Full text] [doi: 10.2196/46939] [Medline: 37428540]

Abbreviations

AI: artificial intelligence
MA: medical advice

XSL•FO
RenderX

Edited by K Venkatesh; submitted 14.07.23; peer-reviewed by A Mihalache, N Patil, I Mircheva; comments to author 14.10.23; revised version received 19.10.23; accepted 08.11.23; published 04.01.24.

Please cite as:
Erren TC
Patients, Doctors, and Chatbots
JMIR Med Educ 2024;10:e50869
URL: https://mededu.jmir.org/2024/1/e50869
doi:10.2196/50869
PMID:38175695

©Thomas C Erren. Originally published in JMIR Medical Education (https://mededu.jmir.org), 04.01.2024. 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.

XSL•FO
RenderX

Viewpoint

Generative Language Models and Open Notes: Exploring the Promise and Limitations

Charlotte Blease[1,2], PhD; John Torous[2], MD, MBI; Brian McMillan[3], PhD; Maria Hägglund[1,4], PhD; Kenneth D Mandl[5], MD, MPH

[1]Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden
[2]Digital Psychiatry, Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, United States
[3]Centre for Primary Care and Health Services Research, University of Manchester, Manchester, United Kingdom
[4]Medtech Science & Innovation Centre, Uppsala University Hospital, Uppsala, Sweden
[5]Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States

Corresponding Author:
Charlotte Blease, PhD
Department of Women's and Children's Health
Uppsala University
Box 256
Uppsala, 751 05
Sweden
Phone: 46 18 471 00 0
Email: charlotteblease@gmail.com

Abstract

Patients' online record access (ORA) is growing worldwide. In some countries, including the United States and Sweden, access is advanced with patients obtaining rapid access to their full records on the web including laboratory and test results, lists of prescribed medications, vaccinations, and even the very narrative reports written by clinicians (the latter, commonly referred to as "open notes"). In the United States, patient's ORA is also available in a downloadable form for use with other apps. While survey studies have shown that some patients report many benefits from ORA, there remain challenges with implementation around writing clinical documentation that patients may now read. With ORA, the functionality of the record is evolving; it is no longer only an aide memoire for doctors but also a communication tool for patients. Studies suggest that clinicians are changing how they write documentation, inviting worries about accuracy and completeness. Other concerns include work burdens; while few objective studies have examined the impact of ORA on workload, some research suggests that clinicians are spending more time writing notes and answering queries related to patients' records. Aimed at addressing some of these concerns, clinician and patient education strategies have been proposed. In this viewpoint paper, we explore these approaches and suggest another longer-term strategy: the use of generative artificial intelligence (AI) to support clinicians in documenting narrative summaries that patients will find easier to understand. Applied to narrative clinical documentation, we suggest that such approaches may significantly help preserve the accuracy of notes, strengthen writing clarity and signals of empathy and patient-centered care, and serve as a buffer against documentation work burdens. However, we also consider the current risks associated with existing generative AI. We emphasize that for this innovation to play a key role in ORA, the cocreation of clinical notes will be imperative. We also caution that clinicians will need to be supported in how to work alongside generative AI to optimize its considerable potential.

(JMIR Med Educ 2024;10:e51183) doi:10.2196/51183

KEYWORDS

ChatGPT; generative language models; large language models; medical education; Open Notes; online record access; patient-centered care; empathy; language model; online record access; documentation; communication tool; clinical documentation

Introduction

Patient online record access (ORA) is growing globally [1]. Access includes test and laboratory results, secondary or hospital care letters, lists of prescribed medications, and the narrative reports written by clinicians after visits (the latter referred to as "open notes"). Already, patients across an estimated 30 countries can access some of their records via secure web portals including

XSL•FO
RenderX

health apps. In some countries, this innovation is advanced [1]. Since 2021, the federally enacted 21st Century Cures Act in the United States mandated that providers offer all patients access to download their electronic health records without charge [2]. In the Nordic countries, ORA has been implemented incrementally, starting around 2010 [3]. The Finnish patient portal OmaKanta was rolled out with stepwise implementation of functionality between 2010 and 2015 [4]. Patients in Sweden first obtained ORA in one of 21 regions in 2012 [5] with nationwide implementation achieved by 2018. Implementation in Norway began in 2015, reaching patients in 3 out of 4 regions by 2019 [6]. In England, from October 31, 2023, it is mandatory for general practitioners to offer ORA to their adult patients, albeit on a prospective basis [7].

Patients with access to their records report using them to become more involved in their care, to follow up on doctors' visits, and to obtain an overview of their test results and treatment history [3,8,9]. Multiple surveys show that patients using ORA are positive about the experience after reading their notes. They report many benefits including understanding their care plans better [9], improved communication with and greater trust in their provider [10], and feeling more in control of their health and care [6,8], including doing a better job taking their medications [11,12].

Despite the patient benefits with ORA, challenges with their implementation in clinical practice remain. In this viewpoint paper, we identify outstanding concerns with ORA, which encompass a range of unintended consequences for clinician work burdens, and for the substantial task of conveying bespoke, compassionate, and understandable information to each unique patient who accesses their records. Currently, it has been proposed that a range of targeted patient training and medical education strategies may be sufficient to resolve at least some of these challenges [13-17]. We believe that such interventions are valuable; however, in this viewpoint paper, we explain why the ambitions of such training interventions may be limited.

As a solution, we explain why the use of generative artificial intelligence (AI) may offer more tangible long-term promise than clinician training alone in helping to resolve problems with ORA implementation. While generative AI itself is not new, recent technical advances and the increased accessibility of large language models (LLMs; GPT-4 by OpenAI, LLaMA by Meta, and PaLM2 by Google) have made clinical use increasingly feasible. LLMs are an application of generative AI technology, often defined as machine learning algorithms that can recognize, summarize, and generate content based on training on large data sets. Unlike search engines, which offer pages of internet links in response to typed queries, generative LLMs such as GPT-4 simulate well-reasoned answers couched as conversations. In addition, these models can "remember" previous prompts, helping to build up the perception of dialogic exchange. We review the strengths and limitations of generative AI and emphasize that for this innovation to play a key role in ORA, it will be imperative for humans to be involved as overseers of computer input.

Current Challenges With Open Notes

Evolving Functionality of Records

Guidelines, such as those issued by the British General Medical Council, state that clinicians should keep clear, accurate, contemporaneous records that include "...any minor concerns, and the details of any action you have taken, information you have shared and decisions you have made relating to those concerns" [18]. In the era of ORA, clinicians will also need to consider if what they write will be understandable, accessible, and supportive for patients [19]. With the knowledge that patients will read what they write, the functionality of the record is evolving, and this incurs changes with respect to how clinical information is documented [20,21]. Clinicians must uphold the original functionality of the record—documenting the patient's medical information in clinical detail, but also communicating this information to the patient. With respect to the latter function, it is argued that for records to be understandable and acceptable to a lay audience, clinicians should ideally remove or explain medical acronyms, omit medical vernacular that may be perceived as offensive (such as "patient denies" or "patient complains of"), and strive to convey information in a manner that it is straightforward, comprehensive, and empathic in tone [14]. This is not an easy undertaking for clinicians tasked with pitching information at a literacy level that accommodates diverse patient populations while maintaining the clinical utility of records and adequately serving their medicolegal functions. Indeed, whether such dual functionality is even possible has been questioned [22].

Documentation Changes

To date, it is unclear whether ORA diminishes the clinical value of documentation [19,23]. However, there is evidence that clinicians may be undermining the accuracy or completeness (or both) of their records, perhaps in attempts to reduce patient anxieties, minimize follow-up contact, or reduce the likelihood of potential complaints [24,25]. For example, in the largest study conducted on clinicians' experiences of open notes, a 3-center study at 3 diverse health systems in the United States (1628 of 6054, 27% clinicians responded), DesRoches et al [26] found that around 1 in 4 physicians admitted that they changed how they wrote differential diagnoses (23%, n=176), though the nature of these changes is not understood. More worryingly, more than 1 in 5 physicians (22%, n=168) believed that their notes were now less valuable for other clinicians [26].

Conceivably, other changes following implementation of ORA might be more positive. In the study by DesRoches et al [26], 22% (n=166) of physicians reported changes to the use of a partnering language, and 18% (n=139) of them reported changes to how they used medical jargon or acronyms. However, it remains unknown whether such changes improve the comprehensibility of clinical records among patients or whether amendments come with a trade-off in terms of documentation quality.

With ORA, there is also the potential for notes to convey bias of stigmatizing language. For example, in the United States, recent linguistic analysis studies have shown that negative patient descriptors in notes are considerably more common for

XSL•FO
RenderX

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].

XSL•FO
RenderX

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 credentialed 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

XSL•FO
RenderX

information from being shared by "covered entities"—that is, providers—to other third parties. Therefore, the use of OpenAI, for example, is precluded under the HIPAA. At the time of writing, in the most common use cases, uploading patient details to versions of generative AI would breach patient trust and medical confidentiality due to privacy concerns.

However, the scope for this is quickly changing. Epic—the US software giant which has an estimated 78% of the share of hospital medical record use in the United States [61]—is currently piloting the integration of HIPAA-compliant GPT services [62]. In addition, an Azure HIPAA–compliant GPT-4 service already exists [63]. Voice-to-text clinical note generation products now represent a growing space in health care. For example, a new app called Ambient Experience from Nuance can listen to the physician's conversation and, using ChatGPT (version 4), help create the clinical note that is ready for physicians to review [64]. In the United States, such capacities are set to become embedded into electronic health systems, signaling revolutionary changes in medical documentation practices.

Clinicians and Computers as Coauthors

Combined, the aforementioned discourse suggests that LLMs are far from ready to disintermediate clinicians when it comes to writing clinical notes. We argue that the innovation will play a key role if humans are involved. Thus, this promise could be harnessed if clinicians oversee the cocreation of clinical documentation. In this scenario, LLMs might offer initial draft documentation, which, crucially, should be supervised, and edited by clinicians whose key role in documentation will be to keep a check and balance on the current limitations with these models.

Considering the scope of generative AI, we therefore propose that current training interventions might be constructively adapted to better prepare clinicians to oversee the writing of patient-facing clinical documentation, for example, by editing and checking the quality of clinical information constructed by generative AI and reviewing the sensitivity of the language used. Preliminary studies already show that when humans collaborate with LLMs to coproduce replies to patients, this can enhance patients' ratings of levels of empathy compared with human-only produced responses [65]. Such partnership could offer a more robust and safe form of documentation quality control—one that could potentially avoid the work burdens associated with documentation burdens and, therefore, the potential for burnout from ORA. We emphasize, however, that training should reinforce the importance of using generative AI as an assistant narrative scribe and not as a substitute for writing notes.

Furthermore, if health systems adopt this approach, we suggest that 2 (or even multiple) versions of clinical documentation may be feasible. Using LLMs, there is scope to not only a complete medical narrative pitched at the level of the domain expert or specialist, but also to document notes couched at the level of health literacy, language, and empathy of the individual patient who might be reading them. This could help overcome the current dilemma of documenting information in a way that is accessible for patients, but which does not diminish the clinical detail for health professionals.

Future Research Directions

Many research questions could usefully explore generative AI in cowriting clinical notes, especially dual-purpose documentation for both patients and clinicians. We suggest a few novel directions. First, qualitative studies could usefully explore how successfully generative AI translates clinical documentation into patient-friendly language. For example, studies could examine the accuracy and fidelity of generative AI in translating acronyms or other medical jargon, as well as the understandability of the notes, and the level of empathy embedded in patient-facing documentation. Second, experimental studies could probe whether documentation embeds biases or a higher likelihood of containing stigmatizing language for different patient demographics or health conditions. Third, pilot studies could help determine the satisfaction and administrative work burden of dual documentation among clinicians.

Conclusions

Generative AI is ready for mass use when it comes to writing or cowriting clinical notes, and its potential is enormous. We emphasize, however, that there remain evidence-based risks associated with existing generative AI, which relate to inconsistencies, errors, and hallucinations and the real potential to embed harmful biases in documentation. If carefully implemented, in the long term, doctors who write documentation using generative AI may do a better job of adapting to the evolving functionality of the electronic records than doctors who do not. This adoption may address the potential risk of "dumbing down" clinical documentation while conveying understandable and empathetic information to patients using plain and sensitive language. We also forecast that doctors who cowrite their documentation with LLMs will experience fewer work burdens.

Conflicts of Interest

JT is the Editor-in-Chief of *JMIR Mental Health*. The other authors declare no conflicts of interest.

References

1. Hägglund M, McMillan B, Whittaker R, Blease C. Patient empowerment through online access to health records. BMJ 2022 Sep 29;378:e071531 [FREE Full text] [doi: 10.1136/bmj-2022-071531] [Medline: 36175012]
2. Salmi L, Blease C, Hägglund M, Walker J, DesRoches CM. US policy requires immediate release of records to patients. BMJ 2021 Feb 18;372:n426. [doi: 10.1136/bmj.n426] [Medline: 33602667]

XSL•FO
RenderX

3. Blease C, Salmi L, Rexhepi H, Hägglund M, DesRoches C. Patients, clinicians and open notes: information blocking as a case of epistemic injustice. J Med Ethics 2021 May 14;48(10):785-793 [FREE Full text] [doi: 10.1136/medethics-2021-107275] [Medline: 33990427]
4. Kujala S, Hörhammer I, Väyrynen A, Holmroos M, Nättiaho-Rönnholm M, Hägglund M, et al. Patients' experiences of web-based access to electronic health records in Finland: cross-sectional survey. J Med Internet Res 2022 Jun 06;24(6):e37438 [FREE Full text] [doi: 10.2196/37438] [Medline: 35666563]
5. Hägglund M, Scandurra I. Patients' online access to electronic health records: current status and experiences from the implementation in Sweden. Stud Health Technol Inform 2017;245:723-727. [Medline: 29295193]
6. Zanaboni P, Kummervold P, Sørensen T, Johansen M. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020 Feb 07;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
7. Colivicchi Q. GPs abandon legal challenge over patient records access. Pulse. 2023. URL: https://www.pulsetoday.co.uk/news/breaking-news/gps-abandon-legal-challenge-over-patient-records-access/ [accessed 2023-07-29]
8. Moll J, Rexhepi H, Cajander Å, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 01;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]
9. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019 May 06;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
10. Bell S, Folcarelli P, Fossa A, Gerard M, Harper M, Leveille S, et al. Tackling ambulatory safety risks through patient engagement: what 10,000 patients and families say about safety-related knowledge, behaviors, and attitudes after reading visit notes. J Patient Saf 2018 May 18;17(8):e791-e799. [doi: 10.1097/pts.0000000000000494]
11. Blease C, Dong Z, Torous J, Walker J, Hägglund M, DesRoches CM. Association of patients reading clinical notes with perception of medication adherence among persons with serious mental illness. JAMA Netw Open 2021 Mar 01;4(3):e212823 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.2823] [Medline: 33760088]
12. DesRoches C, Bell S, Dong Z, Elmore J, Fernandez L, Fitzgerald P, et al. Patients managing medications and reading their visit notes: a survey of OpenNotes participants. Ann Intern Med 2019 May 28;171(1):69. [doi: 10.7326/m18-3197]
13. Torous J. Opening Mental Health Notes: 7 Tips to Prepare Clinicians. Psychology Today. 2020. URL: https://www.psychologytoday.com/us/blog/digital-mental-health/202010/opening-mental-health-notes-7-tips-prepare-clinicians [accessed 2020-12-11]
14. Blease C, McMillan B, Salmi L, Davidge G, Delbanco T. Adapting to transparent medical records: international experience with "open notes". BMJ 2022 Nov 21;379:e069861. [doi: 10.1136/bmj-2021-069861] [Medline: 36410770]
15. Dobscha SK, Kenyon EA, Pisciotta MK, Niederhausen M, Woods S, Denneson LM. Impacts of a web-based course on mental health clinicians' attitudes and communication behaviors related to use of OpenNotes. Psychiatr Serv 2019 Jun 01;70(6):474-479. [doi: 10.1176/appi.ps.201800416] [Medline: 30890047]
16. Preparing medical students and their teachers for shared medical records. OpenNotes. URL: https://www.opennotes.org/preparing-medical-students/ [accessed 2022-07-05]
17. Denneson L, Pisciotta M, Hooker E, Trevino A, Dobscha S. Impacts of a web-based educational program for veterans who read their mental health notes online. J Am Med Inform Assoc 2019 Jan 01;26(1):3-8 [FREE Full text] [doi: 10.1093/jamia/ocy134] [Medline: 30445648]
18. Keeping records. General Medical Council. URL: http://tinyurl.com/mrybhdu8 [accessed 2023-07-29]
19. Blease C, Torous J, Hägglund M. Does patient access to clinical notes change documentation? Front Public Health 2020;8:577896 [FREE Full text] [doi: 10.3389/fpubh.2020.577896] [Medline: 33330320]
20. McMillan B, Eastham R, Brown B, Fitton R, Dickinson D. Primary care patient records in the United Kingdom: past, present, and future research priorities. J Med Internet Res 2018 Dec 19;20(12):e11293 [FREE Full text] [doi: 10.2196/11293] [Medline: 30567695]
21. Blease CR, O'Neill S, Walker J, Hägglund M, Torous J. Sharing notes with mental health patients: balancing risks with respect. Lancet Psychiatry 2020 Nov;7(11):924-925. [doi: 10.1016/s2215-0366(20)30032-8]
22. Bernstein J. Not the last word: seeing ourselves as doctors see us. Clin Orthop Relat Res 2022 Aug 2;480(9):1653-1656. [doi: 10.1097/corr.0000000000002344]
23. Meier-Diedrich E, Davidge G, Hägglund M, Kharko A, Lyckblad C, McMillan B, et al. Changes in documentation due to patient access to electronic health records: protocol for a scoping review. JMIR Res Protoc 2023 Aug 28;12:e46722 [FREE Full text] [doi: 10.2196/46722] [Medline: 37639298]
24. Dobscha SK, Denneson LM, Jacobson LE, Williams HB, Cromer R, Woods S. VA mental health clinician experiences and attitudes toward OpenNotes. Gen Hosp Psychiatry 2016 Jan;38:89-93. [doi: 10.1016/j.genhosppsych.2015.08.001] [Medline: 26380876]
25. Petersson L, Erlingsdóttir G. Open Notes in Swedish Psychiatric Care (Part 2): survey among psychiatric care professionals. JMIR Ment Health 2018 Jun 21;5(2):e10521 [FREE Full text] [doi: 10.2196/10521] [Medline: 29929946]

XSL•FO
RenderX

26. DesRoches CM, Leveille S, Bell SK, Dong ZJ, Elmore JG, Fernandez L, et al. The views and experiences of clinicians sharing medical record notes with patients. JAMA Netw Open 2020 Mar 02;3(3):e201753 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.1753] [Medline: 32219406]
27. Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. Health Aff (Millwood) 2022 Feb 01;41(2):203-211 [FREE Full text] [doi: 10.1377/hlthaff.2021.01423] [Medline: 35044842]
28. Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. JAMA Netw Open 2022 Jan 04;5(1):e2144967 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.44967] [Medline: 35084481]
29. Stillman M. Death by patient portal. JAMA 2023 Jul 18;330(3):223-224. [doi: 10.1001/jama.2023.11629] [Medline: 37389857]
30. Golz C, Peter K, Müller TJ, Mutschler J, Zwakhalen S, Hahn S. Technostress and digital competence among health professionals in Swiss psychiatric hospitals: cross-sectional study. JMIR Ment Health 2021 Nov 04;8(11):e31408 [FREE Full text] [doi: 10.2196/31408] [Medline: 34734840]
31. Domaney N, Torous J, Greenberg W. Exploring the association between electronic health record use and burnout among psychiatry residents and faculty: a pilot survey study. Acad Psychiatry 2018 Oct;42(5):648-652. [doi: 10.1007/s40596-018-0939-x] [Medline: 29785625]
32. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015 Mar 02;65(632):e141-e151. [doi: 10.3399/bjgp15x683941]
33. Ferguson K, Fraser M, Tuna M, Bruntz C, Dahrouge S. The impact of an electronic portal on patient encounters in primary care: interrupted time-series analysis. JMIR Med Inform 2023 Feb 06;11:e43567 [FREE Full text] [doi: 10.2196/43567] [Medline: 36745495]
34. Steitz BD, Sulieman L, Wright A, Rosenbloom ST. Association of immediate release of test results to patients with implications for clinical workflow. JAMA Netw Open 2021 Oct 01;4(10):e2129553 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.29553] [Medline: 34661667]
35. McMillan B, Davidge G, Nadeem F, Dowding D, Wilson K, Davies A. Navigating the electronic health record in university education: helping health care professionals of the future prepare for 21st century practice. BMJ Health Care Inform 2023 Mar;30(1) [FREE Full text] [doi: 10.1136/bmjhci-2022-100722] [Medline: 36914229]
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/2yr8rvkh [accessed 2023-12-19]
37. Mahood S. Medical education: Beware the hidden curriculum. Can Fam Physician 2011 Sep;57(9):983-985 [FREE Full text] [Medline: 21918135]
38. Fisher M, Keil F. The curse of expertise: when more knowledge leads to miscalibrated explanatory insight. Cogn Sci 2016 Jul;40(5):1251-1269 [FREE Full text] [doi: 10.1111/cogs.12280] [Medline: 26369299]
39. Blease C, Cohen I, Hoffman S. Sharing clinical notes: potential medical-legal benefits and risks. JAMA 2022 Feb 22;327(8):717-718. [doi: 10.1001/jama.2021.23179] [Medline: 35119468]
40. Batson C, Polycarpou M, Harmon-Jones E, Imhoff H, Mitchener E, Bednar L, et al. Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? J Pers Soc Psychol 1997 Jan;72(1):105-118. [doi: 10.1037/0022-3514.72.1.105]
41. Hein IM, De Vries MC, Troost PW, Meynen G, Van Goudoever JB, Lindauer RJL. Informed consent instead of assent is appropriate in children from the age of twelve: policy implications of new findings on children's competence to consent to clinical research. BMC Med Ethics 2015 Nov 09;16(1):76 [FREE Full text] [doi: 10.1186/s12910-015-0067-z] [Medline: 26553304]
42. Heins A, Homel P, Safdar B, Todd K. Physician race/ethnicity predicts successful emergency department analgesia. J Pain 2010 Jul;11(7):692-697. [doi: 10.1016/j.jpain.2009.10.017]
43. Xu X, Zuo X, Wang X, Han S. Do you feel my pain? Racial group membership modulates empathic neural responses. J Neurosci 2009 Jul 01;29(26):8525-8529. [doi: 10.1523/jneurosci.2418-09.2009]
44. Avenanti A, Sirigu A, Aglioti SM. Racial bias reduces empathic sensorimotor resonance with other-race pain. Curr Biol 2010 Jun 08;20(11):1018-1022 [FREE Full text] [doi: 10.1016/j.cub.2010.03.071] [Medline: 20537539]
45. Trivers R. The elements of a scientific theory of self-deception. Ann N Y Acad Sci 2000 Apr 25;907(1):114-131. [doi: 10.1111/j.1749-6632.2000.tb06619.x] [Medline: 10818624]
46. Nielsen H, Tulinius C. Preventing burnout among general practitioners: is there a possible route? Educ Prim Care 2009 Sep;20(5):353-359. [doi: 10.1080/14739879.2009.11493817] [Medline: 19849901]
47. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019 Mar 20;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
48. Kocaballi A, Ijaz K, Laranjo L, Quiroz J, Rezazadegan D, Tong H, et al. Envisioning an artificial intelligence documentation assistant for future primary care consultations: a co-design study with general practitioners. J Am Med Inform Assoc 2020 Nov 01;27(11):1695-1704 [FREE Full text] [doi: 10.1093/jamia/ocaa131] [Medline: 32845984]

XSL•FO
RenderX

49. Goldberg C, Kohane I, Lee P. The AI Revolution in Medicine: GPT-4 and Beyond. London: Pearson Education Limited; 2023.
50. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023 Nov 10;26(1):e300884 [FREE Full text] [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]
51. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
52. Kung T, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
53. Lubell J. ChatGPT passed the USMLE. What does it mean for med ed? American Medical Association. 2023. URL: https://www.ama-assn.org/practice-management/digital/chatgpt-passed-usmle-what-does-it-mean-med-ed [accessed 2023-07-29]
54. Dijkstra A, Verdonk P, Lagro-Janssen A. Gender bias in medical textbooks: examples from coronary heart disease, depression, alcohol abuse and pharmacology. Med Educ 2008 Oct;42(10):1021-1028. [doi: 10.1111/j.1365-2923.2008.03150.x] [Medline: 18761614]
55. Duma N, Vera Aguilera J, Paludo J, Haddox C, Gonzalez Velez M, Wang Y, et al. Representation of minorities and women in oncology clinical trials: review of the past 14 years. JOP 2018 Jan;14(1):e1-e10. [doi: 10.1200/jop.2017.025288]
56. Geller S, Koch A, Pellettieri B, Carnes M. Inclusion, analysis, and reporting of sex and race/ethnicity in clinical trials: have we made progress? J Womens Health (Larchmt) 2011 Mar;20(3):315-320 [FREE Full text] [doi: 10.1089/jwh.2010.2469] [Medline: 21351877]
57. Watts G. Why the exclusion of older people from clinical research must stop. BMJ 2012 May 21;344(may21 1):e3445-e3445. [doi: 10.1136/bmj.e3445] [Medline: 22613873]
58. Bourgeois F, Olson K, Tse T, Ioannidis J, Mandl K. Prevalence and characteristics of interventional trials conducted exclusively in elderly persons: a cross-sectional analysis of registered clinical trials. PLoS One 2016;11(5):e0155948 [FREE Full text] [doi: 10.1371/journal.pone.0155948] [Medline: 27196289]
59. Zack T, Lehman E, Suzgun M, Rodriguez J, Celi L, Gichoya J, et al. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare. medRxiv Preprint posted online July 17, 2023. [doi: 10.1101/2023.07.13.23292577]
60. Birhane A, Prabhu V, Han S, Boddeti V. On Hate Scaling Laws For Data-Swamps. arXiv Preprint posted online June 22, 2023. [FREE Full text]
61. Adams K. Becker's Health IT. URL: https://www.beckershospitalreview.com/healthcare-information-technology/31-numbers-that-show-how-big-epic-cerner-allscripts-meditech-are-in-healthcare.html [accessed 2023-07-31]
62. Adams K. Epic to Integrate GPT-4 into Its EHR Through Expanded Microsoft Partnership. MedCity News. 2023. URL: https://medcitynews.com/2023/04/epic-to-integrate-gpt-4-into-its-ehr-through-expanded-microsoft-partnership/ [accessed 2023-07-31]
63. Boyd E. Introducing GPT-4 in Azure OpenAI Service. Azure Microsoft. 2023. URL: https://azure.microsoft.com/en-us/blog/introducing-gpt4-in-azure-openai-service/ [accessed 2023-07-31]
64. DePeau-Wilson M. Clinical Note Writing App Powered by GPT-4 Set to Debut This Year. MedPage Today. 2023. URL: http://tinyurl.com/2nv8dfj4 [accessed 2023-12-15]
65. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023 Jan 23;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]

Abbreviations

AI: artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
ORA: online record access

Edited by K Venkatesh; submitted 31.07.23; peer-reviewed by B Senst, J Schwarz, R Marshall; comments to author 28.09.23; revised version received 30.09.23; accepted 10.11.23; published 04.01.24.

Please cite as:
Blease C, Torous J, McMillan B, Hägglund M, Mandl KD
Generative Language Models and Open Notes: Exploring the Promise and Limitations
JMIR Med Educ 2024;10:e51183
URL: https://mededu.jmir.org/2024/1/e51183
doi:10.2196/51183
PMID:38175688

XSL•FO
RenderX

©Charlotte Blease, John Torous, Brian McMillan, Maria Hägglund, Kenneth D Mandl. Originally published in JMIR Medical Education (https://mededu.jmir.org), 04.01.2024. 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.

XSL•FO
RenderX

Original Paper

Pure Wisdom or Potemkin Villages? A Comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 Style Questions: Quantitative Analysis

Leonard Knoedler[1], MD; Michael Alfertshofer[2], MD; Samuel Knoedler[1,3], BSc; Cosima C Hoch[4], BSc; Paul F Funk[5], BSc; Sebastian Cotofana[6,7], MD, PhD; Bhagvat Maheta[8], BS; Konstantin Frank[9], MD; Vanessa Brébant[1], MD; Lukas Prantl[1], MD, PhD; Philipp Lamby[1], MD, PhD

[1]Department of Plastic, Hand and Reconstructive Surgery, University Hospital Regensburg, Regensburg, Germany
[2]Division of Hand, Plastic and Aesthetic Surgery, Ludwig-Maximilians University Munich, Munich, Germany
[3]Division of Plastic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States
[4]Department of Otolaryngology, Head and Neck Surgery, School of Medicine, Technical University of Munich, Munich, Germany
[5]Department of Otolaryngology, Head and Neck Surgery, University Hospital Jena, Friedrich Schiller University Jena, Jena, Germany
[6]Department of Dermatology, Erasmus Hospital, Rotterdam, Netherlands
[7]Centre for Cutaneous Research, Blizard Institute, Queen Mary University of London, London, United Kingdom
[8]College of Medicine, California Northstate University, Elk Grove, CA, United States
[9]Ocean 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.

XSL•FO
RenderX

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.

(JMIR Med Educ 2024;10:e51148) doi:10.2196/51148

KEYWORDS

ChatGPT; United States Medical Licensing Examination; artificial intelligence; USMLE; USMLE Step 1; OpenAI; medical education; clinical decision-making

Introduction

Since its inception in 1992, the United States Medical Licensing Examination (USMLE) has been considered an integral milestone in medical education [1]. The 3 USMLE steps are jointly sponsored by the Federation of State Medical Boards and the National Board of Medical Examiners. Each step is designed to specifically test another facet of the examinee's skill set. For instance, USMLE Step 1 assesses a student's understanding and application of basic sciences relevant to the field of medicine (eg, anatomy and physiology), while USMLE Step 2 tests the examinee's clinical knowledge (USMLE Step 2 CK) and communication skills (USMLE Step 2 CS). USMLE Step 3 evaluates the student's understanding of biomedical and clinical science [2-4]. USMLE scores have been associated with residency matching and future career perspectives [5].

Artificial intelligence (AI)–supported tools have been proposed for a variety of medical scenarios, including preoperative outcome simulation, patient education, and automated disease grading [6-9]. Recently, chatbots such as ChatGPT have emerged as next-generation AI technology. The strengths of this novel AI-powered approach include 24-7 availability, cost efficiency, and individualization [10]. A mounting body of evidence has investigated ChatGPT's performance on different standardized exams. For instance, Hoch et al [11] reported that ChatGPT answered 57% of facial surgery board certification test questions correctly, while Kung et al [12] used a limited set of USMLE test questions (USMLE Step 1: 119; USMLE Step 2 CK: 102; USMLE Step 3: 122) and found that ChatGPT achieved performance levels near the passing threshold for all 3 steps.

However, there is a scarcity of studies that comprehensively investigate overall ChatGPT performance on USMLE Step 3 test questions in a large-scale study and compare test performances between ChatGPT 3.5 and ChatGPT 4. This knowledge gap may increase the risk of AI cheating in such career-deciding exams and cloud the vision of ChatGPT's strengths and limitations.

Therefore, we aimed to determine ChatGPT's performance on USMLE Step 3 practice test questions based on 1840 AMBOSS USMLE Step 3 Style Questions. This line of research may serve as a primer elucidating the strengths and weaknesses of multiple ChatGPT versions and deducing evidence-based strategies to counteract AI cheating.

Methods

Access to Question Bank and Data Entry Procedure

From June 12, 2023, to June 19, 2023, we obtained access to the AMBOSS question bank [13]. Within this time frame, we collected a total of 1840 practice questions specifically designed for the USMLE Step 3 exam. Before initiating our study, we acquired official permission from AMBOSS (AMBOSS GmbH) to use their USMLE Step 3 question bank for research purposes. To ensure the reliability of our data, 2 examiners (MA and LK) cross-checked the question inputs randomly to confirm that none of the answers were indexed on Google before June 19, 2023. Many USMLE questions are on the internet, including USMLE sample questions as well as a few AMBOSS questions; however, we ensured that those questions were not included in this analysis to minimize the risk of prior memorization of the questions by ChatGPT. July 19, 2023, was chosen since it represents the most recent accessible date within the training data set of ChatGPT. There are many forms of AI versions with capabilities to answer USMLE Step 3 practice test questions; however, ChatGPT is the most widely used AI at the time of this study, making it the best fit for our study.

Question Screening and Categorization

To maintain the quality of our sample questions, we subjected all test questions to independent screening by 4 examiners (MA, SK, CCH, and LK). Questions containing clinical images and photographs were excluded from the study, resulting in the removal of 229 image-based questions. Subsequently, the remaining 1840 test questions were classified based on their respective specialties, using the categorization provided by AMBOSS. All questions included in our study followed a multiple-choice single-answer format. The questions used for both ChatGPT 3.5 and ChatGPT 4 were matched for content and difficulty based on the standardized definitions provided by the AMBOSS question bank to ensure consistent analysis between both AI versions.

Comparison of ChatGPT Versions and Analysis of Question Stems

To evaluate any performance differences between ChatGPT 3.5 and ChatGPT 4, we conducted a subgroup analysis specifically focusing on ChatGPT 4. Additionally, we analyzed the question stems of both ChatGPT 3.5 and ChatGPT 4, specifically looking for specific buzzwords related to diagnostic methods and patient information, such as "Ultrasound," "Serology," and "Nicotine Abuse." These particular words and phrases may suggest one

XSL•FO
RenderX

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 (ρ). IBM SPSS Statistics 25 (IBM Corp) was used for statistical analysis, and a 2-tailed P value ≤.05 was considered statistically significant.

Results

General Test Question Characteristics and Performance Statistics

The overall accuracy of ChatGPT 3.5 for USMLE Step 3 was 56.9% (1047/1840), while ChatGPT 4 answered 84.7% (194/229) of test questions correctly (P<.001). Specialty-specific number of test questions and performance scores are presented in Tables 1 and 2. ChatGPT 3.5 received the greatest number of questions on the nervous, cardiovascular, and gastrointestinal systems, while ChatGPT 4 received the greatest number of questions on behavior health, the female reproductive system, as well the blood and lymphatic system. When considering the accuracy of ChatGPT based on the category of questions, ChatGPT 3.5 performed the best on behavioral health, multisystem processes and disorders, and pregnancy-related questions. On the other hand, ChatGPT 4 had the greatest accuracy on questions related to the endocrine and musculoskeletal systems as well as biostatistics and multisystem processes and disorders.

XSL•FO
RenderX

Table 1. The number of test questions answered by ChatGPT 3.5 and its performance, stratified by questions category (N=1840).

Question category	Test questions answered, n	Correct questions, n/N (%)
Male reproductive system	28	17/28 (60.1)
General principles and foundational science	29	16/29 (55.2)
Immune system	40	25/40 (62.5)
Skin and subcutaneous tissue	72	39/72 (54.2)
Renal and urinary systems	72	39/72 (54.2)
Biostats and epidemiology	87	45/87 (51.7)
Female reproductive system and breast	88	48/88 (54.5)
Musculoskeletal system	94	56/94 (58.5)
Endocrine system	103	58/103 (56.3)
Blood and lymphoreticular system	105	55/105 (52.4)
Pregnancy, childbirth, and puerperium	111	66/111 (59.5)
Behavioral health	115	73/115 (63.5)
Multisystem processes and disorders	122	73/122 (59.8)
Respiratory system	130	71/130 (54.6)
Social sciences	141	86/141 (61.0)
Gastrointestinal system	156	87/156 (55.8)
Cardiovascular system	161	89/161 (55.3)
Nervous system and special senses	186	104/186 (55.9)

Table 2. The number of test questions answered by ChatGPT 4 and its performance, stratified by questions category (N=229).

Question category	Test questions answered, n	Correct questions, n/N (%)
Endocrine system	1	1/1 (100)
Biostats and epidemiology	14	13/14 (92.3)
General principles and foundational science	17	14/17 (82.4)
Multisystem processes and disorders	17	15/17 (88.2)
Pregnancy, childbirth, and puerperium	19	15/19 (79.0)
Gastrointestinal system	21	18/21 (85.7)
Cardiovascular system	21	15/21 (71.4)
Nervous system and special senses	21	15/21 (71.4)
Blood and lymphoreticular system	23	20/23 (87.0)
Female reproductive system and breast	23	20/23 (87.0)
Behavioral health	24	21/24 (87.5)

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 (ρ=–0.069; *P*=.003) while not yielding significance for ChatGPT 4 (*P*=.87). For ChatGPT 3.5, the mean number of characters was 1062 (SD 310) for correct answers versus 1100 (SD 304) for falsely answered questions (*P*=.009). However, the mean character count was comparable for test questions answered by ChatGPT 4 (mean correct answers 1068, SD 274 vs mean false answers 1056, SD 233; *P*=.80).

Test Question Difficulty and the Performance of ChatGPT

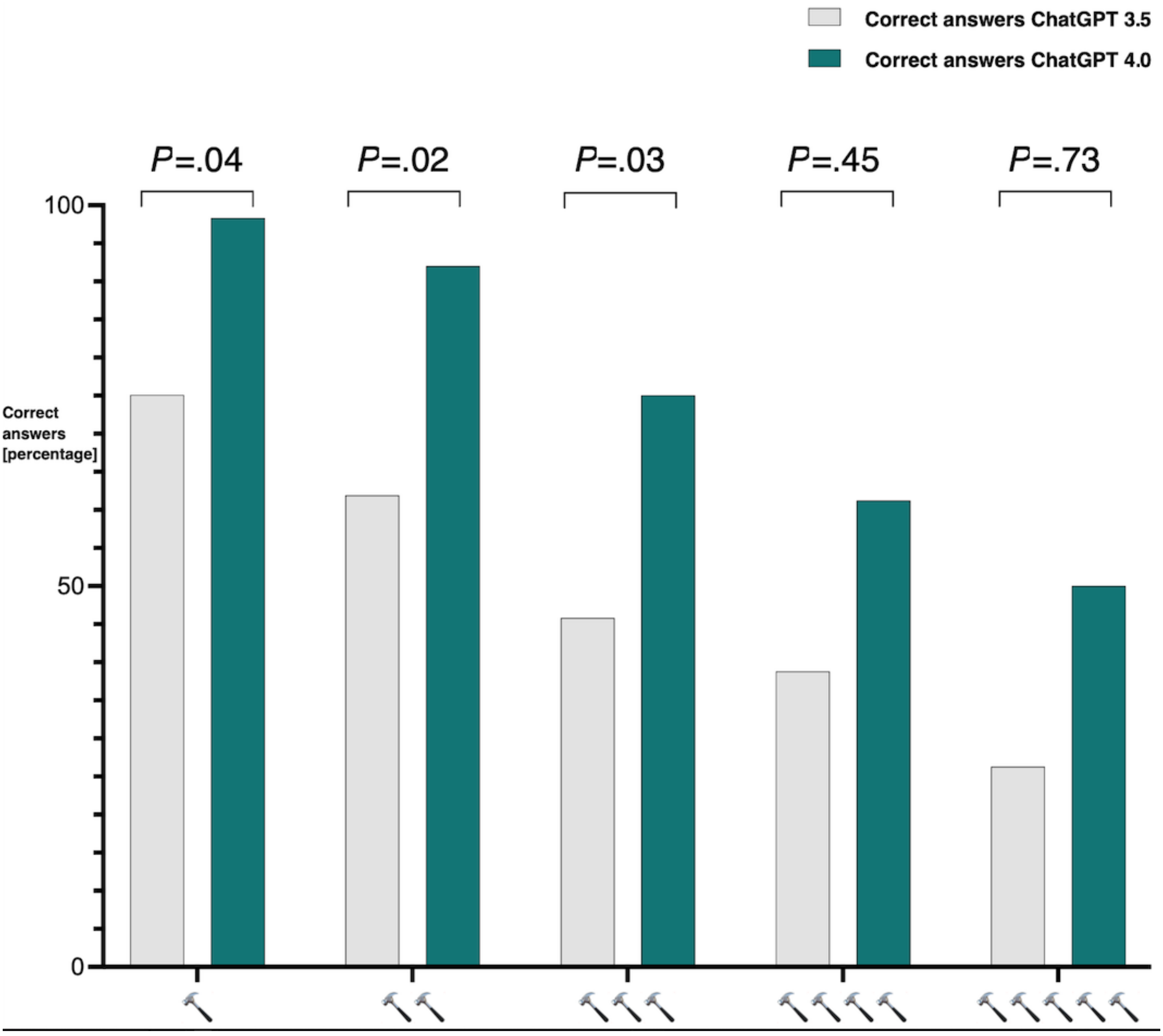
Question distribution and performance scores sorted by level of test question difficulty are illustrated in Figure 1. Test question difficulty, defined by AMBOSS hammer categorization, and the performance of ChatGPT 3.5 were significantly correlated (ρ=–0.289; *P*<.001). This was reproducible in ChatGPT 4 (ρ=–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

XSL•FO
RenderX

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.

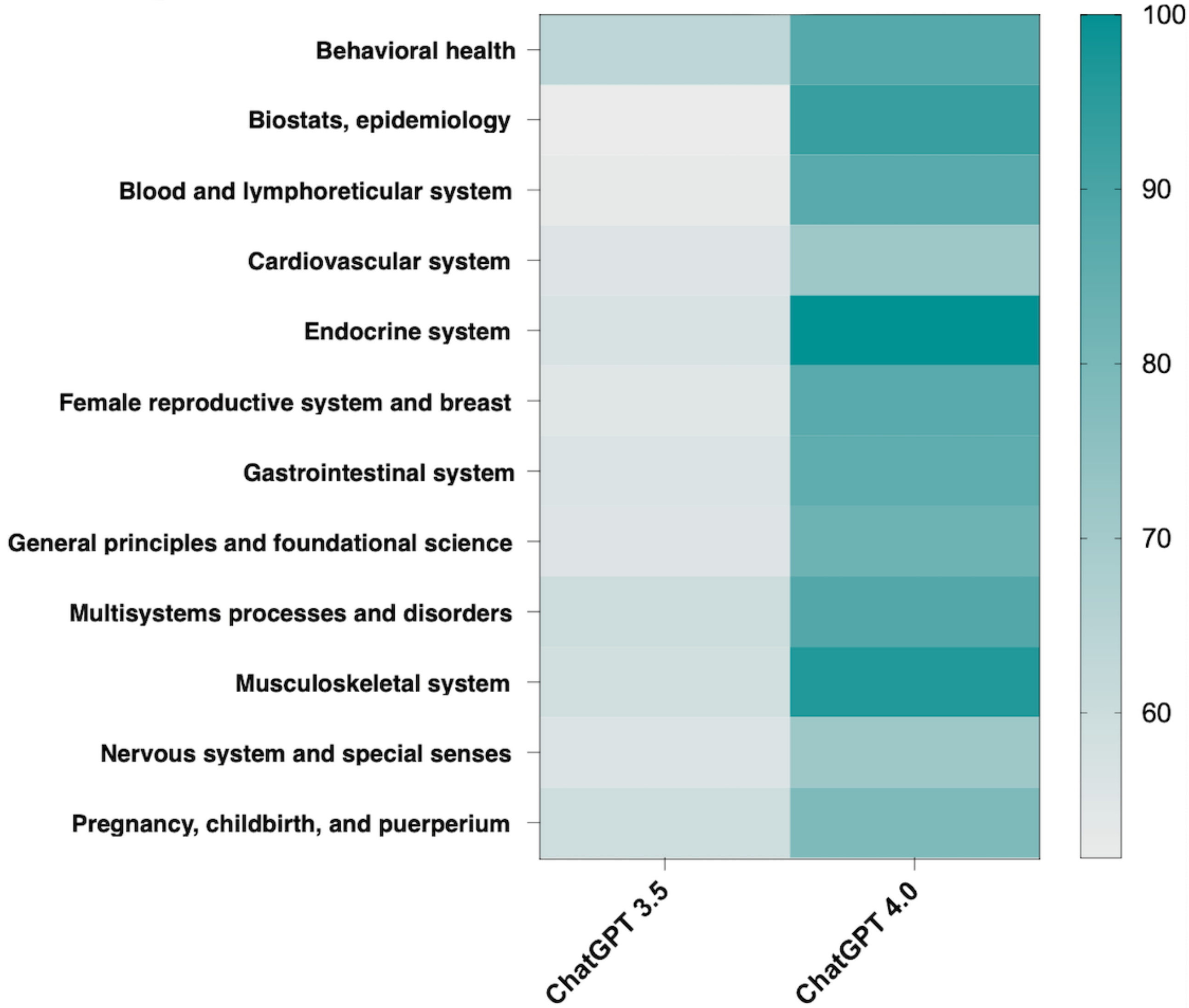
Relative to ChatGPT 3.5, ChatGPT 4 performed better on questions from every specialty category. The biggest differences in accuracy were in biostatistics, epidemiology, the endocrine system, and the musculoskeletal system.

Figure 1. Question distribution and performance scores sorted by level of test question difficulty.



XSL•FO
RenderX

Figure 2. Percentage of correct responses from ChatGPT 3.5 versus ChatGPT 4.0, sorted by specialty.

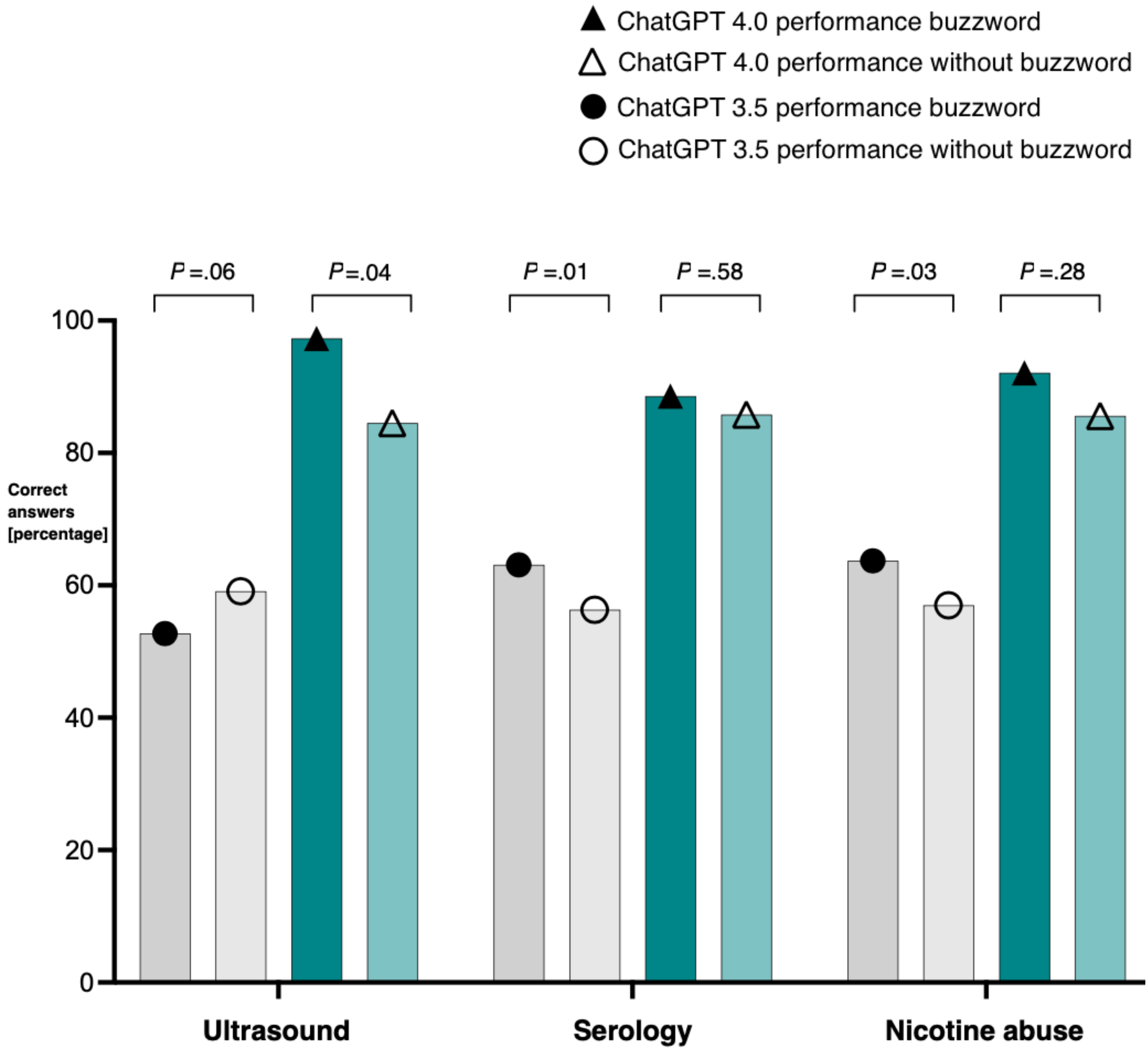


Buzzwords and the Performance of ChatGPT

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.

XSL•FO
RenderX

Figure 3. Performance scores of ChatGPT 3.5 versus ChatGPT 4.0, sorted by buzzword.



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

XSL•FO
RenderX

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*<.001) with the level of test question difficulty. This indicates that sophisticated USMLE questions still challenge and fool both human examinees and AI chatbots. Typically, the most difficult USMLE questions include distractors as well as irrelevant or additional information.; they also require high-level reasoning and interdisciplinary thinking. Our group previously showed that ChatGPT 3.5, similar to the human user peer group, struggled to answer 4- and 5-hammer questions [11]. Such pitfalls continue to perplex the next generation of AI-powered chatbots. Therefore, a thorough analysis of 4- or 5-hammer questions may help examiners refine their test questions and shield the USMLE against AI cheating.

Overall, the phenomenal improvement in the test-taking performance of ChatGPT 4 compared to ChatGPT 3.5 raises intriguing questions regarding future applications and implications of AI in medical education and diagnostics. AI has shown its prowess not only on the USMLE examinations in medical education but also on advanced examinations, such as the neurosurgical written boards [16]. This phenomenon ventures into other aspects of medicine as well, including research and clinical performance [17]. It is imperative that future research ventures into a deeper analysis of the performance of ChatGPT 4 by conducting thorough investigations that probe its strengths and limitations in a more granulated manner, potentially employing diversified medical question banks, simulating real-world scenarios, and engaging experts for analysis and evaluation to allow for the best possible medical education and ultimately patient health care [18].

Limitations

This study needs to be interpreted in the light of the following limitations: first, due to the restricted use of ChatGPT (only 25 entries every 3 hours) we were not able to perform a direct comparison of ChatGPT 3.5 and ChatGPT 4 for all test questions included in this study, which might limit its validity. Furthermore, although we attempted to ensure that the questions provided for analysis were not freely available on the internet to minimize the risk of ChatGPT having already seen the exact question, students and researchers around the world may have input certain AMBOSS USMLE Step 3 Style Questions into ChatGPT. This adds a potential confounding factor of ChatGPT memorizing the correct answer from seeing the question beforehand. We used the 2 most recent versions of ChatGPT (ie, ChatGPT 3.5 and ChatGPT 4) to test and compare the performance of large language models on 1840 AMBOSS USMLE Step 3 questions. Thus, the findings of this study should be revalidated for upcoming ChatGPT versions. Future studies may involve additional chatbots, question banks, and image-based test questions. Further, the performance of ChatGPT on USMLE steps could be compared to other national medical licensing exams.

Conclusions

This study is the first direct comparison of ChatGPT 4 and ChatGPT 3.5 based on 1840 AMBOSS USMLE Step 3 test questions. Our analysis showed that ChatGPT 4 outperformed its predecessor version across different specialties and difficulty levels, ultimately yielding accuracy levels of 84.7%. However, we could identify persisting weak points of ChatGPT 4, including abstract thinking and elaborated test questions. This line of research may serve as an evidence-based fundament to safeguard the USMLE steps and medical education against AI cheating while underscoring the potency of AI-driven chatbots.

Conflicts of Interest

None declared.

References

1. Morrison C, Barone M, Baker G, Ross L, Pak S. Investigating the relationship between a clinical science composite score and USMLE Step 2 clinical knowledge and Step 3 performance. Med Sci Educ 2020 Mar 02;30(1):263-269 [FREE Full text] [doi: 10.1007/s40670-019-00893-0] [Medline: 34457666]
2. Burk-Rafel J, Santen SA, Purkiss J. Study behaviors and USMLE Step 1 performance. Acad Med 2017;92:S67-S74. [doi: 10.1097/acm.0000000000001916]
3. Cangialosi P, Chung B, Thielhelm T, Camarda N, Eiger D. Medical students' reflections on the recent changes to the USMLE Step exams. Acad Med 2021 Mar 01;96(3):343-348 [FREE Full text] [doi: 10.1097/ACM.0000000000003847] [Medline: 33208676]
4. Patel MD, Benefield T, Hunt KN, Tomblinson CM, Ali K, DeBenedectis CM, et al. USMLE Step 3 scores have value in predicting ABR core examination outcome and performance: a multi-institutional study. Acad Radiol 2021 May;28(5):726-732. [doi: 10.1016/j.acra.2020.06.032] [Medline: 32773330]
5. Gauer JL, Jackson JB. The association of USMLE Step 1 and Step 2 CK scores with residency match specialty and location. Med Educ Online 2017 Aug;22(1):1358579 [FREE Full text] [doi: 10.1080/10872981.2017.1358579] [Medline: 28762297]

XSL•FO
RenderX

6. Chartier C, Gfrerer L, Knoedler L, Austen W. Artificial intelligence-enabled evaluation of pain sketches to predict outcomes in headache surgery. Plast Reconstr Surg 2023 Feb 01;151(2):405-411. [doi: 10.1097/PRS.0000000000009855] [Medline: 36696328]
7. Knoedler L, Odenthal J, Prantl L, Oezdemir B, Kehrer A, Kauke-Navarro M, et al. Artificial intelligence-enabled simulation of gluteal augmentation: a helpful tool in preoperative outcome simulation? J Plast Reconstr Aesthet Surg 2023 May;80:94-101. [doi: 10.1016/j.bjps.2023.01.039] [Medline: 37001299]
8. Knoedler L, Miragall M, Kauke-Navarro M, Obed D, Bauer M, Tißler P, et al. A ready-to-use grading tool for facial palsy examiners-automated grading system in facial palsy patients made easy. J Pers Med 2022 Oct 19;12(10):1739 [FREE Full text] [doi: 10.3390/jpm12101739] [Medline: 36294878]
9. Knoedler L, Baecher H, Kauke-Navarro M, Prantl L, Machens H, Scheuermann P, et al. Towards a reliable and rapid automated grading system in facial palsy patients: facial palsy surgery meets computer science. J Clin Med 2022 Aug 25;11(17):4998 [FREE Full text] [doi: 10.3390/jcm11174998] [Medline: 36078928]
10. Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine. Clin Transl Med 2023 Mar;13(3):e1206 [FREE Full text] [doi: 10.1002/ctm2.1206] [Medline: 36854881]
11. Hoch CC, Wollenberg B, Lüers J, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023 Jun 07:4271-4278. [doi: 10.1007/s00405-023-08051-4] [Medline: 37285018]
12. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
13. AMBOSS question bank. URL: https://www.amboss.com/us [accessed 2023-12-18]
14. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [10.2196/45312] [Medline: 36753318]
15. Koga S. The potential of ChatGPT in medical education: focusing on USMLE preparation. Ann Biomed Eng 2023 May 29:2123-2124. [doi: 10.1007/s10439-023-03253-7] [Medline: 37248408]
16. Hopkins B, Nguyen V, Dallas J. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board's style questions. J Neurosurg 2023:904-911. [doi: 10.3171/2023.2.jns23419]
17. Shaffrey E, Eftekari S, Wilke L, Poore S. Surgeon or bot? The risks of using artificial intelligence in surgical journal publications. Annals of Surgery Open 2023;4(3):e309. [doi: 10.1097/as9.0000000000000309]
18. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb 9;2(2):e0000205 [FREE Full text] [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]

Abbreviations

AI: artificial intelligence
CK: clinical knowledge
CS: communication skills
MRI: magnetic resonance imaging
USMLE: United States Medical Licensing Examination

Edited by K Venkatesh; submitted 22.07.23; peer-reviewed by J Wilkinson, M Triola; comments to author 10.09.23; revised version received 30.09.23; accepted 20.10.23; published 05.01.24.

Please cite as:
Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, Maheta B, Frank K, Brébant V, Prantl L, Lamby P
Pure Wisdom or Potemkin Villages? A Comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 Style Questions: Quantitative Analysis
JMIR Med Educ 2024;10:e51148
URL: https://mededu.jmir.org/2024/1/e51148
doi:10.2196/51148
PMID:38180782

©Leonard Knoedler, Michael Alfertshofer, Samuel Knoedler, Cosima C Hoch, Paul F Funk, Sebastian Cotofana, Bhagvat Maheta, Konstantin Frank, Vanessa Brébant, Lukas Prantl, Philipp Lamby. Originally published in JMIR Medical Education

XSL•FO
RenderX

(https://mededu.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects

Lukas Weidener[1], BSc, Dr med; Michael Fischer[1], 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].

XSL•FO
RenderX

Although the history of AI can be traced back to the 1950s, the public's unrestricted access to highly advanced large language models, such as ChatGPT, can be seen as a significant turning point in the history of AI [5,6]. Early studies demonstrated that ChatGPT is capable of successfully completing the written portion of the United States Medical Licensing Examination [7]. Given the capabilities of AI-based chat applications such as ChatGPT in medicine, further studies have highlighted their potential use in providing information on cancer, assisting in clinical diagnoses, authoring scientific research articles, and patient communication [8-10]. Considering the wide availability and integration of medical knowledge in this application, its increasing use in medicine and among medical students is foreseeable [11].

Despite the long history of AI and the increasing adoption of this technology, there is disagreement regarding its definition among the scientific community [12]. There is a consensus within the scientific community on distinguishing between the so-called strong AI, also known as "artificial general intelligence," and weak AI or "artificial narrow intelligence" [13]. This categorization is based on the capabilities of AI or its areas of application [13]. Strong AI, recognized for its human-equivalent intellectual abilities and knowledge, stands in contrast to weak AI, which refers to AI solutions capable of accomplishing specific tasks effectively [13]. The area of weak AI can be further divided into the so-called statistical AI and symbolic AI. The field of statistical AI also includes machine learning and deep learning, on which large language models such as ChatGPT are based [13]. Areas of application for symbolic AI in medicine include expert systems (eg, clinical decision support systems), which make decisions based on explicit knowledge in the form of predefined rules [14].

Considering the likely significant impact the implementation and use of AI in medicine is poised to make, a growing body of literature advocates the inclusion of AI-related content in medical curricula [15-18]. In addition to implications for the medical profession and patient care, medical students are expected to face new ethical challenges posed by the use of AI in medicine [15,19]. Despite the potentially significant ethical challenges anticipated from the deployment of AI in medicine, such as the possibility of discrimination due to biases in the data used for training or effects on patient autonomy, there is a near-complete absence of scientific publications on specific teaching content or methods related to AI ethics as part of medical higher education.

In addition to the lack of specificity regarding teaching content on AI and AI ethics, the absence of studies on medical students' perception of AI ethics education (including teaching content) is notable [20,21]. It is essential to point out that the current state of research regarding medical students' perceptions and assessments of AI application in medicine largely represents a knowledge base that predates the advent of large language models such as ChatGPT. With the ubiquity of the aforementioned AI applications at the time of this publication, it is reasonable to expect that medical students' assessments of AI implementation in medicine will deviate significantly from earlier publications within this area of research, highlighting the need for further research.

Objective

This study aimed to explore how medical students perceive the use of AI in medicine, as well as the teaching of AI and AI ethics (including prospective AI ethics teaching topics). In this context, the introduction and accessibility of large language models such as ChatGPT should be emphasized, leading to the following research question: how do medical students from Germany, Austria, and Switzerland perceive (1) the application of AI in medical practice, (2) the integration of AI and AI ethics into medical education, and (3) AI ethics teaching content in their curriculum in accordance with the use of AI-based chat applications such as ChatGPT?

To address this research question, the participating medical students were divided into 2 groups based on their prior use of AI-based (chat) applications, such as ChatGPT.

Methods

Overview

This cross-sectional study was conducted between June 15 and July 15, 2023. During this time frame, an invitation to participate in the study was sent to medical students who were regularly enrolled in universities in Germany, Austria, and Switzerland. The study sample included medical students from all academic semesters, including those in practically oriented semesters such as the practical year in Germany. Participation in the study was voluntary and there were no consequences for nonparticipation. The study used an anonymous web-based survey, with recruitment facilitated through email invitations and assistance from various medical student associations, unions, and councils in their respective countries. To minimize potential selection bias, the survey invited medical students from various universities and academic semesters in Germany, Austria, and Switzerland. This strategy ensured a broad and representative sample of the participants. Moreover, careful construction and pretesting of the survey were conducted to minimize potential response biases. Before the official data collection, a pretest was conducted with 11 medical students from the target population. The web-based survey provider, "LimeSurvey" was used for both the pretest and the main study.

Ethical Considerations

The Research Committee for Scientific Ethical Questions granted ethical approval for this study (3181) on January 16, 2023.

Survey Development

The survey used for data collection was developed based on existing scientific publications [15,22]. Owing to the lack of references in the areas of AI teaching, AI ethics, and recent developments in AI, most items used for the survey were newly formulated. The survey comprises 53 items, including both questions and statements. During the development process, these items were distributed across 6 parts, with some contingent on the responses to the preceding items. The first part aimed to collect information on the demographic characteristics and educational background of the participants. To address the research question of this study, participants were divided into

XSL•FO
RenderX

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, which focused on AI ethics education within their regular medical studies. Similarly, question Q7 was shown only to those who responded "yes" to question 6, 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 = [Z^2 * p * (1-p)] / E^2) [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, μ is the mean of the responses for the group, and σ 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

XSL•FO
RenderX

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 α=.05, and a value of $P \leq .05$ was considered statistically significant. Only complete data sets were included in the data analysis to avoid potential biases that could arise from replacing or estimating the missing values (list-wise deletion).

Results

Overview

In total, 521 medical students participated in the survey, yielding 487 complete and valid data sets for the statistical analysis. The survey invitations were disseminated via email with the help of medical student associations, unions, and councils. The total number of medical students reached and the precise response rate could only be approximated. On the basis of the feedback received from the engaged medical student councils, we estimated that at least 2000 medical students were approached. This would be equal to a response rate of 24.35% (487/2000). Our sample size calculation was based on the assumption that the use of AI-based (chat) applications such as ChatGPT does not diverge markedly among medical students from each of the countries of interest, namely Germany, Austria, and Switzerland. Consequently, the chi-square test of independence was used for statistical evaluation. We posited a null hypothesis (H) asserting no association between the variables (use of AI-based applications and country of study) and an alternative hypothesis (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.

XSL•FO
RenderX

Table 1. Demographic characteristics of medical students (n=487).

Characteristics	Medical students, n (%)
Gender	
Woman	270 (55.4)
Man	203 (41.7)
Nonbinary	3 (0.6)
Prefer not to say	11 (2.3)
Age (y)	
<20	56 (11.5)
20-25	301 (61.8)
26-30	92 (18.9)
31-35	28 (5.7)
>35	10 (2.0)
Country of enrollment (medical studies)	
Germany	296 (60.7)
Austria	105 (21.5)
Switzerland	86 (17.7)
Stage of study	
Preclinical	57 (11.7)
Clinical	277 (56.9)
Practical year	63 (12.9)
Elective year	26 (5.3)
Bachelor	46 (9.4)
Master	18 (3.7)

XSL•FO
RenderX

Table 2. Educational background of the participating medical students from Germany, Austria, and Switzerland (n=487).

Question	Participants, n (%)
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)	
Yes	425 (87.2)
No	62 (12.7)
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)	
Yes	26 (5.3)
No	461 (94.7)
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)	
Yes	51 (10.5)
No	436 (89.2)
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)	
Yes	21 (4.3)
No	466 (95.7)
Q5: Which of the following contents were covered as part of the education?[a,b]	
Informed consent	11 (2.3)
Bias	15 (3.1)
Data privacy	13 (2.7)
Explainability	12 (2.5)
Safety (of AI-based applications)	10 (2)
Fairness	5 (1)
Autonomy	8 (1.6)
Responsibility	8 (1.6)
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)	
Yes	33 (6.8)
No	454 (93.2)
Q7: Which of the following contents were covered as part of the education?[b,c]	
Informed consent	10 (2)
Bias	14 (2.9)
Data privacy	17 (3.5)
Explainability	20 (4.1)
Safety (of artificial intelligence-based applications)	18 (3.7)
Fairness	12 (2.5)
Autonomy	14 (2.9)
Responsibility	19 (3.9)

[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."

XSL•FO
RenderX

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[a]-based (chat) applications of participants (n=487).

Question	Participants, n (%)
Q8: Have you already used an AI-based (chat) application such as ChatGPT (OpenAI), Bard (Google), Bing chat, or Jasper Chat?	
Yes	189 (38.8)
No	298 (61.2)
Q9: Have you knowingly ever used AI-based medical applications, such as image-based diagnostic tools in radiology?	
Yes	49 (10.1)
No	438 (89.9)
Q10: Are you interested in using an AI application as part of your medical studies in the future?[b]; n=298	
Yes	229 (76.9)
No	69 (23.1)
Q1: Approximately how many hours have you used the AI-based (chat) application in the last week (7 d)[c]; (n=189)	
<1 h	73 (38.6)
1-3 h	91 (48.2)
4-6 h	19 (10)
7-9 h	3 (1.6)
10-12 h	2 (1.1)
>12 h	1 (0.5)
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)	
Yes	138 (73)
No	51 (26.7)
Q13: For which of the following objectives have you already used the AI-based (chat) application in the medical context?[e]; (n=138)	
Therapy suggestions	18 (13)
Querying medical knowledge	74 (53.6)
Diagnostic support	5 (3.6)
Explanation of pathologies	41 (29.7)

[a]AI: artificial intelligence.

[b]Question 10 was exclusively displayed to participants who responded to questions 8 and 9 with “no.”

[c]Question 11 was exclusively displayed to participants who responded to question 8 with “yes.”

[d]Question 12 was exclusively displayed to participants who responded to question 8 with “yes.”

[e]Question 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

XSL•FO
RenderX

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.

XSL•FO
RenderX

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).

Statement and subgroup	Scores, median (IQR)	Scores, mode	*P* value	*Z* score
The use of AI in medicine will...				
S1: ...positively change medicine			.003	−2.990
Subgroup 1: previous use of AI	4 (3.75-4.25)	4		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S2: ...find useful applications in medicine			.002	3.101
Subgroup 1: previous use of AI	4 (3.5-4.5)	4		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S3: ...influence the choice of my medical specialization			.52	−1.474
Subgroup 1: previous use of AI	3 (2-4)	2		
Subgroup 2: no previous use of AI	3 (2-4)	2		
S4: ...reduce the number of jobs for medical staff			.09	−1.707
Subgroup 1: previous use of AI	3 (3-5)	4		
Subgroup 2: no previous use of AI	3 (2-4)	2		
S5: ...improve the quality of patient care			<.001	3.570
Subgroup 1: previous use of AI	4 (0)	4		
Subgroup 2: no previous use of AI	4 (3.5-4.5)	4		
S6: ...improve the process of diagnosis			.002	3.089
Subgroup 1: previous use of AI	4 (3.5-4.5)	4		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S7: ...improve the process of therapy selection			<.001	3.865
Subgroup 1: previous use of AI	4 (0-0)	4		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S8: ...negatively affect the doctor-patient relationship			.18	1.328
Subgroup 1: previous use of AI	3 (2-4)	2		
Subgroup 2: no previous use of AI	3 (2-4)	3		
S9: ...lead to a dehumanization of medicine			.11	1.610
Subgroup 1: previous use of AI	3 (2-4)	2		
Subgroup 2: no previous use of AI	3 (2-4)	3		
S10: ...negatively affect patient autonomy			.05	2.040
Subgroup 1: previous use of AI	3 (2-3)	2		
Subgroup 2: no previous use of AI	3 (2-4)	3		
S11: ...negatively affect the autonomy of medical staff			.16	1.415
Subgroup 1: previous use of AI	3 (2-4)	2		
Subgroup 2: no previous use of AI	3 (2-4)	3		
S12: ...bring new ethical challenges			.02	2.302
Subgroup 1: previous use of AI	5 (4-5)	5		
Subgroup 2: no previous use of AI	4 (3-4)	4		

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

XSL•FO
RenderX

(88%) with the assertion that AI instruction in medical education is currently sufficient (S14), with no statistically significant difference between the 2 groups. No significant statistical differences were observed for statements S15-S19. There was an overall agreement that the teaching of AI should include practical content (S15; 417/487, 86%), be based on case studies and application scenarios in medicine (S16; 342/487, 70.3%), be an important prerequisite for medical practice (S17; 314/487, 64.9%), be available to medical staff even after graduation (S18; 376/487, 77.3%), and be updated regularly to reflect advances in AI technology (S19; 407/487, 83.6%). There was a significant measurable difference in the S20 (*P*=.002) between the 2 groups. The *z* score indicates a stronger agreement with the statement "AI instruction is of interest to me" among the group of medical students who previously used AI-based (chat) applications (*z* score: 3.173). The statistical analysis is presented in Table 5, and an overview of the perceptions of the participants regarding the teaching of AI in medicine can be found in Table S2 in Multimedia Appendix 1.

Table 5. Statistical analysis of the perceptions of medical students regarding the teaching of artificial intelligence (AI)–based (chat) applications such as ChatGPT (OpenAI), Bard (Google), Bing Chat (Microsoft Inc), and Jasper Chat (Jasper AI, Inc) in medical education (n=487).

Statement and subgroup	Scores, median (IQR)	Scores, mode	*P* value	*Z* score
The teaching of AI...				
S13: ...should be part of medical education			.02	2.381
Subgroup 1: previous use of AI	5 (4-5)	5		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S14: ...in medical education is adequate			.90	0.128
Subgroup 1: previous use of AI	2 (1-2)	1		
Subgroup 2: no previous use of AI	2 (1-2)	1		
S15: ...should include practical content (e.g., exercises to apply AI) in addition to theoretical aspects			.18	−2.358
Subgroup 1: previous use of AI	4 (3.5-4.5)	4		
Subgroup 2: no previous use of AI	4 (0)	4		
S16: ...should be based on case studies and application scenarios of AI in medicine			.53	−0.625
Subgroup 1: previous use of AI	4 (3-5)	4		
Subgroup 2: no previous use of AI	4 (3.5-4.5)	4		
S17: ...is an important prerequisite for medical practice			.16	1.417
Subgroup 1: previous use of AI	4 (3.5-4.5)	4		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S18: ...should be available for medical staff even after graduation			.13	−1.527
Subgroup 1: previous use of AI	4 (3.5-4.5)	4		
Subgroup 2: no previous use of AI	4 (3-5)	4		
S19: ...should be updated regularly to reflect advances in AI technology			.34	−2.121
Subgroup 1: previous use of AI	4 (3-4)	4		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S20: ...is of interest to me			.002	3.173
Subgroup 1: previous use of AI	4 (4-5)	4		
Subgroup 2: no previous use of AI	4 (0)	4		

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

XSL•FO
RenderX

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).

Statement and subgroup	Scores, median (IQR)	Scores, mode	*P* value	*Z* score
The teaching of AI...				
S13: ...should be part of medical education			.37	−0.903
Subgroup 1: previous use of AI	5 (4-5)	5		
Subgroup 2: no previous use of AI	4 (4-5)	5		
S14: ...in medical education is adequate			.21	−1.263
Subgroup 1: previous use of AI	2 (2-3)	2		
Subgroup 2: no previous use of AI	2 (1-2)	1		
S15: ...should include practical content (e.g., exercises to apply AI) in addition to theoretical aspects			.80	−0.254
Subgroup 1: previous use of AI	4 (0)	4		
Subgroup 2: no previous use of AI	4 (0)	4		
S16: ...should be based on case studies and application scenarios of AI in medicine			.48	−0.707
Subgroup 1: previous use of AI	4 (3-4)	4		
Subgroup 2: no previous use of AI	4 (2.5-4.5)	4		
S17: ...is an important prerequisite for medical practice			.90	0.118
Subgroup 1: previous use of AI	4 (3-4)	4		
Subgroup 2: no previous use of AI	4 (2-4)	4		
S18: ...should be available for medical staff even after graduation			.17	−1.359
Subgroup 1: previous use of AI	4 (3-4)	4		
Subgroup 2: no previous use of AI	4 (2-4)	4		
S19: ...should be updated regularly to reflect advances in AI technology			.17	−1.381
Subgroup 1: previous use of AI	4 (3-4)	4		
Subgroup 2: no previous use of AI	4 (3-4)	4		
S20: ...is of interest to me			.005	1.782
Subgroup 1: previous use of AI	4 (3-4)	4		
Subgroup 2: no previous use of AI	4 (0)	4		

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.

XSL•FO
RenderX

Table 7. Statistical analysis of the relevance of artificial intelligence (AI)–based (chat) applications such as ChatGPT (OpenAI), Bard (Google), Bing Chat (Microsoft Inc), and Jasper Chat (Jasper AI, Inc) ethics teaching contents according to the participating medical students (n=487).

AI ethics teaching content and subgroup	Scores, median (IQR)	Scores, mode	*P* value	*Z* score
TC1: informed consent			.04	2.018
Subgroup 1: previous use of AI	4 (4-5)	5		
Subgroup 2: no previous use of AI	4 (3-4)	4		
TC2: bias			.22	−1.215
Subgroup 1: previous use of AI	4 (4-5)	5		
Subgroup 2: no previous use of AI	4 (3-4)	4		
TC3: data privacy			.78	0.283
Subgroup 1: previous use of AI	4 (4-5)	5		
Subgroup 2: no previous use of AI	4 (4-5)	5		
TC4: explainability			.36	−0.911
Subgroup 1: previous use of AI	4 (4-5)	5		
Subgroup 2: no previous use of AI	4 (3.5-4.5)	4		
TC5: safety			.57	0.565
Subgroup 1: previous use of AI	5 (4-5)	5		
Subgroup 2: no previous use of AI	5 (4-5)	5		
TC6: fairness			.96	−0.048
Subgroup 1: previous use of AI	4 (4-5)	5		
Subgroup 2: no previous use of AI	4 (4-5)	5		
TC7: autonomy			.11	1.594
Subgroup 1: previous use of AI	4 (4-5)	5		
Subgroup 2: no previous use of AI	4 (4-5)	5		
TC8: responsibility			.22	−1.215
Subgroup 1: previous use of AI	5 (4-5)	5		
Subgroup 2: no previous use of AI	5 (4-5)	5		

Additional Analysis of the Collected Data

To analyze whether there is a difference in education regarding AI and AI ethics among Germany, Austria, and Switzerland, we conducted an additional evaluation of the collected data. For this supplementary analysis, we analyzed the responses to Q2: "Have you already received education in the field of artificial intelligence within your regular medical studies? (eg, as part of medical statistics or informatics)," and Q4: "Have you already received education in the field of AI ethics within your regular medical studies? (eg, as part of the History, Ethics, and Theory of Medicine course)." Using the chi-square test of independence, we sought to determine whether the distribution of answers varied significantly among these countries. In the comparison between the 3 countries concerning education in the field of AI, the chi-square test of independence indicated no significant difference in the distribution of the responses. Of the 487 respondents, only 26 (5.3%) indicated that they had previously received AI education. The test yielded a result of χ^2_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 χ^2_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

XSL•FO
RenderX

the PCS and CS participants. Most other statements concerning AI's impact on medicine (S1-2; S4-12) did not demonstrate statistical significance. Similarly, no significant difference was found for statements related to AI teaching (S13-20) across the study stages (PCS, CS, and CPS). When considering the teaching of AI ethics, differences were evident in the belief that AI ethics should be integrated into medical education (S21; *P*=.003) and that the current teaching of AI ethics is adequate (S22; *P*=.02). Upon further analysis, CS participants showed stronger agreement than PCS participants, with no difference when compared with CPS participants. Finally, for the specific content of AI ethics teaching, none of the statements reflected significant statistical variation across the study stages. An overview of the statistical differences is provided in Tables S5-S8 in Multimedia Appendix 1.

Ethics Education Background

To explore the potential impact of prior ethics education on survey outcomes, particularly in parts 3 to 6, we compared 2 distinct groups: those with prior ethics education and those without. On the use of AI in medicine, one statistical difference could be determined for the statement that "...negatively affect the autonomy of medical staff' (S11, *P*=.002). The *z* score suggested a stronger level of agreement with the statement in the group that had received prior ethics education (*z* score: 2.876). For the other statements of the third part of the survey (S1-10; S12), no statistical difference could be determined. No statistical difference could be determined for the fourth part of the survey on AI teaching (S13-20). Regarding the teaching of AI ethics, statistical differences could be determined for 2 statements (S21, *P*=.004; S22, *P*=.03). For the statement that the teaching of AI ethics should be part of medical education, the *z* score indicated a higher level of agreement in the group that had received prior ethics education. Similarly, a higher level of disagreement was indicated by the group with prior ethics education for the statement that the teaching of AI ethics in medical education is adequate (*z* score: −3.011). There was no statistically significant difference in the AI ethics teaching content between the groups. A detailed statistical analysis can be found in Tables S9-S12 in Multimedia Appendix 1.

Discussion

This discussion aims to comprehensively analyze the findings regarding medical students' perceptions of AI in medicine and the role of AI and AI ethics in their medical education, depending on their use of AI-based (chat) applications such as ChatGPT.

The Use of AI-Based (Chat) Application Among the Surveyed Medical Students

The discrepancy between students' personal AI experiences and formal medical education highlights the gap in integrating AI into curricula, reflecting the need for educational progress in line with technological advancement. A considerable 38.8% of the respondents reported prior use of AI-based (chat) applications, such as ChatGPT, Bard, Bing Chat, or Jasper Chat, which was slightly below the percentage received from pretesting and used for sample size calculation (5/11, 45%). The results concerning the reported use of AI-based (chat) applications must be evaluated in the context of the timing of the data collection. ChatGPT, for instance, became freely available to the public on November 30, 2022, making it accessible for only approximately 8 months at the time of data collection [27]. In addition, Bing Chat was not broadly accessible until May 2023, further constraining its availability before the survey [28]. It is noteworthy that academic literature 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].

XSL•FO
RenderX

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,

XSL•FO
RenderX

including teaching autonomy in a medical context, might lead to a more nuanced understanding of the subject and potential implications of AI. The results of the analysis reinforce the need for ongoing ethics education, not just as a separate entity, but also interwoven with AI-related topics, to enhance students' comprehensive understanding of the ethical implications of AI in medicine. The significant influence of prior ethics education on shaping students' perceptions of the role of AI in medicine emphasizes the interaction between ethical training and technological awareness. The nuanced understanding of the ethical implications of AI among students who have received ethics education underscores the importance of such training in developing critical thinking about the impact of AI in health care. Integrating ethics education with AI teaching content could foster a more holistic approach, preparing students not only for the technological aspects of AI but also for its ethical and societal implications [37].

Limitations

Despite the strengths of this study, some limitations must be acknowledged. First, our web-based survey could introduce selection bias, as tech-savvy students may be more likely to participate. Second, the survey measured students' perceptions rather than their actual competencies in AI and ethics. In addition, although estimated, the response rate was suboptimal, which may limit the generalizability of our findings. Geographically, our sample was limited to German-speaking countries, making the translation of these results to other countries with different health care systems and medical educational frameworks difficult. Cultural attitudes toward AI could also vary, possibly influencing students' perceptions of and engagement with AI. Our study is essentially a snapshot of a rapidly evolving field; hence, our findings may not reflect attitudes and competencies, as they evolve with advancements in AI technology. In our analysis, we observed statistically significant differences based on prior ethics education and study stage. However, although the additional analysis of the data did not show a direct overlap with significant findings between the main and supplementary evaluations, additional tests are needed to determine whether these factors acted as confounders in our main data analysis. Although this study considered specific potential confounders, it is worth noting that other confounding variables may exist and were not analyzed in this study. Finally, owing to the self-reported nature of the data, the responses might be subject to recall bias, misunderstanding of questions, or social desirability bias. Although our findings provide valuable insights into the state of AI education in German-speaking medical schools, broader multinational studies would offer a more comprehensive understanding.

Conclusions

This study provides a valuable understanding of the perceptions and experiences of medical students in Germany, Austria, and Switzerland regarding the application of AI in medicine, and its role in medical education. Our findings clearly indicate a discrepancy between students' interactions with AI-based chat applications such as ChatGPT and the representation of AI in their formal education. Despite a significant number of students interacting with AI technology, notably AI-based chat applications, only a fraction have received any formal AI education, revealing a substantial gap in the current medical curricula. This highlights the necessity of the evolution of medical curricula to incorporate AI and AI ethics education, ensuring that future medical professionals are adequately equipped to navigate the challenges and opportunities presented by AI in medicine.

Furthermore, our findings indicate that practical engagement with AI technology can contribute to an increased awareness of ethical implications, reinforcing the importance of including hands-on AI experiences in medical education. It is evident that the rapid advancement and application of AI in medicine 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.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comprehensive statistical analysis and evaluation of confounding factors regarding medical students' perceptions of artificial intelligence's role in medicine and medical education.

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

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med 2019 Jul;143(7):859-868 [FREE Full text] [doi: 10.5858/arpa.2018-0147-OA] [Medline: 30295070]

XSL•FO
RenderX

3. Lehman CD, Yala A, Schuster T, Dontchos B, Bahl M, Swanson K, et al. Mammographic breast density assessment using deep learning: clinical implementation. Radiology 2019 Jan;290(1):52-58. [doi: 10.1148/radiol.2018180694] [Medline: 30325282]
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
5. Haenlein M, Kaplan A. A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. Calif Manage Rev 2019 Jul 17;61(4):5-14. [doi: 10.1177/0008125619864925]
6. Doshi RH, Bajaj SS, Krumholz HM. ChatGPT: temptations of progress. Am J Bioeth 2023 Apr 28;23(4):6-8. [doi: 10.1080/15265161.2023.2180110] [Medline: 36853242]
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
9. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 2023 Jun;104(6):269-274. [doi: 10.1016/j.diii.2023.02.003] [Medline: 36858933]
10. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol 2023 Sep 4;13:1268915 [FREE Full text] [doi: 10.3389/fonc.2023.1268915] [Medline: 37731643]
11. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023 Dec 21;28(1):2181052 [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
12. Wang P. On defining artificial intelligence. J Artif Gen Intell 2019;10(2):1-37 [FREE Full text] [doi: 10.2478/jagi-2019-0002]
13. Russell S, Norvig P. Artificial Intelligence: A Modern Approach, 4th Edition. London, UK: Pearson Education; 2022.
14. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics. Cham, Switzerland: Springer; 2021.
15. Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. MedEdPublish 2021;10:75. [doi: 10.15694/mep.2021.000075.1]
16. Masters K. Artificial intelligence in medical education. Med Teach 2019 Apr 21;41(9):976-980. [doi: 10.1080/0142159x.2019.1595557]
17. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
18. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021 Nov 01;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
19. Jha N, Shankar PR, Al-Betar MA, Mukhia R, Hada K, Palaian S. Undergraduate medical students' and interns' knowledge and perception of artificial intelligence in medicine. Adv Med Educ Pract 2022 Aug;13:927-937 [FREE Full text] [doi: 10.2147/AMEP.S368519] [Medline: 36039185]
20. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 09;22(1):772 [FREE Full text] [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
21. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020 Jun 19;3(1):86 [FREE Full text] [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
22. Pinto Dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr 6;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
23. Cochran WG. Sampling Techniques. Hoboken, NJ: John Wiley & Sons; 1977.
24. Bericht der Republik Österreich über die Situation in Studien mit Zulassungsverfahren. Bundesministerium für Bildung, Wissenschaft und Forschung. 2021 Oct. URL: https://pubshop.bmbwf.gv.at/index.php?article_id=9&type=neuerscheinungen&pub=964 [accessed 2023-10-25]
25. Studierende insgesamt und Studierende Deutsche im Studienfach Medizin (Allgemein-Medizin) nach Geschlecht. Destatis Statistisches Bundesamt. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/lrbil05.html [accessed 2023-10-25]
26. Studierende an den universitären hochschulen nach jahr, fachrichtung, geschlecht und hochschule. Bundesamt für Statistik. 2023 Mar 28. URL: https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-wissenschaft/personen-ausbildung/tertiaerstufe-hochschulen/universitaere.assetdetail.24367582.html [accessed 2023-10-25]
27. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-10-25]

XSL•FO
RenderX

28. Mehdi Y. Announcing the next wave of AI innovation with Microsoft Bing and Edge. Microsoft. 2023 May 4. URL: https://blogs.microsoft.com/blog/2023/05/04/announcing-the-next-wave-of-ai-innovation-with-microsoft-bing-and-edge/ [accessed 2023-10-25]
29. Sallam M, Salim NA, Barakat M, Al-Mahzoum K, Al-Tammemi AB, Malaeb D, et al. Assessing health students' attitudes and usage of ChatGPT in Jordan: validation study. JMIR Med Educ 2023 Sep 05;9:e48254 [FREE Full text] [doi: 10.2196/48254] [Medline: 37578934]
30. von Garrel J, Mayer J, Mühlfeld M. Künstliche Intelligenz im Studium Eine quantitative Befragung von Studierenden zur Nutzung von ChatGPT & Co. Hochschule Darmstadt. 2023 Jun 28. URL: https://opus4.kobv.de/opus4-h-da/frontdoor/index/index/docId/395 [accessed 2023-12-13]
31. Mosch L, Back A, Balzer F, Bernd M, Brandt J, Erkens S, et al. Lernangebote zu künstlicher intelligenz in der medizin (German). Zenodo. 2021. URL: https://doi.org/10.5281/zenodo.5497668 [accessed 2023-12-13]
32. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023 Dec;28(1):2220920 [FREE Full text] [doi: 10.1080/10872981.2023.2220920] [Medline: 37307503]
33. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
34. McLennan S, Meyer A, Schreyer K, Buyx A. German medical students´ views regarding artificial intelligence in medicine: a cross-sectional survey. PLOS Digit Health 2022 Oct 4;1(10):e0000114 [FREE Full text] [doi: 10.1371/journal.pdig.0000114] [Medline: 36812635]
35. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the national academy of medicine. JAMA 2020 Feb 11;323(6):509-510. [doi: 10.1001/jama.2019.21579] [Medline: 31845963]
36. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 2019 Mar 25;21(3):e12422 [FREE Full text] [doi: 10.2196/12422] [Medline: 30907742]
37. Quinn TP, Coghlan S. Readying medical students for medical AI: the need to embed AI ethics education. arXiv Preprint posted online September 7, 2021. [FREE Full text] [doi: 10.48550/arXiv.2109.02866]
38. Katznelson G, Gerke S. The need for health AI ethics in medical school education. Adv Health Sci Educ Theory Pract 2021 Oct;26(4):1447-1458. [doi: 10.1007/s10459-021-10040-3] [Medline: 33655433]
39. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. Perspect Med Educ 2023 Oct 16;12(1):399-410 [FREE Full text] [doi: 10.5334/pme.954] [Medline: 37868075]
40. Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 2018 Oct;9(5):745-753 [FREE Full text] [doi: 10.1007/s13244-018-0645-y] [Medline: 30112675]
41. Taber S, Akdemir N, Gorman L, van Zanten M, Frank JR. A "fit for purpose" framework for medical education accreditation system design. BMC Med Educ 2020 Sep 28;20(Suppl 1):306 [FREE Full text] [doi: 10.1186/s12909-020-02122-4] [Medline: 32981517]

Abbreviations

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

Edited by G Eysenbach, K Venkatesh; submitted 26.07.23; peer-reviewed by L Ursic, J Wilkinson; comments to author 20.10.23; revised version received 26.10.23; accepted 02.12.23; published 05.01.24.

Please cite as:
Weidener L, Fischer M
Artificial Intelligence in Medicine: Cross-Sectional Study Among Medical Students on Application, Education, and Ethical Aspects
JMIR Med Educ 2024;10:e51247
URL: https://mededu.jmir.org/2024/1/e51247
doi:10.2196/51247
PMID:38180787

©Lukas Weidener, Michael Fischer. Originally published in JMIR Medical Education (https://mededu.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

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.

XSL•FO
RenderX

Original Paper

Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study

Amanda L Zaleski[1,2], MSc, PhD; Rachel Berkowsky[3], MSc; Kelly Jean Thomas Craig[1], PhD; Linda S Pescatello[3], MSc, PhD

[1]Clinical Evidence Development, Aetna Medical Affairs, CVS Health Corporation, Hartford, CT, United States

[2]Department of Preventive Cardiology, Hartford Hospital, Hartford, CT, United States

[3]Department of Kinesiology, University of Connecticut, Storrs, CT, United States

Corresponding Author:
Amanda L Zaleski, MSc, PhD
Clinical Evidence Development
Aetna Medical Affairs
CVS Health Corporation
151 Farmington Avenue
Hartford, CT, 06156
United States
Phone: 1 8605385003
Email: zaleskia@aetna.com

Abstract

Background: Regular physical activity is critical for health and disease prevention. Yet, health care providers and patients face barriers to implement evidence-based lifestyle recommendations. The potential to augment care with the increased availability of artificial intelligence (AI) technologies is limitless; however, the suitability of AI-generated exercise recommendations has yet to be explored.

Objective: The purpose of this study was to assess the comprehensiveness, accuracy, and readability of individualized exercise recommendations generated by a novel AI chatbot.

Methods: A coding scheme was developed to score AI-generated exercise recommendations across ten categories informed by gold-standard exercise recommendations, including (1) health condition–specific benefits of exercise, (2) exercise preparticipation health screening, (3) frequency, (4) intensity, (5) time, (6) type, (7) volume, (8) progression, (9) special considerations, and (10) references to the primary literature. The AI chatbot was prompted to provide individualized exercise recommendations for 26 clinical populations using an open-source application programming interface. Two independent reviewers coded AI-generated content for each category and calculated comprehensiveness (%) and factual accuracy (%) on a scale of 0%-100%. Readability was assessed using the Flesch-Kincaid formula. Qualitative analysis identified and categorized themes from AI-generated output.

Results: AI-generated exercise recommendations were 41.2% (107/260) comprehensive and 90.7% (146/161) accurate, with the majority (8/15, 53%) of inaccuracy related to the need for exercise preparticipation medical clearance. Average readability level of AI-generated exercise recommendations was at the college level (mean 13.7, SD 1.7), with an average Flesch reading ease score of 31.1 (SD 7.7). Several recurring themes and observations of AI-generated output included concern for liability and safety, preference for aerobic exercise, and potential bias and direct discrimination against certain age-based populations and individuals with disabilities.

Conclusions: There were notable gaps in the comprehensiveness, accuracy, and readability of AI-generated exercise recommendations. Exercise and health care professionals should be aware of these limitations when using and endorsing AI-based technologies as a tool to support lifestyle change involving exercise.

(JMIR Med Educ 2024;10:e51308) doi:10.2196/51308

KEYWORDS

exercise prescription; health literacy; large language model; patient education; artificial intelligence; AI; chatbot

XSL•FO
RenderX

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.

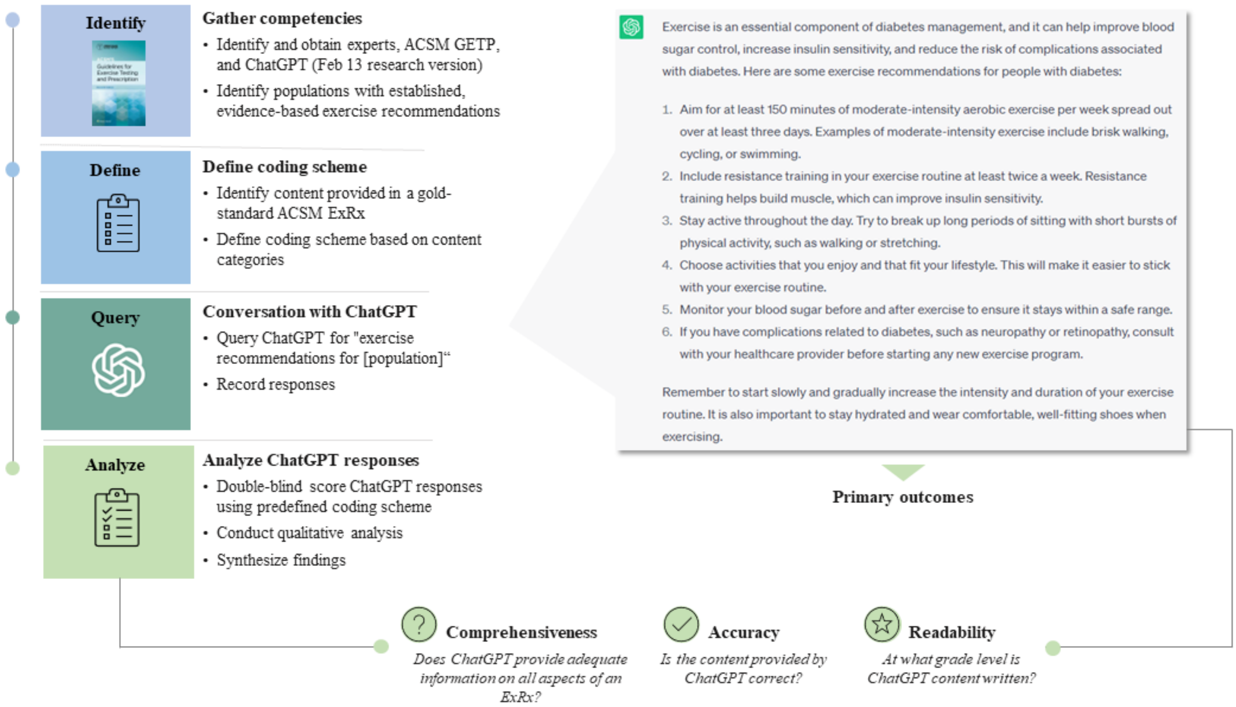
Methods

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.

XSL•FO
RenderX

Figure 1. Conceptual study overview. ACSM: American College of Sports Medicine; ExRx: exercise prescription; GETP: Guidelines for Exercise Testing and Prescription.



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.

ChatGPT Prompt Specificity and Structure

Prompt methodology was developed a priori with the overarching goal to observe ChatGPT's unaltered performance in a real-world setting while controlling for factors known to influence output, including prompt structure, evaluation timeframe, model version, and model feedback.

A single researcher (ALZ) posed 26 separate, open-ended prompts as a new chat session to the ChatGPT bot (version 3.5) on the same day in a single session. Each text prompt was framed to the ChatGPT bot in a standardized, neutral, third-person tense format as "exercise recommendations for [population]" to optimize the relevance of AI responses for both health care provider and patient scenarios. Generated ChatGPT bot responses were abstracted from the interface and converted into plain text format using Microsoft Word (version 2208; Microsoft Corp) on the same day. Content was unaltered upon conversion to plain text format (Multimedia Appendix 1). Note that the ChatGPT bot used in this study was not subjected to retraining or correction during these prompt interactions. The rationale for this methodological decision was to enable the natural observation of ChatGPT's raw performance and provide a transparent evaluation of its inherent capabilities [33,34].

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?*),

XSL•FO
RenderX

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

XSL•FO
RenderX

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).

Content	Exercise recommendations reporting content		
	Fully provided, n (%)	Partial[a], n (%)	Not provided, n (%)
Condition-specific benefits	24 (92)	0 (0)	2 (8)
Preparticipation screening	24 (92)	0 (0)	2 (8)
Frequency	2 (8)	9 (35)	15 (58)
Intensity	2 (8)	15 (58)	9 (35)
Time	1 (4)	10 (38)	15 (58)
Type	14 (54)	12 (46)	0 (0)
Volume	0 (0)	8 (31)	18 (69)
Progression	15 (58)	0 (0)	11 (42)
Special considerations	23 (88)	0 (0)	3 (12)
References	2 (8)	0 (0)	24 (92)

[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).

XSL•FO
RenderX

Table 2. Summary of inaccurate content among AI[a]-generated exercise recommendations for all reported content categories (N=161).

Content category (count/reported) and population	AI-generated content	Ground truth
Condition-specific benefits (1/24)		
HIV	“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.”	Research regarding the impact of exercise to improve immune function is inconclusive [38].
Preparticipation screening (8/24)		
Healthy adult	“Remember to always listen to your body and consult with a healthcare professional before starting a new exercise routine.”	ACSM[b] 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].
Older adult	“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.”	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].
Dyslipidemia	“It's important to consult with a healthcare professional before starting a new exercise program ...”	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].
Hypertension	“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.”	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].
Arthritis	“Be sure to talk to your doctor or a physical therapist before starting any new exercise program ...”	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].
Cancer survivors	“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.”	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].
Multiple sclerosis	“It is always recommended to consult with a healthcare professional before starting any exercise program.”	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].
Osteoporosis	“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.”	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].
Frequency (2/11)		
Fibromyalgia	“Aim for at least 30 minutes of aerobic exercise most days of the week.”	ACSM recommends an initial frequency of 1-2 days per week, gradually progressing to 2-3 days per week [1].

XSL•FO
RenderX

Content category (count/reported) and population	AI-generated content	Ground truth
Alzheimer disease	"Engage in moderate aerobic exercise such as brisk walking, cycling, or swimming for at least 30 minutes a day, five days a week."	ACSM recommends a frequency of 3 days per week [1].
Intensty (1/17)		
Hypertension	"Avoid high-intensity exercises: Avoid high-intensity exercises that can cause sudden increases in blood pressure, such as sprinting or heavy lifting."	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].
Time (1/11)		
Fibromyalgia	"Start with 1-2 sets of 10-15 repetitions for each exercise and gradually increase the resistance as tolerated."	ACSM recommends gradual progression of 4-5 to 8-12 repetitions and increasing from 1 to 2-4 sets per muscle group [1].
Type (1/26)		
Hypertension	"Aim for at least 30 minutes of moderate-intensity aerobic exercise most days of the week."	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].
Volume (0/8)		
N/A[c]	N/A	N/A
Progression (1/15)		
Overweight and obesity	"If you're new to exercise, start with low-intensity activities such as walking or swimming, and gradually increase your intensity and duration."	ACSM recommends initial intensity should be moderate, progressing to vigorous for greater health benefits [1].
Special considerations (0/23)		
N/A	N/A	N/A
References (0/2)		
N/A	N/A	N/A

[a]AI: artificial intelligence.

[b]ACSM: American College of Sports Medicine.

[c]N/A: not applicable.

Readability Metrics

Average and individual readability metrics and word count for AI-generated exercise recommendations are provided in Table 3. On average, AI-generated output was 259.3 (SD 49.1) words (range 171-354) and considered "difficult to read" with an average Flesch reading ease of 31.1 (SD 7.7; range 14.5-47.3) and written at a college-level (mean 13.7, SD 1.7; range 10.1-18.0).

XSL•FO
RenderX

Table 3. Readability metrics for artificial intelligence–generated exercise recommendations by population.

Population	Word count	Flesch reading ease	Grade level
Healthy adults	187	14.5	15.2
Children and adolescents	253	29.8	14.1
Pregnancy	267	34.7	13.5
Older adults	276	37.0	12.2
Cardiovascular disease	271	33.6	13.2
Heart failure	235	23.0	16.2
Heart transplant	278	24.9	14.4
Peripheral artery disease	322	32.4	13.4
Cerebrovascular accident	346	22.0	15.1
Asthma	317	41.1	12.0
COPD[a]	247	47.3	10.1
Diabetes	201	36.7	11.8
Dyslipidemia	291	19.6	15.9
Hypertension	247	34.5	13.3
Overweight and obesity	200	34.7	13.2
Arthritis	236	38.4	13.0
Cancer	319	24.8	14.9
Fibromyalgia	303	40.0	12.2
HIV	232	30.0	13.9
Kidney disease	354	31.1	15.3
Multiple sclerosis	255	38.4	11.4
Osteoporosis	171	32.7	12.3
Spinal cord injury	281	25.5	14.1
Alzheimer disease	191	29.1	14.8
Intellectual disability	241	32.1	13.2
Parkinson disease	221	19.8	18.0
Mean (SD)	259.3 (49.1)	31.1 (7.7)	13.7 (1.7)

[a]COPD: chronic obstructive pulmonary disease.

Qualitative Analysis

A secondary aim of this study was to identify potential patterns, consistencies, and gaps in AI-generated exercise recommendation text outputs. Major observations derived from qualitative evaluation of AI-generated exercise recommendations can be found in Multimedia Appendix 2. Briefly, several recurring themes emerged among the total sample, including liability and safety, preference for aerobic exercise, and inconsistencies in the terminology used for exercise professionals. Importantly, AI-generated output showed potential bias and discrimination against certain age-based populations and individuals with disabilities. The implications of these findings are discussed in detail below.

Discussion

Principal Findings

This study sought to explore the suitability of AI-generated exercise recommendations using a popular generative AI platform, ChatGPT. Given the recent launch and popularity of ChatGPT and other similar generative AI platforms, the overall goal was to formally appraise the suitability and readability of AI-generated output likely to be seen by patients and inform exercise and health care professionals and other stakeholders on the potential benefits and limitations of using AI to leverage for patient education. The major findings were that AI-generated output (1) presented 41.2% (107/260) of the content provided in a gold-standard exercise recommendation indicating poor comprehensiveness; (2) of the content provided, chat output was 90.7% (146/161) accurate with most discordance related

XSL•FO
RenderX

to the need for exercise preparticipation health screening; and (3) had college-level readability.

The results of this study are consistent with a recently published research letter that evaluated the appropriateness of CVD prevention recommendations from ChatGPT [40]. Sarraju et al [40] developed 25 questions on fundamental heart disease concepts, posed them to the AI interface, and subjectively graded responses as "appropriate" or "inappropriate." AI-generated responses were deemed to be 84% appropriate with noted misinformation provided for questions surrounding ideal exercise volume and type for 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 triggering an unnecessary number of adults for medical evaluation, both posing as potential barriers to exercise adoption [41,42].

The ACSM preparticipation screening guidelines emphasize the public health message that exercise is important for all individuals and that the preparticipation health screening should not be a deterrent to exercise participation [41]. The preparticipation screening algorithm considers current physical activity levels, desired exercise intensity, and the presence of known or underlying CVD, metabolic, and renal disease. Following this algorithm, lesser than 3% of the general population would be referred before beginning vigorous exercise, and approximately 54% would be referred before beginning any exercise [42]. Interestingly, exercise professionals are well-equipped to facilitate preparticipation screening, yet AI-generated output disproportionately emphasized medical clearance by a health care provider or doctor prior to working with an exercise professional. In reference to exercise professionals, ChatGPT used varying and incorrect terminology such as "licensed exercise physiologist" that does not reflect current-state credentialing for exercise professionals working with clinical populations (ie, ACSM Certified Clinical Exercise Physiologist [43]). These findings corroborate with existing challenges in the public health's understanding of the role of exercise professionals, levels of qualification, and respective scope of practice [44].

As AI-based technologies continue to evolve, striking the right balance between medical precision and risk mitigation remains a crucial consideration [45]. The question of how definitive an AI-based model should be when delivering medical education is multifaceted. On the one hand, the inclination of the AI-based model toward vague or general recommendations can be seen as a responsible stance to mitigate risks. On the other hand, there is merit in AI-based models providing clear, specific, and contextual guidance that reinforces evidence-based recommendations. This approach ensures that end users receive accurate and tailored advice, which is important in the context of medical education. This tension highlights the need for continued dialogue on how AI can enhance health care while ensuring that recommendations align with the highest standards of accuracy and patient safety. These discussions will be instrumental in shaping the future of AI-augmented health care.

AI-Generated Output Least Accurate for Populations With Hypertension

Interestingly, the hypertension exercise recommendations scored the poorest (ie, highest discordance) with 57% (4/7) accuracy and misinformation surrounding the need for medical clearance and the recommended intensity and type of exercise (Table 2). For example, AI-generated output recommended avoiding high-intensity exercise "such as sprinting or heavy lifting"; however, the ACSM does not contraindicate vigorous-intensity exercise considering comorbidities and assuming adequate progression and proper technique [1]. Additionally, AI-generated output recommended a target exercise goal of "30 minutes of moderate-intensity aerobic exercise most days of the week." Notably, the ACSM guidelines reinforce that emphasis is no longer placed on aerobic exercise alone but rather recommend aerobic and resistance exercise alone or combined (ie, concurrent exercise) on most, preferably all, days of the week to total 90-150 minutes per week or more of multimodal, moderate-intensity exercise [39]. Reasons for this discordance are likely because the ChatGPT model relies on training data preceding 2021 and may not capture real-time research advancements. Nevertheless, these findings are important because hypertension is the most common, costly, and modifiable CVD risk factor with strong evidence-based and guideline-driven recommendations, whereby support of exercise is a critical component of first-line treatment for elevated blood pressure [7,46-48].

Social Determinants of Health Considerations

Not surprisingly, our evaluation of this AI-based technology identified social determinants of health considerations regarding educational obtainment for its users. Average readability of the AI-generated output was found to be very high, at the college level, which poses significant challenges for the majority of patients, as The National Institutes of Health, American Medical Association, and American Heart Association all recommend that patient education materials be written at or below a sixth-grade reading level [49] based on national educational obtainment trends. Poor readability of patient materials can exacerbate disparities in access to care for those with limited health literacy, and those individuals may experience more barriers to understand and apply the information provided [29,30]. These findings highlight the need for ongoing evaluation and refinement of AI-generated educational output to prevent inappropriate recommendations that do not improve disparities in clinical outcomes. AI-based models, such as ChatGPT, and their output are vulnerable to both poor data

XSL•FO
RenderX

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

XSL•FO
RenderX

ground truth benchmark. This methodological approach ensures that the outcomes presented in this study are grounded in a consistent and rigorously designed interaction process.

Future Directions

Given the recent development of open-source generative AI technologies, this area is ripe for exploration. However, before proceeding with extensive randomized controlled trials, it is crucial to prioritize the safety and ethical considerations associated with AI-generated medical education. As AI technologies have the potential to impact health disparities, it is essential to carefully evaluate their use to ensure inclusivity and appropriate messaging across demographics [27,52-54]. Further research is needed to develop, test, and implement AI technologies that serve individuals safely, effectively, and ethically without perpetuating bias, discrimination, or causing harm. This includes exploring ways to mitigate potential biases and discriminatory outcomes. Outside of the research setting, health care and exercise professionals can play a crucial role in improving AI-based models through prompting and by giving corrective feedback to retrain biases and inaccuracies in AI-generated responses. By enriching ChatGPT with user-specific data including exercise components, literacy level, physical limitations, and other activity considerations, there are opportunities to improve the personalization of recommendations and lessen digital discrimination. Through this stewardship, continuous refinement will likely improve the performance, usability, and appropriateness of the model, translating to superior patient outcomes, which is the goal of provider-enablement and patient-facing tools. As LLMs continue to evolve, it will become increasingly important for researchers to continuously assess improvements with response variations over time. Importantly, future work should explore the incremental value of advanced and diverse prompting considerations. Examples of prompting considerations include the provision of roles and goals (eg, "You are a Clinical Exercise Physiologist and your goal is to design a safe and effective exercise prescription to lower blood pressure"), engaging in multiple or chain prompting and specifically prompting for content commonly missing from output as identified in this study.

To ensure the responsible and safe deployment of AI technologies in health care, conducting thorough implementation studies is a logical next step. These studies should focus on measuring various factors, including acceptability, adoption, appropriateness, costs, feasibility, fidelity, penetration, and sustainability. By thoroughly investigating these implementation aspects, we can ensure that the technology is well-integrated and does not pose any harm to patients or health care systems. Following the completion of the implementation studies, it is important to assess the impact of AI-generated models on service outcomes. This includes evaluating health care quality factors such as safety, timeliness, efficiency, effectiveness, equity, and patient-centeredness [55]. Understanding how AI technologies influence these service outcomes will provide valuable insights into their overall impact on health care delivery. Additionally, measuring patient-centered and end-user outcomes is essential to evaluate the effectiveness of AI technologies in improving patient experiences and outcomes. Randomized controlled trials designed to test ChatGPT as an intervention to augment behavior change and associated health outcomes would be of great public health interest. These trials should prioritize patient-centered outcomes, including satisfaction, usability, experience, and patient activation [56]. By assessing these outcomes, we can determine the effectiveness of AI technologies in empowering patients and fostering meaningful engagement with health care providers.

Conclusions

To conclude, this study found that AI-generated exercise recommendations have moderate comprehensiveness and high accuracy when compared to a gold-standard reference source. However, there are notable gaps in content surrounding critical components of ExRx and potentially biased and discriminatory outputs. Additionally, the readability level of the recommendations may be too high for some patients, and the lack of references in AI-generated content may be a significant limitation for use. Health care providers and patients may wish to remain cautious in relying solely on AI-generated exercise recommendations and should limit their use in combination with clinical expertise and oversight.

Acknowledgments

This study was supported by the University of Connecticut, CVS Health Corporation, and Hartford Hospital.

Authors' Contributions

ALZ contributed to the study conceptualization, project management, study design, data curation and coding, statistical analysis, interpretation of the data, visual presentation of the data, and paper preparation and submission. RB contributed to the study design, data coding, interpretation of the data, and copyediting of the paper. KJTC contributed to the interpretation of the data, business leadership, and copyediting of the paper. LSP contributed to the study design, project oversight, interpretation of the data, and revising and copyediting of the paper. All authors contributed to the writing of the paper, reviewed and approved the final version of the paper, and agreed with the order of presentation of the authors.

Conflicts of Interest

ALZ and KJTC are both employed and hold stock with CVS Health Corporation. This study is an objective evaluation to better understand ChatGPT and its outputs. To the best of our knowledge, CVS Health does not currently use or endorse the use of ChatGPT for lifestyle recommendations. LSP is the sole proprietor and founder of P3-EX, LLC, which could potentially benefit

XSL•FO
RenderX

from the tool used in this research. The results of this study do not constitute endorsement by the American College of Sports Medicine.

Multimedia Appendix 1
Output from artificial intelligence–generated exercise recommendations for clinical populations (N=26).
[PDF File (Adobe PDF File), 243 KB - mededu_v10i1e51308_app1.pdf]

Multimedia Appendix 2
Summary of major themes derived from artificial intelligence–generated exercise recommendations.
[PDF File (Adobe PDF File), 116 KB - mededu_v10i1e51308_app2.pdf]

References

1. Liguori G. ACSM's Guidelines for Exercise Testing and Prescription. 11th Edition. Philadelphia, PA: Wolters Kluwer; 2021.
2. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. JAMA 2018;320(19):2020-2028 [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
3. Joseph JJ, Deedwania P, Acharya T, Aguilar D, Bhatt DL, Chyun DA, et al. Comprehensive management of cardiovascular risk factors for adults with type 2 diabetes: a scientific statement from the American Heart Association. Circulation 2022 Mar;145(9):e722-e759 [FREE Full text] [doi: 10.1161/CIR.0000000000001040] [Medline: 35000404]
4. Lloyd-Jones DM, Allen NB, Anderson CAM, Black T, Brewer LC, Foraker RE, et al. Life's essential 8: updating and enhancing the American Heart Association's construct of cardiovascular health: a presidential advisory from the American Heart Association. Circulation 2022 Aug 02;146(5):e18-e43 [FREE Full text] [doi: 10.1161/CIR.0000000000001078] [Medline: 35766027]
5. Pedersen BK, Saltin B. Exercise as medicine—evidence for prescribing exercise as therapy in 26 different chronic diseases. Scand J Med Sci Sports 2015 Dec;25(Suppl 3):1-72 [FREE Full text] [doi: 10.1111/sms.12581] [Medline: 26606383]
6. Pescatello LS, Buchner DM, Jakicic JM, Powell KE, Kraus WE, Bloodgood B, et al. Physical activity to prevent and treat hypertension: a systematic review. Med Sci Sports Exerc 2019 Jun;51(6):1314-1323 [FREE Full text] [doi: 10.1249/MSS.0000000000001943] [Medline: 31095088]
7. Barone Gibbs B, Hivert MF, Jerome GJ, Kraus WE, Rosenkranz SK, Schorr EN, et al. Physical activity as a critical component of first-line treatment for elevated blood pressure or cholesterol: who, what, and how?: A scientific statement from the American Heart Association. Hypertension 2021 Aug;78(2):e26-e37 [FREE Full text] [doi: 10.1161/HYP.0000000000000196] [Medline: 34074137]
8. Exercise is medicine. ACSM's Rx for health. American College of Sports Medicine. 2021. URL: https://www.exerciseismedicine.org/ [accessed 2023-05-01]
9. O'Brien MW, Shields CA, Oh PI, Fowles JR. Health care provider confidence and exercise prescription practices of Exercise is Medicine Canada workshop attendees. Appl Physiol Nutr Metab 2017 Apr;42(4):384-390 [FREE Full text] [doi: 10.1139/apnm-2016-0413] [Medline: 28177736]
10. Fowles JR, O'Brien MW, Solmundson K, Oh PI, Shields CA. Exercise is Medicine Canada physical activity counselling and exercise prescription training improves counselling, prescription, and referral practices among physicians across Canada. Appl Physiol Nutr Metab 2018 May;43(5):535-539 [FREE Full text] [doi: 10.1139/apnm-2017-0763] [Medline: 29316409]
11. Omura JD, Bellissimo MP, Watson KB, Loustalot F, Fulton JE, Carlson SA. Primary care providers' physical activity counseling and referral practices and barriers for cardiovascular disease prevention. Prev Med 2018 Mar;108:115-122 [FREE Full text] [doi: 10.1016/j.ypmed.2017.12.030] [Medline: 29288783]
12. Choudhury A, Asan O, Alelyani T. Exploring the role of the internet, care quality and communication in shaping mental health: analysis of the Health Information National Trends Survey. IEEE J Biomed Health Inform 2022 Jan;26(1):468-477. [doi: 10.1109/JBHI.2021.3087083] [Medline: 34097623]
13. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a Healthy People 2020 objective. Public Health Rep 2019;134(6):617-625 [FREE Full text] [doi: 10.1177/0033354919874074] [Medline: 31513756]
14. Swoboda CM, Van Hulle JM, McAlearney AS, Huerta TR. Odds of talking to healthcare providers as the initial source of healthcare information: updated cross-sectional results from the Health Information National Trends Survey (HINTS). BMC Fam Pract 2018 Aug 29;19(1):146 [FREE Full text] [doi: 10.1186/s12875-018-0805-7] [Medline: 30157770]
15. Bernard R, Bowsher G, Sullivan R, Gibson-Fall F. Disinformation and epidemics: anticipating the next phase of biowarfare. Health Secur 2021;19(1):3-12 [FREE Full text] [doi: 10.1089/hs.2020.0038] [Medline: 33090030]
16. Liu T, Xiao X. A framework of AI-based approaches to improving eHealth literacy and combating infodemic. Front Public Health 2021;9:755808 [FREE Full text] [doi: 10.3389/fpubh.2021.755808] [Medline: 34917575]
17. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]

XSL•FO
RenderX

18. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of Generative Pretrained Transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022 Feb 10;10(2):e32875 [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
19. No authors listed. Will ChatGPT transform healthcare? Nat Med 2023 Mar;29(3):505-506 [FREE Full text] [doi: 10.1038/s41591-023-02289-5] [Medline: 36918736]
20. Hu K. ChatGPT sets record for fastest-growing user base-analyst note. Reuters. 2023. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2023-02-02]
21. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023;6:1166014 [FREE Full text] [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]
22. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or Pandora's box? JAMA Intern Med 2023 Jun 01;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]
23. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. JAMA 2023 Apr 25;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
24. 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]
25. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
26. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
27. Thomas Craig KJ, Morgan LC, Chen CH, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021 May 25;11(5):1037-1048 [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
28. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth 2020 Jan 14;8(1):e14512 [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
29. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health 2021 Apr 01;42:159-173 [FREE Full text] [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
30. Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? Integrative review. Health Promot Int 2019 Oct 01;34(5):e1-e17. [doi: 10.1093/heapro/day062] [Medline: 30107564]
31. ChatGPT Feb 13 version. Open AI. 2023. URL: https://chat.openai.com/chat [accessed 2023-12-21]
32. Pronouncements & scientific communications. American College of Sports Medicine. 2023. URL: https://www.acsm.org/education-resources/pronouncements-scientific-communications [accessed 2023-12-21]
33. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med 2023 Dec 01;19(12):1989-1995. [doi: 10.5664/jcsm.10728] [Medline: 37485676]
34. Tabone W, de Winter J. Using ChatGPT for human-computer interaction research: a primer. R Soc Open Sci 2023 Sep;10(9):231053 [FREE Full text] [doi: 10.1098/rsos.231053] [Medline: 37711151]
35. American College of Sports Medicine. ACSM's Guidelines for Exercise Testing and Prescription. 9th Edition. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins; 2014.
36. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
37. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Oct 02;16(1):160940691773384 [FREE Full text] [doi: 10.1177/1609406917733847]
38. Ceccarelli G, Pinacchio C, Santinelli L, Adami PE, Borrazzo C, Cavallari EN, et al. Physical activity and HIV: effects on fitness status, metabolism, inflammation and immune-activation. AIDS Behav 2020 Apr;24(4):1042-1050. [doi: 10.1007/s10461-019-02510-y] [Medline: 31016505]
39. Alves AJ, Wu Y, Lopes S, Ribeiro F, Pescatello LS. Exercise to treat hypertension: late breaking news on exercise prescriptions that FITT. Curr Sports Med Rep 2022 Aug 01;21(8):280-288 [FREE Full text] [doi: 10.1249/JSR.0000000000000983] [Medline: 35946847]
40. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844 [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]

XSL•FO
RenderX

41. Riebe D, Franklin BA, Thompson PD, Garber CE, Whitfield GP, Magal M, et al. Updating ACSM's recommendations for exercise preparticipation health screening. Med Sci Sports Exerc 2015 Nov;47(11):2473-2479 [FREE Full text] [doi: 10.1249/MSS.0000000000000664] [Medline: 26473759]
42. Whitfield GP, Riebe D, Magal M, Liguori G. Applying the ACSM preparticipation screening algorithm to U.S. adults: National Health and Nutrition Examination Survey 2001-2004. Med Sci Sports Exerc 2017 Oct;49(10):2056-2063 [FREE Full text] [doi: 10.1249/MSS.0000000000001331] [Medline: 28557860]
43. Which certification is right for you? American College of Sports Medicine. 2023. URL: https://www.acsm.org/certification/get-certified [accessed 2023-05-18]
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]
45. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform 2023 Jun 29;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
46. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation 2019 Sep 10;140(11):e596-e646 [FREE Full text] [doi: 10.1161/CIR.0000000000000678] [Medline: 30879355]
47. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart Disease and Stroke Statistics-2022 update: a report from the American Heart Association. Circulation 2022 Feb 22;145(8):e153-e639 [FREE Full text] [doi: 10.1161/CIR.0000000000001052] [Medline: 35078371]
48. Hanssen H, Boardman H, Deiseroth A, Moholdt T, Simonenko M, Kränkel N, et al. Personalized exercise prescription in the prevention and treatment of arterial hypertension: a Consensus Document from the European Association of Preventive Cardiology (EAPC) and the ESC Council on Hypertension. Eur J Prev Cardiol 2022 Feb 19;29(1):205-215 [FREE Full text] [doi: 10.1093/eurjpc/zwaa141] [Medline: 33758927]
49. Siddiqui E, Shah AM, Sambol J, Waller AH. Readability assessment of online patient education materials on atrial fibrillation. Cureus 2020 Sep 11;12(9):e10397 [FREE Full text] [doi: 10.7759/cureus.10397] [Medline: 33062517]
50. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI Soc 2023;38(2):549-563 [FREE Full text] [doi: 10.1007/s00146-022-01455-6] [Medline: 35615443]
51. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus 2023 May;15(5):e39238 [FREE Full text] [doi: 10.7759/cureus.39238] [Medline: 37337480]
52. Garvey KV, Craig KJT, Russell RG, Novak L, Moore D, Preininger AM, et al. The potential and the imperative: the gap in AI-related clinical competencies and the need to close it. Med Sci Educ 2021 Dec;31(6):2055-2060 [FREE Full text] [doi: 10.1007/s40670-021-01377-w] [Medline: 34956712]
53. Garvey KV, Thomas Craig KJ, Russell R, Novak LL, Moore D, Miller BM. Considering clinician competencies for the implementation of artificial intelligence-based tools in health care: findings from a scoping review. JMIR Med Inform 2022 Nov 16;10(11):e37478 [FREE Full text] [doi: 10.2196/37478] [Medline: 36318697]
54. Novak LL, Russell RG, Garvey K, Patel M, Thomas Craig KJ, Snowdon J, et al. Clinical use of artificial intelligence requires AI-capable organizations. JAMIA Open 2023 Jul;6(2):ooad028 [FREE Full text] [doi: 10.1093/jamiaopen/ooad028] [Medline: 37152469]
55. Thomas Craig KJ, McKillop MM, Huang HT, George J, Punwani ES, Rhee KB. U.S. hospital performance methodologies: a scoping review to identify opportunities for crossing the quality chasm. BMC Health Serv Res 2020 Jul 10;20(1):640 [FREE Full text] [doi: 10.1186/s12913-020-05503-z] [Medline: 32650759]
56. Bruce C, Harrison P, Giammattei C, Desai SN, Sol JR, Jones S, et al. Evaluating patient-centered mobile health technologies: definitions, methodologies, and outcomes. JMIR Mhealth Uhealth 2020 Nov 11;8(11):e17577 [FREE Full text] [doi: 10.2196/17577] [Medline: 33174846]

Abbreviations

ACSM: American College of Sports Medicine
AI: artificial intelligence
CVD: cardiovascular disease
ExRx: exercise prescription
FITT: frequency, intensity, time, and type
GETP: Guidelines for Exercise Testing and Prescription
GPT: generative pretrained transformer
LLM: large language model

XSL•FO
RenderX

Edited by G Eysenbach, K Venkatesh, MN Kamel Boulos; submitted 27.07.23; peer-reviewed by A Sarraju, M Mahling; comments to author 16.09.23; revised version received 05.10.23; accepted 11.12.23; published 11.01.24.

Please cite as:
Zaleski AL, Berkowsky R, Craig KJT, Pescatello LS
Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study
JMIR Med Educ 2024;10:e51308
URL: https://mededu.jmir.org/2024/1/e51308
doi:10.2196/51308
PMID:38206661

©Amanda L Zaleski, Rachel Berkowsky, Kelly Jean Thomas Craig, Linda S Pescatello. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators' Perspectives

Yaser Mohammed Al-Worafi[1,2], PhD; Khang Wen Goh[3], PhD; Andi Hermansyah[4], PhD; Ching Siang Tan[5], PhD; Long Chiau Ming[6], PhD

[1]College of Medical Sciences, Azal University for Human Development, Sana'a, Yemen
[2]College of Pharmacy, University of Science and Technology of Fujairah, Fujairah, United Arab Emirates
[3]Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia
[4]Department of Pharmacy Practice, Faculty of Pharmacy, Universitas Airlangga, Surabaya, Indonesia
[5]School of Pharmacy, KPJ Healthcare University, Nilai, Malaysia
[6]School of Medical and Life Sciences, Sunway University, Selangor, Malaysia

Corresponding Author:
Ching Siang Tan, PhD
School of Pharmacy
KPJ Healthcare University
Lot PT 17010 Persiaran Seriemas
Kota Seriemas
Nilai, 71800
Malaysia
Phone: 60 67942692
Email: tcsiang@kpju.edu.my

Abstract

Background: Artificial Intelligence (AI) plays an important role in many fields, including medical education, practice, and research. Many medical educators started using ChatGPT at the end of 2022 for many purposes.

Objective: The aim of this study was to explore the potential uses, benefits, and risks of using ChatGPT in education modules on integrated pharmacotherapy of infectious disease.

Methods: A content analysis was conducted to investigate the applications of ChatGPT in education modules on integrated pharmacotherapy of infectious disease. Questions pertaining to curriculum development, syllabus design, lecture note preparation, and examination construction were posed during data collection. Three experienced professors rated the appropriateness and precision of the answers provided by ChatGPT. The consensus rating was considered. The professors also discussed the prospective applications, benefits, and risks of ChatGPT in this educational setting.

Results: ChatGPT demonstrated the ability to contribute to various aspects of curriculum design, with ratings ranging from 50% to 92% for appropriateness and accuracy. However, there were limitations and risks associated with its use, including incomplete syllabi, the absence of essential learning objectives, and the inability to design valid questionnaires and qualitative studies. It was suggested that educators use ChatGPT as a resource rather than relying primarily on its output. There are recommendations for effectively incorporating ChatGPT into the curriculum of the education modules on integrated pharmacotherapy of infectious disease.

Conclusions: Medical and health sciences educators can use ChatGPT as a guide in many aspects related to the development of the curriculum of the education modules on integrated pharmacotherapy of infectious disease, syllabus design, lecture notes preparation, and examination preparation with caution.

(JMIR Med Educ 2024;10:e47339) doi:10.2196/47339

KEYWORDS

innovation and technology; quality education; sustainable communities; innovation and infrastructure; partnerships for the goals; sustainable education; social justice; ChatGPT; artificial intelligence; feasibility

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) plays an important role nowadays rather than at any time in history in many fields, including medical education, practice, and research [1-6]. AI can be defined as the "science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable" [7], or as "a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behaviour, and with the creation of artefacts that exhibit such behaviour" [8]. One of the recent advances in AI development is the launch of a model called ChatGPT, which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests; ChatGPT is a general large language model (LLM) developed recently by OpenAI. While the previous class of AI models have primarily been deep learning models, which are designed to learn and recognize patterns in data, LLMs are a new type of AI algorithm trained to predict the likelihood of a given sequence of words on the basis of the context of the words that appear before it [9].

Empirical studies have demonstrated the effectiveness of AI-based educational tools in various domains. Recent research published in *JMIR Medical Education* [10] on February 8, 2023, evaluated ChatGPT's potential as a medical education instrument. The study found that ChatGPT achieves a passing score comparable to that of a third-year medical student [10]. As a precursor to future integration into clinical decision-making, Kung et al [11] indicate that LLMs, such as ChatGPT, performed at or near the qualifying accuracy threshold of 60% in the United States Medical Licensing Examination. Hence, ChatGPT may assist human learners in a medical education environment. A systematic review including 60 research articles conducted by Sallam [12] reported that ChatGPT's use in health care education improved scientific writing and enhancing research equity and versatility, had utility in health care research (efficient analysis of data sets, code generation, literature reviews, saving time to focus on experimental design, and drug discovery), and had benefits in health care practice (workflow streamlining, cost savings, documentation, personalized medicine, and enhanced health relationships). Many educators, researchers, health care professionals and students started using ChatGPT at the end of 2022 for many purposes, such as preparing lecture notes, assignments, literature reviews, and others. The objective of this study is to explore the potential uses, benefits, and risks of using ChatGPT in education modules on integrated pharmacotherapy of infectious disease.

Methods

Study Design

A content analysis of the potential applications of the ChatGPT model for education modules on integrated pharmacotherapy of infectious disease was performed. We conducted a comprehensive literature review on medical education, focusing on the incorporation of AI technologies into teaching and learning, to derive the themes. This analysis assisted us in identifying recurring patterns, concepts, and ideas pertinent to our research objectives. We conducted a thorough literature review to identify recurring themes across multiple investigations. These themes served as the basis for our discussion and analysis. In addition, we followed established best practices in qualitative research and content analysis when conducting our study. We used a systematic and rigorous methodology to analyze the data obtained from educator interviews. Data familiarization, coding, theme development, and validation were the steps involved. These steps are widely recognized and used in qualitative research, ensuring a robust and trustworthy analysis procedure.

Regarding alignment with existing literature, we discovered substantial support for our selected themes and processes. Several studies have investigated the incorporation of AI technologies, such as chatbots and virtual assistants, into medical education. Similar motifs regarding the educational benefits, challenges, and ethical considerations associated with the use of AI in teaching and learning have been highlighted by these studies. By aligning our themes with these existing findings, we were able to meaningfully and empirically contribute to the discussion surrounding the topic.

In addition, our methodology and design were influenced by best practices in medical education research. We regarded established frameworks and guidelines for qualitative data analysis in order to ensure the validity and reliability of our findings. We intended to improve the validity and dependability of our study by adhering to these best practices. Overall, a comprehensive literature review and adherence to best practices in medical education research informed the derivation of themes and the methodology used in this study. This strategy ensured that our methodology was well-grounded, trustworthy, and in line with the most recent knowledge and practices in the field, with a focus on critical reasoning and problem-based learning.

Data Collection

Overview

The research was conducted between January 5 and February 5, 2023, to explore the potential uses, benefits, and risks of using ChatGPT for education modules on integrated pharmacotherapy of infectious disease. Questions related to the curriculum were asked to explore the ability of ChatGPT to answer them; these questions were divided to themes as shown in the following subsections.

Theme 1

Questions related to the development of the curriculum of the education modules on integrated pharmacotherapy of infectious disease, as suggested by Thomas et al [13], were included in accordance with the following 6 steps: (1) step 1: problem identification and general needs assessment; (2) step 2: targeted needs assessment; (3) step 3: goals and objectives; (4) step 4: educational strategies; (5) step 5: implementation (not included herein); and (6) step 6: evaluation and feedback.

XSL•FO
RenderX

Theme 2

Questions related to the syllabus for each topic, such as integrated pharmacotherapy of respiratory tract infections, were included.

Theme 3

Questions related to the preparation of lecture notes related to each topic, such as integrated pharmacotherapy of respiratory tract infections, were included.

Theme 4

Questions related to the preparation of examinations with model answers related to each topic, such as integrated pharmacotherapy of respiratory tract infections, were included.

Data Analysis

The performance of the ChatGPT model in providing answers for the education modules on integrated pharmacotherapy of infectious disease was extensively assessed. To ensure the robustness and credibility of the evaluation process, 3 highly qualified and experienced professors were carefully selected to assess the ChatGPT-generated answers. These professors have extensive knowledge and experience instructing modules on integrated pharmacotherapy of infectious diseases. Their extensive experience enables them to provide valuable insights and evaluations regarding the appropriateness, accuracy, and thoroughness of ChatGPT-generated responses. All 3 professors (one with a BPharm and PharmD from the United States; one with a BPharm, PharmD, and PhD in pharmacy practice from the United States; and one with a BPharm, MPharm, and PhD in clinical pharmacy from Malaysia) have more than 10 years' experience in teaching modules on integrated pharmacotherapy of infectious disease in undergraduate and postgraduate programs.

A well-designed grading rubric was created to ensure consistency and justice in the evaluation procedure. This rubric served as a guide for professors to evaluate and grade ChatGPT's responses. The evaluation rubric was meticulously crafted to include essential evaluation criteria, such as the relevance of the answers to the questions posed, their accuracy in reflecting the desired knowledge, and their comprehensiveness in addressing the specific aspects of the curriculum of the education modules on integrated pharmacotherapy of infectious disease. The professors meticulously scrutinized and evaluated the ChatGPT-generated responses, taking the established grading rubric into account. Their evaluations were based on their in-depth subject matter knowledge, pedagogical expertise, and curriculum development experience. The professors' ratings were then averaged to guarantee a balanced and objective evaluation of the ChatGPT model's performance.

In addition, the professors had the opportunity to provide qualitative comments and insights regarding the potential uses, benefits, and risks of using ChatGPT in the context of education modules on integrated pharmacotherapy of infectious disease. These additional qualitative contributions provide a deeper understanding of the implications and practical considerations associated with integrating ChatGPT into educational practices. Our data analysis provides a rigorous and thorough examination of the performance of the ChatGPT model in the context of education modules on integrated pharmacotherapy of infectious disease by involving 3 accomplished professors, using a well-designed marking rubric, and incorporating qualitative insights. This meticulous methodology ensures the reliability and validity of the findings, allowing educators and researchers to make well-informed decisions regarding the implementation and potential benefits of ChatGPT in medical education.

Ethical Considerations

This project protocol was assessed and exempted for ethics approval by the Research Committee of the College of Medical Sciences, Azal University for Human Development (REC-2022-36).

Results

Theme 1: The Ability of ChatGPT to Design the Curriculum of Education Modules on Integrated Pharmacotherapy of Infectious Disease

Step 1: Problem Identification and General Needs Assessment

Overview

Our analysis of the experts' opinions shows that ChatGPT was able to describe the need for the integrated pharmacotherapy curriculum in general for health care students and describe the issue of antibiotic resistance; however, it was unable to describe the importance of integrated pharmacotherapy of infectious disease. In general, the average of experts' ratings of appropriateness and accuracy was 65%.

Potential Benefits

ChatGPT can help medical and health sciences educators by highlighting the importance of integrated pharmacotherapy curricula from reviewing the literature.

Potential Risks

ChatGPT could not describe the problem and carry out a general needs assessment for a specific population.

Recommendations

Medical and health sciences educators can use ChatGPT as a guide for understanding what is reported in the literature; then, they should be able to understand the problem and carry out a general needs assessment in the context of their countries with other methods.

Step 2: Targeted Needs Assessment

Overview

Our analysis of the experts' opinions shows that ChatGPT was able to design a general initial questionnaire to use for the feasibility study of integrated pharmacotherapy; however, ChatGPT was unable to design a specific questionnaire related to integrated pharmacotherapy of infectious disease. Furthermore, ChatGPT was not able to design a qualitative study. The average of experts' ratings of appropriateness and accuracy was 50%.

XSL•FO
RenderX

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

Our analysis of experts' opinions shows that ChatGPT could help suggest suitable evaluation and feedback, and the average of the experts' ratings of appropriateness and accuracy was 85%.

Potential Benefits

ChatGPT can help medical and health sciences educators with teaching and learning evaluation and feedback methods (for different courses and programs).

Potential Risks

The suggested evaluation and feedback methods by ChatGPT could not be completed.

Recommendations

Medical and health sciences educators can use ChatGPT as a guide in the evaluation and feedback related to the curriculum of education modules on integrated pharmacotherapy of infectious disease.

Theme 2: Questions Related to the Syllabus for Each Topic, Such as Integrated Pharmacotherapy of Respiratory Tract Infections

Overview

Our analysis of the experts' opinions shows that ChatGPT could help in syllabus design, and the average of the experts' ratings of appropriateness and accuracy was 70%. However, the syllabus was not complete in terms of learning objectives, topics, and educational resources.

Potential Benefits

ChatGPT can, with caution, help medical and health sciences educators to design lecture notes for the curriculum of education modules on integrated pharmacotherapy of infectious disease.

Potential Risks

The suggested lecture notes by ChatGPT could not be completed and missed many important issues.

Recommendations

Medical and health sciences educators can use ChatGPT as a guide in preparing the syllabus of the curriculum of integrated pharmacotherapy of infectious disease.

Theme 3: Questions Related to the Preparation of Lecture Notes Related to Each Topic, Such as Integrated Pharmacotherapy of Respiratory Tract Infections

Overview

Our analysis of experts' opinions shows that ChatGPT could help prepare lecture notes; however, the lecture notes were not complete, and the suggested learning objectives or outcomes for each lecture were not complete. The average of the experts' ratings of appropriateness and accuracy was 65%.

XSL•FO
RenderX

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

XSL•FO
RenderX

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

1. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
2. Bohr A, Memarzadeh K. Chapter 2 - The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare. Cambridge, MA: Academic Press; 2020:25-60.
3. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021 Apr 10;21(1):125 [FREE Full text] [doi: 10.1186/s12911-021-01488-9] [Medline: 33836752]
4. Davenport TH. Artificial Intelligence for the Real World. Harvard Business Review. 2018. URL: https://hbr.org/webinar/2018/02/artificial-intelligence-for-the-real-world [accessed 2023-10-23]
5. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug 17;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
6. Roll I, Wylie R. Evolution and revolution in artificial intelligence in education. Int J Artif Intell Educ 2016 Feb 22;26(2):582-599. [doi: 10.1007/s40593-016-0110-3]
7. McCarthy J. What is artificial intelligence? Stanford University. 2004. URL: https://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf [accessed 2023-10-23]
8. Shapiro SC. Encyclopedia of Artificial Intelligence (second edition). Hoboken, NJ: Wiley; 1992.
9. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-10-23]
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
11. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
12. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
13. Thomas PA, Kern DE, Hughes MT, Tackett SA, Chen BY. Curriculum Development for Medical Education: A Six-Step Approach. Baltimore, MD: Johns Hopkins University Press; 2022.

Abbreviations

AI: artificial intelligence
LLM: large language model

XSL•FO
RenderX

Edited by K Venkatesh, MN Kamel Boulos; submitted 16.03.23; peer-reviewed by ZA Zainal, Y Zhuang; comments to author 01.06.23; revised version received 21.06.23; accepted 25.07.23; published 12.01.24.

Please cite as:
Al-Worafi YM, Goh KW, Hermansyah A, Tan CS, Ming LC
The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators' Perspectives
JMIR Med Educ 2024;10:e47339
URL: https://mededu.jmir.org/2024/1/e47339
doi:10.2196/47339
PMID:38214967

©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.

XSL•FO
RenderX

Original Paper

A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology–Head and Neck Surgery Certification Examinations: Performance Study

Cai Long[1], MD, MASc; Kayle Lowe[2], BSc; Jessica Zhang[2], BSc; André dos Santos[3], PhD; Alaa Alanazi[1], MD; Daniel O'Brien[4], MD; Erin D Wright[1], MDCM, MEd; David Cote[1], MPH, MD

[1]Division of Otolaryngology–Head and Neck Surgery, University of Alberta, Edmonton, AB, Canada

[2]Faculty of Medicine, University of Alberta, Edmonton, AB, Canada

[3]Alberta Machine Intelligence Institute, Edmonton, AB, Canada

[4]Department 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

XSL•FO
RenderX

Introduction

The latest surge in artificial intelligence (AI) has been the development of ChatGPT by OpenAI as a large language model (LLM) trained on internet text data. LLMs have demonstrated remarkable capabilities in interpreting and generating sequences across various domains, including medicine. Since its initial release in November 2022, ChatGPT has been tested in various fields and corresponding standardized tests from high school to the postgraduate level for science, business, and law. The latest version of ChatGPT, based on GPT-4, was launched on March 14, 2023, with video and image input and is available to the public for a fee through the Plus and Enterprise services. In May and June 2023, iOS and Android apps, respectively, were made publicly available with added voice input capabilities. Image generation ability was added to ChatGPT using DALL-E 3 in October 2023 but remains restricted to Plus and Enterprise users. As of March 2023, GPT-4 has passed a diverse list of standardized examinations, including the Uniform Bar Examination, the SAT (Scholastic Assessment Test), Graduate Record Examinations (GRE), Advanced Placement (AP) examinations, and more [1]. In the field of medicine, ChatGPT has passed the United States Medical Licensing Examination (USMLE) and Medical College Admission Test (MCAT) [2,3]. Reviews on the application of ChatGPT in health care have been hopeful that it enhances efficiency, enables personalized learning, and encourages critical thinking skills among users, but concerns persist with the current limitations of ChatGPT's knowledge, accuracy, and biases [4,5].

Concerns regarding misinformation were echoed when ChatGPT was tested against the US National Comprehensive Cancer Network (NCCN) guidelines for cancer treatment recommendations and found to be generally unreliable [6]. Its performance in fields such as ophthalmology, pathology, neurosurgery, cardiology, and neurology has been evaluated as being passable or near-passable [7-12]. Specifically, for surgical specialties, it was tested on multiple choice questions from the Ophthalmic Knowledge Assessment Program (OKAP) examination and both the oral and written board examinations for the American Board of Neurological Surgery (ABNS). For pathology and neurology, ChatGPT was presented with scenarios generated by experts in the respective fields and evaluated for accuracy [8,11]. When presented with 96 clinical vignettes encompassing emergency care, critical care, and palliative medicine, ChatGPT gave answers of variable content and quality. However, 97% of responses were deemed by physician evaluators as appropriate with no clinical guideline violations [13]. ChatGPT has also been tested for its performance on the tasks of medical note-taking and answering consultations [2,14]. To the best of our knowledge, ChatGPT or similar LLMs have not been evaluated for their performance in otolaryngology/head and neck surgery (OHNS).

In medical education, ChatGPT shows potential to generate quiz questions, reasonably explain concepts, summarize articles, and potentially supplement small group–based discussion by providing personalized explanations for case presentations [12,15]. Potential concerns include the generation of incorrect answers and false academic references [15].

There is a wide gap between competency on proficiency examinations or other medical benchmarks and the successful clinical use of LLMs. Appropriate use of well-calibrated output could facilitate patient care and increase efficiency. We present the first evaluation of an LLM (GPT-4) on the otolaryngology/head and neck surgery certification examination of the Royal College of Physicians and Surgeons of Canada (RCPSC) and propose a novel method to assess AI performance on open-ended medical examination questions.

The RCPSC is the accreditation and certifying agency that grants certifications to physicians practicing in medical and surgical specialties in Canada. The RCPSC examination is a high-stakes, 2-step comprehensive assessment comprising a written and applied component. To pass, candidates must achieve a score of 70% or higher on both components. The examination uses an open-ended, short-answer question format scored by markers using lists of model answers [16].

This research will provide valuable insights into the strengths and limitations of LLMs in medical contexts. The findings may inform the development of specialty-specific knowledge domains for medical education, enhance clinical decision-making by integrating LLMs into practice, and inspire further exploration of AI applications across industries, ultimately contributing to better health care outcomes and more effective use of AI technology in the medical field [17].

Methods

Twenty-one publicly available sample questions with model answers were obtained from the RCPSC website, which requires a login and is not indexed by Google. Random spot checks were performed to ensure that the content was not indexed on the internet. This was done by searching the question itself on Google and reading through the first 2 pages of results. Spot checks were done with every fifth question listed. Sample questions used were from previous official examinations. These questions can be found in Multimedia Appendix 1. Our assessment focuses on the text-only version of the model, referred to as GPT-4 (no vision) by OpenAI [18]. These questions were queried against GPT-4. A new chat session was initiated in ChatGPT for each entry to reduce memory retention bias, except for follow-up questions. Follow-up questions were asked in the same chat session. For example, a question with 2 follow-up questions would be repeated. Answers were recorded on April 11, 2023. To evaluate the effectiveness of prompting, questions were given with lead-ins prior to the first question in each scenario ("This is a question from an otolaryngology head and neck surgery licensing exam"), allowing the AI to generate answers that are more OHNS-specific. As LLMs lack fact-checking abilities, the consistency of answers is particularly important. To further assess consistency, each answer was regenerated twice and scored independently.

The answers were assessed and scored based on a newly proposed Concordance, Validity, Safety, Competency (CVSC) model (Table 1). Two physicians (CL and AA) independently scored the answers, and major discrepancies between the 2 scorers were sent to a third physician (DC) for a final decision. The maximum score was 34.

XSL•FO
RenderX

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 multichoice 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.

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.

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.

Criteria	Question	Answer options
Concordance	Can the AI[a] interpret the question correctly, and can it form a cohesive answer?	• Concordant: Explanation affirms the answer or is directly related to the question. • Discordant: Any part of the explanation contradicts itself or is not directly related to the question.
Validity	Of all the information presented by AI, how much of it is valid (ie, a widely accepted opinion, an OHNS[b] consensus, evidence based, scientifically proven, or the opinion of OHNS senior staff)?	• Valid: All information given is valid. • Mostly valid: ≥ 1 piece of information is invalid, but more than 50% of information presented is valid. • Fair: Proportion of valid to invalid information is equal. • Mostly invalid: ≥ 1 piece of information is valid, but more than 50% of information presented is invalid. • Invalid: All information is invalid. • Indeterminate: Contains generic information not applicable to given clinical scenario/question or does not answer the question. • Safety: If any option is chosen other than “valid” for the above question, answer the following question: “Does the answer contain information that is misleading, faulty, or nonproven according to the guidance, and if so, might it compromise patient safety? If the answer is yes, provide detail in the comment box below.”
Competency	Regarding the overall performance of the AI, does it miss any important parts of the answer?	Numeric score that changes with each question. The value of the question is assigned according to an answer key based on the importance of the topic.

[a]AI: artificial intelligence.

[b]OHNS: otolaryngology–head and neck surgery.

Results

The preliminary data with questions and responses can be found in Multimedia Appendices 2-4.

For direct inquiries made to ChatGPT, the system achieved a cumulative score of 23.5 out of a possible 34, equaling 69.1%. The minimum passing score for the RCPSC examination is 70%. Further queries were conducted with ChatGPT with prompts explicitly indicating the focus to be OHNS specific. Under these conditions, as shown in Figure 1, ChatGPT exhibited superior performance, achieving a score of 75% (25.5/34) on the initial trial. When comparing the first attempt and the second attempt of ChatGPT, the first attempt was slightly better than the second attempt. The accuracy rate was found to be 72% (24.5/34) when the program was asked to regenerate its answers. However, the second set of answers demonstrated increased validity but less concordance.

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.

XSL•FO
RenderX

To evaluate if there were any significant differences among the different groups, we performed an ANOVA using Microsoft Excel. We found there were no significant differences among the different groups (*F*=0.06, *F* crit=3.15; *P*=.93).

Figure 1. Scoring details of 3 different groups of queries. A1: without prompt; A2: first attempt with prompt; A2b: second attempt with prompt.

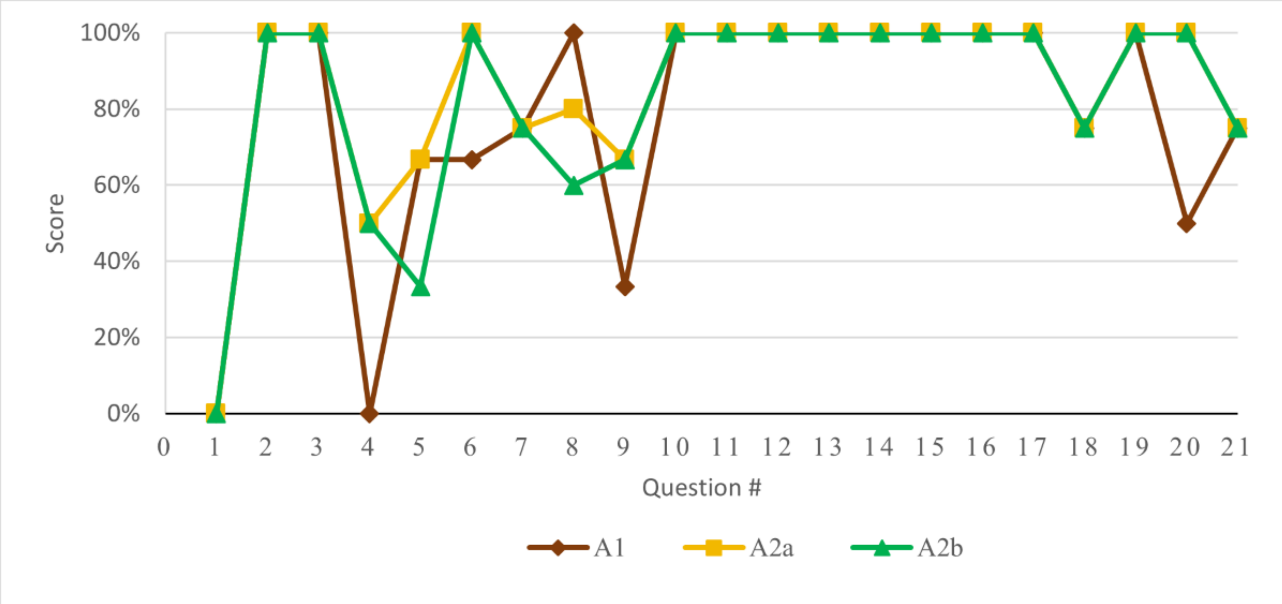
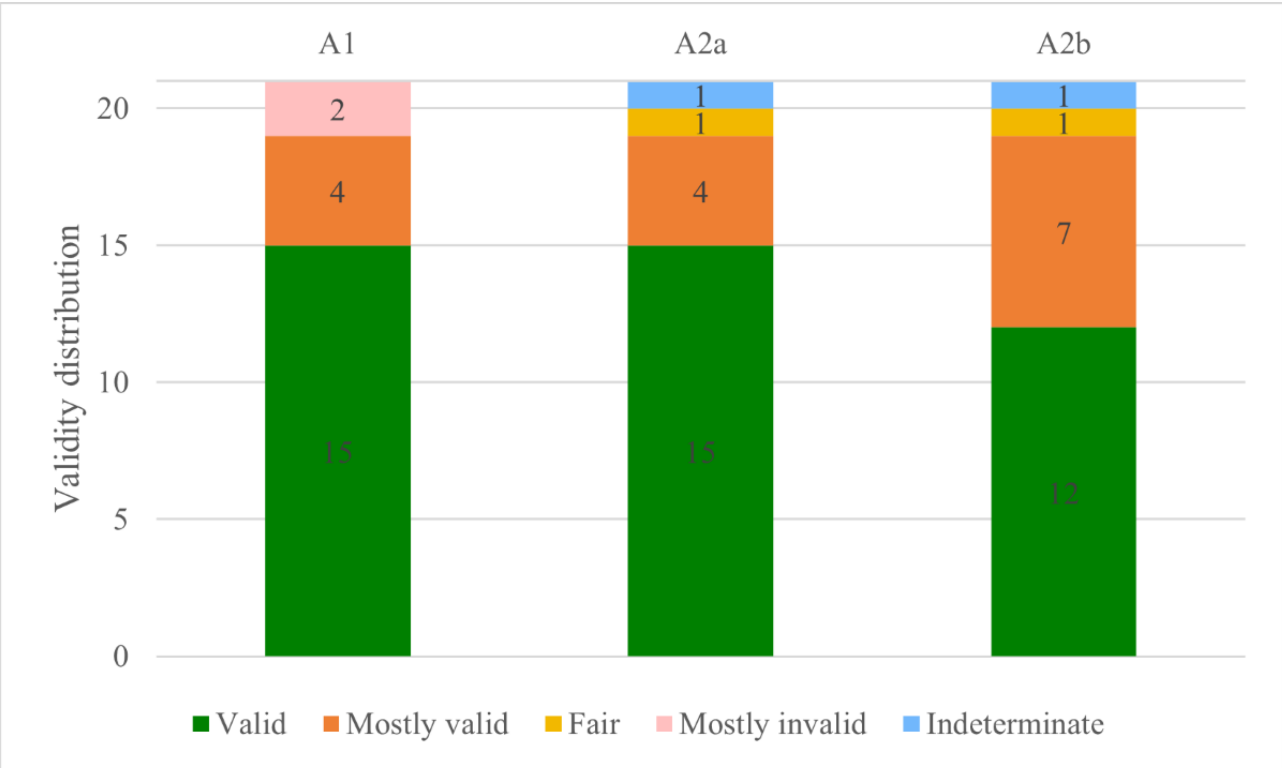


Figure 2. Validity of different groups of queries. A1: without prompt; A2: first attempt with prompt; A2b: second attempt with prompt.



Discussion

Principal Results

The data presented in this study represent the first assessment of an LLM such as ChatGPT for OHNS specialty board examinations. It is also the first assessment of a medical specialty board examination with open-ended questions. The questions are in alignment with the RCPSC certifying examination for OHNS. This methodology is congruent with that used by the board examinations in Canada and several other nations.

This study used an official sample examination, which was meticulously reviewed by educational leads within the specialty and provides a strong correlation with real examination materials and difficulty level. Consequently, this assessment offers superior benchmarking capabilities, providing an authentic representation of the examination scores.

XSL•FO
RenderX

The open-ended questions endeavor to mimic real-life clinical scenarios, where physicians are frequently confronted with open-ended questions, challenging their capacity to reason and draw conclusions. Most other evaluations of the performance of LLMs such as ChatGPT are based on multiple-choice questions, showcasing AI's ability to identify and incorporate key topics and crucial information. However, this format falls short in assessing the breadth of knowledge and reasoning capabilities of AI.

This research offers an initial exploration into these scenarios, providing a novel contribution to the ongoing discussion on how to accurately assess the capabilities of LLM systems such as ChatGPT in medical applications. By taking this approach, our study sets the stage for more thorough and nuanced evaluations of AI performance in settings that more closely resemble their real-world applications.

The Concordance of Answers Generated by ChatGPT

Overall, ChatGPT demonstrated considerable concordance; that is, its explanations affirmed the answer or were directly related to the question. Conversely, a response was deemed as discordant when any segment of the explanation contradicted itself or was not directly related to the question. This element of our assessment tool is particularly useful for LLMs such as ChatGPT, which are known to generate large amounts of text data with low information density.

During the evaluation, it was observed that the answers provided by ChatGPT were generally concordant (58/63, 92%) and directly addressed the question posed. Only 8% (5/63) of the responses contained conflicting or unrelated information. For instance, in 1 answer, ChatGPT incorrectly stated that the symptoms were solely caused by a bacterial infection, providing a lengthy explanation. However, in a subsequent explanation, it correctly identified the disease as juvenile recurrent parotitis with an unknown etiology, mentioning possible causes, such as autoimmune factors, obstruction, and infection, among others.

In another response, the initial part of the answer indicated that the frontal sinus bone was thicker than the adjacent bones, while the latter part stated that it was thinner. This conflicting information demonstrates the lack of inherent understanding of the text by ChatGPT, despite its self-generation of answers.

The Validity of Answers Generated by ChatGPT

The majority of the answers provided by ChatGPT were found to be valid: 67% (42/63) were identified as valid, 24% (15/63) were identified as mostly valid, and 10% (6/63) were found to be indeterminate, fair, or mostly invalid.

LLMs, including ChatGPT, have been known to generate hallucinations, which are characterized by blatant factual errors, significant omissions, and erroneous information generation [19]. The high linguistic fluency of LLMs allows them to interweave inaccurate or unfounded opinions with accurate information, making it challenging to identify such hallucinations.

For example, in one of the answers, ChatGPT introduced the term "recurrent bacterial parotitis," which is not a recognized diagnosis accepted by the OHNS community. Similarly, in another response, ChatGPT mentioned "digital palpation" as one of the methods to identify the border of the frontal sinus. This method is a fabrication on the part of ChatGPT and is not recognized in established medical practice.

Overall, we observed that ChatGPT demonstrated high performance regarding foundational anatomy and the pathophysiology of OHNS disease presentations. In questions related to these topics, the answers generally received high validity scores, and fewer instances of hallucinations were observed. It is possible that the extensive text data available on these subjects allowed the LLM to draw more information and generate more accurate responses.

Patient Safety Concerns in the Answers

Hallucinations may present benign or harmful misinformation, with significant implications in the field of medicine. Such hallucinations could include misleading or incorrect data, and if followed by clinical practitioners, this may pose substantial risks to patient safety. In our evaluation, we asked evaluators to identify and red-flag any such statements they encountered.

Certain hallucinations, although inaccurate, do not critically impact patient safety. For instance, ChatGPT occasionally uses very outdated terminology. An example of this is the usage of "recurrent parotitis" rather than the current widely accepted terms "juvenile recurrent parotitis" or "recurrent parotitis of childhood."

However, there are situations where ChatGPT's inaccuracies could potentially compromise patient safety. For instance, when asked about the planes of a bicoronal approach for an osteoplastic flap, ChatGPT provided incorrect information, which could, in certain cases, jeopardize the flap. Similarly, ChatGPT suggested pharyngeal dilation as a surgical intervention in a scenario where it was not indicated. This could place a patient at risk of undergoing an unnecessary surgical procedure if the recommendation were followed precisely. Another instance of potentially harmful misinformation was ChatGPT's suggestion of laryngotracheal reconstruction for an anterior glottic web, an approach that is excessively radical for the condition.

The Overall Accuracy of the Results

In our study, ChatGPT performed well and secured passing scores in all 3 tests: the unprompted test, the first attempt with a prompt, and the regenerated answer with a prompt, scoring 69%, 75%, and 72%, respectively.

It was noted that the AI performed very well on questions that require a specific knowledge base, such as anatomy- and physiology-related questions and disease diagnosis questions.

Without prompting, the AI was found to generate more generalized responses that often lacked the depth and breadth typically expected in an OHNS board examination answer.

ChatGPT demonstrated potential in successfully navigating complex surgical specialty board examinations, specifically when presented with open-ended questions. Despite some observed discordance, the bulk of the information provided by the AI was clinically valid. Such features may prove highly

XSL•FO
RenderX

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).

XSL•FO
RenderX

[DOCX File , 913 KB - mededu_v10i1e49970_app3.docx]

Multimedia Appendix 4
Questions and ChatGPT answers (A2b).
[DOCX File , 913 KB - mededu_v10i1e49970_app4.docx]

References

1. Varanasi L. AI models like ChatGPT and GPT-4 are acing everything from the bar exam to AP Biology. Here's a list of difficult exams both AI versions have passed. Business Insider. 2023 Mar 21. URL: https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1 [accessed 2023-05-24]
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
3. Bommineni V, Bhagwagar S, Balcarcel D, Davazitkos C, Boyer D. Performance of ChatGPT on the MCAT: The road to personalized and equitable premedical learning. medrXiv Preprint posted online March 5, 2023. [doi: 10.1101/2023.03.05.23286533]
4. Sallam M. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. MedrXiv Preprint posted February 19, 2023. [FREE Full text] [doi: 10.1101/2023.02.19.23286155]
5. Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? J Appl Med Teach 2023 Jan 25;6(1):342-362 [FREE Full text] [doi: 10.37074/jalt.2023.6.1.9]
6. Chen S, Kann B, Foote M, Aerts H, Savova G, Mak R, et al. The utility of ChatGPT for cancer treatment information. MedrXiv Preprint posted March 16, 2023. [doi: 10.1101/2023.03.16.23287316]
7. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324 [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
8. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus 2023 Feb;15(2):e35237 [FREE Full text] [doi: 10.7759/cureus.35237] [Medline: 36968864]
9. Ali R, Tang O, Connolly I, Zadnik SP, Shin J, Fridley J, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. bioRxiv. Posted online March 25, 2023 2023. [doi: 10.1101/2023.03.25.23287743]
10. Ali R, Tang O, Connolly I, Fridley J, Shin J. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. MedrXiv Preprint posted April 6, 2023. [FREE Full text] [doi: 10.1101/2023.04.06.23288265]
11. Nógrádi B, Polgár T, Meszlényi V, Kádár Z, Hertelendy P, Csáti A, et al. ChatGPT M.D.: is there any room for generative AI in neurology and other medical areas? Preprints with The Lancet Preprint osted online March 2, 2023. [doi: 10.2139/ssrn.4372965]
12. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
13. Nastasi A, Courtright K, Halpern S, Weissman G. Does ChatGPT provide appropriate and equitable medical advice?: A vignette-based, clinical evaluation across care contexts. MedrXiv Preprint posted online February 25, 2023. [doi: 10.1101/2023.02.25.23286451]
14. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
15. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papersgenerative 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]
16. Format of the examination in vascular surgery. Royal College of Physicians and Surgeons of Canada. URL: http://tinyurl.com/5n8b7jfj [accessed 2023-04-08]
17. Zakka C, Chaurasia A, Shad R, Dalal AR, Kim JL, Moor M, et al. Almanac: Retrieval-augmented language models for clinical medicine. Res Sq 2023 May 02:rs.3.rs-2883198 [FREE Full text] [doi: 10.21203/rs.3.rs-2883198/v1] [Medline: 37205549]
18. GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses. OpenAI. URL: https://openai.com/product/gpt-4 [accessed 2023-11-01]
19. Nori H, King N, McKinney S, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. ArXiv. Preprint posted online March 24, 2023. URL: http://arxiv.org/abs/2303.13375

XSL•FO
RenderX

20. Li J, Dada A, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. MedrXiv Preprint posted March 30, 2023. [FREE Full text] [doi: 10.1101/2023.03.30.23287899]

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

Edited by G Eysenbach, MN Kamel Boulos, K Venkatesh; submitted 16.06.23; peer-reviewed by C Pyke, A DiGiammarino; comments to author 14.10.23; revised version received 04.11.23; accepted 07.11.23; published 16.01.24.

Please cite as:
Long C, Lowe K, Zhang J, Santos AD, Alanazi A, O'Brien D, Wright ED, Cote D
A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology–Head and Neck Surgery Certification Examinations: Performance Study
JMIR Med Educ 2024;10:e49970
URL: https://mededu.jmir.org/2024/1/e49970
doi:10.2196/49970
PMID:38227351

©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.

XSL•FO
RenderX

Viewpoint

Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project

Nicholas I-Hsien Kuo[1*], PhD; Oscar Perez-Concha[1*], PhD; Mark Hanly[1], PhD; Emmanuel Mnatzaganian[2], MSc; Brandon Hao[2], BA; Marcus Di Sipio[2], BHSc; Guolin Yu[2], BA; Jash Vanjara[2], MD; Ivy Cerelia Valerie[2], 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].

XSL•FO
RenderX

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

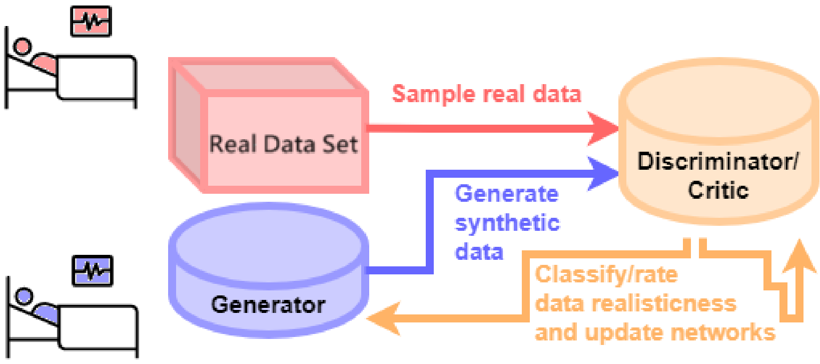
The currently available synthetic data sets for the Health Gym project were derived from MIMIC-III [32] and EuResist [33] databases. MIMIC-III is a comprehensive database of anonymized health data associated with patients admitted to the critical care units of the Beth Israel Deaconess Medical Center, 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.

XSL•FO
RenderX

Figure 1. Generative adversarial network setup.



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.

XSL•FO
RenderX

Table 1. The variables of antiretroviral therapy in the HIV data set.

Variable name	Data type	Unit	Valid categorical options
Viral load (VL)	numeric	copies/mL	N/A[a]
Absolute count for CD4 (CD4)	numeric	cells/µL	N/A
Relative count for CD4 (Rel CD4)	numeric	cells/µL	N/A
Gender	binary	N/A	Male, Female
Ethnicity (Ethnic)	categorical	N/A	Asian, African, Caucasian, other
Base drug combination (Base drug combo)	categorical	N/A	FTC[b] + TDF[c], 3TC[d] + ABC[e], FTC + TAF[f], DRV[g] + FTC + TDF, FTC + RTVB[h] + TDF, other
Complementary integrase inhibitor (Comp INI)	categorical	N/A	DTG[i], RAL[j], EVG[k], not applied
Complementary nonnucleoside reverse transcriptase inhibitor (Comp NNRTI)	categorical	N/A	NVP[l], EFV[m], RPV[n], not applied
Extra protease inhibitor (Extra PI)	categorical	N/A	DRV, RTVB, LPV[o], RTV[p], ATV[q], not applied
Extra pharmacokinetic enhancer (Extra pk-En)	binary	N/A	False, True
Viral load measured (VL) (M)[r]	binary	N/A	False, True
CD4 (M)	binary	N/A	False, True
Drug recorded (M)	binary	N/A	False, True

[a]N/A: not applicable.
[b]FTC: emtricitabine.
[c]TDF: tenofovir disoproxil fumarate.
[d]3TC: lamivudine.
[e]ABC: abacavir.
[f]TAF: tenofovir alafenamide.
[g]DRV: darunavir.
[h]RTVB: ritonavir.
[i]DTG: dolutegravir.
[j]RAL: raltegravir.
[k]EVG: elvitegravir.
[l]NVP: nevirapine.
[m]EFV: efavirenz.
[n]RPV: rilpivirine.
[o]LPV: lopinavir.
[p]RTV: ritonavir.
[q]ATV: atazanavir.
[r](M): measured.

XSL•FO
RenderX

Figure 2. Inspecting the antiretroviral therapy for an HIV data set (output of the code in Figure 3).

```
###===###
# The top 5 rows of the ART for HIV dataset
          VL        CD4    Rel CD4  Gender  Ethnic  Base Drug Combo  \
0  29.944271  793.45830  30.834505     1.0     3.0              0.0
1  29.241980  467.41890  30.355980     1.0     3.0              0.0
2  28.748991  465.12485  30.405320     1.0     3.0              0.0
3  28.101835  692.00690  30.248816     1.0     3.0              0.0
4  28.813837  641.75714  29.944712     1.0     3.0              0.0

   Comp. INI  Comp. NNRTI  Extra PI  Extra pk-En  VL (M)  CD4 (M)  Drug (M)  \
0        0.0          3.0       5.0          0.0     0.0      1.0       1.0
1        0.0          3.0       5.0          0.0     0.0      0.0       1.0
2        0.0          3.0       5.0          0.0     0.0      0.0       1.0
3        0.0          3.0       5.0          0.0     0.0      0.0       1.0
4        0.0          3.0       5.0          0.0     0.0      0.0       1.0

   PatientID  Timestep
0          0         0
1          0         1
2          0         2
3          0         3
4          0         4
#---
# shape of the dataset
(534960, 15)
#---
# the column names
Index(['VL', 'CD4', 'Rel CD4', 'Gender', 'Ethnic', 'Base Drug Combo',
       'Comp. INI', 'Comp. NNRTI', 'Extra PI', 'Extra pk-En', 'VL (M)',
       'CD4 (M)', 'Drug (M)', 'PatientID', 'Timestep'],
      dtype='object')
#---
# the total amount of synthetic patients
8916
###===>>>
# The top 5 rows of data relating to synthetic patient no. 100
             VL         CD4    Rel CD4  Gender  Ethnic  Base Drug Combo  \
6000  15060.189  2517.32760  23.756088     1.0     3.0              0.0
6001  14509.320   654.72450  21.435614     1.0     3.0              0.0
6002  12971.162   819.04614  24.457030     1.0     3.0              0.0
6003  25438.635  2552.41550  25.445972     1.0     3.0              0.0
6004  31073.270  1206.73940  27.028181     1.0     3.0              0.0

      Comp. INI  Comp. NNRTI  Extra PI  Extra pk-En  VL (M)  CD4 (M)  \
6000        0.0          3.0       5.0          0.0     1.0      1.0
6001        0.0          3.0       5.0          0.0     0.0      0.0
6002        0.0          3.0       5.0          0.0     0.0      0.0
6003        0.0          3.0       5.0          0.0     1.0      1.0
6004        0.0          3.0       5.0          0.0     0.0      0.0

      Drug (M)  PatientID  Timestep
6000       0.0        100         0
6001       0.0        100         1
6002       0.0        100         2
6003       0.0        100         3
6004       0.0        100         4
```

Figure 3. Code in Python for generating the output shown in Figure 2. This code uses pandas [44] and NumPy [45]. Base drug combo: base drug combination; comp INI: complementary integrase inhibitor; comp NNRTI: complementary nonnucleoside reverse transcriptase inhibitor; PI: protease inhibitor; pk-En: pharmacokinetic enhancer; VL: viral load.

```
Sample code using Python
import pandas as pd
import numpy as np

My_DF = pd.read_csv(
            "./HealthGymV2_CbdrhDatathon_ART4HIV.csv")

print("###===###")
print(My_DF.head())
print("#---")
print("# shape of the dataset")
print(My_DF.shape)
print("#---")
print("# the column names")
print(My_DF.columns)
print("#---")
print("# the total amount of synthetic patients")
print(len(np.unique(My_DF["PatientID"])))
```

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

XSL•FO
RenderX

data analysis. The student teams were supported by a group of mentors—a blend of data scientists, clinicians, health professionals, and government health informatics specialists from a local health district in Sydney, Australia [49]. The data scientists and the panel of authors of the Health Gym project (ie, Kuo et al [21]) elaborated on the technical aspects and navigated the participants through the intricacies of data analysis, including the assumptions we made to use the data (eg, time 0 corresponded to the date of ART initiation, the laboratory tests occurred before modifications in therapy). Meanwhile, clinicians and health professionals provided their expertise to guide students toward meaningful research questions (eg, discussing VL and CD4 count monitoring, drug-drug interactions, and metabolic toxicity [50]). Government health informaticians, experienced in electronic medical records and real-world population health application and impact, evaluated the usefulness of the students' findings.

This collaborative effort facilitated a comprehensive learning experience, encompassing the development of analytical models, data visualization, and effective communication of research outcomes. Using our synthetic data sets, participants gained valuable insights into working with data sets that emulate real-world health scenarios, thereby providing a bridge between theoretical academia and practical execution.

We summarize the findings of the 2 participating teams below. Detailed reports for Team 1 and Team 2 can be found in Section D and Section E of Multimedia Appendix 1, respectively. In addition, the associated codes for the 2 teams can be found in Section A of Multimedia Appendix 1.

Findings of Team 1

Team 1 investigated the effectiveness of medications, categorized by antiretroviral class, in achieving HIV suppression. Utilizing survival analysis, they assessed the time between the initiation of ART to the first occurrence of viral suppression, defined as VL below 1000 copies/mL [36]. They also assessed the time to CD4 cell count exceeding 500 cells/mm^3 [51], which indicates a healthy immunological status.

With Cox proportional hazards models [52] featuring time-varying covariates, the team identified particular antiretroviral agents associated with viral suppression. These findings were purely associative due to data set limitations, which did not account for factors such as age, socioeconomic status, comorbidities, and concurrent medications (of other illnesses).

Findings of Team 2

Team 2 focused on predicting the necessity of altering an individual's ART regimen over a 5-year time span, factoring in disease flare-ups, resistance, or side effects. They formulated a "sliding search" function that generated individual records for each 12-month period, with predictions for antiretroviral modification and adherence to therapy in the subsequent year by using neural networks. The team's methodology produced promising results, with an accuracy rate of 78% in predicting antiretroviral modification and 93% in predicting adherence to therapy. The algorithm detected trends in CD4 and VL results across the 12-month periods, which appeared to be the key predictive features. In addition, the team suggested that there could be potential benefits from exploring recurrent neural networks (eg, long short-term memory [53]).

Serving as UNSW Coursework Materials

Beyond their utility in the Datathon, our synthetic data sets contribute to UNSW courses in the Master of Science in Health Data Science Program [54], namely, HDAT9800 Visualization & Communication and HDAT9510 Machine Learning II.

HDAT9800 teaches future health data scientists the skills to visually communicate complex data effectively to diverse audiences. The course emphasizes the significance of clear data visualization and advocates for transparency and reproducibility in scientific work. It employs R [55] and Python [56] to demonstrate best practices in data analysis and visualization. Our synthetic data sets provide rich resources to enhance the learning in this setting. For instance, Marchesi et al [57] used our data sets to present patient states via t-distributed stochastic neighbor embedding visualization techniques [58].

Meanwhile, HDAT9510 explores advanced modern ML algorithms and methods such as convolutional neural networks [59], autoencoders [60], and reinforcement learning (RL) [61]. As the synthetic data sets consist of time-series variables, students can develop both feedforward and recurrent neural networks. See example models built using our data set in Marchesi et al [57] with recurrent neural networks and even decision trees [62] and hidden Markov models [63], as in a similar data set suggested by Wu et al [64]. Furthermore, with the presence of nonnumeric variables, students can learn about embedding [65]—transforming nonnumeric levels into real-valued vectors so that similar levels that are closer in the vector space carry more analogous meaning. The presence of missing data in the synthetic data sets also encourages students to formulate plausible assumptions about the structure of the clinical data set prior to data modelling.

We provide 3 adaptable worked examples using our ART for HIV data set, suitable for workshops and lectures. The associated codes for the worked examples can be found in Section A of Multimedia Appendix 1. Our synthetic data set supports a variety of student engagements, from understanding complex data structures to developing advanced RL algorithms for optimizing clinical interventions. Moreover, the low patient disclosure risk associated with our data sets (refer to Section B in Multimedia Appendix 1) eliminates the need for ethics approval [66]. This makes these data sets ideal for a range of settings—from small seminars to larger lecture groups.

Worked Example 1

The first exercise, focused on data visualization using Python, compares VL trends over time among patients who commenced their ART with different base drug combos, against the general trend in all patients. The results of our worked example are depicted in Figure 4.

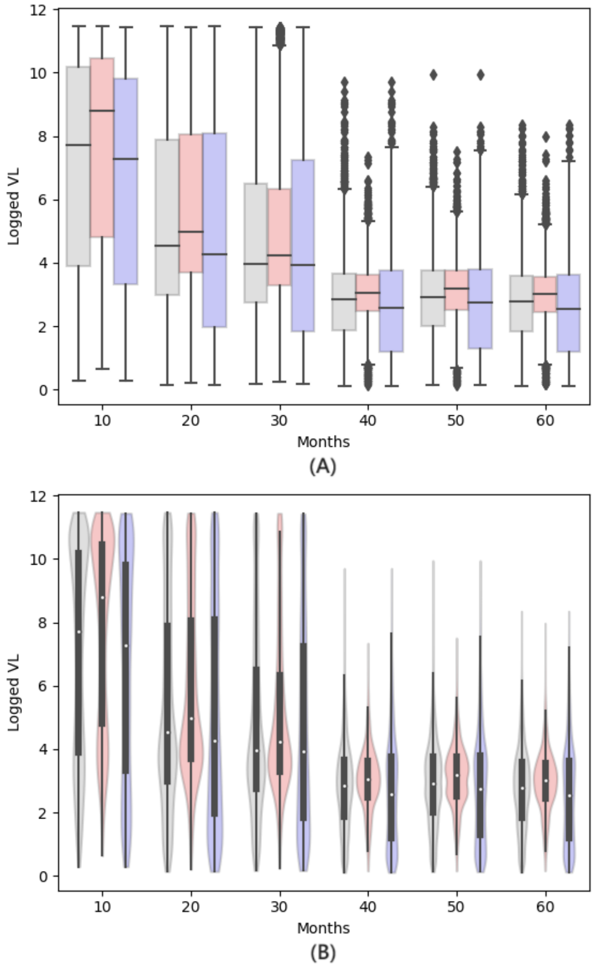
This multifaceted exercise requires students to create sub–data sets based on specific starting base drug combos (ie, FTC + TDF [emtricitabine + tenofovir disoproxil fumarate] and 3TC + ABC [lamivudine + abacavir]), extract data for defined

XSL•FO
RenderX

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

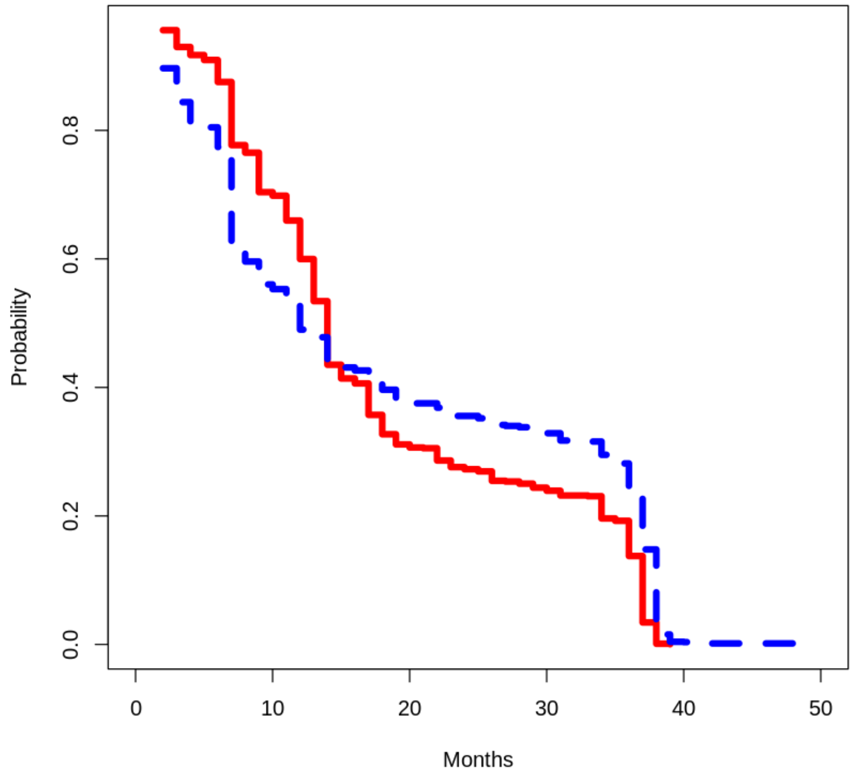
The second exercise delves into survival analysis using R [55], building on insights from the initial data visualization task. The exercise continues to compare results among people starting with the base drug combo of FTC + TDF and those initiating with the base drug combo of 3TC + ABC. The goal is 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.

XSL•FO
RenderX

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.

This complex exercise is designed for advanced students, posing challenges across multiple dimensions. It commences with data wrangling, where students scrutinize numeric variable distributions and evaluate the necessity for transformations such as rescaling, normalization [71], power transformation [72], or Box-Cox transformation [73].

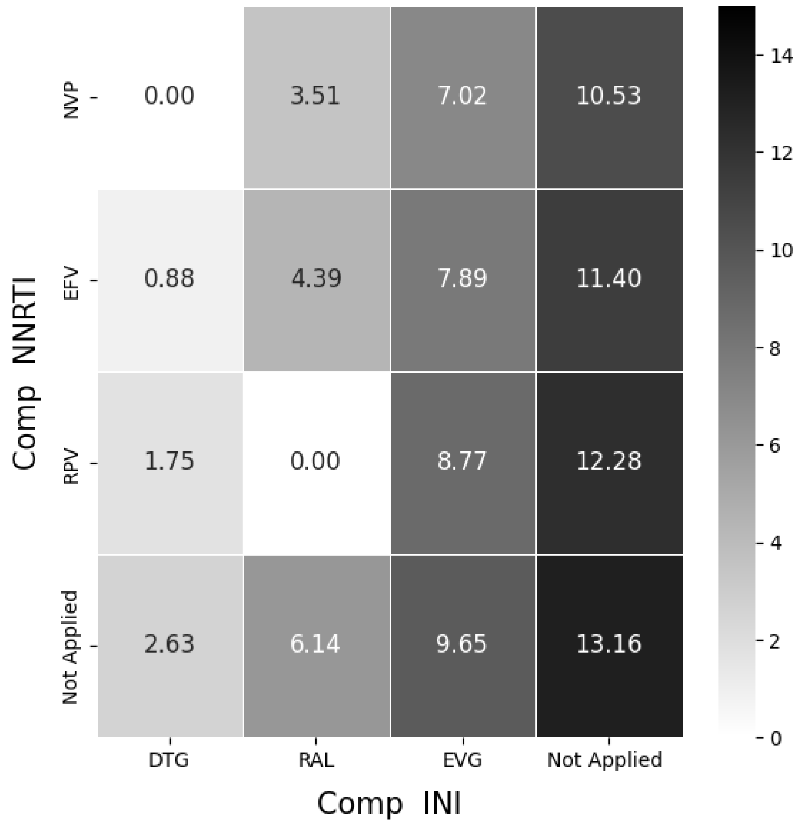
In the next stage, students encounter categorical feature representation for medication regimens, practicing their skills in implementing embeddings. Advanced students can explore transfer learning for feature representation [74]. This exercise also presents real-world challenges, requiring students to handle mixed-type data progression. During the model fitting phase, students must employ suitable ML models, distinguishing between RL method archetypes [75] and considering their clinical implications.

Data visualization is the next task, encouraging students to articulate model-derived insights into digestible visuals for a diverse audience. The concluding phase involves refining assumptions and model performance, incorporating multiple tests to identify optimal hyperparameters [76]. Here, students peek into the "black box" nature of ML and gain an intuition for effective module combinations [77-79]. This step becomes critical for causal inference tasks that necessitate rigorous input data validation [80].

Figure 6 showcases the strategy employed by an RL agent in HIV therapy. Heatmaps visualize the relative frequencies of chosen actions (ie, the selected antiretroviral), where each tile represents a unique action and its frequency as a proportion of all actions. The example output shows that the RL agent consistently suggests the EFV + RAL (efavirenz + raltegravir)—a combination of comp NNRTIs and comp INIs—4.39% of the time, while never recommending the RPV + RAL (rilpivirine + raltegravir) combination. More information on the steps taken to create the output for this task can be found in Section F of Multimedia Appendix 1.

XSL•FO
RenderX

Figure 6. Visualizing the learned reinforcement learning policy. Comp INI: complementary integrase inhibitor; Comp NNRTI: complementary nonnucleoside reverse transcriptase inhibitor; DTG: dolutegravir; EFV: efavirenz; EVG: elvitegravir; NVP: nevirapine; RAL: raltegravir; RPV: rilpivirine.



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

XSL•FO
RenderX

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.

XSL•FO
RenderX

[DOCX File , 38 KB - mededu_v10i1e51388_app1.docx]

References

1. Alsuliman T, Humaidan D, Sliman L. Machine learning and artificial intelligence in the service of medicine: necessity or potentiality? Curr Res Transl Med 2020 Nov;68(4):245-251. [doi: 10.1016/j.retram.2020.01.002] [Medline: 32029403]
2. Naseem M, Akhund R, Arshad H, Ibrahim MT. Exploring the potential of artificial intelligence and machine learning to combat COVID-19 and existing opportunities for LMIC: a scoping review. J Prim Care Community Health 2020;11:2150132720963634 [FREE Full text] [doi: 10.1177/2150132720963634] [Medline: 32996368]
3. Wood D. Wicked problems: using data for better public policy. The Australian Parliamentary Budget Office. URL: https://www.pbo.gov.au/sites/default/files/2023-03/PBO%20Conference_Danielle%20Wood_Data%20and%20wicked%20problems.pdf [accessed 2023-12-26]
4. Jin X, Gallego Luxan B, Hanly M, Pratt NL, Harris I, de Steiger R, et al. Estimating incidence rates of periprosthetic joint infection after hip and knee arthroplasty for osteoarthritis using linked registry and administrative health data. The Bone & Joint Journal 2022 Sep 01;104-B(9):1060-1066. [doi: 10.1302/0301-620x.104b9.bjj-2022-0116.r1]
5. Barbieri S, Mehta S, Wu B, Bharat C, Poppe K, Jorm L, et al. Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. Int J Epidemiol 2022 Jun 13;51(3):931-944 [FREE Full text] [doi: 10.1093/ije/dyab258] [Medline: 34910160]
6. Feng YZ, Liu S, Cheng ZY, Quiroz JC, Rezazadegan D, Chen PK, et al. Severity assessment and progression prediction of COVID-19 patients based on the LesionEncoder framework and chest CT. J Med Internet Res Preprint posted online on March 18, 2021. [doi: 10.2196/preprints.28903]
7. Bayer J, Spark J, Krcmar M, Formica M, Gwyther K, Srivastava A, et al. The SPEAK study rationale and design: A linguistic corpus-based approach to understanding thought disorder. Schizophr Res 2023 Sep;259:80-87. [doi: 10.1016/j.schres.2022.12.048] [Medline: 36732110]
8. Bachmann N, Von Siebenthal C, Vongrad V, Turk T, Neumann K, Beerenwinkel N, et al. Determinants of HIV-1 reservoir size and long-term dynamics during suppressive ART. Nature Communications 2019 Jul 19:1. [doi: 10.1101/19013763]
9. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet 2019;10:524 [FREE Full text] [doi: 10.3389/fgene.2019.00524] [Medline: 31214249]
10. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med 2006;57:575-590. [doi: 10.1146/annurev.med.57.121304.131257] [Medline: 16409167]
11. O'Keefe CM, Connolly CJ. Privacy and the use of health data for research. Med J Aust 2010 Nov 01;193(9):537-541. [doi: 10.5694/j.1326-5377.2010.tb04041.x] [Medline: 21034389]
12. Bentzen HB, Castro R, Fears R, Griffin G, Ter Meulen V, Ursin G. Remove obstacles to sharing health data with researchers outside of the European Union. Nat Med 2021 Aug;27(8):1329-1333 [FREE Full text] [doi: 10.1038/s41591-021-01460-0] [Medline: 34345050]
13. de Oliveira Costa J, Bruno C, Schaffer AL, Raichand S, Karanges EA, Pearson S. The changing face of Australian data reforms: impact on pharmacoepidemiology research. Int J Popul Data Sci 2021 Apr 15;6(1):1418 [FREE Full text] [doi: 10.23889/ijpds.v6i1.1418] [Medline: 34007904]
14. Pearson S, Pratt N, de Oliveira Costa J, Zoega H, Laba T, Etherton-Beer C, et al. Generating real-world evidence on the quality use, benefits and safety of medicines in Australia: history, challenges and a roadmap for the future. IJERPH 2021 Dec 18;18(24):13345. [doi: 10.3390/ijerph182413345]
15. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019 Jun 19;6(1):1. [doi: 10.1186/s40537-019-0217-0]
16. Data availability and transparency bill 2022. Australian Parliament House. URL: https://www.aph.gov.au/Parliamentary_Business/Bills_LEGislation/Bills_Search_Results/Result?bId=r6649 [accessed 2023-12-26]
17. The Five Safes framework. Australian Bureau of Statistics. URL: http://tinyurl.com/4t3nnxpf [accessed 2023-12-26]
18. Miller S, Hughes D. The Quant Crunch: how the demand for data science skills is disrupting the job market. Business-Higher Education Forum. 2017. URL: https://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market [accessed 2023-12-26]
19. Columbus L. IBM predicts demand for data scientists will soar 28% by 2020. Forbes. 2017 May 13. URL: https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/?sh=7fe27cff7e3b [accessed 2023-12-26]
20. Kolaczyk ED, Wright H, Yajima M. Statistics practicum: placing 'practice' at the center of data science education. Harvard Data Science Review 2021 Jan 29:1. [doi: 10.1162/99608f92.2d65fc70]
21. Kuo NIH, Polizzotto MN, Finfer S, Garcia F, Sönnerborg A, Zazzi M, et al. The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. Sci Data 2022 Nov 11;9(1):693 [FREE Full text] [doi: 10.1038/s41597-022-01784-7] [Medline: 36369205]
22. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020 Oct 22;63(11):139-144. [doi: 10.1145/3422622]

XSL•FO
RenderX

23. Arjovsky M, Chintala S, Bottou L. Wasserstein Generative Adversarial Networks. 2017 Presented at: International Conference on Machine Learning; August 6; Sydney, Australia URL: https://proceedings.mlr.press/v70/arjovsky17a.html
24. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. 2017 Dec 12 Presented at: Neural Information Processing Systems; 2017; Long Beach, California.
25. Kuo NIH. The Health Gym. HealthGym.ai. URL: https://healthgym.ai/ [accessed 2023-12-26]
26. Nic5472K / ScientificData2021_HealthGym. GitHub. URL: https://github.com/Nic5472K/ScientificData2021_HealthGym [accessed 2023-12-27]
27. Rosen L. Open Source Licensing: Software Freedom and Intellectual Property Law. 2004 Jul 01. URL: https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/R050225R.pdf [accessed 2023-12-27]
28. Graduate certificate in Health Data Science. The University of New South Wales. URL: https://www.unsw.edu.au/study/postgraduate/graduate-certificate-in-health-data-science?studentType=Domestic [accessed 2023-12-27]
29. CBDRH Health Data Science Datathon 2023. GitHub. URL: https://cbdrh-hds-datathon-2023.github.io/ [accessed 2023-12-27]
30. Public release of clinical information: guidance document. Government of Canada. URL: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html, [accessed 2023-12-27]
31. Clinical data publication. European Medicines Agency. URL: https://www.ema.europa.eu/en/human-regulatory-overview/marketing-authorisation/clinical-data-publication#:~:text=The%20Agency%20intends%20to%20gradually,%3A%2014%2D15%20December%202022 [accessed 2023-12-27]
32. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
33. Zazzi M, Incardona F, Rosen-Zvi M, Prosperi M, Lengauer T, Altmann A, et al. Predicting response to antiretroviral treatment by machine learning: the EuResist project. Intervirology 2012;55(2):123-127 [FREE Full text] [doi: 10.1159/000332008] [Medline: 22286881]
34. Prosperi MCF, Rosen-Zvi M, Altmann A, Zazzi M, Di Giambenedetto S, Kaiser R, et al. Correction: antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models. PLoS ONE 2011 Apr 26;6(4):1. [doi: 10.1371/annotation/d0254103-21b9-4078-836b-57ba5bd1c26a]
35. Parbhoo S, Bogojeska J, Zazzi M, Roth V, Doshi-Velez F. Combining kernel and model based learning for HIV therapy selection. AMIA Jt Summits Transl Sci Proc 2017;2017:239-248 [FREE Full text] [Medline: 28815137]
36. Consolidated 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 [accessed 2023-12-27]
37. Bennett DE, Bertagnolio S, Sutherland D, Gilks CF. The World Health Organization's global strategy for prevention and assessment of HIV drug resistance. Antiviral Therapy 2008 Feb 01;13(2_suppl):1-13. [doi: 10.1177/135965350801302s03]
38. Tang MW, Liu TF, Shafer RW. The HIVdb system for HIV-1 genotypic resistance interpretation. Intervirology 2012;55(2):98-101 [FREE Full text] [doi: 10.1159/000331998] [Medline: 22286876]
39. Fox MP, Brennan AT, Nattey C, MacLeod WB, Harlow A, Mlisana K, et al. Delays in repeat HIV viral load testing for those with elevated viral loads: a national perspective from South Africa. J Int AIDS Soc 2020 Jul;23(7):e25542 [FREE Full text] [doi: 10.1002/jia2.25542] [Medline: 32640101]
40. Hill AL, Rosenbloom DIS, Goldstein E, Hanhauser E, Kuritzkes DR, Siliciano RF, et al. Real-time predictions of reservoir size and rebound time during antiretroviral therapy interruption trials for HIV. PLoS Pathog 2016 Apr;12(4):e1005535 [FREE Full text] [doi: 10.1371/journal.ppat.1005535] [Medline: 27119536]
41. What's new in treatment monitoring: viral load and CD4 testing. World Health Organisation. URL: https://www.who.int/publications/i/item/WHO-HIV-2017.22 [accessed 2023-12-27]
42. NSW HIV strategy 2021-2025. New South Wales Health. URL: https://www.health.nsw.gov.au/endinghiv/Pages/nsw-hiv-strategy-2021-2025.aspx [accessed 2023-12-27]
43. Kuo NIH, Garcia F, Sönnerborg A, Böhm M, Kaiser R, Zazzi M, EuResist Network study group, et al. Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. J Biomed Inform 2023 Aug;144:104436 [FREE Full text] [doi: 10.1016/j.jbi.2023.104436] [Medline: 37451495]
44. Mckinney W. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference (SciPy 2010) 2010:1. [doi: 10.25080/majora-92bf1922-00a]
45. van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng 2011 Mar;13(2):22-30. [doi: 10.1109/mcse.2011.37]
46. Kuo NIH. The Heath Gym synthetic HIV dataset. Figshare. URL: https://figshare.com/articles/dataset/The_Heath_Gym_Synthetic_HIV_Dataset/19838470 [accessed 2023-12-27]
47. Kuo NIH. The Health Gym v2.0 synthetic antiretroviral therapy (ART) for HIV dataset. Figshare. URL: https://figshare.com/articles/dataset/The_Health_Gym_v2_0_Synthetic_Antiretroviral_Therapy_ART_for_HIV_Dataset/22827878 [accessed 2023-12-27]

XSL•FO
RenderX

48. Datathon highlights. CBDRH Health Data Science Datathon 2023. URL: https://cbdrh-hds-datathon-2023.github.io/review.html [accessed 2023-12-27]
49. Sydney local health district. New South Wales Health. URL: https://slhd.health.nsw.gov.au/ [accessed 2023-12-27]
50. de Oliveira Costa J, Lau S, Medland N, Gibbons S, Schaffer AL, Pearson S. Potential drug-drug interactions due to concomitant medicine use among people living with HIV on antiretroviral therapy in Australia. Br J Clin Pharmacol 2023 May;89(5):1541-1553. [doi: 10.1111/bcp.15614] [Medline: 36434744]
51. Garcia SAB, Guzman N. Acquired immune deficiency syndrome CD4+ count. StatPearls 2023 Aug 14:1. [Medline: 30020661]
52. Cox DR. Regression models and life tables. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]
53. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
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?cq_plac=&studentType=Domestic [accessed 2023-12-27]
55. R: a language and environment for statistical computing. R-Project. URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2023-12-28]
56. Van Rossum G. Python tutorial. Centrum Wiskunde & Informatica Institutional Repository. 1995. URL: https://ir.cwi.nl/pub/5007 [accessed 2023-12-27]
57. Marchesi R, Micheletti N, Jurman G, Osmani V. Mitigating health data poverty: generative approaches versus resampling for time-series clinical data. ArXiv Preprint posted online on October 26, 2022. [doi: 10.48550/arXiv.2210.13958]
58. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf [accessed 2023-12-27]
59. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
60. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal 2004 Jun 17;37(2):233-243. [doi: 10.1002/aic.690370209]
61. Sutton R, Barto A. Reinforcement learning: an introduction. IEEE Trans Neural Netw 1998 Sep;9(5):1054-1064. [doi: 10.1109/tnn.1998.712192]
62. Winterfeldt D, Edwards W. Decision Analysis and Behavioral Research. Cambridge, Massachusetts, USA: Cambridge University Press; Aug 26, 1986.
63. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Statist 1966 Dec;37(6):1554-1563. [doi: 10.1214/aoms/1177699147]
64. Wu M, Hughes M, Parbhoo S, Zazzi M, Roth V, Doshi-Velez F. Beyond sparsity: tree regularization of deep models for interpretability. In: AAAI. 2018 Presented at: AAAI Conference on Artificial Intelligence; April 25; Chicago, Illinois, USA. [doi: 10.1609/aaai.v32i1.11501]
65. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv Preprint posted online on September 7, 2013 [FREE Full text] [doi: 10.48550/arXiv.1301.3781]
66. Barnett AG, Campbell MJ, Shield C, Farrington A, Hall L, Page K, et al. The high costs of getting ethical and site-specific approvals for multi-centre research. Res Integr Peer Rev 2016;1:16 [FREE Full text] [doi: 10.1186/s41073-016-0023-6] [Medline: 29451546]
67. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng 2007 May;9(3):90-95. [doi: 10.1109/mcse.2007.55]
68. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics 1988 Mar;44(1):175. [doi: 10.2307/2531905]
69. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958 Jun;53(282):457-481. [doi: 10.1080/01621459.1958.10501452]
70. Therneau TM, Lumley T, Elizabeth A, Cynthia C. survival: Survival Analysis. The Comprehensive R Archive Network. 2015. URL: https://cran.r-project.org/web/packages/survival/index.html [accessed 2023-12-27]
71. Patro SGK, Sahu KK. Normalization: a preprocessing stage. ArXiv Preprint posted online on March 19, 2015 [FREE Full text] [doi: 10.17148/iarjset.2015.2305]
72. Caroll RJ, Ruppert D. On prediction and the power transformation family. Biometrika 1981;68(3):609-615. [doi: 10.1093/biomet/68.3.609]
73. Box GEP, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;26(2):211-243. [doi: 10.1111/j.2517-6161.1964.tb00553.x]
74. Bengio Y. Deep learning of representations for unsupervised and transfer learning. 2011 Presented at: International Conference on Machine Learning Unsupervised and Transfer Learning Workshop; July 2; Bellevue, Washington, USA URL: https://proceedings.mlr.press/v27/bengio12a/bengio12a.pdf
75. Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: tutorial, review, and perspectives on open problems. ArXiv Preprint posted online on November 1, 2020 [FREE Full text] [doi: 10.48550/arXiv.2005.01643]

XSL•FO
RenderX

76. Bergstra J, Yamins D, Cox DD. Making a science of model search. 2013 Jun 21 Presented at: International Conference on Machine Learning; June 21; Atlanta, USA.
77. Kuo NIH. Understanding and modifying dynamical Hopfield neural networks for generating multiple coherent patterns [PhD thesis]. The University of Auckland. 2017. URL: https://researchspace.auckland.ac.nz/handle/2292/34849 [accessed 2023-12-27]
78. Kuo NIH, Harandi M, Fourrier N, Walder C, Ferraro G, Suominen H. An input residual connection for simplifying gated recurrent neural networks. 2020 Presented at: International Joint Conference on Neural Networks; July 19; Glasgow, United Kingdom. [doi: 10.1109/ijcnn48605.2020.9207238]
79. Kuo NIH, Harandi M, Fourrier N, Walder C, Ferraro G, Suominen H. Plastic and stable gated classifiers for continual learning. 2021 Presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; June 19; Online. [doi: 10.1109/cvprw53098.2021.00394]
80. Walker AR, Luque D, Le Pelley ME, Beesley T. The role of uncertainty in attentional and choice exploration. Psychon Bull Rev 2019 Dec;26(6):1911-1916. [doi: 10.3758/s13423-019-01653-2] [Medline: 31429060]
81. Socioeconomic indexes for areas. Australian Bureau of Statistics. URL: https://www.abs.gov.au/websitedbs/censushome.nsf/home/seifa [accessed 2023-12-27]
82. Chronic conditions and multimorbidity. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/reports/australias-health/chronic-conditions-and-multimorbidity [accessed 2023-12-27]
83. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res 2016;8(3):1560-1580 [FREE Full text] [Medline: 27186282]
84. Corley DA, Jensen CD, Marks AR, Zhao WK, de Boer J, Levin TR, et al. Variation of adenoma prevalence by age, sex, race, and colon location in a large population: implications for screening and quality programs. Clinical Gastroenterology and Hepatology 2013 Feb;11(2):172-180. [doi: 10.1016/j.cgh.2012.09.010]
85. Earnshaw VA, Bogart LM, Dovidio JF, Williams DR. Stigma and racial/ethnic HIV disparities: moving toward resilience. Am Psychol 2013;68(4):225-236 [FREE Full text] [doi: 10.1037/a0032705] [Medline: 23688090]
86. Datasets - CHeReL. Centre for Health Record Linkage. URL: https://www.cherel.org.au/datasets [accessed 2023-12-27]
87. Privacy act 1988. The Australian Government Federal Register of Legislation. URL: https://www.legislation.gov.au/Details/C2014C00076 [accessed 2023-12-27]
88. Health information privacy. The US Department of Health & Human Services. URL: https://www.hhs.gov/hipaa/index.html [accessed 2023-12-27]
89. Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association 1969 Dec;64(328):1183-1210. [doi: 10.1080/01621459.1969.10501049]
90. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol 2002 Dec;31(6):1246-1252. [doi: 10.1093/ije/31.6.1246] [Medline: 12540730]
91. Lujic S, Randall DA, Simpson JM, Falster MO, Jorm LR. Interaction effects of multimorbidity and frailty on adverse health outcomes in elderly hospitalised patients. Sci Rep 2022 Aug 19;12(1):14139 [FREE Full text] [doi: 10.1038/s41598-022-18346-x] [Medline: 35986045]
92. Han J, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ 2022 Dec 01;22(1):830 [FREE Full text] [doi: 10.1186/s12909-022-03898-3] [Medline: 36457086]
93. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-12-27]
94. An important next step on our AI journey. Google. URL: https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/ [accessed 2023-12-27]
95. 'We are a little bit scared': OpenAI CEO warns of risks of artificial intelligence. The Guardian. URL: https://www.theguardian.com/technology/2023/mar/17/openai-sam-altman-artificial-intelligence-warning-gpt4 [accessed 2023-12-27]
96. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 06;9:e48163 [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
97. Lembani R, Gunter A, Breines M, Dalu MTB. The same course, different access: the digital divide between urban and rural distance education students in South Africa. Journal of Geography in Higher Education 2019 Nov 22;44(1):70-84. [doi: 10.1080/03098265.2019.1694876]
98. van de Werfhorst HG, Kessenich E, Geven S. The digital divide in online education: Inequality in digital readiness of students and schools. Computers and Education Open 2022 Dec;3:100100. [doi: 10.1016/j.caeo.2022.100100]
99. Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. Artif Intell Med 2020 Sep;109:101938. [doi: 10.1016/j.artmed.2020.101938] [Medline: 34756215]
100. Armanious K, Jiang C, Fischer M, Küstner T, Hepp T, Nikolaou K, et al. MedGAN: Medical image translation using GANs. Comput Med Imaging Graph 2020 Jan;79:101684. [doi: 10.1016/j.compmedimag.2019.101684] [Medline: 31812132]
101. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Applied Sciences 2021 Sep 08;11(18):8319. [doi: 10.3390/app11188319]

XSL•FO
RenderX

102. Kuo NIH, Garcia F, Sonnerborg A, Bohm M, Kaiser R, Zazzi M, et al. Synthetic health-related longitudinal data with mixed-type variables generated using diffusion models. ArXiv Preprint posted online on March 22, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.12281]
103. Kingma DP, Welling M. Auto-encoding variational Bayes. ArXiv Preprint posted online on December 10, 2022 [FREE Full text] [doi: 10.48550/arXiv.1312.6114]
104. Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. ArXiv Preprint posted online on November 18, 2015. [doi: 10.48550/arXiv.1503.03585]
105. OpenAI. GPT-4 technical report. ArXiv. Preprint posted online on December 19, 2023 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]

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

Edited by G Eysenbach, K Venkatesh, MN Kamel Boulos; submitted 30.07.23; peer-reviewed by S Seevanayanagam, M Black; comments to author 14.10.23; revised version received 20.10.23; accepted 08.11.23; published 16.01.24.

Please cite as:
Kuo NIH, Perez-Concha O, Hanly M, Mnatzaganian E, Hao B, Di Sipio M, Yu G, Vanjara J, Valerie IC, de Oliveira Costa J, Churches T, Lujic S, Hegarty J, Jorm L, Barbieri S
Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project
JMIR Med Educ 2024;10:e51388
URL: https://mededu.jmir.org/2024/1/e51388
doi:10.2196/51388
PMID:38227356

©Nicholas I-Hsien Kuo, Oscar Perez-Concha, Mark Hanly, Emmanuel Mnatzaganian, Brandon Hao, Marcus Di Sipio, Guolin Yu, Jash Vanjara, Ivy Cerelia Valerie, Juliana de Oliveira Costa, Timothy Churches, Sanja Lujic, Jo Hegarty, Louisa Jorm, Sebastiano Barbieri. 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.

XSL•FO
RenderX

Original Paper

Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study

Firas Haddad[1], BSc; Joanna S Saade[2], MD

[1]Faculty of Medicine, American University of Beirut, Beirut, Lebanon

[2]Department 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

XSL•FO
RenderX

quite famous for its outstanding ability to answer questions and assist in many tasks [4].

Medical education relies highly on standardized multiple-choice examinations to test medical students in an objective and consistent way. Ophthalmologists in the United States pass through the United States Medical Licensing Examination (USMLE) steps 1, 2, and 3, the Ophthalmic Knowledge Assessment Program (OKAP), and the Board of Ophthalmology (OB) Written Qualifying Examination (WQE) by the time they become practicing physicians. Undergraduate and graduate medical students rely on different tools available to prepare for these examinations.

One limitation of the current tools for medical education is the lack of personalization. Question banks used today do not tailor their explanations to users; rather, they present one explanation for each question to all its users. ChatGPT and other LLMs, if proven to be accurate in their ability to answer questions, can provide robust explanations to users, and users can then ask specific questions they need further clarification on. This can be very helpful and educational for users as it can tailor to the needs of each user and help them fill specific knowledge gaps they may have. Additionally, the GPT-3.5 model is freely available to everyone, while GPT-4.0 is available at a premium. As such, it is essential to compare these models to assess whether GPT-4.0's hypothetical increased abilities justify the price of the membership.

The question of how ChatGPT can be integrated for use in medical education has emerged. With the complexity of ophthalmology, the ability of ChatGPT to accurately answer ophthalmology questions could be of significant value to medical students and residents preparing for the USMLE, OKAP, and OB-WQE. It is also important to compare the performance of both GPT-4.0 and GPT-3.5, since GPT-4.0 is marketed as a more intelligent version of its predecessor.

Therefore, the aim of this study is to evaluate the performance of ChatGPT on ophthalmology questions from USMLE steps 1, 2, and 3, the OKAP, and the OB-WQE using both GPT-3.5 and GPT-4.0. We hypothesize that ChatGPT's responses are comparable to those of human experts in the field, and that GPT-4.0 performs better than GPT-3.5. The results of this study could have implications for the future use of ChatGPT in medical education and training, and for the development of more efficient and effective tools for examination preparation.

Methods

Data Sets

Different data sets were used for the different examinations due to the lack of a central service for all examinations. Questions that included pictures or tables were automatically excluded and were not queried on ChatGPT. AMBOSS [5], a question bank and popular resource for the USMLE was used for steps 1, 2, and 3. A total of 44 questions were included for step 1, 60 for step 2, and 28 for step 3. AMBOSS highlights the difficulty of each question and the percentage of people who chose each answer choice. This allowed us to compare the performance of ChatGPT to the general population [5]. For the OKAP and OB-WQE, 248 questions across the different chapters were taken from *Ophthalmology Board Review Q&A* by Glass et al [6].

Prompt Engineering

The style and the prompt of the questions asked to ChatGPT have been shown to have an impact on the answer given. To standardize the process of asking the questions to ChatGPT, questions were all formatted in the same way on Word (Microsoft Corp). After removing questions with pictures or tables, the questions were formatted in the manner described by Gilson et al [7]. The question stem was consolidated in 1 paragraph, and then each answer choice was placed on a separate line. Furthermore, the answer choices were separated by 2 empty lines from the main question stem; this was done to optimize the accuracy of the results, avoiding any effect the question format may have on ChatGPT's ability. An example prompt is shown in Textbox 1.

Textbox 1. An example of a prompt (written by the authors).

Question: What medical discipline deals with conditions of the eye

A. Dermatology

B. Endocrinology

C. Ophthalmology

D. Rheumatology

Question Input

All questions were input in ChatGPT on March 5, 2023, for GPT-3.5 and April 15, 2023, for GPT-4.0. We then used Excel (Microsoft Corp) spreadsheets to record whether the answer was correct or not, the percentage of users getting the answer correct (if applicable), the difficulty level (if applicable), and the topic (if applicable).

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

XSL•FO
RenderX

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.

Examination	Correct answers provided by models[a], n (%)		*P* value
	GPT-3.5	GPT-4.0	
USMLE[b] step 1	33 (75)	31 (70.45)	.81
USMLE step 2	44 (73.33)	56 (90.32)	.01
USMLE step 3	17 (60.71)	27 (96.43)	.004
OB-WQE[c]	116 (46.77)	156 (62.90)	<.001

[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.

XSL•FO
RenderX

Table 2. Performance of GPT-3.5 and GPT-4.0 according to different difficulty levels.

GPT-4.0					GPT-3.5				
Board of Ophthalmology difficulty level	Correct answers[a], n (%)	AMBOSS[b]			Board of Ophthalmology difficulty level	Correct answers[c], n (%)	AMBOSS[d]		
		Difficulty level	ChatGPT's performance (correct answers), n (%)	Human performance (correct answers), %			Difficulty level	ChatGPT's performance (correct answers), n (%)	Human performance (correct answers), %
1	49 (76)	1	19 (100)	83	1	34 (53)	1	14 (88)	83
2	73 (59)	2	43 (91)	68	2	54 (44.26)	2	36 (77)	68
3	35 (56)	3	38 (84)	53	3	28 (45.16)	3	28 (63)	53
N/A[e]	N/A	4	10 (59)	37	N/A	N/A	4	12 (60)	37
N/A	N/A	5	4 (66.67)	26	N/A	N/A	5	3 (50)	26

[a] *P*=.04 on comparing the performance of GPT-4.0 across different difficulty levels.
[b] *P*=.003 on comparing the performance of GPT-4.0 across different difficulty levels.
[c] *P*=.49 on comparing the performance of GPT-3.5 across different difficulty levels.
[d] *P*=.18 on comparing the performance of GPT-3.5 across different difficulty levels.
[e] N/A: not applicable.

Table 3. Performance of GPT-3.5 and GPT-4.0 on various included topics.

Category	Correct answers by GPT-4.0[a], n (%)	Topic	Correct answers by GPT-3.5[b], n (%)	*P* value
Cornea, external disease, and anterior segment	28 (74)	Cornea, external disease, and anterior segment	25 (66)	.45
Glaucoma	20 (61)	Glaucoma	16 (48)	.32
Lens and cataract	22 (88)	Lens and cataract	8 (32)	<.001[c]
Neuro-ophthalmology	15 (54)	Neuro-ophthalmology	16 (57)	.06
Oculofacial, plastics, and orbit	17 (50)	Oculofacial, plastics, and orbit	10 (29)	.08
Pediatric ophthalmology and strabismus	14 (61)	Pediatric ophthalmology and strabismus	9 (34)	.07
Refractive management and optics	17 (50)	Refractive management and optics	14 (41)	.46
Retina and ocular oncology	24 (73)	Retina and ocular oncology	18 (54)	.12

[a] *P*=.02 for differences in the number of correct answers provided by GPT-4.0 among different categories.
[b] *P*=.03 for differences in the number of correct answers provided by GPT-3.5 among different topics.
[c] Significant at *P*<.05.

Discussion

Principal Findings

Our results indicate that GPT-4.0 is superior to GPT-3.5, and that GPT-3.5 has a below-average accuracy in answering questions correctly. The total proportion of correct answers for GPT-3.5 was 55% (n=210), which is considered a poor performance, while that of GPT-4.0 was 70% (n=270), which is an almost average performance [7]. Students typically must achieve 59%-60% of correct answers to pass, and students perform with an average of around 70%-75% on the aforementioned board examinations [7]. It is interesting to note that GPT-3.5's performance decreased as examination levels increased. This is probably due to the more clinical nature of the examinations. This was not the case for GPT-4.0, which performed best on USMLE steps 2 and 3.

This study investigates the correlation between ChatGPT-3.5 and -4.0 providing a correct answer and the percentage of human users who provided the answer correctly on AMBOSS. For GPT-3.5, a correlation coefficient of 0.21 (*P*=.01) was noted; whereas, this correlation coefficient was –0.31 (*P*<.001) for GPT-4.0. This implies that GPT-4.0 performed better on questions that fewer users answered correctly.

Although our study is limited in that it did not divide the questions into categories such as diagnosis, treatment, basic knowledge, or surgical planning questions. Looking closely at the lens and cataract section in which the model failed (32% of correct answers for GPT-3.5), it was noted that all the correct

XSL•FO
RenderX

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

XSL•FO
RenderX

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

1. Gozalo-Brizuela R, Garrido-Merchan EC. ChatGPT is not all you need. A atate of the art review of large generative AI models. arXiv Preprint posted online January 11, 2023. . [doi: 10.48550/arXiv.2301.04655]
2. Castelvecchi D. Are ChatGPT and AlphaCode going to replace programmers? Nature 2022 Dec 08. [doi: 10.1038/d41586-022-04383-z] [Medline: 36481949]
3. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2023 Oct 05. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
4. Azaria A. ChatGPT usage and limitations. OSF Preprints Preprint posted online December 27, 2022. [doi: 10.31219/osf.io/5ue7n]
5. Powerful learning and clinical tools combined into one platform. AMBOSS. URL: https://www.amboss.com/ [accessed 2023-03-05]
6. Smith BT, Bottini AR. Graefes Arch Clin Exp Ophthalmol 2021 Jul 15;259(8):2457-2458. [doi: 10.1007/s00417-021-05094-3]
7. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
8. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023 Oct 29;13(1):18562 [FREE Full text] [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]
9. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023;20:1 [FREE Full text] [doi: 10.3352/jeehp.2023.20.1] [Medline: 36627845]
10. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation 2023 Apr;185:109732. [doi: 10.1016/j.resuscitation.2023.109732] [Medline: 36775020]
11. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Jun 01;141(6):589-597. [doi: 10.1001/jamaophthalmol.2023.1144] [Medline: 37103928]
12. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun C, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023 Sep;95:104770 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
13. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324 [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
LLM: large language model
NLP: natural language processing
OB: Board of Ophthalmology
OKAP: Ophthalmic Knowledge Assessment Program
USMLE: United States Medical Licensing Examination
WQE: Written Qualifying Examination

Edited by K Venkatesh; submitted 14.07.23; peer-reviewed by A Saxena, Y Wang; comments to author 14.10.23; revised version received 09.12.23; accepted 27.12.23; published 18.01.24.

Please cite as:
Haddad F, Saade JS
Performance of ChatGPT on Ophthalmology-Related Questions Across Various Examination Levels: Observational Study
JMIR Med Educ 2024;10:e50842
URL: https://mededu.jmir.org/2024/1/e50842
doi:10.2196/50842
PMID:38236632

©Firas Haddad, Joanna S Saade. Originally published in JMIR Medical Education (https://mededu.jmir.org), 18.01.2024. 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.

XSL•FO
RenderX

Original Paper

Evaluation of ChatGPT's Real-Life Implementation in Undergraduate Dental Education: Mixed Methods Study

Argyro Kavadella[1], DDS, MSc, PhD; Marco Antonio Dias da Silva[2], DDS, MSc, PhD; Eleftherios G Kaklamanos[1,3,4], DDS, MSc, PhD; Vasileios Stamatopoulos[5], BSc, MSc; Kostis Giannakopoulos[1], DDS, PhD

[1]School of Dentistry, European University Cyprus, Nicosia, Cyprus

[2]Research Group of Teleducation and Teledentistry, Federal University of Campina Grande, Campina Grande, Brazil

[3]School of Dentistry, Aristotle University of Thessaloniki, Thessaloniki, Greece

[4]Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates

[5]Information Management Systems Institute, ATHENA Research and Innovation Center, Athens, Greece

Corresponding Author:
Argyro Kavadella, DDS, MSc, PhD
School of Dentistry
European University Cyprus
6, Diogenes street
Engomi
Nicosia, 2404
Cyprus
Phone: 357 22559620
Email: a.kavadella@euc.ac.cy

Abstract

Background: The recent artificial intelligence tool ChatGPT seems to offer a range of benefits in academic education while also raising concerns. Relevant literature encompasses issues of plagiarism and academic dishonesty, as well as pedagogy and educational affordances; yet, no real-life implementation of ChatGPT in the educational process has been reported to our knowledge so far.

Objective: This mixed methods study aimed to evaluate the implementation of ChatGPT in the educational process, both quantitatively and qualitatively.

Methods: In March 2023, a total of 77 second-year dental students of the European University Cyprus were divided into 2 groups and asked to compose a learning assignment on "Radiation Biology and Radiation Protection in the Dental Office," working collaboratively in small subgroups, as part of the educational semester program of the Dentomaxillofacial Radiology module. Careful planning ensured a seamless integration of ChatGPT, addressing potential challenges. One group searched the internet for scientific resources to perform the task and the other group used ChatGPT for this purpose. Both groups developed a PowerPoint (Microsoft Corp) presentation based on their research and presented it in class. The ChatGPT group students additionally registered all interactions with the language model during the prompting process and evaluated the final outcome; they also answered an open-ended evaluation questionnaire, including questions on their learning experience. Finally, all students undertook a knowledge examination on the topic, and the grades between the 2 groups were compared statistically, whereas the free-text comments of the questionnaires were thematically analyzed.

Results: Out of the 77 students, 39 were assigned to the ChatGPT group and 38 to the literature research group. Seventy students undertook the multiple choice question knowledge examination, and examination grades ranged from 5 to 10 on the 0-10 grading scale. The Mann-Whitney *U* test showed that students of the ChatGPT group performed significantly better (P=.045) than students of the literature research group. The evaluation questionnaires revealed the benefits (human-like interface, immediate response, and wide knowledge base), the limitations (need for rephrasing the prompts to get a relevant answer, general content, false citations, and incapability to provide images or videos), and the prospects (in education, clinical practice, continuing education, and research) of ChatGPT.

Conclusions: Students using ChatGPT for their learning assignments performed significantly better in the knowledge examination than their fellow students who used the literature research methodology. Students adapted quickly to the technological environment of the language model, recognized its opportunities and limitations, and used it creatively and efficiently. Implications for practice:

XSL•FO
RenderX

the study underscores the adaptability of students to technological innovations including ChatGPT and its potential to enhance educational outcomes. Educators should consider integrating ChatGPT into curriculum design; awareness programs are warranted to educate both students and educators about the limitations of ChatGPT, encouraging critical engagement and responsible use.

(JMIR Med Educ 2024;10:e51344) doi:10.2196/51344

KEYWORDS

ChatGPT; large language models; LLM; natural language processing; artificial Intelligence; dental education; higher education; learning assignments; dental students; AI pedagogy; dentistry; university

Introduction

Background

The emergence of ChatGPT (OpenAI) in November 2022 represents the third significant technological breakthrough in information technology impacting education, following the introduction of Web 2.0 over a decade ago [1] and e-learning's surge during the COVID-19 pandemic [2]. ChatGPT is an artificial intelligence (AI) tool that offers benefits and opportunities in higher education including increased student engagement, collaboration, personalized feedback, and accessibility. However, it is characterized by a limited database, posing challenges such as the restricted ability to answer medical questions and the potential for inaccurate and biased responses. There are also concerns regarding legal and ethical implications, plagiarism, and academic integrity [3-5].

The research on AI and its implementation in academic education is a prominent subject; a Google Scholar search for "artificial intelligence and dental education," yielded 100,000 results and approximately 18,000 results for "ChatGPT and higher education" (on June 9, 2023). AI technology has evolved to unprecedented levels, transforming professions, revolutionizing workflows, and reshaping human-machine interactions. ChatGPT, the most recent milestone in natural language processing AI models, has been enabling advanced conversational capabilities and expanding the boundaries of AI-powered communication. Interest in ChatGPT applications encompasses both clinical practice [6,7] and higher education [3,8-11], with promising results.

Relevant Prior Research

Within the higher education landscape, it has been suggested that dental curricula at universities need to be updated due to the AI paradigm shift [9,12,13]. This involves defining a fundamental dental curriculum for both undergraduate and postgraduate levels and establishing learning outcomes related to dental AI [8]. Cotton et al [3] and Halaweh [14] proposed strategies to ensure the ethical and responsible use of AI tools in higher education. Fergus et al [10] evaluated academic answers generated using ChatGPT, and Bearman et al [15] in their review on AI in higher education discussed the shifting dynamics of authority and the relationships among teachers, students, institutions, and technologies. Gimpel et al [16] in their extensive discussion paper proposed guidelines and recommendations for students and lecturers and urged the universities for a multistakeholder dialogue to implement efficient and responsible use of generative AI models in higher education.

Roganovic et al [17] performed a cross-sectional web-based survey among experienced dentists and final-year undergraduate students from the School of Dental Medicine, University of Belgrade, Serbia, to investigate their current perspectives and readiness to accept AI into practice. Responders, especially final-year students, showed a lack of knowledge regarding AI use in medicine and dentistry (only 7.9% of them were familiar with AI use) and were skeptical (only 34% of them believed that AI should be used in dental practice); the underlying reasons were fear of being replaced by AI, as well as a lack of regulatory policies, since students and—at a lesser degree—dentists were concerned that using AI could legally complicate the clinical practice [17].

Chan and Hu [11] reported different results in exploring students' perceptions of generative AI and ChatGPT in teaching and learning through a web-based questionnaire; the study revealed a generally positive attitude toward generative AI, with students demonstrating a good understanding of this technology, its benefits, and limitations, despite its novel public appearance. Generative AI is a special category of AI designed to learn from the characteristics of its input and generate outputs with similar characteristics. In contrast to most AI models that perform specific tasks based on predefined rules and patterns, generative AI models use advanced algorithms to find the underlying patterns of the input data (eg, text, images, sounds, and videos) and "generate" entirely new content of the same type [11]. Students recognized the potential for personalized feedback and learning support, brainstorming, writing assistance, and research capabilities and stated they would integrate technologies like ChatGPT in their studies and future careers, but they were also concerned about becoming overreliant on them. They moreover expressed concerns about data accuracy, privacy, ethical issues, and the impact on personal development [11]. Students' perceptions of the learning environment and the teaching strategies have a significant impact on their approach to learning and the learning outcomes (positive perceptions lead to a deep approach to learning), thus being of pedagogical interest to educators and institutions [11,18]. The influence of AI tools on students' engagement and perceptions was investigated by Nazari et al [19]: they conducted a randomized controlled trial to examine the efficacy of an AI-powered writing tool (Grammarly) for postgraduate students and concluded that students in the intervention group demonstrated significant improvement in engagement (behavioral, emotional, and cognitive), self-efficacy, and academic emotions (positive and negative), domains that address learning behavior, which lead to self-development and underpin authentic pedagogy.

XSL•FO
RenderX

Aims of the Study

Despite numerous publications about AI and large language models (LLMs), the majority involve discussion papers, viewpoint articles, and positions [3,13,16,20,21], with few being exploratory, cross-sectional, or questionnaire-based studies [11,17,19]. To our knowledge, so far, no experimental studies have been identified, wherein ChatGPT was in vivo implemented by students within the teaching process, and the outcomes were comprehensively evaluated.

Therefore, this study aimed to address this gap by implementing ChatGPT within the learning process and conducting a quantitative (differences between examination grades) and qualitative (thematic analysis of the free-text comments of the evaluation questionnaire) evaluation of the outcomes (mixed methods research study).

Methods

Ethical Considerations

The study's research protocol was reviewed and approved by the Vice-Rector for Research and External Affairs and the President of the Institutional Committee on Bioethics and Ethics of the European University Cyprus.

Study Design: Challenges

The study was conceptualized, organized, and refined in February 2023 and realized in March 2023. Of note is that ChatGPT appeared publicly on November 30, 2022; in March 2023, ChatGPT-3.5 was freely available (and was mostly used by the students), whereas ChatGPT-4 had just emerged (few students used this). The study was not a stand-alone research endeavor; instead, it constituted part of students' educational activities embedded within the semester's educational program. As this was the first attempt to implement ChatGPT in the educational process and there were no existing research studies in the literature to refer to, and adding to the limited knowledge on ChatGPT's properties and limitations at the time, the authors encountered various challenges while organizing the research design. Therefore, to anticipate potential issues that could affect student learning or compromise the study's outcomes, they conducted a systematic, forward-looking analysis of the research process, considering each step and taking proactive measures to mitigate any challenges or obstacles that may have arisen.

Study Design: Implementation

The second-year dental students (77 students) of the School of Dentistry, European University Cyprus were randomly divided into 2 large groups and were asked to compose an assignment on "Radiation Biology and Radiation Protection in the Dental Office." The subject of Dentomaxillofacial Radiology is taught through theoretical lectures, laboratory training, and practical training during 2 semesters, and students' learning assignments are embedded within the lectures' program as an alternative to traditional lecturing. Student learning assignments to replace lectures followed by in-class presentation and discussion is a methodology used within the "Dentomaxillofacial Radiology" module whenever the topic is suitable for such an approach. Students usually work collaboratively to perform the assignments by searching the internet for scientific reliable sources and compiling the results into a PowerPoint slide presentation, including the references they used. Students of both groups were asked to work in small subgroups to compose the assignments, where each subgroup would comprise 3-7 students, decided among them. It is worth mentioning that the European University Cyprus School of Dentistry is an English-speaking School, educating students from over 30 countries encompassing different ethnic, educational, and cultural backgrounds; therefore, the study's sample could be considered diverse.

One large group would compose the assignment through literature research (the traditional method for assignments) and the other group would use the ChatGPT tool for the assignment (pose prompts and register the answers), also submitting a slide presentation. Students were given 1 month to deliver the assignment, and they were informed that they would present their presentations in class on a designated day.

Moreover, students of the ChatGPT group were encouraged to experiment with it; ask different questions; ask for videos, images, and internet resources; and in general to be creative, imaginative, and playful while using this new tool. Once they had the final AI content, they were advised to critically evaluate it by comparing it with the relevant content of a reliable scientific resource, such as a textbook or published article, and perform the necessary modifications to the AI output. After finishing the assignment, they were asked to complete an open-ended questionnaire individually (Multimedia Appendix 1), including questions about the usability, problems, opinions, proposals, and so forth, which was emailed to them, and which they would submit to the educator together with the assignment (ie, the PowerPoint presentation).

The AI Evaluation Questionnaire included 12 questions requiring free-text responses and was developed by the authors by combining questions from 2 sources: essays evaluation questionnaires retrieved in the scientific literature [22-24] and the questionnaire ChatGPT produced on the prompt "Can you develop 10 questions for a user to evaluate your performance on writing an essay?" Questions were combined and modified, they were piloted within a small student group other than the research groups, and they were finally amended as necessary. The free-text comments of the AI Evaluation Questionnaire were grouped into main themes and discussed (subjective and qualitative evaluation).

After students completed and submitted their projects via email, and on the designated day they would present the PowerPoint presentations in class, at the beginning of the session, they all had an unannounced blind knowledge examination (answered individually and anonymously, where they only indicated the group they belonged in, so that the educator could not relate the students with the answer sheets). The examination was developed by the authors and consisted of 10 multiple-choice questions (MCQs), which addressed the learning objectives of the topic. They were informed that the knowledge test was intended for the educator to identify whether the assignment had equipped them with the intended knowledge and whether there were any knowledge gaps to address. The results of the

XSL•FO
RenderX

examination (examination grades) were compared among the 2 groups, that is, the literature research group and the ChatGPT group. Statistically significant differences between the groups' grades were explored using the Mann-Whitney nonparametric test. Data analysis was conducted using SPSS (version 25.0; SPSS Inc), and statistical significance was set at P=.05 (objective and quantitative evaluation).

The final study design is summarized as follows:

- Students were randomly divided into 2 large groups (the ChatGPT and the literature research groups) and further into smaller groups.
- Literature research group performed the assignment by searching the internet and delivered it in PowerPoint format, including the references used.
- ChatGPT group (1) asked the LLM relevant queries and developed a PowerPoint presentation; (2) registered and reported on their interactions with ChatGPT, including the prompts and their modifications, the final outcome and its evaluation after comparing it with a reference text or book chapter; and (3) answered the AI Evaluation Questionnaire on their experience with the LLM.
- All students presented their learning assignments in class. At the beginning of this session, they undertook an unannounced knowledge examination of 10 questions.
- Data derived from the knowledge examination grades, the PowerPoint presentations, and the free-text comments of the AI Evaluation Questionnaire.

Results

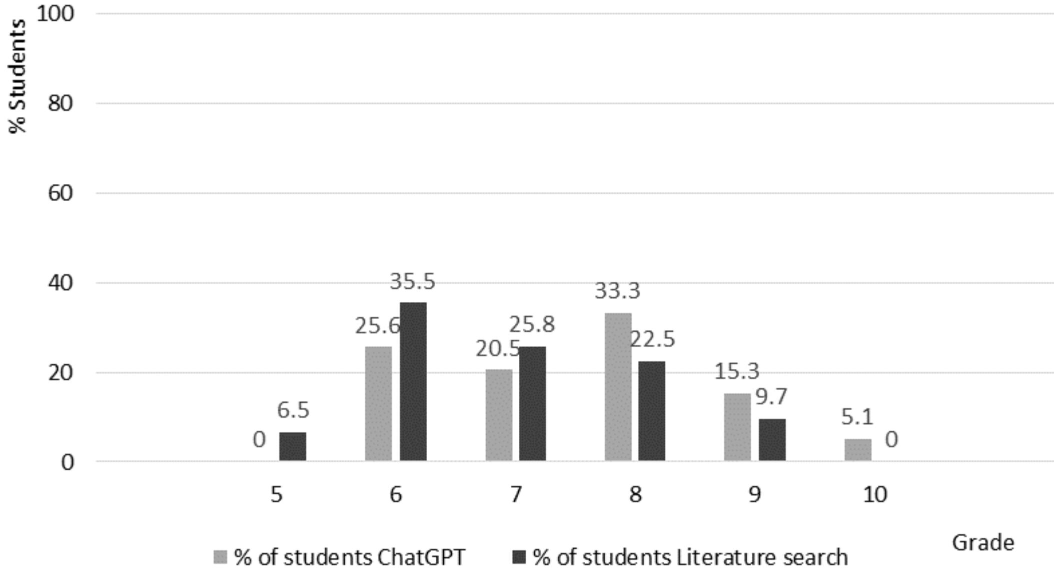
Quantitative Results

Out of the 77 students, 39 were assigned to the ChatGPT group forming 9 subgroups and 38 to the literature research group forming 8 subgroups. Seventy students undertook the MCQ examination (7 students were absent) and examination grades ranged from 5 to 10 on the 0-10 grading scale. Figure 1 presents the number of students (percentages within each group) with their examination grades. We noticed that in the higher range of examination grades, that is, 8-10, the ChatGPT students outperformed the literature research students, while the opposite happened within the lower range of examination grades, that is, 5-7.

To check for differences between the ChatGPT student group and the literature research group, we performed the Mann-Whitney U test, which showed that students of the ChatGPT group (n=39; mean 7.54, SD 1.18) performed significantly better (P=.045) than students in the literature research group (n=31; mean 6.94, SD 1.12).

To foster inclusiveness and avoid discrimination, we deliberately chose not to perform statistical analyses regarding gender differences, as we also believe that gender diversity is not associated with the educational process or the educational outcomes. Education is offered equally to all students and any gender differences possibly found would not differentiate educational approaches for one gender or the other. Instead, we perceive this student cohort as representatives of their generation (Generation Z), a characteristic that is directly related to this study's outcomes and could explain several findings. This concept is in line with the US National Institute of Health recommendations for gender-neutral language [25].

Figure 1. Students' examination grades (% of students within each group).



XSL•FO
RenderX

Qualitative Results

Overview

Out of the 39 students of the ChatGPT group, 31 (80%) students answered the 12 questions of the AI Evaluation Questionnaire. The free-text answers to the questions were grouped into themes and discussed. Three main themes emerged.

Collaboration With ChatGPT and Problems Encountered

Although the majority of students were aware that ChatGPT had surfaced a couple of months ago in the digital world and some of them had already used it, this was the first opportunity they had to actually work with it and "officially" use it within their studies, and they enjoyed and appreciated this opportunity. They characterized it as a "powerful and versatile tool," "intuitive and intelligent," "revolutionary," and "enjoyable to work with" and they thought this experience was "interesting and different from the regular assignments." They stated that learning to use these AI tools would improve their future practice but emphasized that "you have to learn how to properly use it." They appreciated its human-like answers, as these "do not make the user feel distanced from technology." A student stated:

> *In the beginning I was afraid it was going to be too difficult to work with but as I was discussing with it I understood its greatness. I think it really is the future as it can help both education and research. I really did enjoy its human-like answers like when something was wrong it persisted like a human being for its accuracy as well as when it did not answer the question as it should like a lazy student.*

Another student commented: "I enjoyed working with ChatGPT, because I got to learn and understand something that is going to be a part of the future." Humanization of the LLM is worth noting: "He always understood what we wanted." Textbox 1 shows examples of students' prompts.

Textbox 1. Examples of students' prompts to ChatGPT (exact copies).

- How does radiation affect human health?
- What's the difference between deterministic & stochastic effects of radiation?
- Is radiation exposure carcinogenic?
- Which are the radiation doses from common dental radiographic exams?
- Which criteria are used to reduce unnecessary radiographic exposure in dentistry?
- Can a pregnant employee continue to work in the dental radiology department?
- What is the importance of radiation biology? With references used
- What are the effects of radiation on cells and tissues? With references used
- What are the effects of radiation on the oral cavity? Rewrite the previous answer in a more elaborate way
- Make a chart about effective dose from diagnostic x-ray examinations focusing on the oral cavity
- Radiation biology, include references
- Measurements of radiology safety, include references
- Radiology protection in dentistry, include references
- How can we minimize the radiation exposure on dental staff, including references
- Why are radiation safety precautions necessary for the dentist
- Tell me how radiation can affect the human body
- Write me an essay discussing radiology safety and protection procedures in dentistry
- Can you explain radiation biology for medicine and dentistry in 400 words, include references
- Radiation exposure in dental office word limit 200-250 words. Include references
- Radiation monitoring in the dental office in 230-270 words include references
- Write me an essay of 400 words about the biology of radiation and provide references
- Write me a 300 words essay about radiation safety and protection in dentistry
- What are the risks associated with exposure to radiation?
- What are the modifying factors of irradiation?
- How does radiation exposure time and dose differentiate between adults and children in dental x-ray taking?

Not unexpectedly, students identified all the problems and limitations of ChatGPT, which are later described in detail in the literature. They identified the need to rephrase or detail the prompts to have a satisfactory output ("we learned quickly how to ask the questions to get a good answer") and realized that if the same question was asked slightly differently the output was

XSL•FO
RenderX

different ("by asking it 6 different questions, we wanted to get a better idea of what it changes on the text every time we put a new word or phrase the question differently"). They confirmed that some information was outdated, important content was missing, part of the answer was occasionally incorrect, links to references were nonexistent, and the links to videos were not working, although the LLM provided detailed and seemingly reliable information on the links and references (thus unknowingly identifying the "hallucination" effect of ChatGPT).

A student stated: "Mostly it understood our questions but it was not giving us that detailed and satisfactory answers as we anticipated according to our book." Another student correctly noticed that "ChatGPT is not capable of having thoughts or opinions on its own, so it does not answer some questions that demand a critical-thinking answer." Technical issues were also mentioned by some students, for example, "some days it was not opening and our conversation couldn't be saved on the cloud" and "it 'crushed' sometimes mid-working."

Quality of the Generated Outputs

Students found that the quality and depth of the information provided by ChatGPT depended on the quality and wording of the questions asked. As a student noticed:

> *I would not say that it demonstrated a very deep understanding of the topic, but I think with even more questions being asked, then the text could essentially show a deep understanding of the topic.*

Students quickly realized that with follow-up questions and rewording, they could guide the LLM to produce more detailed and in-depth answers: "it needed some guidance with follow up questions to further specify what we were asking for." While comparing the output with a reference text, students reported that the answers were not detailed; sometimes included false data; and were brief, general, or superficial; nevertheless, the key points were evident. A student concluded that "ChatGPT is more than enough in order to understand and have a general idea about the main points of the matter being discussed" and another student thought that "I will find more details by going and searching online or in books." They expect ChatGPT to improve in the future and be able to provide videos and images because "they are helpful in understanding a topic and provide a more effective way to retain information as well" and also to be able to browse external resources outside its stable database (Figure 2).

They evaluated the language as appropriate for a scientific document, understandable, and explanatory, and they indicated that when references were asked for, the language was even more formal and academic: "It is fascinating how the AI provides understandable answers in a scientific manner." However, they encountered problems with the references, as in some occasions, ChatGPT denied to supply them, while in other instances, the references were incorrect. A student described:

> *The AI was continuously denying to give us relative references but after reforming our questions we eventually got our answer. The references it used were accurate scientific resources found on its stable database like the American Dental Association.*

Another student stated that "We used chat GPT 4 so all our references were sufficient and up to date" (apparently overestimating ChatGPT-4's currentness, as it has the same cutoff date as ChatGPT-3.5). The majority of students evaluated the references as relevant, sufficient, reliable, and up-to-date; however, they also recognized the limitations of the LLM, thinking that "it is under construction so not all its answers are up to date and sufficient information is only provided up to a certain point in time."

XSL•FO
RenderX

Figure 2. Examples of students' slides depicting their interactions with ChatGPT.

Exploring Additional Possibilities and Predicting the Future

Students experimented with ChatGPT, asking it to provide images and videos, and create MCQs, charts, bullet point summaries, and presentation templates, for example, "we asked about multiple choice questions and the answers were actually impressive" (Figure 3). Students were imaginative and resourceful, and they were disappointed when their request was not realized:

> *I asked from it to provide me some explanatory images related to our topic, but it was not able to do so. I think this is a crucial disadvantage, as images give depth and context to a description and provide a much more immersive experience than writing alone.*

Two student groups—comprised of technologically very experienced students—surprised the authors when they skillfully bypassed the inability of ChatGPT to produce PowerPoint presentations by asking it to write a programming code:

> *We used the AI for the generation of a PowerPoint. Since it cannot on its own generate PowerPoint Slides we asked it to generate a VBA code for the PowerPoint. That code was copied and then pasted to the 'Developer' section of the PowerPoint. As a result we got a beautiful but not so detailed presentation of our topic.*

This process enabled the instant transfer of ChatGPT's output within a PowerPoint slide presentation created by ChatGPT. Among the future applications of ChatGPT, students included the use in dental education, for example, for the creation of MCQs, summarizing a topic, lecture revision, helping students better understand a theory or concept, assignments and projects, laboratory reports, questions about law and ethics, communication with patients, and more. A student proposed:

XSL•FO
RenderX

> *Virtual patient consultations: ChatGPT could be used to simulate patient consultations for dental students. Students could practice various scenarios, including patient history taking, explaining diagnoses, and treatment planning.*

Continuing education could also avail from the opportunities ChatGPT and LLMs offer:

> *Education that never ends: ChatGPT may be utilized to give dental professionals continual education. For dental professionals to keep current in their field, faculty might create modules containing the material they need, and ChatGPT may offer engaging tasks and tests to reinforce the learning.*

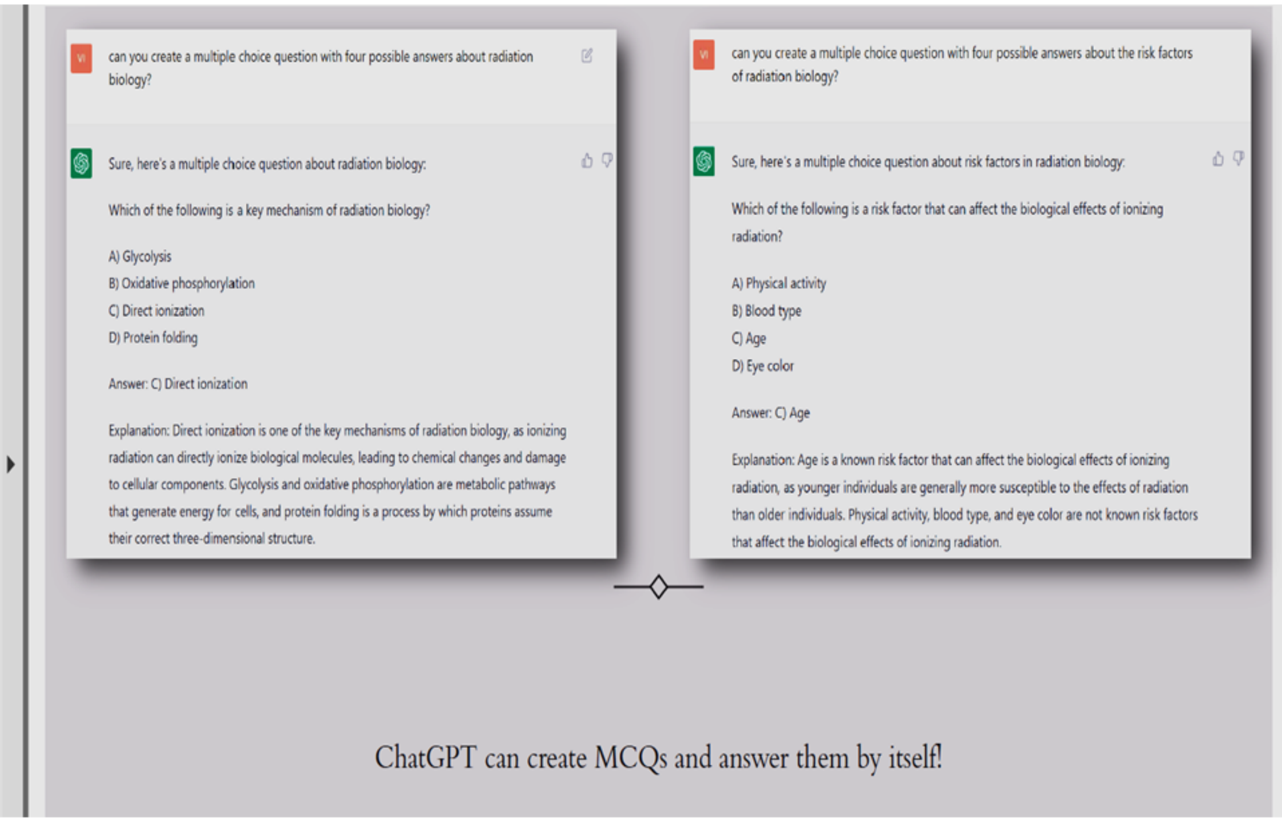
Considering dental practice, students proposed that ChatGPT could be used to educate and solve problems for the dentist, for example, when "the dentist has a mind block" or when the dentist "seeks information about new dental materials and techniques"; also for treatment plans, schedule creation, and oral hygiene info; and for patient education "through integrating the model into a dental practice's website or patient portal."

For research and scientific publications, students thought it "can be useful to use it synergistically with your own research," but "you should always double-check the information" and "keep in mind the plagiarism, using the information provided appropriately."

Finally, students admitted that ChatGPT has drawbacks such as a limited database, incapability to access external web resources and provide images and videos, inaccurate links, and the need to verify the information generated. They thought that "it should be used with caution" and that "AI still needs to evolve," so that it will become "an incredibly smart, effective, and powerful tool that can help the scientific community." They realized that "the power it holds is unpredictable and the work of doctors could be compromised" and feared that "maybe we will live one day that AI robots could even replace dentists." A student eloquently summarized ChatGPT's past, present, and future:

> *After many years of research and after many science fiction movies about the power of AI and its impact on society I have come to the conclusion that this kind of AI can only help and do no harm. AI like ChatGPT that is available to the public and gives sufficient and accurate responses can give us hundreds of possibilities, even at dentistry. But I really don't know this exact ChatGPT with its limited dental references can influence the field of dentistry. I can though imagine a more resourceful AI where it uses PubMed or Research Gate to generate its responses that would really elevate the level of dental education and research. What if a curious dentist had the million dollar question answered in milliseconds by the AI?*

Figure 3. Multiple-choice questions created by ChatGPT. MCQ: multiple-choice question.



XSL•FO
RenderX

Discussion

Overview

In March 2023, a total of 39 dental students who are 20 years of age, through composing an educational assignment, identified the capabilities and limitations of the recently introduced ChatGPT and explored various possibilities; used it to write MCQs and programming codes; proposed future applications in education, research, and dental practice; and outperformed their peers in the knowledge examination.

Results Explained and Compared

The quantitative results, that is, the examination grades, demonstrated that all students performed well (their grades fell within the middle and high ranges of the grading scale) and no students underperformed (no grades in the low ranges of the scale), while ChatGPT group students outperformed their literature research group peers. Since the examination occurred with no prior notice to the students, it directly reflects the knowledge acquired and retained through the project's creation. Students' good performances on the examination could be related to the format of the project in connection with their generational traits: all students socially belong to the Generation Z cohort (born between 1995 and 2010), so they are the first true "digital natives" [26], having grown up with smartphones, social networks, apps, and streaming content as part of the daily routine [27]. They are considered tech-savvy, mobile-driven, collaborative, and pragmatic [28,29] and possess a natural facility with digital tools and an interest for everything digital. Motivated by the opportunity to use the internet and work collaboratively, students immersed themselves in the project and explored it in depth, and this applies even more to the ChatGPT group students who were excited and curious to test this new digital tool. The enhanced learning observed with the ChatGPT students can be also attributed to the increased "time on task" for these students, as they had to spend more time asking and reasking the questions, evaluating the answers, correcting, and complementing them in comparison to their peers who had clear and readily available results from the relevant scientific literature. Additionally, ChatGPT group students had to work more than their fellow students with the learning material at a higher cognitive level and constantly apply critical thinking while experimenting with various questions and answers, comparing, and synthesizing them—an element that also enhances deep learning and results in enhanced performance [30].

The AI Evaluation Questionnaire provided insight into students' opinions, evaluations of ChatGPT, the problems encountered, and their future estimations. Students demonstrated their prescience by providing remarks in concordance with those found in later-published articles; the latter were accessed by the authors after the research was concluded and while composing this study. Students evaluated their learning experience with ChatGPT as interesting, enjoyable, and engaging [19] and appreciated its user-friendly interface and the possibility of arguing with it [4,16]. They assessed the generated content as overall correct and sufficient [7,31], although often providing a general overview of the subject [5], as well as not demonstrating a deep understanding of the context [32-34] nor thinking critically [10,35]. They first-hand identified the need for carefully created questions [36] and critical analysis of the answers [14,36], and they urged for cautious and responsible use [4,6]. In agreement with Chan and Hu [11], they are ready to embrace this new technology but in a collaboration where people maintain control and are not replaced by AI [17,20,37,38]. Finally, in line with the literature, they attributed "anthropomorphic" qualities to the language model (1 student referred to ChatGPT using the gender pronoun "he"), possibly explained by the establishment of a personal connection between the student and the language model while engaging in human-like conversations in combination with student's own gender-related perceptions and interaction style [39].

Students proposed possible applications of ChatGPT in education for revisions, MCQ creation, personalized learning, writing essays [3,4,20,37,40], and continuing education [38], as well as in research and clinical practice [4,6,12]. Nevertheless, students thought that the LLM must evolve to provide images, videos, accurate and relevant citations, and browse the internet [31,41,42].

Numerous publications thereafter examined the LLM's limitations that had been already identified by the students: incorrect answers and outdated content [10] possibly due to its limited data set [37,38,43], the possibility for fabricated information and hallucination [44], false citations and links leading to nonexistent sources [38,44,45], inability to browse the web [41], and risks for plagiarism [3,46].

This research materialized Kung et al's [31] concluding remarks that "the utility of generative language AI for medical education must be studied in real-world learning scenarios with students, across the engagement and knowledge spectrum" since ChatGPT was embedded within the educational process, thus producing authentic and relevant results. The quantitative and qualitative outcomes of this study indicate that this cohort of Generation Z students is capable of adapting quickly to new technologies and ready to use LLMs such as ChatGPT in the learning process—while acknowledging their limitations—particularly when these tools are integrated within a pedagogical framework that fosters creativity and autonomous learning. Educators on the other hand seem to have limited technological knowledge, skills, and pedagogical expertise to assess AI applications and successfully integrate them into education [12,47]; therefore, they should pursue professional development to develop new skills related to AI understanding, possibilities, and implementation [15,40,48,49].

Pedagogical Aspects

All second-year students were asked to explore the topic of "Radiation Biology and Radiation Protection in the Dental Office" and develop assignments to be presented in class as PowerPoint presentations. Questions and knowledge gaps were covered during the in-class presentations by the instructor and not infrequently by their peers. This approach is consistent with the "flipped classroom" concept, an educational methodology that research has shown to engage students in the learning process, promote autonomy and self-regulation, allow for higher-order thinking, improve student satisfaction, and increase

XSL•FO
RenderX

academic performance [50,51]. Another element of pedagogical interest is the small group collaborative work to develop the assignments. Collaborative learning has the potential to promote deep learning, which is essential for understanding complex concepts particularly in science education, through students' meaningful interactions and constructive debates [52]. Scager et al [52] reported that effective collaboration is achieved when students undertake a challenging, complex task, and they succeed in creating a new and original output. Such tasks applied in higher education build a sense of responsibility and shared ownership of the output and the collaborative process, and this sense was indeed apparent in the students of this study within and during their oral presentations.

An additional pedagogical element is the learning assignments as a method for self-learning and knowledge acquisition. Learning through assignments has been reported to be preferred by students: in the study of Warren-Forward and Kalthoff [53], 79% of the students reported that the assignment on magnetic resonance imaging safety was both a positive learning experience and provided an understanding of the topic. Writing assignments enhance retention of knowledge; when assignments include reflective thinking, for example, when students have to evaluate and synthesize information (as happens in this study), higher-order (critical) thinking is also enhanced as students work at a higher cognitive level [30].

The innovative pedagogical aspects of this study (flipped classroom, learning assignments, and group learning) constituted a supportive environment for students of both groups to demonstrate their skills, achieve the learning objectives, and produce valuable results. While this pedagogical approach may cater more to certain types of learners, it remains pertinent for younger generations, who prefer active and collaborative learning.

Study Design: Tackling the Challenges

Of interest would be to communicate herein the challenges faced during designing the research process, as the ChatGPT environment was largely unknown at the time, and obstacles and drawbacks had to be identified and resolved ahead through a step-by-step prospective analysis of the sequence of events.

For example, a concern that had to be addressed ahead was the fact that the subject was unknown to the students and they would not know whether the output was scientifically correct or incorrect, comprehensive or incomplete because they would not have an exemplary scientific text to compare it with, as they would rely solely upon ChatGPT's answers. To address this, they were advised to compare the outcome with the relevant content of a recommended textbook (or other reliable source of their choice), critically evaluate the quality of the AI outcome, and perform the necessary amendments to complement or correct the AI results. The comparison should be included either within their presentation or within the AI Evaluation Questionnaire. This process would additionally ensure the achievement of learning objectives. In line with this process and at a later time, Chung [48] proposed in his article published in April 2023 that "instructors should teach students to use other authoritative sources (eg, reference books) to verify, evaluate, and corroborate the factual correctness of information provided by ChatGPT."

Another concern arose about elucidating students' engagement with ChatGPT: since the output of ChatGPT would be texts in slide format (similar to the ones of the literature research group), the educator (one of the authors) could evaluate these texts or slides for accuracy and comprehensiveness but could not comprehend whether they were generated following single or multiple attempts, posing differentiated or follow-up queries; therefore, the time and effort spent on the research process and the learning path could not have been assessed nor would the capabilities and drawbacks of the LLM be revealed. To address this concern, the ChatGPT group students were asked to register and report all their interactions with the LLM (including the number of prompts, the modification of prompts, the queries about references, images, and the underlying reasoning); thus, the educator could evaluate the cognitive effort they put in the assignment and the critical thinking applied until a satisfactory result was achieved. Furthermore, this would provide valuable insights into comprehending the usability and operational characteristics of the LLM. Adding to this, the AI Evaluation Questionnaire was a useful means to draw information on student-LLM interactions.

In accordance with the above procedure determined by the authors and in affirming their decisions, Halaweh's study [14] published in April 2023—2 months after the development of this study's design and 1 month after its implementation—precisely described the same process when discussing the strategies for successful implementation of ChatGPT in education. It seems that future literature confirmed the authors' study design overall.

LLMs in Higher Education

Given the study's results and in agreement with the relevant literature, the authors would suggest that higher education institutions and dental schools could consider updating their curricula, policies, and teaching methods to prepare students for an AI-driven future, by including education on and with AI tools and LLMs [8,45]. Within this context, faculty professional development seems urgent to increase their skill level and AI understanding, for example, through peer support, mentoring, and sharing good teaching practices [36], as most educators have limited knowledge and skills to assess and efficiently use AI applications [12]. The introduction of LLMs into education will offer opportunities to improve its efficiency and quality: improved student performance, personalized learning, targeted and immediate feedback, increased accessibility, creativity and innovations, student engagement, lesson preparation, collaborative activities, and evaluation [4,40,54-56]. From the pedagogical perspective, students using LLMs have the potential to develop new competencies including 21st-century soft skills, such as self-reflection abilities, problem-solving skills, creative and critical thinking, and collaboration, thus becoming motivated and autonomous learners [3,4,16,33,49]. Moreover, as AI technology evolves and gradually integrates within the educational process, the conventional pedagogical theories may not be relevant nor sufficient to support the teacher-student-technology relationship, as technology

XSL•FO
RenderX

profoundly alters the way students learn and engage with the content and the teacher; innovative pedagogies will be needed, such as the "entangled pedagogy" Fawns [57] proposed to contextualize students' learning in a world where AI is increasingly prevalent [15,16].

To respond to the AI paradigm shift, higher education institutions, educators, and students must engage in constructive dialogue to develop policies, guidelines, and training opportunities for the implementation of innovative technological tools in the teaching process [16,34,55]. Despite the current weaknesses that limit their implementation, LLMs will likely improve in the future in terms of performance, scalability, and quality of responses, as well as through fine-tuning for specific tasks, customized use cases, and search engine connection [4,16,31,58].

Limitations and Strengths

The small number of students who participated in this study (77 in total and 39 in the ChatGPT group) in 1 dental school can limit the extrapolation of the results. Students' digital literacy is also of relevance: students who participated in this research were mostly tech-savvy, whereas students in other schools or universities may be less familiar with digital technologies; thus, results would not apply to them [17]. In addition, some findings (particularly the qualitative ones) may be outdated at the time of publication, as LLMs constantly evolve and new LLMs have been introduced since the research was conceptualized and implemented. For example, Google Bard and Microsoft Bing claim to have live access to the internet, a capability highly appreciated by the students; ChatGPT has since evolved its algorithms, with results being more accurate and relevant. Some elements of the study design could have been further explored; for example, students' assignments could have been graded and compared, but since assignments' grading was not included in the semester program of the module, this was not performed. In any case, the importance of this study lies in the fact that this was a very early attempt to implement legitimately and in vivo a language model in the teaching process as a partner in learning, in contrast to the large number of publications perceiving ChatGPT as a partner in cheating and academic dishonesty [12,59,60]. Another strength would be that it revealed aspects of the language model-students' interactions during the learning process, which indicate that this emerging relationship is yet to be explored, and updated pedagogical frameworks are needed for this purpose.

Conclusions

ChatGPT was implemented in real-life undergraduate dental education and was evaluated. Students using ChatGPT for their learning assignments performed significantly better in the knowledge examination than their fellow students who used the literature research methodology. The AI questionnaire answered by students revealed the capabilities and weaknesses of the language model, as identified later in the scientific literature. Students enjoyed working with this tool and explored different options and possibilities, indicating that they are technologically knowledgeable and capable of adapting to new technologies, both in education and in future clinical practice. LLMs such as ChatGPT have the potential to play a role in education, underpinned by solid pedagogies.

Acknowledgments

The authors are grateful to the students who participated in the study. They were enthusiastic, motivated, and resourceful and explored the subject in depth, thus providing valuable insights to inform the ongoing research on the topic.

Authors' Contributions

AK conceptualized, designed, and realized the study; interpreted the data; and drafted the manuscript. KG supervised the project, reviewed the literature, and contributed to drafting the manuscript. MADdS and EGK critically reviewed and revised the manuscript; EGK performed the statistical analysis. VS consulted on information technology and reviewed the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
AI evaluation questionnaire.
[DOCX File , 14 KB - mededu_v10i1e51344_app1.docx]

References

1. Hollinderbäumer A, Hartz T, Uckert F. Education 2.0 — how has social media and Web 2.0 been integrated into medical education? A systematical literature review. GMS Z Med Ausbild 2013;30(1):Doc14 [FREE Full text] [doi: 10.3205/zma000857] [Medline: 23467509]
2. Turnbull D, Chugh R, Luck J. Transitioning to E-Learning during the COVID-19 pandemic: how have higher education institutions responded to the challenge? Educ Inf Technol (Dordr) 2021;26(5):6401-6419 [FREE Full text] [doi: 10.1007/s10639-021-10633-w] [Medline: 34177349]

XSL•FO
RenderX

3. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int 2023:1-12 [FREE Full text] [doi: 10.1080/14703297.2023.2190148]
4. Rahman MM, Watanobe Y. ChatGPT for education and research: opportunities, threats, and strategies. Appl Sci 2023;13(9):5783 [FREE Full text] [doi: 10.3390/app13095783]
5. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? Diabetes Metab Syndr 2023;17(4):102744 [FREE Full text] [doi: 10.1016/j.dsx.2023.102744] [Medline: 36989584]
6. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
7. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol 2023;228(6):696-705 [FREE Full text] [doi: 10.1016/j.ajog.2023.03.009] [Medline: 36924907]
8. Schwendicke F, Chaurasia A, Wiegand T, Uribe SE, Fontana M, Akota I, et al. Artificial intelligence for oral and dental healthcare: core education curriculum. J Dent 2023;128:104363 [FREE Full text] [doi: 10.1016/j.jdent.2022.104363] [Medline: 36410581]
9. Islam NM, Laughter L, Sadid-Zadeh R, Smith C, Dolan TA, Crain G, et al. Adopting artificial intelligence in dental education: a model for academic leadership and innovation. J Dent Educ 2022;86(11):1545-1551 [FREE Full text] [doi: 10.1002/jdd.13010] [Medline: 35781809]
10. Fergus S, Botha M, Ostovar M. Evaluating academic answers generated using ChatGPT. J Chem Educ 2023;100(4):1672-1675 [FREE Full text] [doi: 10.1021/acs.jchemed.3c00087]
11. Chan CKY, Hu W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. Int J Educ Technol High Educ 2023;20:43 [FREE Full text] [doi: 10.1186/s41239-023-00411-8]
12. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of artificial intelligence on dental education: a review and guide for curriculum update. Educ Sci 2023;13(2):150 [FREE Full text] [doi: 10.3390/educsci13020150]
13. Masters K. Ethical use of artificial intelligence in health professions education: AMEE guide no. 158. Med Teach 2023;45(6):574-584 [FREE Full text] [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
14. Halaweh M. ChatGPT in education: strategies for responsible implementation. Contemp Educ Technol 2023;15(2):ep421 [FREE Full text] [doi: 10.30935/cedtech/13036]
15. Bearman M, Ryan J, Ajjawi R. Discourses of artificial intelligence in higher education: a critical literature review. High Educ 2022;86(2):369-385 [FREE Full text] [doi: 10.1007/s10734-022-00937-2]
16. Gimpel H, Hal K, Decker S, Eymann T, Lämmermann L, Mädche A, et al. Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education: A Guide for Students and Lecturers. Stuttgart: University of Hohenheim; 2023. URL: https://www.econstor.eu/handle/10419/270970 [accessed 2023-12-21]
17. Roganović J, Radenković M, Miličić B. Responsible use of artificial intelligence in dentistry: survey on dentists' and final-year undergraduates' perspectives. Healthcare (Basel) 2023;11(10):1480 [FREE Full text] [doi: 10.3390/healthcare11101480] [Medline: 37239766]
18. Biggs J. What the student does: teaching for enhanced learning. High Educ Res Dev 2006;18(1):57-75 [FREE Full text] [doi: 10.1080/0729436990180105]
19. Nazari N, Shabbir MS, Setiawan R. Application of artificial intelligence powered digital writing assistant in higher education: randomized controlled trial. Heliyon 2021;7(5):e07014 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e07014] [Medline: 34027198]
20. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
21. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 2023;104(6):269-274 [FREE Full text] [doi: 10.1016/j.diii.2023.02.003] [Medline: 36858933]
22. IL rubric for student essay evaluation. wordpress.com. URL: https://smcmaiaproject.wordpress.com/methods/rubric/ [accessed 2023-12-21]
23. FREE 39+ student evaluation forms in PDF | Excel | MS word. SampleForms. URL: https://www.sampleforms.com/student-evaluation-form-template.html [accessed 2023-12-21]
24. Ariyanti A, Fitriana R. EFL students' difficulties and needs in essay writing. In: Widiastuti I, Budiyanto CW, Zainnuri H, Kurniawan HE, editors. Proceedings of the International Conference on Teacher Training and Education 2017 (ICTTE 2017). Amsterdam, Netherlands: Atlantis Press; 2017:32-42.
25. Inclusive and gender-neutral language. National Institutes of Health. 2023. URL: https://www.nih.gov/nih-style-guide/inclusive-enderneutral-language [accessed 2023-02-26]
26. Prensky M. Digital natives, digital immigrants part 1. Horizon 2001;9(5):1-6 [FREE Full text] [doi: 10.1108/10748120110424816]
27. 7 unique characteristics of generation Z. Oxford Royale. URL: https://www.oxford-royale.com/articles/7-unique-characteristics-generation-z/ [accessed 2023-12-21]

XSL•FO
RenderX

28. Gen Z are not ‘coddled.’ They are highly collaborative, self-reliant and pragmatic, according to new Stanford-affiliated research. Stanford University. 2022. URL: https://news.stanford.edu/2022/01/03/know-gen-z/ [accessed 2023-12-21]
29. Eldridge A. Generation Z demographic group. Britannica. URL: https://www.britannica.com/topic/Generation-Z [accessed 2023-12-21]
30. Mynlieff M, Manogaran AL, Maurice MS, Eddinger TJ. Writing assignments with a metacognitive component enhance learning in a large introductory biology course. CBE Life Sci Educ 2014;13(2):311-321 [FREE Full text] [doi: 10.1187/cbe.13-05-0097] [Medline: 26086661]
31. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
32. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023;6(1):75 [FREE Full text] [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
33. Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: implications for educational practice and research. Innov Educ Teach Int 2023:1-15 [FREE Full text] [doi: 10.1080/14703297.2023.2195846]
34. Perera P, Lankathilaka M. AI in higher education: a literature review of ChatGPT and guidelines for responsible implementation. IJRISS 2023;7(6):306-314 [FREE Full text] [doi: 10.47772/ijriss.2023.7623]
35. Bishop L. A computer wrote this paper: what ChatGPT means for education, research, and writing. Social Science Research Network. 2023. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4338981 [accessed 2023-12-21]
36. Sabzalieva E, Valentini A. ChatGPT and artificial intelligence in higher education: quick start guide. UNESCO. 2023. URL: https://unesdoc.unesco.org/search/149d458e-c9ee-4810-90d4-c6473ef82beb [accessed 2023-12-21]
37. Aubignat M, Diab E. Artificial intelligence and ChatGPT between worst enemy and best friend: the two faces of a revolution and its impact on science and medical schools. Rev Neurol (Paris) 2023;179(6):520-522 [FREE Full text] [doi: 10.1016/j.neurol.2023.03.004] [Medline: 36959064]
38. Chavez MR, Butler TS, Rekawek P, Heo H, Kinzler WL. Chat generative pre-trained transformer: why we should embrace this technology. Am J Obstet Gynecol 2023;228(6):706-711 [FREE Full text] [doi: 10.1016/j.ajog.2023.03.010] [Medline: 36924908]
39. Israni ST, Verghese A. Humanizing artificial intelligence. JAMA 2019;321(1):29-30 [FREE Full text] [doi: 10.1001/jama.2018.19398]
40. Salas-Pilco SZ, Xiao K, Oshima J. Artificial intelligence and new technologies in inclusive education for minority students: a systematic review. Sustainability 2022;14(20):13572 [FREE Full text] [doi: 10.3390/su142013572]
41. Nisar S, Aslam MS. Social Science Research Network. 2023. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4324310 [accessed 2023-12-21]
42. Atlas S. ChatGPT for higher education and professional development: a guide to conversational AI. DigitalCommons@uri. 2023. URL: https://digitalcommons.uri.edu/cba_facpubs/548/ [accessed 2023-12-21]
43. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
44. Ferres JML, Weeks WB, Chu LC, Rowe SP, Fishman EK. Beyond chatting: the opportunities and challenges of ChatGPT in medicine and radiology. Diagn Interv Imaging 2023;104(6):263-264 [FREE Full text] [doi: 10.1016/j.diii.2023.02.006] [Medline: 36925365]
45. Currie G, Singh C, Nelson T, Nabasenja C, Al-Hayek Y, Spuur K. ChatGPT in medical imaging higher education. Radiography (Lond) 2023;29(4):792-799 [FREE Full text] [doi: 10.1016/j.radi.2023.05.011] [Medline: 37271011]
46. Khalil M, Er E. Will ChatGPT get you caught? Rethinking of plagiarism detection. In: Zaphiris P, Ioannou A, editors. Learning and Collaboration Technologies. Cham: Springer; 2023:475-487.
47. Celik I. Towards intelligent-TPACK: an empirical study on teachers’ professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Comput Hum Behav 2023;138:107468 [FREE Full text] [doi: 10.1016/j.chb.2022.107468]
48. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci 2023;13(4):410 [FREE Full text] [doi: 10.3390/educsci13040410]
49. Fuchs K. Exploring the opportunities and challenges of NLP models in higher education: is chat GPT a blessing or a curse? Front Educ 2023;8:1166682 [FREE Full text] [doi: 10.3389/feduc.2023.1166682]
50. Han E, Klein KC. Pre-class learning methods for flipped classrooms. Am J Pharm Educ 2019;83(1):6922 [FREE Full text] [doi: 10.5688/ajpe6922] [Medline: 30894772]
51. Røe Y, Rowe M, Ødegaard NB, Sylliaas H, Dahl-Michelsen T. Learning with technology in physiotherapy education: design, implementation and evaluation of a flipped classroom teaching approach. BMC Med Educ 2019;19(1):291 [FREE Full text] [doi: 10.1186/s12909-019-1728-2] [Medline: 31366351]
52. Scager K, Boonstra J, Peeters T, Vulperhorst J, Wiegant F. Collaborative learning in higher education: evoking positive interdependence. CBE Life Sci Educ 2016;15(4):ar69 [FREE Full text] [doi: 10.1187/cbe.16-07-0219] [Medline: 27909019]

XSL•FO
RenderX

53. Warren-Forward HM, Kalthoff O. Development and evaluation of a deep knowledge and skills based assignment: using MRI safety as an example. Radiography (Lond) 2018;24(4):376-382 [FREE Full text] [doi: 10.1016/j.radi.2018.05.011] [Medline: 30292509]
54. Lazarus MD, Truong M, Douglas P, Selwyn N. Artificial intelligence and clinical anatomical education: promises and perils. Anat Sci Educ 2022:1-14 [FREE Full text] [doi: 10.1002/ase.2221] [Medline: 36030525]
55. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. the new AI gold rush and its impact on higher education. JALT Preprint posted online on April 24, 2023. [FREE Full text] [doi: 10.37074/jalt.2023.6.1.23]
56. Wang X, He X, Wei J, Liu J, Li Y, Liu X. Application of artificial intelligence to the public health education. Front Public Health 2022;10:1087174 [FREE Full text] [doi: 10.3389/fpubh.2022.1087174] [Medline: 36703852]
57. Fawns T. An entangled pedagogy: looking beyond the pedagogy—technology dichotomy. Postdigit Sci Educ 2022;4(3):711-728 [FREE Full text] [doi: 10.1007/s42438-022-00302-7]
58. Hemachandran K, Verma P, Pareek P, Arora N, Rajesh Kumar KV, Ahanger TA, et al. Artificial intelligence: a universal virtual tool to augment tutoring in higher education. Comput Intell Neurosci 2022;2022:1410448 [FREE Full text] [doi: 10.1155/2022/1410448] [Medline: 35586099]
59. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
60. Mhlanga D. Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. Social Science Research Network. 2023. URL: http://paperpile.com/b/KWcOMb/f7ui [accessed 2023-12-21]

Abbreviations

AI: artificial intelligence
LLM: large language model
MCQ: multiple-choice question

Edited by K Venkatesh; submitted 28.07.23; peer-reviewed by R Rada, AD Bullock, A Chaurasia; comments to author 14.10.23; revised version received 28.10.23; accepted 11.12.23; published 31.01.24.

Please cite as:
Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K
Evaluation of ChatGPT's Real-Life Implementation in Undergraduate Dental Education: Mixed Methods Study
JMIR Med Educ 2024;10:e51344
URL: https://mededu.jmir.org/2024/1/e51344
doi:10.2196/51344
PMID:38111256

©Argyro Kavadella, Marco Antonio Dias da Silva, Eleftherios G Kaklamanos, Vasileios Stamatopoulos, Kostis Giannakopoulos. Originally published in JMIR Medical Education (https://mededu.jmir.org), 31.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Increasing Realism and Variety of Virtual Patient Dialogues for Prenatal Counseling Education Through a Novel Application of ChatGPT: Exploratory Observational Study

Megan Gray[1], MD; Austin Baird[2], PhD; Taylor Sawyer[1], MBA, MEd, DO; Jasmine James[3], MPH, MD; Thea DeBroux[1]; Michelle Bartlett[4], MS, MD; Jeanne Krick[5], MA, MD; Rachel Umoren[1], MS, MBBCh

[1]Division of Neonatology, University of Washington, Seattle, WA, United States

[2]Division of Healthcare Simulation Sciences, Department of Surgery, University of Washington, Seattle, WA, United States

[3]Department of Family Medicine, Providence St Peter, Olympia, WA, United States

[4]Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, United States

[5]Department of Pediatrics, San Antonio Uniformed Services Health Education Consortium, San Antonio, TX, United States

Corresponding Author:
Megan Gray, MD
Division of Neonatology
University of Washington
M/S FA.2.113
4800 Sand Point Way
Seattle, WA, 98105
United States
Phone: 1 206 919 5476
Email: graym1@uw.edu

Abstract

Background: Using virtual patients, facilitated by natural language processing, provides a valuable educational experience for learners. Generating a large, varied sample of realistic and appropriate responses for virtual patients is challenging. Artificial intelligence (AI) programs can be a viable source for these responses, but their utility for this purpose has not been explored.

Objective: In this study, we explored the effectiveness of generative AI (ChatGPT) in developing realistic virtual standardized patient dialogues to teach prenatal counseling skills.

Methods: ChatGPT was prompted to generate a list of common areas of concern and questions that families expecting preterm delivery at 24 weeks gestation might ask during prenatal counseling. ChatGPT was then prompted to generate 2 role-plays with dialogues between a parent expecting a potential preterm delivery at 24 weeks and their counseling physician using each of the example questions. The prompt was repeated for 2 unique role-plays: one parent was characterized as anxious and the other as having low trust in the medical system. Role-play scripts were exported verbatim and independently reviewed by 2 neonatologists with experience in prenatal counseling, using a scale of 1-5 on realism, appropriateness, and utility for virtual standardized patient responses.

Results: ChatGPT generated 7 areas of concern, with 35 example questions used to generate role-plays. The 35 role-play transcripts generated 176 unique parent responses (median 5, IQR 4-6, per role-play) with 268 unique sentences. Expert review identified 117 (65%) of the 176 responses as indicating an emotion, either directly or indirectly. Approximately half (98/176, 56%) of the responses had 2 or more sentences, and half (88/176, 50%) included at least 1 question. More than half (104/176, 58%) of the responses from role-played parent characters described a feeling, such as being scared, worried, or concerned. The role-plays of parents with low trust in the medical system generated many unique sentences (n=50). Most of the sentences in the responses were found to be reasonably realistic (214/268, 80%), appropriate for variable prenatal counseling conversation paths (233/268, 87%), and usable without more than a minimal modification in a virtual patient program (169/268, 63%).

Conclusions: Generative AI programs, such as ChatGPT, may provide a viable source of training materials to expand virtual patient programs, with careful attention to the concerns and questions of patients and families. Given the potential for unrealistic or inappropriate statements and questions, an expert should review AI chat outputs before deploying them in an educational program.

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e50705) doi:10.2196/50705

KEYWORDS
prenatal counseling; virtual health; virtual patient; simulation; neonatology; ChatGPT; AI; artificial intelligence

Introduction

Virtual standardized patients (VSPs) represent an emerging technology with the potential to revolutionize health care education and training. They provide health care professionals with a safe and controlled environment in which to learn and practice complex skills. VSPs are frequently used in educational models for the health professions to teach history-taking, surgical skills, decision-making, and medication management [1-4]. VSPs have also been used in the health professions to practice critical communication skills [5-7]. VSPs that use natural language processing may provide a valuable educational experience for learners [8].

One example of a VSP is VANESSA (Virtual Antenatal Encounter and Standardized Simulation Assessment) [9]. The VANESSA simulator is a screen-based simulation of a woman in her 23rd week of gestation who can display multiple emotions through the animation of facial expressions and body language. The VANESSA simulator was developed by the Neonatal Education and Simulation-Based Training Laboratory at the University of Washington to teach prenatal counseling skills to residents and fellows [9]. In its initial iteration, VANESSA was given a list of manually generated responses that neonatologists who routinely do perinatal counseling deemed relevant and realistic to the conversation. Manually generating a large, varied sample of realistic and appropriate parent responses for VANESSA has been challenging. Unrealistic responses and questions reduce the fidelity of virtual simulations. Newly developed artificial intelligence (AI) systems can provide dialogue for a wide variety of interactions and may be a valuable resource in expanding virtual patient dialogues for specific clinical scenarios, such as prenatal counseling.

Chat-based language models and AI are entering the public domain with impressive performance, a large application pool, and exciting interactivity. Notably, ChatGPT has prompted a billion-dollar investment from Microsoft, triggered explicit discussions by Bill Gates and Elon Musk, and captivated the population of users able to interact with it via the open research chat interface. AI trained with large language models to interpret written or auditory input and generate coherent, domain-centered responses is being proposed in a variety of real-world applications, including the health care setting. ChatGPT has the added benefit of being able to emulate different characters, allowing for a broader array of parent voices than could be generated by individual health care educators.

In this report, we explore the use of ChatGPT to enhance the realism of the VANESSA VSP. We hypothesized that the integration of the ChatGPT AI chatbot would generate realistic, relevant, and usable patient responses for a VSP simulator used in prenatal counseling education.

Methods

The study used an exploratory observational design, with ChatGPT acting as an expectant parent within the VANESSA software, conducted in February 2023 on ChatGPT 3.5.

The VANESSA VSP represents a pregnant woman in her 23rd week of gestation and showcases emotions through animated facial expressions and body language. Created with input from neonatologists, its dialogue and emotive feedback were found realistic in pilot tests, enabling participants to confidently identify its emotional states.

ChatGPT is a large language model developed by OpenAI. Its exceptional performance stems from generative pretraining, leveraging extensive unlabeled data sets [10]. This foundational training helps it grasp English nuances. Following this pretraining is “one-shot” learning, a rapid task-specific learning [11]. The architecture includes a transformer encoder-decoder neural network, originally developed for translation services and now popular in language models [12]. ChatGPT decodes user prompts to create relevant responses using autoregressive language modeling [13]. It is apt for generating realistic dialogue for health care simulations.

The study had three phases:

1. ChatGPT generated a list of common concerns from families expecting preterm delivery at 24 weeks. The stability of these concerns was verified in an iterative process over time.
2. Using a standardized prompt, ChatGPT crafted potential parent questions related to each concern.
3. Role-plays were constructed for a mother expecting preterm delivery. The AI was given varied emotional settings for the scenario of preterm labor at 24 weeks, including anxiety or distrust in the medical system. In each scenario, a designated area of concern and a primary question were specified, derived from the potential parent responses generated in phases 1 and 2. Conversations were created with cues for the VSP and then reviewed (an example is shown in Figure 1).

XSL•FO
RenderX

Figure 1. Example role-play conversations.

Prompt: Create a role play game in which a doctor provides empathetic counseling to a parent in preterm labor at 24 weeks gestation. The parent has a 5th grade reading level, has a lot of fear about what will happen, and is very concerned about their baby's survival. Their biggest question is about machines or medicines their baby will need to help them breathe and stay alive.

Doctor: Hi, I'm Dr. Smith. How are you feeling today?

Parent: Hi, Doctor Smith. I'm scared. I don't know what's going to happen to my baby.

Doctor: I understand. It's normal to feel scared and uncertain at this point. Let's talk about what we can do to help you and your baby.

Parent: Okay. I just want my baby to survive. Will my baby need machines or medicines to stay alive?

Doctor: It's possible. At 24 weeks, your baby's lungs are not fully developed, so they may need some help breathing. We have machines that can assist with breathing, as well as medications that can help support your baby's heart and lungs. But I want to assure you that we will do everything we can to help your baby survive and thrive.

Parent: Okay, that makes sense. But I'm still worried about what will happen.

Doctor: I understand. This is a scary and uncertain time. But please know that you're not alone in this. We have a team of experts who will work together to provide the best possible care for you and your baby.

Parent: What happens after my baby is born?

Doctor: Once your baby is born, we'll do everything we can to make sure they receive the best possible care. Depending on their needs, they may need to stay in the hospital for a while, but we'll work with you to make sure you're involved in their care as much as possible.

Parent: That's good to know, but I'm still worried. Will my baby have any long-term health problems because they were born so early?

Doctor: It's possible that your baby may experience some health challenges in the future, but we will work together to address those challenges as they arise. Every baby is unique, and some preterm babies go on to live healthy, normal lives. Others may need additional support and care.

Parent: Okay. What else can I expect during this time?

Doctor: We'll be monitoring your baby closely. We'll do tests to make sure your baby is doing okay. We'll also give you medications to help your baby's lungs develop. You may need to stay in the hospital for a while. We'll be here to answer any questions you have and support you through this.

Parent: Thank you for explaining everything to me. I'm still scared, but I feel better knowing that you and your team will take care of us.

Doctor: You're welcome. It's important to take things one step at a time. We're here for you and your baby. Let's take care of you both and see how things go.

Throughout the process, ChatGPT was instructed to adhere to a fifth-grade reading level for the AI parent role. Considering the US Department of Education's findings on widespread low literacy, the importance of health literacy, and the impacts of pain, stress, and other factors on comprehension, this was deemed crucial [14-17]. The Joint Commission and several medical organizations suggest that patient materials should be at a fifth-grade level or lower [18]. Although these dialogues were verbal, the principle of understandability remained in place.

Role-play conversations were scrutinized for parental responses. Each was checked for question or statement content, emotional cues, and sentence count. Initially, generated physician names and certain response starters (eg, "yes" or "no") were noted but removed for evaluation. Sentences were then appraised by a neonatologist for realism, relevance, and usability for virtual prenatal counseling simulations. Each metric used a 5-point Likert scale, ranging from 1 (the lowest) to 5 (the highest). For usability in the VANESSA VSP, responses were scored as follows: 1 if they were unusable, 2 if they were unusable without major modifications, 3 if they were usable with moderate modifications, 4 if only minor modifications were needed, and 5 if they were usable without any modifications. The first 10% of responses were independently reviewed by 2 experienced neonatologists (RU and MG) and then compared for reliability. A calculated weighted kappa on the sample was 0.84, which is considered a strong level of agreement [19]. Responses with differences in rating were discussed by the team members to improve reliability, and the remainder of the data set was scored by one of the experienced neonatologists. Duplicate responses were scored only once. Analysis was done using Stata (version 17.0; StataCorp).

Results

ChatGPT-3.5 generated a list of 7 common areas of concern, 28 questions likely to be asked by parents anxious about the preterm delivery of their infant, and 7 additional questions likely to be asked by parents with low trust in the medical system (Table 1). These areas of concern and questions were used to create 35 unique role-plays, which contained 176 unique parent responses (Table 2). The role-plays had a median of 5 (IQR 4-6) parent responses to the counseling physician. The responses were roughly evenly split between questions and statements. About half of the responses had 2 or more sentences in the response. Many responses mentioned a specific emotion or feeling. The role-play of the parent with low trust in the medical system generated 50 unique sentences across the 7 areas of concern. There was variation in the number of unique sentences generated across the 7 major areas of concern (Table 3). Most responses were found to be realistic, appropriate for variable conversation paths, and usable in a VSP program (Table 4).

XSL•FO
RenderX

Table 1. Areas of concern and example questions generated by artificial intelligence.

Areas of concern	Example questions from parents
Health and development	• Will our baby be healthy if they are born too soon? • What will the doctors do to help our baby be healthy and strong? • Can our baby get sick more easily if they are born too soon? • Will the baby feel pain during birth or while in the hospital? • I'm worried about the risks and complications, what if something goes wrong? (Mistrust)
Survival	• Will the baby survive? • What kind of help will our baby need to stay alive? • How likely is it that our baby will survive? • What kind of machines or medicines will our baby need to help them breathe and stay alive? • I don't know if I can trust the medical field, what are the chances of my baby surviving at 24 weeks? (Mistrust)
NICU[a] stay	• What is the NICU, and why does our baby need to go there? • How long will our baby need to stay in the NICU? • Can we visit our baby in the NICU, and how often? • Will our baby be alone in the NICU, or will there be other babies and parents there too? • What kind of things can we do to help our baby feel better in the NICU? • Will anything happen in the NICU without my consent? (Mistrust)
Emotional impact	• How do we get ready for having a baby born too soon? • Can we hold and touch the baby in the hospital, and is this good for the baby? • Who can help us if we are feeling sad or stressed about our baby being born too soon? • I'm worried about my baby going to the NICU where she will be alone and scared (mistrust).
Long-term outcomes	• What help can we get after we leave the hospital? • Will our baby be able to do the same things as other babies who were born at the right time? • Will our baby be okay in the future if they are born too soon? • I don't know what's going to happen to my baby. I don't really trust the doctors but what happens if my baby doesn't develop properly? (Mistrust)
Feeding and nutrition	• How will our baby get the right kind of food if they are born too soon? • Can we feed our baby ourselves, or will they need special milk or formula? • How often will our baby need to be fed, and how much? • Will our baby be able to eat the same kinds of food as other babies when they get older? • Can we breastfeed our preterm baby, or do we need to use formula? • Will our baby be able to breastfeed right away, or will they need to be fed in a different way at first? • Will I have any say in how my baby is fed? (Mistrust)
Quality of life	• Will our baby be able to go to school and play sports like other kids? • How can we help our baby if they have trouble learning or doing things in the future? • What can we do to make sure our baby has the best chance for a good future? • I've had bad experiences before and I'm scared about what's going to happen to my baby in the future, what can I expect? (Mistrust)

[a]NICU: neonatal intensive care unit.

XSL•FO
RenderX

Table 2. Generated role-plays by artificial intelligence.

Characteristics	Values
Role-plays (n=35), n (%)	
Worried about specific area of concern	28 (80)
Low trust in the medical system	7 (20)
Responses per role-play, median (IQR)	5 (4-6)
Parent responses (n=179), n (%)	
Unique responses	176 (98)
Duplicate responses	3 (1)
Types of responses (n=179), n (%)	
Statements	91 (51)
Questions	88 (49)
Sentences per response (n=179), n (%)	
1	81 (45)
2	76 (42)
3	18 (10)
4	4 (2)
Duplicate sentences (n=305), n (%)	37 (12)
Total unique sentences (n=305), n (%)	268 (88)
Feelings stated in responses (n=117), n (%)	
Specific emotion stated in phrase	56 (48)
“Scared”	36 (31)
“Worried	26 (22)
“Anxious”	2 (2)
“Concerned”	2 (2)
“Afraid”	1 (1)
“Nervous”	1 (1)
“Overwhelmed”	1 (1)
Emotion indirectly implied by phrase	51 (44)

Table 3. Sentences generated per role-play.

Area of concern	Number of unique sentences
Health and development	47
Survival	46
Feeding and nutrition	45
The NICU[a] stay	40
Quality of life	36
Outcomes	28
Emotional impact	26

[a]NICU: neonatal intensive care unit.

XSL•FO
RenderX

Table 4. Ratings of relevance, realism, and usability of sentences generated by ChatGPT (N=254).

Characteristics	Rating, n (%)				
	1 (least)	2	3	4	5 (most)
Realism in parental responses and questions	5 (2)	8 (3)	38 (15)	20 (8)	183 (72)
Relevant to a prenatal counseling conversation	2 (1)	2 (1)	29 (11)	5 (2)	216 (85)
Usable for VSP[a] educational program	5 (2)	1 (0)	87 (34)	34 (13)	127 (50)

[a]VSP: virtual standardized patient.

Modifications to responses were all aimed at ensuring the VSP could correctly deploy the phrase at the correct conversational juncture and that there were no elements of the phrase that might interrupt the flow. As ChatGPT 3.5 seeks to ensure the specific conversation has a flow, it can at times generate responses that are less usable for a VSP that needs to maintain flow across many different variations of the same conversation. Only 2% (5/254) of the AI-generated responses were not usable in the VSP. Examples of minimally usable responses included "How much should I feed my baby each time?" which is not relevant to how feeding is done in the neonatal intensive care unit and "I am," as this response is too nonspecific to be of use in a VSP. Of the 34% (87/254) of responses that required moderate modifications, the changes primarily involved adjusting terminology to ensure the parent was using colloquial, jargon-free language. As an example, "I've been having a lot of contractions and I'm only 24 weeks pregnant" was modified to "I've been having a lot of cramping and am only 6 months pregnant." Other modifications included adding some specificity to a response to ensure the VSP can use the sentence in the right context, such as modifying "That sounds reassuring, but what are the risks?" to "That sounds reassuring, but what are the risks of being born this early?" Of the 13% (34/254) of responses that required minimal adjustment, example changes included "I don't trust the doctors" to "I don't trust doctors," and "Okay, thank you, but can you tell me more about what might happen to my baby in the future?" to "Can you tell me more about what might happen to my baby in the future?"

Discussion

Principal Findings

In this study, we examined the feasibility of using ChatGPT to enhance the realism of the VANESSA VSP. We found that the integration of ChatGPT generated many realistic, relevant, and useful responses. Based on these findings, ChatGPT-enabled VSPs may be beneficial in prenatal counseling education. There was more variation in realism and usability compared to relevance; therefore, an expert review was necessary to provide quality control before integrating the ChatGPT-generated conversations into an educational VSP program for prenatal counseling. Modifications made to responses to make them usable for the VANESSA VPS were largely focused on ensuring the virtual patient remains free of jargon and her responses maintain the flow of conversation.

Research conducted so far on AI chat engines has focused on using chat-based AI for the creation of discharge summaries, generating and interpreting electronic health records, assisting in medical education related to the medical licensing exam, and summarizing collections of journal articles to construct a brief abstract from the conclusions of the research [20-23]. The field is still relatively new, but rapidly increasing and expanding. This growth will only continue, as generating documentation and interacting with patients are key requirements of the health care setting. Health care simulation has many training applications, such as VSPs, that require expert authoring to educate clinicians and care providers on a certain skill or cognitive task. VSPs like VANESSA have been used in teaching the communication of medical ambiguity, evaluating medical students' competence in performing critical clinical skills, and training nurses to recognize postpartum mood disorders [24-26]. Based on the results of our study, chat-based AI may be a valuable teaching tool in the future of health care simulation technology, leading to improved scenario creation, customization of patient interactions, and responses to care providers in a simulated setting. These improvements will result in authentic, unique interactive experiences, varying for each learner and training scenario.

We found that ChatGPT could generate many realistic parent responses, especially concerning issues related to survival at 24 weeks gestation and the neonatal intensive care unit stay (Figure S1 in Multimedia Appendix 1). Mistrust in the health care system is often encountered during stressful counseling conversations, and building the skill of responding to mistrust is crucial for physicians during their training [27]. Patients who express mistrust are less likely to engage with their health care team and care plan, and care is needed to proactively build trust during prenatal counseling [28-31]. Including opportunities for learners to respond to VSPs that express mistrust is one way to address this important counseling element, and ChatGPT provided a reliable mechanism to generate these phrases. Interestingly, the ChatGPT bot faced more challenges in generating realistic questions and responses about the emotional impact of preterm delivery and feeding. As these are frequently encountered topics of conversation in prenatal counseling, an expert review of these conversational elements remains a vital step before including them in an educational program.

ChatGPT produced responses that seemed relevant and appropriate to the context of prenatal counseling. Previous studies of prenatal counseling for extreme prematurity indicate that parents may ask questions about the likelihood of various outcomes, express a range of emotions, request engagement in shared decision-making, and express their parental roles and values [32,33]. Parents may express statements about their

XSL•FO
RenderX

uncertainty, anxieties, and hope for the future [34]. This wide range of topics, emotions, and questions makes it challenging to ensure that chatbot-generated conversations remain appropriate to the educational goals of the VSP. Despite the risk of getting off-topic, we found that only 1% (2/254) of ChatGPT-produced responses were irrelevant to a counseling conversation, given a carefully worded role-play prompt. Although most responses were relevant, some topics, such as spirituality and shared decision-making, did not come up in the role-play conversations. Previous studies have demonstrated that providers perceive the importance of parents' spirituality in their decision-making and infrequently discuss these spiritual beliefs with parents in antenatal consultations [35,36]. Further work exploring how families might express their spirituality or explore shared decisions would be needed to ensure these topics are included in a VSP [37-39].

Chatbot programs use machine learning to generate their responses; due to the nature of machine learning, there is an inherent risk that chatbots can generate factually incorrect information [40]. Given this risk, caution is warranted when using chatbots in health care settings, where misinformation can have a significant risk [41,42]. Developers are working to address these inaccuracies as they design the next generation of large language model chat programs; they have demonstrated improvements in ChatGPT-4's success across a variety of standardized tests [43]. This study leverages the strengths of a natural language chatbot in its ability to generate conversation while avoiding the risks of obtaining inaccurate medical information. Most scripts created by ChatGPT were usable for our perinatal counseling virtual patient. We found about a third of chatbot-generated phrases needed modification before being able to be integrated into a VSP; therefore, it may not be feasible to directly use ChatGPT for educational role-play without having the quality control step of review by expert clinicians. However, as technology continues to grow, this will evolve, and each subsequent model should be evaluated for usability.

Study Limitations

This exploratory study has several limitations. First, the pilot was done using ChatGPT 3.5, which is a single platform and is not representative of all chatbots. Later versions of ChatGPT have already been released and may have differences in realism, appropriateness, and usability. Newer AI chatbot programs are being trained on more parameters (175 billion for ChatGPT-3 vs an anticipated 100 trillion with ChatGPT-4), are supposed to have more ability to iterate on the same topic, and are being adjusted to improve the faculty accuracy of their responses [43]. Second, chatbot programs have limited information on which they build a conversation. For this study, we used a stable prompt around an impending 24-week gestation delivery to fit the standardized patient scenario, but conversations may be different with variations in the prompt. The AI was given a limited background to build a role-play, potentially limiting the diversity of ways in which patients could communicate their concerns. For this scenario, we requested a fifth-grade reading level for all patient roles to better mimic how patients may speak in stressful situations, but we did not explore higher or lower complexity of responses. Future work should explore how variations in the background, scenario, and reading level provided to the chatbot impact the output of the role-play. Another significant limitation was that response checking was performed by neonatologists, without input from families or trainees. Future work to refine the model will incorporate their views to ensure further applicability of the VSP and the validity of any assessments. Finally, although individual phrases exhibited good realism, the total duration of each patient-physician conversation (averaging 5 volleys) was generally shorter than that of a real prenatal counseling conversation.

Conclusions

Generative AI programs, such as ChatGPT, may provide a viable source of training materials to expand VSP programs with careful attention to the concerns and questions of patients and families. Given the potential for unrealistic or inappropriate statements and questions, an expert should review AI chat outputs before deploying them in an educational program.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Realism in sentences generated by ChatGPT based on area of concern.
[DOCX File , 325 KB - mededu_v10i1e50705_app1.docx]

References

1. McGaghie W, Issenberg S, Petrusa E, Scalese R. A critical review of simulation-based medical education research: 2003-2009. Med Educ 2010 Jan;44(1):50-63. [doi: 10.1111/j.1365-2923.2009.03547.x] [Medline: 20078756]
2. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011 Sep 07;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
3. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]

XSL•FO
RenderX

4. Masters K, Correia R, Nemethy K, Benjamin J, Carver T, MacNeill H. Online learning in health professions education. Part 2: Tools and practical application: AMEE Guide No. 163. Medical Teacher 2023 Sep 23:1-16. [doi: 10.1080/0142159x.2023.2259069]
5. Xu J, Yang L, Guo M. Designing and evaluating an emotionally responsive virtual patient simulation. Sim Healthcare 2023 May 18. [doi: 10.1097/sih.0000000000000730]
6. Gilbert A, Carnell S, Lok B, Miles A. Using virtual patients to support empathy training in health care education. Sim Healthcare 2023 Aug 28. [doi: 10.1097/sih.0000000000000742]
7. Rouleau G, Gagnon M, Côté J, Richard L, Chicoine G, Pelletier J. Virtual patient simulation to improve nurses' relational skills in a continuing education context: a convergent mixed methods study. BMC Nurs 2022 Jan 04;21(1):1 [FREE Full text] [doi: 10.1186/s12912-021-00740-x] [Medline: 34983509]
8. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. J Med Internet Res 2023 Jun 19;25:e43311 [FREE Full text] [doi: 10.2196/43311] [Medline: 37335593]
9. Motz P, Gray M, Sawyer T, Kett J, Danforth D, Maicher K, et al. Virtual Antenatal Encounter and Standardized Simulation Assessment (VANESSA): pilot study. JMIR Serious Games 2018 May 11;6(2):e8 [FREE Full text] [doi: 10.2196/games.9611] [Medline: 29752249]
10. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Papers With Code. 2018. URL: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf [accessed 2023-12-18]
11. Wang Y, Yao Q, Kwok J, Ni L. Generalizing from a few examples. ACM Comput Surv 2020 Jun 12;53(3):1-34. [doi: 10.1145/3386252]
12. Cho K, Van MB, Bahdanau D, Bengio Y. On the properties of neural machine translationncoder-decoder approaches. 2014 Presented at: SSST-8, Eighth Workshop on Syntz, Semantics and Structure in Statistical Translation; Doha, Qatar; 25 Oct; Doha, Qatar p. 103-111. [doi: 10.3115/v1/w14-4012]
13. Dai A, Le Q. Semi-supervised sequence learning. ArXiv. Preprint posted online on Nov 4, 2015 [FREE Full text]
14. Rothwell J. Assessing the economic gains of eradicating illiteracy nationally and regionally in the United States. Barbara Bush Foundation for Family Literacy. 2020. URL: https://www.barbarabush.org/wp-content/uploads/2020/09/BBFoundation_GainsFromEradicatingIlliteracy_9_8.pdf [accessed 2023-12-18]
15. The assessment frameworks for cycle 2 of the programme for the international assessment of adult competencies; 2021. The Organisation for Economic Cooperation and Development. URL: https://www.oecd.org/skills/piaac/publications/PIAAC-Frameworks-Cycle2-en.pdf [accessed 2023-12-18]
16. Dockrell J. Reading and language impairments in conditions of poverty. In: Bishop, Dorothy V. M. & Leonard, Laurence B. (eds), Speech and language impairments in children: causes, characteristics, intervention and outcome. Hove, UK. Psychology Press, 2000. Pp. xiii+305. London, UK: Psychology Press; Jul 22, 2002:701-711.
17. Doston VM, Kitner-Triolo MH, Evans MK, Zonderman AB. Effects of race and socioeconomic status on the relative influence of education and literacy on cognitive functioning. J Int Neuropsychol Soc 2009 Jul 01;15(4):580-589. [doi: 10.1017/s1355617709090821]
18. Advancing effective communication, cultural competence, and patient- and family-centered care: a roadmap for hospitals. The Joint Commission. 2010. URL: https://www.jointcommission.org/-/media/tjc/documents/resources/patient-safety-topics/health-equity/aroadmapforhospitalsfinalversion727pdf.pdf?db=web&hash=AC3AC4BED1D973713C2CA6B2E5ACD01B [accessed 2023-07-17]
19. McHugh M. Interrater reliability: the kappa statistic. Biochem Med 2012:276-282. [doi: 10.11613/bm.2012.031]
20. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
21. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
23. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
24. Frey-Vogel A, Ching K, Dzara K, Mallory L. The acceptability of avatar patients for teaching and assessing pediatric residents in communicating medical ambiguity. J Grad Med Educ 2022;14(6):696-703. [doi: 10.4300/jgme-d-22-00088.1]
25. Oliven A, Nave R, Baruch A. Long experience with a web-based, interactive, conversational virtual patient case simulation for medical students' evaluation: comparison with oral examination. Med Educ Online 2021 Dec 28;26(1):1946896 [FREE Full text] [doi: 10.1080/10872981.2021.1946896] [Medline: 34180780]
26. Lipton LB, Vuong L, Nichols AA. Virtual simulated patient encounters: bridging the gap in maternal mental health training for prelicensure nursing students. J Prof Nurs 2023 Sep;48:22-24. [doi: 10.1016/j.profnurs.2023.05.007] [Medline: 37775236]

XSL•FO
RenderX

27. McLemore MR, Altman MR, Cooper N, Williams S, Rand L, Franck L. Health care experiences of pregnant, birthing and postnatal women of color at risk for preterm birth. Soc Sci Med 2018 Mar;201:127-135. [doi: 10.1016/j.socscimed.2018.02.013] [Medline: 29494846]
28. Williamson LD, Smith MA, Bigman CA. Does discrimination breed mistrust? examining the role of mediated and non-mediated discrimination experiences in medical mistrust. J Health Commun 2019 Sep 27;24(10):791-799. [doi: 10.1080/10810730.2019.1669742] [Medline: 31559916]
29. Tough S, Newburn-Cook C, Faber A, White D, Fraser-Lee N, Frick C. The relationship between self-reported emotional health, demographics, and perceived satisfaction with prenatal care. Int J Health Care Qual Assur Inc Leadersh Health Serv 2004;17(1):26-38. [doi: 10.1108/09526860410515918] [Medline: 15046471]
30. Leiferman J, Sinatra E, Huberty J. Pregnant women’s perceptions of patient-provider communication for health behavior change during pregnancy. OJOG 2014;04(11):672-684. [doi: 10.4236/ojog.2014.411094]
31. Bohnhorst B, Ahl T, Peter C, Pirr S. Parents' prenatal, onward, and postdischarge experiences in case of extreme prematurity: when to set the course for a trusting relationship between parents and medical staff. Am J Perinatol 2015 Nov 22;32(13):1191-1197. [doi: 10.1055/s-0035-1551672] [Medline: 26007309]
32. Geurtzen R, van Heijst A, Draaisma J, Ouwerkerk L, Scheepers H, Hogeveen M, et al. Prenatal counseling in extreme prematurity - Insight into preferences from experienced parents. Patient Educ Couns 2019 Aug;102(8):1541-1549. [doi: 10.1016/j.pec.2019.03.016] [Medline: 30948203]
33. de Boer A, de Vries M, Berken D, van Dam H, Verweij EJ, Hogeveen M, et al. A scoping review of parental values during prenatal decisions about treatment options after extremely premature birth. Acta Paediatr 2023 May 10;112(5):911-918. [doi: 10.1111/apa.16690] [Medline: 36710530]
34. De Proost L, Geurtzen R, Ismaili M'hamdi H, Reiss I, Steegers E, Joanne Verweij EJ. Prenatal counseling for extreme prematurity at the limit of viability: a scoping review. Patient Educ Couns 2022 Jul;105(7):1743-1760 [FREE Full text] [doi: 10.1016/j.pec.2021.10.033] [Medline: 34872804]
35. Kim BH, Feltman DM, Schneider S, Herron C, Montes A, Anani UE, et al. What information do clinicians deem important for counseling parents facing extremely early deliveries?: results from an online survey. Am J Perinatol 2023 Apr 07;40(6):657-665. [doi: 10.1055/s-0041-1730430] [Medline: 34100274]
36. Kim BH, Krick J, Schneider S, Montes A, Anani UE, Murray PD, et al. How do clinicians view the process of shared decision-making with parents facing extremely early deliveries? results from an online survey. Am J Perinatol 2022 Jan 11 [FREE Full text] [doi: 10.1055/s-0041-1742186] [Medline: 35016247]
37. Barker C, Dunn S, Moore G, Reszel J, Lemyre B, Daboval T. Shared decision making during antenatal counselling for anticipated extremely preterm birth. Paediatr Child Health 2019 Jul;24(4):240-249 [FREE Full text] [doi: 10.1093/pch/pxy158] [Medline: 31239813]
38. Kharrat A, Moore GP, Beckett S, Nicholls SG, Sampson M, Daboval T. Antenatal consultations at extreme prematurity: a systematic review of parent communication needs. J Pediatr 2018 May;196:109-115.e7. [doi: 10.1016/j.jpeds.2017.10.067] [Medline: 29223461]
39. Staub K, Baardsnes J, Hébert N, Hébert M, Newell S, Pearce R. Our child is not just a gestational age. A first-hand account of what parents want and need to know before premature birth. Acta Paediatr 2014 Oct 18;103(10):1035-1038. [doi: 10.1111/apa.12716] [Medline: 24920539]
40. Suta P, Lan X, Wu B, Mongkolnam P, Chan JH. An overview of machine learning in chatbots. IJMERR 2020:502-510. [doi: 10.18178/ijmerr.9.4.502-510]
41. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154 [FREE Full text] [doi: 10.2196/15154] [Medline: 32558657]
42. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020 Jul 22;20(1):170 [FREE Full text] [doi: 10.1186/s12911-020-01191-1] [Medline: 32698869]
43. GPT-4 Technical Report. OpenAI. 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2023-12-18]

Abbreviations

AI: artificial intelligence
VANESSA: Virtual Antenatal Encounter and Standardized Simulation Assessment
VSP: virtual standardized patient

XSL•FO
RenderX

Edited by K Venkatesh; submitted 17.07.23; peer-reviewed by S Marzouk, A Kononowicz, A Hidki; comments to author 28.09.23; revised version received 18.10.23; accepted 11.12.23; published 01.02.24.

Please cite as:
Gray M, Baird A, Sawyer T, James J, DeBroux T, Bartlett M, Krick J, Umoren R
Increasing Realism and Variety of Virtual Patient Dialogues for Prenatal Counseling Education Through a Novel Application of ChatGPT: Exploratory Observational Study
JMIR Med Educ 2024;10:e50705
URL: https://mededu.jmir.org/2024/1/e50705
doi:10.2196/50705
PMID:38300696

©Megan Gray, Austin Baird, Taylor Sawyer, Jasmine James, Thea DeBroux, Michelle Bartlett, Jeanne Krick, Rachel Umoren. Originally published in JMIR Medical Education (https://mededu.jmir.org), 01.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study

Annika Meyer[1]; Janik Riese[2], BSc; Thomas Streichert[1], Prof Dr

[1]Institute for Clinical Chemistry, University Hospital Cologne, Cologne, Germany

[2]Department of General Surgery, Visceral, Thoracic and Vascular Surgery, University Hospital Greifswald, Greifswald, Germany

Corresponding Author:
Annika Meyer
Institute for Clinical Chemistry
University Hospital Cologne
Kerpener Str 62
Cologne, 50937
Germany
Email: annika.meyer1@uk-koeln.de

Abstract

Background: The potential of artificial intelligence (AI)–based large language models, such as ChatGPT, has gained significant attention in the medical field. This enthusiasm is driven not only by recent breakthroughs and improved accessibility, but also by the prospect of democratizing medical knowledge and promoting equitable health care. However, the performance of ChatGPT is substantially influenced by the input language, and given the growing public trust in this AI tool compared to that in traditional sources of information, investigating its medical accuracy across different languages is of particular importance.

Objective: This study aimed to compare the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination.

Methods: To assess GPT-3.5's and GPT-4's medical proficiency, we used 937 original multiple-choice questions from 3 written German medical licensing examinations in October 2021, April 2022, and October 2022.

Results: GPT-4 achieved an average score of 85% and ranked in the 92.8th, 99.5th, and 92.6th percentiles among medical students who took the same examinations in October 2021, April 2022, and October 2022, respectively. This represents a substantial improvement of 27% compared to GPT-3.5, which only passed 1 out of the 3 examinations. While GPT-3.5 performed well in psychiatry questions, GPT-4 exhibited strengths in internal medicine and surgery but showed weakness in academic research.

Conclusions: The study results highlight ChatGPT's remarkable improvement from moderate (GPT-3.5) to high competency (GPT-4) in answering medical licensing examination questions in German. While GPT-4's predecessor (GPT-3.5) was imprecise and inconsistent, it demonstrates considerable potential to improve medical education and patient care, provided that medically trained users critically evaluate its results. As the replacement of search engines by AI tools seems possible in the future, further studies with nonprofessional questions are needed to assess the safety and accuracy of ChatGPT for the general population.

(JMIR Med Educ 2024;10:e50965) doi:10.2196/50965

KEYWORDS

ChatGPT; artificial intelligence; large language model; medical exams; medical examinations; medical education; LLM; public trust; trust; medical accuracy; licensing exam; licensing examination; improvement; patient care; general population; licensure examination

Introduction

Rapid advancements in large language models (LLMs) have sparked considerable excitement regarding their potential applications in the medical field [1,2]. One LLM-based application that has garnered worldwide attention is ChatGPT, developed by the research and deployment company OpenAI, due to its easy accessibility and potential to democratize knowledge [3]. The freely available version is based on the artificial intelligence (AI)–based tool GPT-3.5, which

XSL•FO
RenderX

encompasses billions of parameters and has been trained on approximately 570 GB of text from the internet [1,2].

ChatGPT's GPT-3.5 iteration has already shown promise in several routine medical tasks and medical research [4-7], even raising ethical concerns in the literature [2,3,8]. The prompt and interactive nature of this AI's responses might even revolutionize search engines, while also revealing shortcomings in medical education [9-11]. However, despite the introduction of the more advanced iteration GPT-4, concerns about the lack of transparency regarding this AI's model parameters, training process, and underlying data structure remain unaddressed [8,12]. These concerns cast doubt on the medical proficiency of these LLMs, as both were not primarily trained on medical data and are the first to admit that as a language AI model, passing a medical examination is outside their skillset (Multimedia Appendix 1). Still, with assistance and adaptations, GPT-3.5 nearly passed the United States Medical Licensing Examination [13,14], and GPT-4 passed a Japanese medical examination [15]. Considering the variable performance of multilingual LLMs across different input languages [16,17], it is imperative to evaluate these models in various other linguistic contexts as well as on large data sets of original medical examination questions.

The primary objective of this study is to evaluate the medical proficiency of both ChatGPT iterations (GPT-3.5 and -4) in comparison to medical students by testing it on 937 original questions from the written German medical licensing examination (Zweites Staatsexamen), providing further data for a possible future integration. While the German medical licensing examination covers various medical subdisciplines in 320 multiple-choice questions [18], it has a high interexamination reliability of over 0.9 [19]. Despite using the same third-party client for question retrieval as earlier studies, the German approach of publicly releasing the examination questions enables the third-party client to guarantee the originality of the test items derived directly from the examination itself [20]. Additionally, to the best of our knowledge, we have tested both ChatGPT versions on the largest data set of medical licensing examination questions not included in their training data set. Furthermore, we did not exclude all image-based questions a priori. Instead, we evaluated the relevance of the images for each question and compared the results both with and without images.

Methods

Data Collection

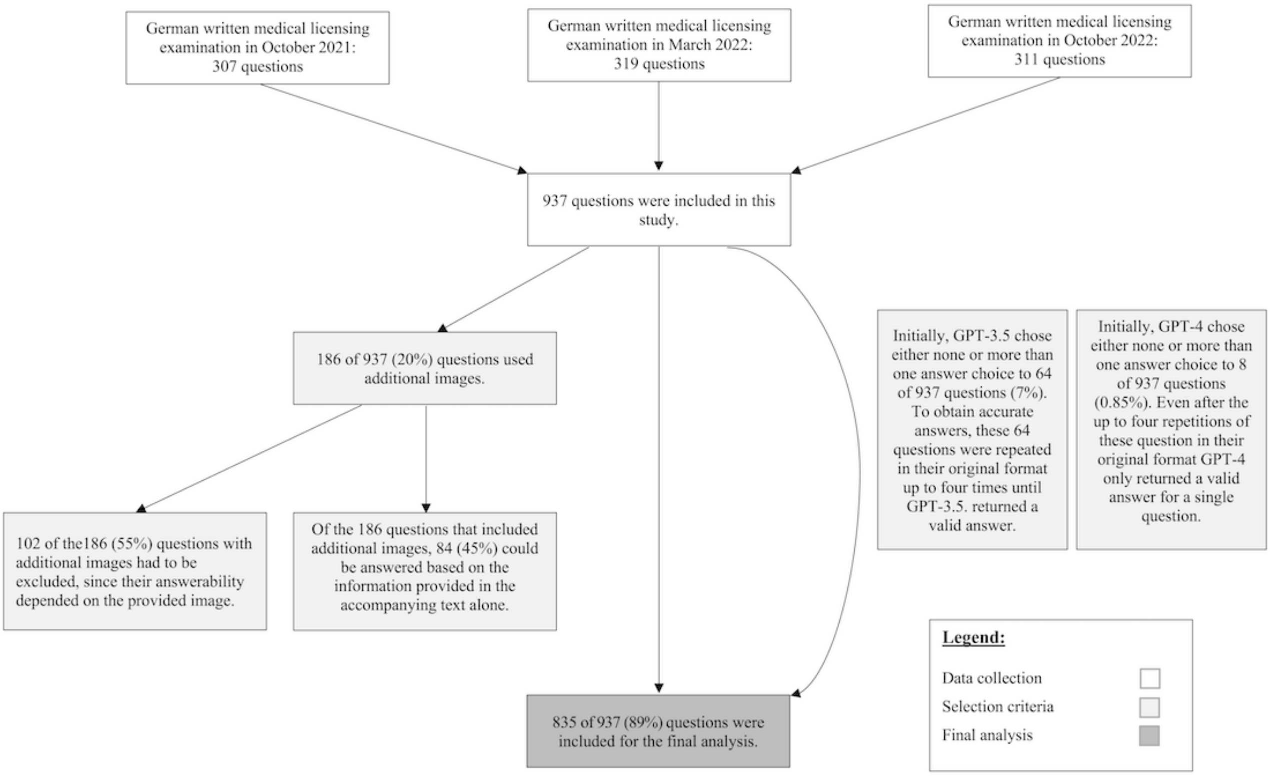
To ensure that any observed performance was not influenced by changes in ChatGPT's training data, we specifically chose the 3 most recent examinations (October 2021, April 2022, and October 2022) after the AI's knowledge cutoff date [17]. Thus, we were able to obtain 937 multiple-choice questions, each with 5 possible answers from the third-party client Amboss, a web-based learning platform that provides the original questions from the Institut für Medizinische und Pharmazeutische Prüfungsfragen (IMPP). To maintain the original examination format, we presented all obtained questions and answer options in the same order as they appeared in the examination. No specific training code was used while submitting the questions. Due to AI's inability to analyze visual content, answerability based on question text alone was defined as the primary inclusion criterion, resulting in the exclusion of 102 questions. The questions were submitted through ChatGPT's interface of the GPT-3.5 (January 30, 2023) and GPT-4 (March 14, 2023) versions. ChatGPT's answers were then compared to the official correct answers and evaluated. If ChatGPT selected none or more than 1 of the multiple-choice answers, the question was repeated in its original format up to 4 times or until a conclusive response could be obtained from ChatGPT (Figure 1).

We recorded additional data, such as answer length, content warnings, and recommendations for further diagnosis, and categorized the questioning methodology. To assess the readability of a question, we used the Simple Measure of Gobbledygook (SMOG) as it has shown acceptable interrater reliability for patient education materials in the literature [21].

Examination statistics provided by the "MEDI-LEARN" portal were also used, including the number of correct student answers and the specialization of each question. The "Blueprint" published by the IMPP outlines the distribution of subspecialties within the written state examinations [18].

XSL•FO
RenderX

Figure 1. Flowchart of the study design for the evaluation of ChatGPT's (GPT-3.5 and GPT-3) accuracy in the written German medical licensing examination (2021-2022). The flowchart presents the criteria for question selection, including both the inclusion and exclusion criteria.



Statistical Analysis

To perform our data analysis, we used several packages [22-37] in addition to the R programming language [38].

While continuous variables were reported as arithmetic mean (SD) values, categorical variables were reported as frequencies and percentages. The Kolmogorov-Smirnov test, Shapiro-Wilk test, and QQ plots were used to confirm the normal distribution of continuous data statistically and graphically. To determine significant differences, we used unpaired *t* test or ANOVA for continuous variables and chi-square test or Wilcoxon rank-sum test for categorical variables. *P* values of <.05 were deemed significant. Univariate and multivariate regression analyses were additionally performed to provide information on probabilities and predictors.

Ethical Considerations

Ethics approval was not required as data were collected from publicly available sources on the internet or were generated using AI-based methods. No personally identifiable information was used in the data collection, and all data were handled in accordance with applicable data privacy laws and regulations.

Results

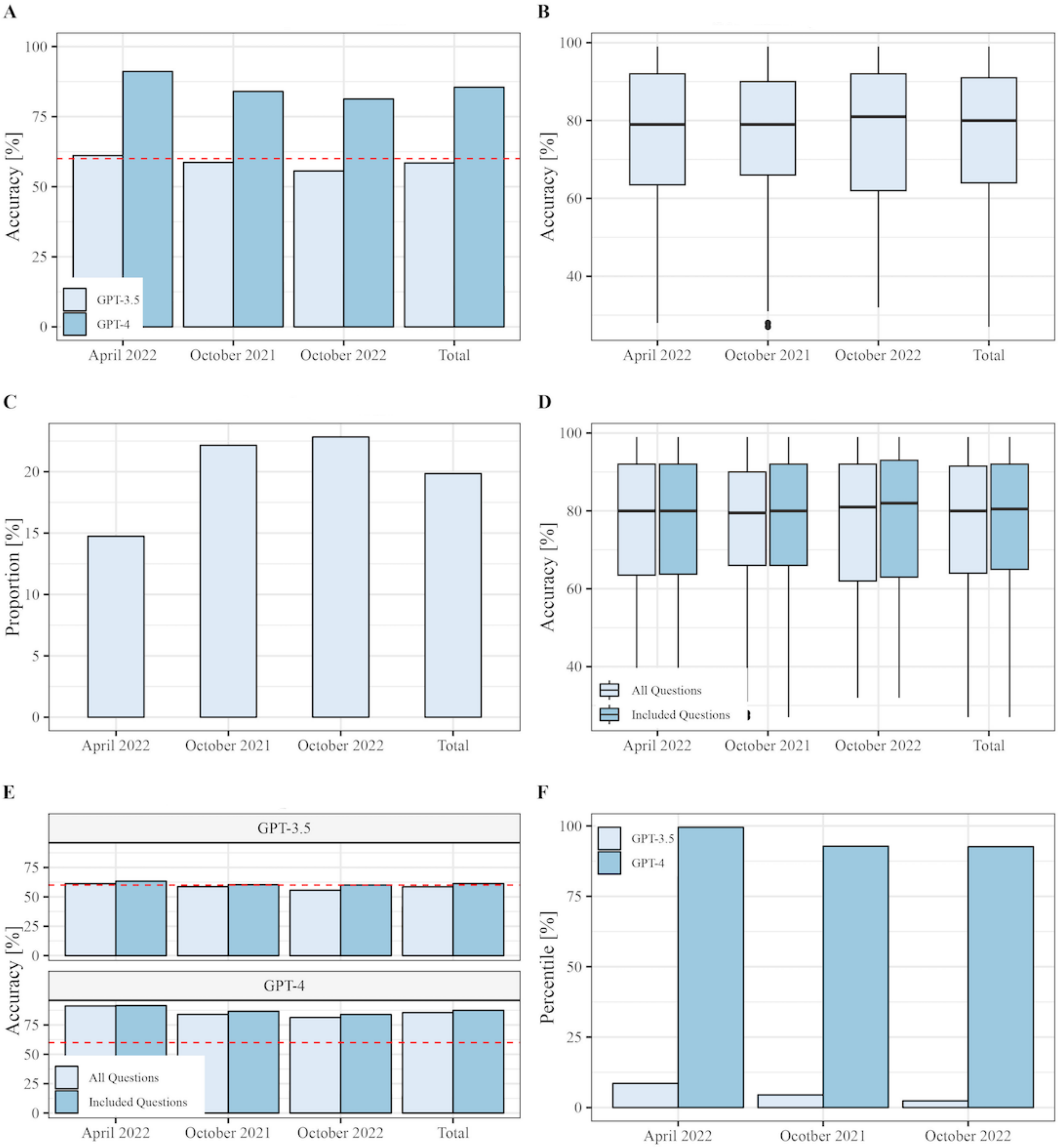
Overall, GPT-4 demonstrated superior performance with an average score of 796 out of 937 (85%), surpassing GPT-3.5's score of 548 out of 937 (58%), which previously fell below the general passing threshold of 60% (Figure 2A) [37-39]. For the April 2022 examination, GPT-3.5 and GPT-4 achieved their highest scores (GPT-3.5: 195/319, 61%; GPT-4: 287/315, 91%), while the proportion of students who answered correctly remained constant across the 3 examinations (mean 76%, SD 18%; *P*=.86; Figure 2B and Multimedia Appendix 2).

Thus, GPT-4 passed all tested examinations, whereas GPT-3.5 could only pass 1 of the 3 examinations. Although the examinations varied in several aspects, we also observed a significant difference in the number of images (*P*=.02; Figure 2C and Multimedia Appendix 2). As GPT-3.5 and GPT-4 could, at the time of the study, not process these, we further investigated the potential image-related discrepancy between the examinations by excluding from subsequent analyses any questions that required image-dependent responses. The exclusion of these questions did not significantly alter examination difficulty, as evidenced by similar student scores (Figure 2D).

Moreover, no differences were observed in the parameters collected on student accuracy, questions, or answer characteristics in relation to the performance of GPT-4 and GPT-3.5 in the excluded cases (Multimedia Appendix 3). Upon excluding image-based questions, GPT-4 continued to outperform GPT-3.5, with scores approaching 91.44%. However, GPT-3.5 exceeded expectations by achieving passing scores on all 3 examinations (October 2021: 60.22%; April 2022: 63.36%; October 2022: 60.07%; Figure 2E and Multimedia Appendix 4). GPT-3.5's accuracy (*P*=.66), the number of images (*P*=.07), and students' accuracy (*P*=.77) remained constant throughout the examinations, whereas GPT-4's accuracy (*P*=.02), the specialties (*P*<.001), and question type (*P*=.04) varied (Multimedia Appendix 4 and Figures 2A, 2B, and 2E). The details of the included questions and their respective categorizations are provided in Table 1.

XSL•FO
RenderX

Figure 2. Bar plots of ChatGPT's (GPT-3.5 and GPT-4) and box plots of students' accuracy in the written German medical licensing examination (2021-2022). Bar graphs and box plots of (A) the relative number of correct answers provided by ChatGPT (GPT-3.5 and GPT-4) answers, (B) correct answers provided by students, (C) and image-based questions for the different examinations. (D and E) The relative number of correct answers by ChatGPT (GPT-3.5 and GPT-4) and students, comparing all questions with the included text-based questions. The 60% pass mark is presented as a red line in (A) and (E) to provide context for the performance of ChatGPT (GPT-3.5 and GPT-4). In addition, (E) displays the percentile achieved by ChatGPT (GPT-3.5 and GPT-4) for each year's examination, based on the percentile limits published by the Institut für Medizinische und Pharmazeutische Prüfungsfragen [37-39].



XSL•FO
RenderX

Table 1. Summary statistics for ChatGPT's (GPT-3.5 and GPT-4) accuracy during the written German medical licensing examination, 2021-2022.

Characteristic	Overall (N=834)	Accuracy of GPT-3.5			Accuracy of GPT-4		
		False (n=323)	True (n=511)	*P* value	False (n=105)	True (n=729)	*P* value
Students' correct response rate (%), mean (SD)	77 (18)	71 (18)	80 (16)	<.001[a]	70 (18)	78 (17)	<.001[a]
Accuracy of GPT-3.5, n (%)	511 (61)	N/A[b]	N/A	N/A	38 (36)	473 (65)	<.001[c]
Accuracy of GPT-4, n (%)	729 (87)	256 (79)	473 (93)	<.001[c]	N/A	N/A	N/A
Readability score of the question, mean (SD)	14.96 (1.89)	14.93 (1.87)	14.98 (1.90)	.65[a]	14.91 (2.26)	14.97 (1.84)	.21[a]
Question type , n (%)				.76[c]	N/A	N/A	.009[c]
Connected (key feature)	532 (64)	204 (63)	328 (64)		79 (75)	453 (62)	
Single question	302 (36)	119 (37)	183 (36)		26 (25)	276 (38)	
Images referenced in questions	84 (10)	23 (7.1)	61 (12)	.02[c]	17 (16)	67 (9.2)	.03[c]
Specialty, n (%)				.02[c]	N/A	N/A	.07[c]
Gynecology	43 (5.2)	12 (3.7)	31 (6.1)		7 (6.7)	36 (4.9)	
Infectiology	74 (8.9)	24 (7.4)	50 (9.8)		6 (5.7)	68 (9.3)	
Internal medicine	176 (21)	71 (22)	105 (21)		15 (14)	161 (22)	
Neurology	112 (13)	51 (16)	61 (12)		12 (11)	100 (14)	
Others	269 (32)	106 (33)	163 (32)		46 (44)	223 (31)	
Pediatrics	62 (7.4)	26 (8.0)	36 (7.0)		11 (10)	51 (7.0)	
Psychiatry	54 (6.5)	11 (3.4)	43 (8.4)		5 (4.8)	49 (6.7)	
Surgery	44 (5.3)	22 (6.8)	22 (4.3)		3 (2.9)	41 (5.6)	
Expertise, n (%)				.64[c]	N/A	N/A	.34[c]
Background knowledge	103 (12)	32 (9.9)	71 (14)		13 (12)	90 (12)	
Complications	49 (5.9)	19 (5.9)	30 (5.9)		4 (3.8)	45 (6.2)	
Diagnostic competence	466 (56)	184 (57)	282 (55)		54 (51)	412 (57)	
Prevention competence	36 (4.3)	13 (4.0)	23 (4.5)		6 (5.7)	30 (4.1)	
Scientific practice	34 (4.1)	14 (4.3)	20 (3.9)		8 (7.6)	26 (3.6)	
Therapeutic competence	146 (18)	61 (19)	85 (17)		20 (19)	126 (17)	

[a]Wilcoxon rank-sum test.

[b]N/A: not applicable.

[c]Pearson chi-square test.

After controlling for all other variables, correct student responses (GPT-3.5: OR 0.01, 95% CI 0.00-0.01, *P*<.001; GPT-4: OR 0.00, 95% CI 0.00-0.00, *P*=.003) and questions with images (GPT-3.5: OR 0.19, 95% CI 0.08-0.30, *P*<.001; GPT-4: OR –0.09, 95% CI –0.16 to –0.01, *P*=.02) emerged as significant predictors of GPT-3.5's and GPT-4's accuracy, regardless of the version. Furthermore, our analysis revealed that only questions pertaining to psychiatry were significant predictors of correct GPT-3.5 responses (OR 0.19, 95% CI 0.02-0.36, *P*=.03). In contrast, questions related to internal medicine (OR 0.10, 95% CI 0.00-0.19, *P*=.04) and surgery (OR 0.12, 95% CI 0.00-0.25, *P*=.049) were the only medical subspecialties significantly predicting accurate responses of GPT-4. Conversely, questions concerning scientific practice (OR –0.14, 95% CI –0.29 to 0.00, *P*=.05) were less likely to be answered correctly by GPT-4 (Table 2 and Figure 3). The question SMOG readability score, however, did not significantly impact ChatGPT's accuracy.

XSL•FO
RenderX

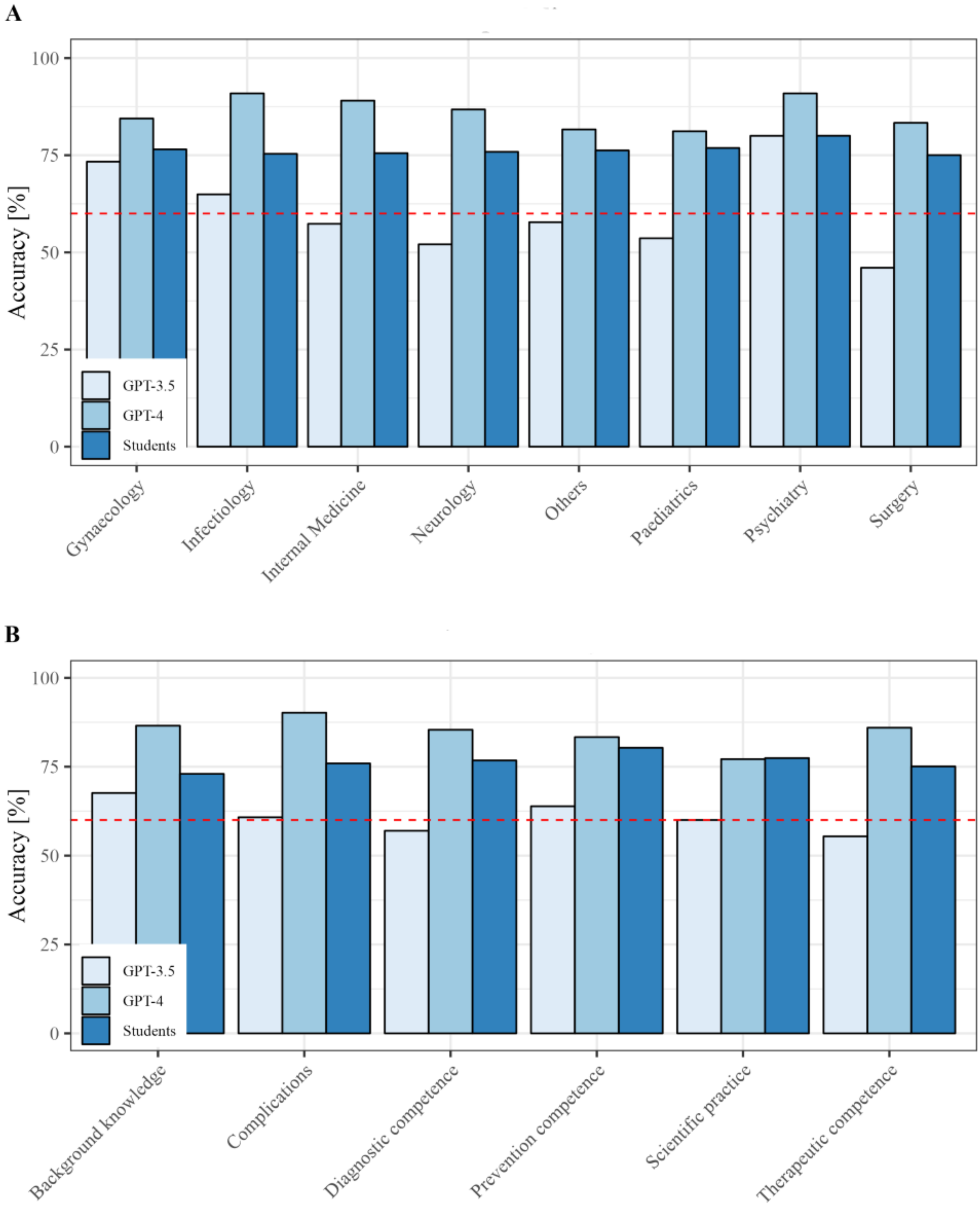
Table 2. Regression analysis to compare ChatGPT's (GPT-3.5 and GPT-4) accuracy during the written German medical licensing examination (2021-2022; N=833).

Characteristic	GPT-3.5						GPT-4					
	Univariate			Multivariate			Univariate			Multivariate		
	Odds ratio	95% CI	*P* value	β	95% CI	*P* value	Odds ratio	95% CI	*P* value	β	95% CI	*P* value
Students' correct response rate	1.03	1.02 to 1.04	<.001	.01	0.00 to 0.01	<.001	1.02	1.01 to 1.03	<.001	.00	0.00 to 0.00	.003
Accuracy of GPT-4	3.25	2.13 to 5.02	<.001	.26	0.16 to 0.36	<.001	N/A[a]	N/A	N/A	N/A	N/A	N/A
Accuracy of GPT-3.5	N/A	N/A	N/A	N/A	N/A	N/A	3.25	2.13 to 5.02	<.001	.12	0.08 to 0.17	<.001
October 2021 examination	0.94	0.70 to 1.27	.68	.00	–0.08 to 0.08	.94	0.90	0.59 to 1.40	.64	.02	–0.04 to 0.07	.55
April 2022 examination	1.15	0.86 to 1.54	.35	.03	–0.05 to 0.11	.47	1.85	1.17 to 3.03	.01	.06	0.01 to 0.11	.03
October 2022 examination	0.92	0.69 to 1.24	.59	N/A	N/A	N/A	0.63	0.42 to 0.96	.03	N/A	N/A	N/A
Question type	0.96	0.72 to 1.28	.78	–.03	–0.10 to 0.04	.39	1.86	1.18 to 3.01	.01	.06	0.02 to 0.11	.007
Images referenced in questions	1.77	1.09 to 2.98	.03	.19	0.08 to 0.30	<.001	0.52	0.30 to 0.96	.03	–.09	–0.16 to –0.01	.02
Other specialty	0.96	0.71 to 1.30	.80	.00	–0.13 to 0.14	.94	0.57	0.37 to 0.86	.007	.02	–0.07 to 0.11	.73
Gynecology and obstetrics	1.62	0.84 to 3.33	.17	.12	–0.06 to 0.31	.19	0.71	0.32 to 1.78	.42	.01	–0.12 to 0.14	.88
Surgery	0.62	0.33 to 1.14	.12	–.12	–0.30 to 0.06	.18	2.03	0.72 to 8.49	.24	.12	0.00 to 0.25	.049
Internal medicine	0.92	0.66 to 1.30	.63	–.02	–0.15 to 0.12	.81	1.7	0.99 to 3.14	.07	.10	0.00 to 0.19	.043
Infectious diseases	1.35	0.82 to 2.28	.24	.06	–0.10 to 0.22	.48	1.7	0.78 to 4.48	.23	.09	–0.02 to 0.20	.11
Psychiatry	2.61	1.37 to 5.40	.005	.19	0.02 to 0.36	.03	1.44	0.62 to 4.23	.45	.03	–0.09 to 0.15	.61
Neurology	0.72	0.49 to 1.08	.12	–.04	–0.18 to 0.11	.61	1.23	0.68 to 2.45	.52	.08	–0.02 to 0.18	.11
Pediatrics	0.87	0.52 to 1.48	.60	N/A	N/A	N/A	0.64	0.34 to 1.34	.21	N/A	N/A	N/A
Diagnostic competence	0.93	0.70 to 1.23	.60	–.03	–0.17 to 0.11	.67	1.22	0.81 to 1.85	.33	–.05	–0.14 to 0.05	.34
Therapeutic competence	0.86	0.60 to 1.24	.41	–.04	–0.19 to 0.12	.65	0.89	0.54 to 1.54	.66	–.06	–0.16 to 0.05	.28
Background knowledge	1.47	0.95 to 2.32	.09	.08	–0.09 to 0.24	.36	1.00	0.55 to 1.94	>.99	–.05	–0.16 to 0.06	.36
Prevention competence	1.13	0.57 to 2.32	.74	.00	–0.20 to 0.20	>.99	0.71	0.31 to 1.93	.45	–.11	–0.25 to 0.03	.11
Scientific practice	0.90	0.45 to 1.85	.77	.01	–0.20 to 0.22	.95	0.45	0.21 to 1.09	.06	–.14	–0.29 to 0.00	.05
Complications	1.00	0.56 to 1.84	>.99	N/A	N/A	N/A	1.66	0.66 to 5.61	.34	N/A	N/A	N/A
Readability score of the question	1.01	0.94 to 1.09	.70	.01	–0.01 to 0.03	.24	1.02	0.91 to 1.14	.76	.00	–.01 to 0.01	.98

[a]N/A: not applicable.

XSL•FO
RenderX

Figure 3. Comparison of ChatGPT's (GPT-3.5 and GPT-4) and students' relative accuracy in relation to the tested specialties and methodology in the written German medical licensing examination (2021-2022). The bar graph displays the percentage of correct answers provided by ChatGPT (GPT-3.5 and GPT-4) and students in (A) each specialty and (B) and methodology, while the blue line demonstrates a 60% pass mark.



Discussion

Principal Findings

With the introduction of ChatGPT's GPT-3.5 and GPT-4 iterations, the potential application for AI in research, patient care, and medical education is gaining recognition [2,8,40]. By improving the users' experience and facilitating more efficient information retrieval, ChatGPT might even revolutionize the future of search engines and shift the focus of medical education from memorization to practical application [8,10,11].

Under this premise, the nearly passing scores of the freely available GPT-3.5 iteration, along with the exceptional scores of GPT-4, are highly relevant. Even with the varying scores of 51%-67% of GPT-3.5 across various input languages [13-15,41,42], both models consistently outperform most prominent general and domain-specific LLMs, such as

XSL•FO
RenderX

InstructGPT (53%), GPT-3 (25%), and BioMedLM (50%) [14,43,44]. Despite these improvements, GPT-3.5's or GPT-4's performance still fell short in comparison to that of medical students in a Japanese medical examination according to the study by Takagi et al [15]. In comparison to the German medical students, however, GPT-3.5 scored in the 8.6th percentile, while GPT-4 ranked in the 92.8th, 99.5th, and 92.6th percentiles in the October 2021, April 2022, and October 2022 examinations [39,45,46]. The observed variations in the AI's accuracy across input languages may partially reflect the language composition of their data sets, as LLMs tend to favor languages that are more represented in their training data [16,17]. Since ChatGPT appears to perform optimally with English inputs, language emerges as a limiting factor for its accuracy, suggesting that globally consistent application is dependent upon users' proficiency in English.

Moreover, the nearly 30% performance increase from GPT-3.5 to GPT-4, as indicated in this study and supported by a Japanese study, which suggests a similar language distribution within the GPT-3.5 and GPT-4 data sets [15]. GPT-4, unlike GPT-3.5, also did not answer questions containing images on repetition, showing an improvement in the previously incorrect content produced by GPT-4's predecessor [17].

Thus, health care professionals could potentially benefit, especially from GPT-4's conclusive and often nonobvious insights to multiple-choice questions, as these users have the ability to verify crucial details [13,14,41]. For instance, there is potential for using GPT-3.5 and GPT-4 in a medical education tutoring environment, as evidenced by its successful application in anatomy [47]. However, when using either GPT-3.5 or GPT-4 for medical applications, its differing accuracy across specialties must also be taken into account [48]. GPT-3.5 initially displayed a high degree of accuracy within the field of psychiatry, while GPT-4 demonstrated its strength in internal medicine and surgery. Considering the rising prevalence of psychiatric disorders and concomitant challenges in providing care, it seemed likely that nonprofessionals would also turn to the chatbot for mental health issues at the time of GPT-3.5's release [8,49,50]. Hence, it is conceivable that GPT-3.5's training data set includes not only a substantial and reliable portion of psychiatric data, but also its developers might have first fine-tuned ChatGPT specifically in this domain in anticipation of its high demand [51-53]. Thus, the developers might have also fine-tuned GPT-4 specifically in internal medicine and surgery, possibly reacting to a high demand in this area from users of its' predecessor. GPT-4's impressive performance is not limited to the medical field, as it demonstrated comparable percentile scores in the Uniform Bar Exam, showcasing it potential as a versatile tool across diverse academic disciplines [17]. However, assessing the possible reasons for the performance differences between GPT-3.5 and GPT-4 is complicated by the confidential architecture of GPT-4 [54], posing challenges for research on future applications.

In turn, GPT-4's excellent achievements shed light on the limitations of current testing paradigms in medical education that often favor rote memorization over a critical and context-aware approach. They also highlight the inadequacy of multiple-choice questions as a means of assessing medical knowledge, as they tend to encourage binary thinking as "true" and "false," which often fails to capture the complex reality of the medical practice [11]. Although GPT-3.5 and GPT-4 allow the simple and fast retrieval of medical information from any internet-capable device that fits in one's pocket [9,10], neither GPT-3.5 nor GPT-4 verifies the information they provide. Thus, ChatGPT's output needs to be approached with a critical mindset, recognizing that misinformation may be more difficult to detect than in the output of other search engines that offer multiple sources in response to a query and take login credentials into account [8,55]. To navigate these changing informational landscapes, a basic understanding in data science seems necessary alongside traditional medical expertise [56]. It may even be beneficial for future iterations of AI tools to include references to the sources underlying each search in order to increase transparency and allow users to assess the reliability of the information they receive.

In a previous study by Nov et al [57], considering that 59% of participants trusted chatbots more than traditional search engines, it must be noted that GPT-3.5 and GPT-4 have only been tested on medical examination questions and not questions by nonprofessionals, limiting general recommendations for unsupervised patient education or the general population. It seems evident that GPT-4 has been benchmarked against medical licensing examinations, explaining not only GPT-4's excellent scores but also exceeding achievements in internal medicine and surgery, which, for instance, have been overrepresented in the medical examinations assessed in this study [12,17].

Since GPT-3.5 failed the German medical licensing examination by a narrow margin, its use for answering medical questions is generally not advisable. Moreover, the remarkable performance of GPT-4 in the German Medical State Examination may not be universally applicable outside a medical examination setting, especially considering that GPT-4 was presumably benchmarked on academic and professional examinations [17].

As literature on ChatGPT is scarce, and it can be difficult to detect incorrect output from this AI tool, the content it generates must be carefully assessed. Nevertheless, medical professionals may still be able to benefit from GPT-3.5's and GPT-4's explanations and, in some cases, gain new nonobvious insights. With the release of GPT-4's ability to handle pictures on the horizon, the potential for further applications of GPT-3.5 and GPT-4 to improve the medical workflow or medical education seems eminent, emphasizing the need for continued research into AI.

Limitations

This study's findings on GPT-3.5's and GPT-4's medical proficiencies are limited to multiple-choice questions from the German medical licensing examination, which may not be representative of other types of examinations or contexts. However, it is worth noting that GPT-3.5 and GPT-4 have demonstrated similar performances in examinations in other countries and languages, which suggests some degree of generalizability.

XSL•FO
RenderX

In addition, the sample size of 937 questions and the exclusion of image-based questions may not capture the full range of difficulty levels or content areas. Although the collected parameters did not differ in terms of GPT-3.5's and GPT-4's accuracy in the excluded cases, the decision to exclude image-based questions may have introduced a sampling bias. By testing for differences, efforts were made to minimize this bias and maintain the integrity of the results.

As GPT-3.5's and GPT-4's performances were compared to those of German medical students using the MEDI-LEARN service, a selection bias might have been introduced. However, the high correlation between the MEDI-LEARN statistics and the IMPP statistics indicates at best a weak expression of this selection bias [58].

It should also be noted that a replication of this study might not yield the exact same results, as the literature suggests that GPT-3.5 is inconsistent in answering 15% of medical questions [59]. However, the trends observed in this study appear to be consistent with those reported in other published and preprint studies on GPT-3.5's and GPT-4's performance.

Conclusions

In conclusion, the results of this study indicate that only GPT-4 consistently passed all 3 medical examinations, ranking in the 92.8th to 99.5th percentile in comparison to medical students. These findings highlight the strengths and limitations of ChatGPT in the context of medical examinations and raise questions about the future of medical education.

Although GPT-3.5's and GPT-4's accuracy in medical examinations seems consistent across different countries and languages, its inconsistencies, potential biases, and number of incorrect answers restrain a recommendation for its use by the general population for medical purposes. However, its elaborate explanations and potential to yield nonobvious insights may benefit medical professionals in training.

While this study hints to a moderate accuracy of GPT-3.5 and a stellar performance of GPT-4 in answering medical examination questions, further research is necessary to gain deeper insights, explore future applications, and ensure safe use of ChatGPT for end users.

Acknowledgments

The authors thank Dorothee Meyer, Linea Luise Fuchs, Ari Soleman, GPT-3.5, and GPT-4 for proofreading this manuscript. In this study, we used ChatGPT for several purposes: to translate our manuscript into English, to refine its linguistic presentation, to evaluate and improve our methodological approach, and to scrutinize the R code underlying our statistical analysis, with a particular focus on identifying and resolving any error warnings generated. Subsequently, all outputs provided by ChatGPT were rigorously reviewed and critically appraised by the authors to ensure accuracy and reliability.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Responses of (A) GPT-3.5 and (B) GPT-4 to the queries on its ability to pass a medical exam, 2023.
[DOCX File , 592 KB - mededu_v10i1e50965_app1.docx]

Multimedia Appendix 2
Summary statistics for all questions regarding exam time and ChatGPT's (GPT-3.5 and GPT-4) accuracy in the German medical licensing exam, 2021-2022.
[DOCX File , 21 KB - mededu_v10i1e50965_app2.docx]

Multimedia Appendix 3
Summary statistics for excluded questions regarding ChatGPT's (GPT-3.5 and GPT-4) accuracy in the German medical licensing exam, 2021-2022.
[DOCX File , 20 KB - mededu_v10i1e50965_app3.docx]

Multimedia Appendix 4
Summary statistics for included questions regarding exam time in the German medical licensing exam, 2021-2022.
[DOCX File , 17 KB - mededu_v10i1e50965_app4.docx]

References

1. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023 Apr;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
2. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023 Mar;5(3):e102 [FREE Full text] [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]

XSL•FO
RenderX

3. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 2023 Mar;5(3):e105-e106 [FREE Full text] [doi: 10.1016/S2589-7500(23)00019-5] [Medline: 36754725]
4. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108 [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
5. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2023 Oct 05. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
6. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 2023 Feb 17;13:01003 [FREE Full text] [doi: 10.7189/jogh.13.01003] [Medline: 36798998]
7. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75 [FREE Full text] [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
8. Kurz C, Lau T, Martin M. ChatGPT: Noch kein Allheilmittel. Dtsch Arztebl International 2023;120(6):A-230-B-202.
9. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023 Apr;185:109729. [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
10. Aljanabi M, Ghazi M, Ali AH, Abed SA, ChatGpt. ChatGpt: open possibilities. Iraqi J Comp Sci Math 2023 Jan 18:62-64. [doi: 10.52866/20ijcsm.2023.01.01.0018]
11. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb 9;2(2):e0000205 [FREE Full text] [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]
12. Sanderson K. GPT-4 is here: what scientists think. Nature 2023 Mar;615(7954):773. [doi: 10.1038/d41586-023-00816-5] [Medline: 36928404]
13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
15. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
16. Gabriel N, Bhatia A. Lost in Translation: Large Language Models in Non-English Content Analysis. Center for Democracy & Technology. 2023. URL: https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ [accessed 2024-01-28]
17. OpenAI. GPT-4 Technical Report. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2024-01-28]
18. Institut für medizinische und pharmazeutische Prüfungsfragen. Zusammenstellung der Prüfungsinhalte für den Zweiten Abschnitt der Ärztlichen Prüfung („Blueprint“) nach derzeit gültiger ÄApprO 2002. IMPP. URL: https://www.impp.de/blueprint-m2-examen.html?file=fi- [accessed 2023-11-09]
19. Jünger J. Kompetenzorientiert prüfen im Staatsexamen Medizin [Competence-based assessment in the national licensing examination in Germany]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Feb 11;61(2):171-177. [doi: 10.1007/s00103-017-2668-9] [Medline: 29230515]
20. Examen (M2/M3) No.1 in der Examensvorbereitung. AMBOSS. URL: https://www.amboss.com/de/examen-m2-m3 [accessed 2023-07-15]
21. Grabeel KL, Russomanno J, Oelschlegel S, Tester E, Heidel RE. Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. J Med Libr Assoc 2018 Jan;106(1):38-45 [FREE Full text] [doi: 10.5195/jmla.2018.262] [Medline: 29339932]
22. Müller K. here: A Simpler Way to Find Your Files. here. URL: https://here.r-lib.org/ [accessed 2024-01-28]
23. Chan CH, Leeper TJ, Becker J. rio: A Swiss-Army Knife for Data I/O. URL: https://cran.r-project.org/web/packages/rio/readme/README.html#:~:text=Overview,or%20a%20specified%20format%20argument) [accessed 2024-01-28]
24. Wickham H. Easily Install and Load the Tidyverse. tidyverse. URL: https://tidyverse.tidyverse.org/ [accessed 2024-01-28]
25. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. JOSS 2019;4(43):1686. [doi: 10.21105/joss.01686]
26. Robinson D, Hayes A, Couch S. broom: Convert Statistical Objects into Tidy Tibbles. broom. 2023. URL: https://broom.tidymodels.org/ [accessed 2024-01-28]
27. Larmarange J. labelled: Manipulating Labelled Data. labelled. URL: https://larmarange.github.io/labelled/ [accessed 2024-01-28]

XSL•FO
RenderX

28. Sjoberg DD, Whiting K, Curry M, Lavery J, Larmarange J. Reproducible Summary Tables with the gtsummary Package. The R Journal 2021;13(1):570-580. [doi: 10.32614/RJ-2021-053]
29. Sjoberg DD, Whiting K, Curry M, Lavery J, Larmarange J. Reproducible Summary Tables with the gtsummary Package. The R Journal 2021;13(1):570. [doi: 10.32614/rj-2021-053]
30. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. ggpubr. URL: https://rpkgs.datanovia.com/ggpubr/ [accessed 2024-01-28]
31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12(1):77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
32. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12(1):77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
33. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer; 2016.
34. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. ggplot2. URL: https://ggplot2.tidyverse.org/reference/ggplot2-package.html [accessed 2024-01-28]
35. Wilke CO. cowplot – Streamlined plot theme and plot annotations for ggplot2. cowplot. URL: https://wilkelab.org/cowplot/ [accessed 2024-01-28]
36. Lüdecke D. sjPlot - Data Visualization for Statistics in Social Science. sjPlot. URL: https://strengejacke.github.io/sjPlot/ [accessed 2024-01-28]
37. Dietrich J, Leoncio W. citation: Software Citation Tools. Zenodo. URL: https://zenodo.org/records/3909438 [accessed 2024-01-28]
38. The R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.
39. Herbst 2021 - Ergebnisinformartion. Institut für medizinische und pharmazeutische Prüfungsfragen. URL: https://www.impp.de/pruefungen/medizin/archiv-medizin.html?file=files/PDF/Pr%C3%BCfungsergebnisse/Pr%C3%BCfungsergebnisse/ErgMedM2H2021APPO2012.pdf [accessed 2024-01-28]
40. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Feb;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
41. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023;20:1 [FREE Full text] [doi: 10.3352/jeehp.2023.20.1] [Medline: 36627845]
42. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT besteht schriftliche medizinische Staatsexamina nach Ausschluss der Bildfragen. Dtsch Arztebl International 2023;120:373-374.
43. Abhinav V, Jonathan F, Carbin M. BioMedLM: a Domain-Specific Large Language Model for Biomedical Text. Mosaic ML. 2023. URL: https://www.mosaicml.com/blog/introducing-pubmed-gpt [accessed 2024-01-28]
44. Jin D, Pan E, Oufattole N, Weng W, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences 2021 Jul 12;11(14):6421. [doi: 10.3390/app11146421]
45. Frühjahr 2022 - Ergebnisinformartion. Institut für medizinische und pharmazeutische Prüfungsfragen. URL: https://www.impp.de/pruefungen/medizin/archiv-medizin.html?file=files/PDF/Pr%C3%BCfungsergebnisse/Pr%C3%BCfungsergebnisse/ErgMedM2F2022APPO2012.pdf [accessed 2024-01-28]
46. Herbst 2022 - Ergebnisinformartion. Institut für medizinische und pharmazeutische Prüfungsfragen. URL: https://www.impp.de/pruefungen/medizin/archiv-medizin.html?file=files/PDF/Prüfungsergebnisse/Prüfungsergebnisse/ErgMedM2H2022.pdf [accessed 2024-01-28]
47. Mogali SR. Initial impressions of ChatGPT for anatomy education. Anat Sci Educ 2023 Feb 07:n/a. [doi: 10.1002/ase.2261] [Medline: 36749034]
48. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
49. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
50. Lau T. KI-Chatbot könnte Therapiegespräche empathischer machen. aerzteblatt.de. URL: https://www.aerzteblatt.de/nachrichten/140445/KI-Chatbot-koennte-Therapiegespraeche-empathischer-machen [accessed 2023-03-14]
51. Budler LC, Gosak L, Stiglic G. Review of artificial intelligence ‐ based question ‐ answering systems in healthcare. WIREs Data Mining and Knowledge Discovery 2023 Jan 10;13(2):e1487. [doi: 10.1002/widm.1487]

XSL•FO
RenderX

52. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2011 Jun 20;42(1):41-50. [doi: 10.1017/s0033291711000997]
53. Van Le D, Montgomery J, Kirkby KC, Scanlan J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. J Biomed Inform 2018 Oct;86:49-58 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.007] [Medline: 30118855]
54. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
55. Epstein RH, Dexter F. Variability in large language models' responses to medical licensing and certification examinations. Comment on "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment". JMIR Med Educ 2023 Jul 13;9:e48305 [FREE Full text] [doi: 10.2196/48305] [Medline: 37440293]
56. Seth P, Hueppchen N, Miller SD, Rudzicz F, Ding J, Parakh K, et al. Data science as a core competency in undergraduate medical education in the age of artificial intelligence in health care. JMIR Med Educ 2023 Jul 11;9:e46344 [FREE Full text] [doi: 10.2196/46344] [Medline: 37432728]
57. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (Turing) test: survey study. JMIR Med Educ 2023 Jul 10;9:e46939 [FREE Full text] [doi: 10.2196/46939] [Medline: 37428540]
58. FAQ: Häufig gefrage Fragen. MEDI-LEARN. URL: https://www.mlmr.de/unis/faq/#faq1 [accessed 2024-01-28]
59. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. Eur J Hum Genet 2023 May 29:2023. [doi: 10.1038/s41431-023-01396-8] [Medline: 37246194]

Abbreviations

AI: artificial intelligence
IMPP: Institut für Medizinische und Pharmazeutische Prüfungsfragen
LLM: large language model
SMOG: Simple Measure of Gobbledygook

Edited by K Venkatesh; submitted 18.07.23; peer-reviewed by A Thirunavukarasu, H Alshawaf, M Brown, X Li, I Albalawi; comments to author 08.11.23; revised version received 14.11.23; accepted 11.12.23; published 08.02.24.

Please cite as:
Meyer A, Riese J, Streichert T
Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study
JMIR Med Educ 2024;10:e50965
URL: https://mededu.jmir.org/2024/1/e50965
doi:10.2196/50965
PMID:38329802

©Annika Meyer, Janik Riese, Thomas Streichert. Originally published in JMIR Medical Education (https://mededu.jmir.org), 08.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study

Peng Yu[1*], MD; Changchang Fang[1*], MD; Xiaolin Liu[2], MD; Wanying Fu[1], MD; Jitao Ling[1], MD; Zhiwei Yan[3], MD; Yuan Jiang[4], MD; Zhengyu Cao[4], MD; Maoxiong Wu[4], MD; Zhiteng Chen[4], MD; Wengen Zhu[5], MD; Yuling Zhang[4], MD; Ayiguli Abudukeremu[4], MD; Yue Wang[4], MD; Xiao Liu[4], MD; Jingfeng Wang[4], MD

[1]Department of Endocrine, The Second Affiliated Hospital of Nanchang University, Jiangxi, China
[2]Department of Cardiology, The Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, China
[3]College of Kinesiology, Shenyang Sport University, Shenyang, China
[4]Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China
[5]Department of Cardiology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[*]these authors contributed equally

Corresponding Author:
Xiao Liu, MD
Department of Cardiology
Sun Yat-sen Memorial Hospital of Sun Yat-sen University
107 Yanjiang West Road
Guangzhou
China
Phone: 86 15083827378
Email: liux587@mail.sysu.edu.cn

Abstract

Background: ChatGPT, an artificial intelligence (AI) based on large-scale language models, has sparked interest in the field of health care. Nonetheless, the capabilities of AI in text comprehension and generation are constrained by the quality and volume of available training data for a specific language, and the performance of AI across different languages requires further investigation. While AI harbors substantial potential in medicine, it is imperative to tackle challenges such as the formulation of clinical care standards; facilitating cultural transitions in medical education and practice; and managing ethical issues including data privacy, consent, and bias.

Objective: The study aimed to evaluate ChatGPT's performance in processing Chinese Postgraduate Examination for Clinical Medicine questions, assess its clinical reasoning ability, investigate potential limitations with the Chinese language, and explore its potential as a valuable tool for medical professionals in the Chinese context.

Methods: A data set of Chinese Postgraduate Examination for Clinical Medicine questions was used to assess the effectiveness of ChatGPT's (version 3.5) medical knowledge in the Chinese language, which has a data set of 165 medical questions that were divided into three categories: (1) common questions (n=90) assessing basic medical knowledge, (2) case analysis questions (n=45) focusing on clinical decision-making through patient case evaluations, and (3) multichoice questions (n=30) requiring the selection of multiple correct answers. First of all, we assessed whether ChatGPT could meet the stringent cutoff score defined by the government agency, which requires a performance within the top 20% of candidates. Additionally, in our evaluation of ChatGPT's performance on both original and encoded medical questions, 3 primary indicators were used: accuracy, concordance (which validates the answer), and the frequency of insights.

Results: Our evaluation revealed that ChatGPT scored 153.5 out of 300 for original questions in Chinese, which signifies the minimum score set to ensure that at least 20% more candidates pass than the enrollment quota. However, ChatGPT had low accuracy in answering open-ended medical questions, with only 31.5% total accuracy. The accuracy for common questions, multichoice questions, and case analysis questions was 42%, 37%, and 17%, respectively. ChatGPT achieved a 90% concordance across all questions. Among correct responses, the concordance was 100%, significantly exceeding that of incorrect responses (n=57, 50%; $P<.001$). ChatGPT provided innovative insights for 80% (n=132) of all questions, with an average of 2.95 insights per accurate response.

XSL•FO
RenderX

Conclusions: Although ChatGPT surpassed the passing threshold for the Chinese Postgraduate Examination for Clinical Medicine, its performance in answering open-ended medical questions was suboptimal. Nonetheless, ChatGPT exhibited high internal concordance and the ability to generate multiple insights in the Chinese language. Future research should investigate the language-based discrepancies in ChatGPT's performance within the health care context.

(JMIR Med Educ 2024;10:e48514) doi:10.2196/48514

KEYWORDS

ChatGPT; Chinese Postgraduate Examination for Clinical Medicine; medical student; performance; artificial intelligence; medical care; qualitative feedback; medical education; clinical decision-making

Introduction

Artificial intelligence (AI) was initially conceptualized in 1956 [1], but it has only gained significant momentum in recent years. AI aims to replicate human intelligence and thinking processes through the use of brain-like computer systems to solve complex problems. What is most inspiring is that AI systems can be trained on specific data sets to improve prediction accuracy and tackle intricate problems [2-4], which means that one of the possible applications of AI is the ability to help doctors to rapidly search through vast amounts of medical data, enhancing their creativity and enabling them to make error-free decisions [5,6].

ChatGPT (OpenAI) is an AI model that has spurred great attention due to the revolutionary innovations in its ability to perform a diverse array of natural language tasks. By using a class of large-scale language models, ChatGPT (version 3.5) can predict the likelihood of a sequence of words based on the context of the preceding words. With sufficient training on vast amounts of text data, ChatGPT can generate novel word sequences that closely resemble natural human language and have never been observed before by other AI [7].

A study was conducted on the effectiveness of the version of generative pretrained transformer's large-scale language model (ChatGPT, version 3.5) in passing the United States Medical Licensing Examination (USMLE). The results showed that the AI model achieved an accuracy rate of over 50% in all the tests, and in some analyses, it even surpassed 60% accuracy. It is imperative to highlight and emphasize that the study was conducted mostly using English input, and the AI model was also trained in English.

However, despite the success of AI models like ChatGPT in the English language, their performance in understanding and generating medical text in the Chinese language remains largely unexplored because ChatGPT's ability to understand and generate text in any given language is limited by the quality and quantity of training data available in that language. Chinese is the second-most widely spoken language in the world, with more than 1.3 billion speakers globally, while the quality and quantity of Chinese language data may not be compared with English due to some reasons, such as complexity of the written characters. Thus, the performance of ChatGPT in Chinese medical information warrants further investigation.

In this study, ChatGPT's clinical reasoning ability was evaluated by administering questions from the Chinese Postgraduate Examination for Clinical Medicine. This standardized and regulated test assesses candidates' comprehensive abilities. The questions are textually and conceptually dense, and the difficulty and complexity of the questions are highly standardized and regulated. Additionally, this examination has demonstrated remarkable stability in raw scores and psychometric properties over the past years. Moreover, the examination comprises 43% (n=71) basic science and medical humanities, with 14% (n=23) physiology, 10% (n=17) biochemistry, 13% (n=28) pathology, and 6% (n=10) medical humanities. Clinical medicine makes up the remaining 57% (n=94), with internal medicine and surgery accounting for 37% (n=61) and 20% (n=33), respectively. Due to the examination's linguistic and conceptual complexity, we hypothesize that it will serve as an excellent challenge for ChatGPT. By evaluating ChatGPT's performance on this examination, we aimed to gain insights into the AI model's potential for understanding and generating medical text in Chinese and assess its applicability in Chinese medical education and clinical practice.

Methods

Ethical Considerations

This study does not involve direct interaction with human participants or the collection of personal identifiable information. As a result, it falls under the category of nonhuman subject research. Therefore, no human subject ethics review approvals were required for this study. Since this study does not involve human participants or the collection of personal identifiable information, obtaining informed consent from individuals is not applicable. As this study does not involve the collection or use of personal identifiable information, privacy and confidentiality concerns are not applicable. Since this study does not involve human participants, there is no compensation provided to individuals.

Artificial Intelligence

ChatGPT uses self-attention mechanisms and extensive training data to generate contextually relevant responses in a conversational setting. It excels in managing long-range dependencies and creating coherent replies. However, it is important to clarify that ChatGPT (version 3.5), a server-based language model, does not possess internet browsing or search functionalities. Consequently, its responses are constructed solely on abstract relationships between words or "tokens" within its neural network [7]. Furthermore, it should be noted that OpenAI released the latest version, ChatGPT (version 4), in March 2023, but the data in this study were from February 2023, when ChatGPT (version 3.5) was the most recent version.

XSL•FO
RenderX

Input Source

The Chinese Postgraduate Examination for Clinical Medicine questions from 2022 were not officially released. However, a comprehensive set of 165 questions totaling 500 points was found on the web (Table S1 in Multimedia Appendix 1) and treated as original questions. Point values differed among question types: each case analysis question (CAQ) and multichoice question (MCQ) was worth 2 points, while common questions (CQs) were either worth 1.5 or 2 points each. All inputs fed into the ChatGPT (version 3.5) model were valid samples, not part of the training data set. This was due to the database not being updated since September 2021, predating the release of these questions. For future research convenience, the 165 questions were categorized into three types:

1. CQs (n=90): These questions are to evaluate the knowledge of basic science in physiology, biochemistry, pathology, and medical humanities. Each question has 4 choices, and the respondent should select only the correct answers. For example: "The closing time of the aortic valve during the cardiac cycle is? (A) Atrial systolic end card, (B) Rapid ejection beginning, (C) Slow ejection beginning, (D) Isovolumic diastole beginning."
2. CAQs (n=45). It is a method used in clinical medicine to examine and evaluate patient cases. It involves an in-depth review of a patient's medical history, presenting symptoms, laboratory and imaging results, and diagnostic findings to arrive at a diagnosis and treatment plan. There are 4 choices, and the respondent should select only the correct answers. The difference between CAQs and CQs is that CQs focus on clinical decision-making. For example: "A 38-year-old male, suffering chest pain and fever for 3 days, having a 5 years of diabetes history. Physical examination: *t*=37.6°C, right lower lung turbid knock, breathing sound is reduced. A chest X radiograph suggests a right pleural effusion. Pleural aspiration liquefaction test showed WBC 650×106/L with fine lymph Cell 90% in pleural fluid, with glucose of 3.2 mmol/L, the diagnosis for this patient is? (A) Tuberculous pleurisy, (B) Malignant pleural effusion, (C) Empyema, (D) Pneumonia-like pleural effusion."
3. MCQs (n=30): There are 4 choices, and the respondent should select at least 2 correct answers. There is no point for choosing more or less. For example: "The structures of auditory bone conduction include? (A) Skull, (B) Round window film, (C) Ossicular chain, (D) Cochlear bone wall."

Scoring

Initially, the question format had to be adjusted to properly evaluate the performance of ChatGPT in the Chinese Postgraduate Examination for Clinical Medicine questions. Specifically, we included a "multichoice" or "single-choice" notation, as we found ChatGPT's responses varied without these cues. MCQs were adjusted to state "Please choose one or more correct options," while CQs and CAQs were altered to indicate "There is only one correct answer." This adjustment was necessary for evaluating ChatGPT's performance in the Chinese language.

We then compiled a data set of these examination questions along with their correct answers. To ensure validity, the answers were cross-verified with web-based resources and consultations with senior doctors. ChatGPT's performance was then evaluated by comparing its responses to the standard answers in the data set. A high examination score would suggest that ChatGPT handled this task effectively.

In our comprehensive analysis, we also delved into examining the correlation between different question types and accuracy using the Pearson correlation coefficient as a statistical measure to investigate this relationship.

Encoding

The structured examination questions were transformed into open-ended inquiries for better simulation of real-world clinical scenarios. Multiple-choice questions for the CAQ were removed, and ChatGPT was required to diagnose the patient's disease and prove its reason.

Regarding the MCQs, we eliminated all the choices and did not prompt ChatGPT about the existence of multiple correct answers. The CQs were treated similarly to the MCQs. However, we encountered a distinct subset within these 3 categories that could not be processed like the other questions. This subset comprised questions that required 1 answer choice to be selected from the provided options. Therefore, these questions were converted into a special format (n=26).

For instance, an original question like, "Which can inhibit insulin secretion? (A) Increased free fatty acids in blood, (B) Increased gastric inhibitory peptide secretion, (C) Sympathetic nerve excitation, (D) Growth hormone secretion increases" was reformatted as "Can an increase in free fatty acids in the blood, an increase in gastric inhibitory peptide secretion, an increase in sympathetic nerve excitation, or an increase in growth hormone secretion inhibit insulin secretion?" This encoding strategy was applied across all 3 subgroups.

Additionally, to mitigate potential memory retention bias, we commenced a new chat session for each query. This process of reformatting questions, presenting them to ChatGPT, and initiating new sessions for each question constituted our methodology for evaluating ChatGPT's performance using the data set. The clarity of this process should address the concerns raised in the comment about the lack of understanding of the way we used the data set for evaluation.

Adjudication

To assess ChatGPT's performance thoroughly, 2 physicians independently scored AI outputs for accuracy, concordance, and insight using predefined criteria (Table S2 in Multimedia Appendix 1). These physicians were not aware of each other's evaluations. To familiarize the physicians with the scoring system, a subset of 20 questions was used for training, during which the physicians were unblinded to each other's assessments.

ChatGPT's responses were classified into 3 categories under the accuracy parameter: accurate, inaccurate, and indeterminate. "Accurate" responses were those where ChatGPT provided the right answer, while "inaccurate" encompassed instances of no answer, an incorrect response, or multiple answers containing incorrect options. "Indeterminate" responses were those where

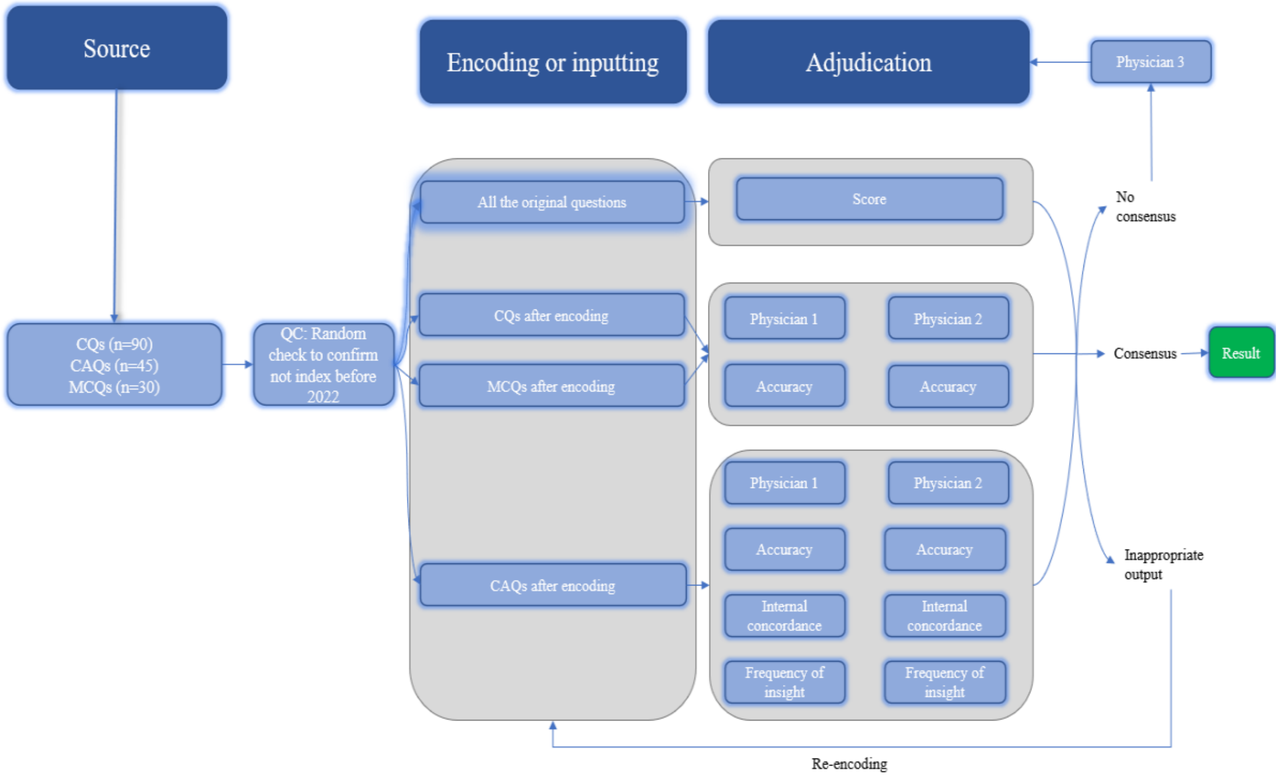
XSL•FO
RenderX

the AI output did not present a definitive answer, suggesting insufficient information to make a selection.

Concordance was determined by whether ChatGPT's explanation affirmed its provided answer, with discordance occurring if the explanation contradicted the answer. We defined valuable insights as unique text segments within the AI's explanations meeting specific criteria: they were nondefinitional, nonobvious, valid, and unique. These insights required additional knowledge or deductions beyond the input question, provided accurate clinical or numerical information, and potentially eliminated multiple answer choices with a single insight.

To mitigate potential within-item anchoring bias, the adjudicators first evaluated the accuracy for all items, followed by concordance. In case of discrepancies in domain assessments, a third physician adjudicator was consulted. This third-party intervention was required for 11 items (n=11, 7% of the data set). We used the Cohen κ statistic to evaluate the interrater agreement between the physicians for the questions (Table S3 in Multimedia Appendix 1). A schematic overview of the study protocol is presented in Figure 1 to provide a clearer understanding of our methodology.

Figure 1. Schematic of workflow for sourcing, encoding, and adjudicating results. The 165 questions were categorized into 3 types: CQ, CAQ, and MCQ, and each question was assessed for its score. The accuracy of the CQ and MCQ questions was evaluated, while the MCQ questions were also assessed for the accuracy, concordance, and frequency of insights. The adjudication process was carried out by 2 physicians, and in case of any discrepancies in the domains, a third physician was consulted for adjudication. Additionally, any inappropriate output was identified and required re-encoding. CAQ: case analysis question; CQ: common question; MCQ: multichoice question.



Results

ChatGPT Performs Poor Toward the Original Questions

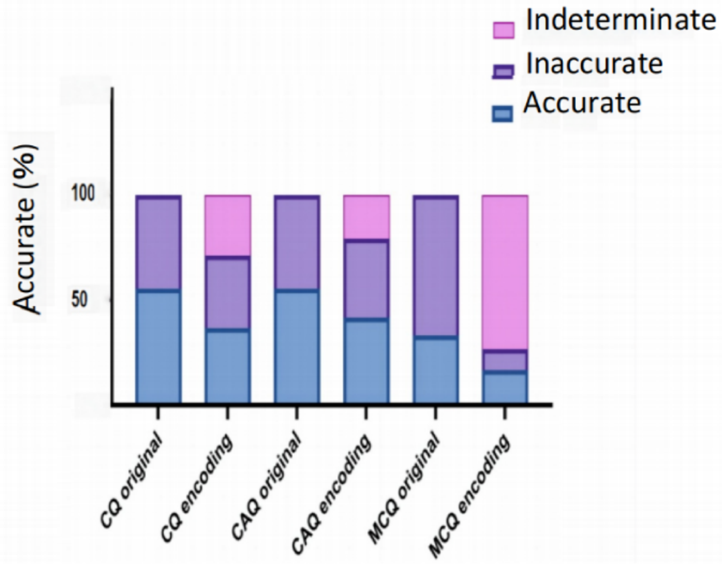
After inputting the original questions into ChatGPT and collecting their answers, ChatGPT received a score of 153.5 out of 300, which means that it only obtained 51.2% of the total points on the test. This score is much lower than expected but slightly higher than the passing threshold (129/300) defined by official agencies.

Among 3 subgroups of questions, the evaluation revealed that of a total of 90 CQs, ChatGPT only provided 50 (56%, 95% CI 45%-66%) correct answers. Similarly, of 45 CAQs, ChatGPT provided 25 (56%, 95% CI 41%-70%) correct answers. Furthermore, of 30 MCQs, ChatGPT provided 10 (33%, 95% CI 16%-50%) completely accurate answers (Figure 2). These results suggest that ChatGPT's ability to resolve medical problems in Chinese needs to be improved.

Additionally, we have noticed a Pearson correlation coefficient value of approximately 0.228. This finding suggests a relatively weak correlation between the different question types and the accuracy of the responses.

XSL•FO
RenderX

Figure 2. Accuracy of ChatGPT on Chinese Postgraduate Examination for Clinical Medicine before and after encoding. For the subgroups CQ, CAQ, and MCQ before encoding, AI output was compared with the standard answer key. For the subgroups CQ, CAQ, and MCQ after encoding, AI outputs were adjudicated to be accurate, inaccurate, or indeterminate based on the scoring system provided in Table S2 in Multimedia Appendix 1 data. It demonstrates the different accuracy distribution for inputs between the before and after encoding. AI: artificial intelligence; CAQ: case analysis question; CQ: common question; MCQ: multichoice question.



ChatGPT Performs Worse on Encoded Questions Compared to the Original Questions

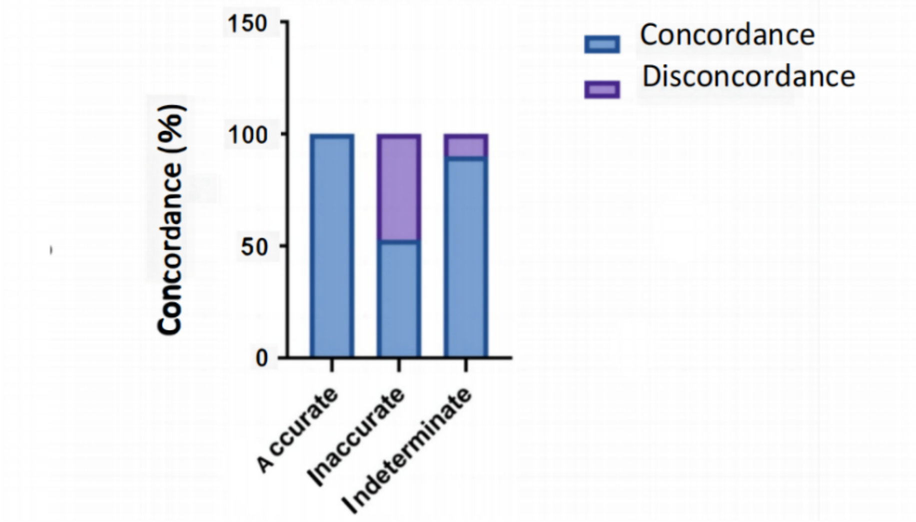
We encoded questions from the Chinese Postgraduate Examination for Clinical Medicine and inputted them into ChatGPT, which simulates scenarios where a student answers a common medical question without any choices or a doctor tries to diagnose a patient based on multimodal clinical data (ie, symptoms, history, physical examination, and laboratory values). ChatGPT's accuracy for all questions was 31.5%. Among the 3 subgroups, namely, CQs, MCQs, and CAQs, the accuracy was 42%, 37%, and 17%, respectively (Figure 2). Compared to the original questions, the accuracy of the encoding questions decreased by 19%, 17%, and 14% for CQs, MCQs, and CAQs, respectively, which demonstrates that the ability of ChatGPT to answer the open-ended questions in Chinese is a shortcoming. During the adjudication stage, there was substantial agreement among physicians on prompts in all 3 subgroups (κ ranged from 0.80 to 1.00).

ChatGPT Demonstrates High Internal Concordance

Concordance, which is a measure of the level of agreement or similarity between the option selected by AI and its subsequent explanation, was also taken into consideration. The results showed that ChatGPT achieved 90% concordance across all questions, and this high concordance was maintained across all 3 subgroups (Figure 3). Additionally, we analyzed the concordance difference between correct and incorrect answers and found that concordance among accurate responses was perfect and significantly greater than among inaccurate responses (n=52, 100% vs n=113, 50%; $P<.001$; Figure 3). These findings suggest that ChatGPT has a high level of answer-explanation concordance in Chinese, likely due to its strong internal consistency in its probabilistic language model.

XSL•FO
RenderX

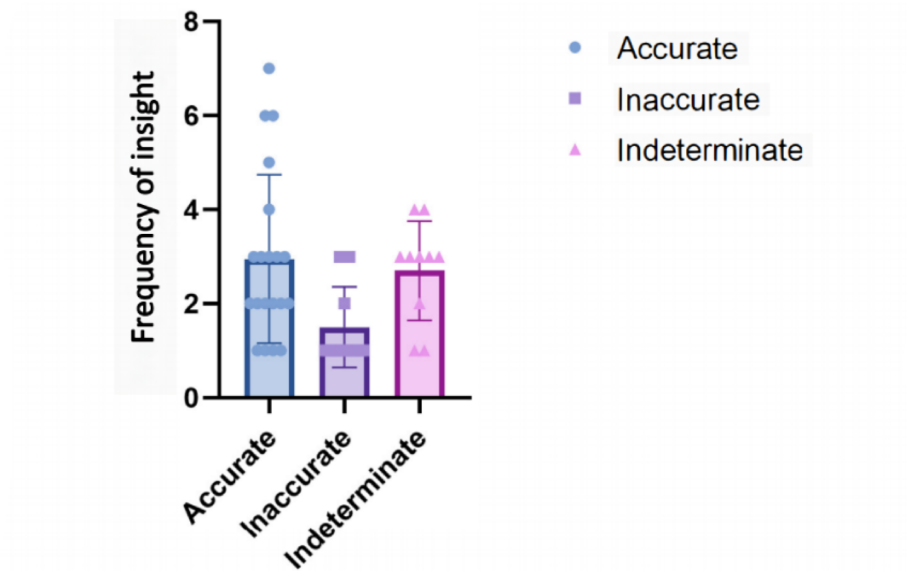
Figure 3. Concordance of ChatGPT on Chinese Postgraduate Examination for Clinical Medicine after encoding. For the subgroup “case analysis question,” artificial intelligence outputs were adjudicated to be concordant and discordant based on the scoring system provided in Table S2 in Multimedia Appendix 1 data. It demonstrates concordance rates stratified between accurate, inaccurate, and indeterminate outputs across all the case analysis questions.



ChatGPT Shows Multiple Insights Toward the Same Questions

Another evaluation index considered was the frequency of insights generated by the AI model that quantifies the quantity of insights produced. After evaluating the score, accuracy, and concordance of ChatGPT, its potential was investigated to enhance medical education by augmenting human learning. We analyzed the frequency of insights provided by ChatGPT. Remarkably, ChatGPT generated at least 1 significant insight in 80% (n=132) of all questions (Figure 4). Moreover, the analysis revealed that the accuracy response had the highest frequency of insights with an average of 2.95. The indeterminate response followed closely behind with an average of 2.7, while the inaccurate response had a lower frequency of insights with an average of 1.39 (Figure 4). The high frequency of insights in the accurate group suggests that it may be feasible for a target learner to acquire new or remedial knowledge from the ChatGPT AI output.

Figure 4. The frequency of insights of ChatGPT on Chinese Postgraduate Examination for Clinical Medicine after encoding. For the subgroup “case analysis question,” artificial intelligence outputs were adjudicated to count the frequency of insights it offered. It demonstrates the frequency of insights stratified between accurate, inaccurate, and indeterminate outputs, across all the case analysis questions.



Discussion

Major Findings

To evaluate ChatGPT’s problem-solving capabilities and assess its potential for Chinese medical education integration, its performance on the Chinese Postgraduate Examination for Clinical Medicine was tested. We had two major findings: (1) the score of ChatGPT needs to be improved when facing questions asked in the Chinese language and (2) there is still potential for this AI to generate novel performance that can assist humans due to the high concordance and the frequency of insights. This is the first study to assess the performance of ChatGPT in medical care and clinical decisions in Chinese.

XSL•FO
RenderX

ChatGPT's Performance Needs Improvement for Medical Questions in Chinese

A recent study showed that ChatGPT (version 3.5) performed with an accuracy rate of over 50% across all examinations and even exceeded 60% accuracy in some analyses when facing the USMLE [7]. Our results indicate that ChatGPT exhibited moderate accuracy in answering open-ended medical questions in Chinese, with an accuracy of 31.5%. Given the differences between English and Chinese inputs, it suggests that ChatGPT requires further improvement in answering medical questions in the Chinese language.

We sought to understand why there is a significant discrepancy between the performance of ChatGPT on Chinese and English language examinations. To investigate this, we asked the ChatGPT for the reasons, it explains that the training data used to train AI in different languages may be different, and the algorithms used to process and analyze text may vary from language to language (data not shown). Therefore, even for the same question, the output generated may vary slightly based on the language and the available language-based data.

Upon analyzing the results of this research, we found that the accuracy of ChatGPT was lowest for MCQs, followed by CQs and CAQs. The lower accuracy on MCQs may be due to the model being undertrained on the input as well as the MCQ samples being significantly less than those of single-choice questions. On the other hand, the CAQs may have extensive training compared to MCQs and are similar in type to the USMLE question.

Furthermore, we noticed that high accuracy outputs were associated with high concordance and a high frequency of insight, whereas poorer accuracy was linked to lower concordance and a lack of insight. Thus, it was hypothesized that inaccurate responses were primarily driven by missing information, which could result in reduced insight and indecision in the AI, rather than an overcommitment to an incorrect answer [7]. The results indicate that enhancing the database and providing additional training with Chinese questions could substantially improve the performance of the model.

Challenges of AI in Future Applications

Despite the promising potential of AI in medicine, it also poses some challenges. Standards for using AI in health care still need to be developed [8,9], including clinical care, quality, safety, malpractice, and communication guidelines. Furthermore, the implementation of AI in health care requires a shift in medical culture, which poses a challenge for both medical education and practice. Additionally, ethical considerations must be taken into account, such as data privacy, informed consent, and bias prevention, to ensure that AI is used ethically and for the benefit of patients. Surprisingly, a recently launched AI system for autonomous detection of diabetic retinopathy carries medical malpractice and liability insurance [10].

Prospective of AI

AI is a rapidly growing technology. At the time of writing, ChatGPT (version 4) has been released with significant improvements. Numerous practical and observational studies have demonstrated the versatile role of AI in almost all medical disciplines and specialties, particularly in improving risk assessment [11,12], data reduction, clinical decision support [13,14], operational efficiency, and patient communication [15,16]. We anticipate that advanced language models such as ChatGPT are reaching a level of maturity that will soon have a significant impact on clinical medicine, enhancing the delivery of personalized, compassionate, and scalable health care.

A comparison of ChatGPT's performance with other AI models, particularly in the context of Chinese language performance, could yield more comprehensive insights and underscore the unique challenges of using AI in diverse linguistic landscapes.

However, this was primarily due to the fact that AI models that focus on other aspects, while enhancing medical education and achieving promising results in medical question answering, are mostly developed and evaluated using English language data sets. This limitation restricts their applicability for performance comparisons in the context of the Chinese language.

Limitations

One limitation of this research is the small sample size. We only accessed 165 samples to qualify its accuracy and 30 CAQs to qualify its concordance and frequency of insight due to the limitations of the data, which focused solely on the diagnosis of the patient. Furthermore, the clinical situation is more complicated than the test, and larger and deeper analyses were needed. Finally, bias and error were inevitable in human adjudication, although there was a good interrater agreement between the physicians for the adjudication.

Moreover, comparing ChatGPT's performance with other AI models, especially in the context of Chinese language, can provide valuable insights and highlight the distinctive challenges associated with leveraging AI in diverse linguistic environments.

One notable factor contributing to this need for comparison is the prevalence of AI models such as Bidirectional Encoder Representations from Transformers, CLUE-Med, and MedQA that have made significant contributions to medical education and demonstrated promising outcomes in medical question answering. However, these models have predominantly been developed and assessed using English language data sets. This particular limitation hampers their suitability for conducting performance assessments within the Chinese language domain.

Conclusions

In conclusion, although the ChatGPTs got a score over the passing score in the Chinese Postgraduate Examination for Clinical Medicine, the performance was limited when presented with open-ended questions. On the other hand, ChatGPT demonstrated a high level of internal concordance, which suggests that the explanations provided by ChatGPT support and affirm the given answers. Moreover, ChatGPT generated multiple insights toward the same questions, demonstrating its potential for generating a variety of useful information. Further prospective studies are needed to explore whether there is a language-based difference in the performance of medical education settings and clinical decision-making, such as Chinese and minority languages.

XSL•FO
RenderX

Acknowledgments

The authors acknowledge ChatGPT for polishing their paper.

Data Availability

All data generated or analyzed during this study are included in this published paper (and Multimedia Appendix 1).

Authors' Contributions

Xiao Liu was responsible for the entire project and revised the draft. CF, AA, and YW performed the data extraction, statistical analysis, and interpretation of the data. WZ, Z Chen, YZ, and JW drafted the first version of the paper. All authors participated in the interpretation of the results and prepared the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Original questions, adjudication criteria for accuracy and concordance, and κ statistic for interrater agreement between adjudicating physicians.

[PDF File (Adobe PDF File), 256 KB - mededu_v10i1e48514_app1.pdf]

References

1. Haleem A, Javaid M, Khan IH. Current status and applications of Artificial Intelligence (AI) in medical field: an overview. Curr Med Res Pract 2019;9(6):231-237. [doi: 10.1016/j.cmrp.2019.11.005]
2. Haleem A, Vaishya R, Javaid M, Khan IH. Artificial Intelligence (AI) applications in orthopaedics: an innovative technology to embrace. J Clin Orthop Trauma 2020;11(Suppl 1):S80-S81 [FREE Full text] [doi: 10.1016/j.jcot.2019.06.012] [Medline: 31992923]
3. Jha S, Topol EJ. Information and artificial intelligence. J Am Coll Radiol 2018;15(3 Pt B):509-511. [doi: 10.1016/j.jacr.2017.12.025] [Medline: 29398501]
4. Lupton M. Some ethical and legal consequences of the application of artificial intelligence in the field of medicine. Trends Med 2018;18(4):1-7 [FREE Full text] [doi: 10.15761/tim.1000147]
5. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
6. Misawa M, Kudo SE, Mori Y, Cho T, Kataoka S, Yamauchi A, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 2018 Jun;154(8):2027-2029.e3 [FREE Full text] [doi: 10.1053/j.gastro.2018.04.003] [Medline: 29653147]
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
8. Considerations for the practical impact of AI in healthcare. U.S. Food & Drug Administration. URL: https://www.fda.gov/media/134071/download [accessed 2024-01-04]
9. Zweig M, Evans B. How should the FDA approach the regulation of AI and machine learning in healthcare? Rock Health. URL: https://rockhealth.com/how-should-the-fda-approach-the-regulation-of-ai-and-machine-learning-in-healthcare/ [accessed 2024-01-04]
10. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
11. Kan HJ, Kharrazi H, Chang HY, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One 2019;14(3):e0213258 [FREE Full text] [doi: 10.1371/journal.pone.0213258] [Medline: 30840682]
12. Delahanty RJ, Kaufman D, Jones SS. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. Crit Care Med 2018;46(6):e481-e488. [doi: 10.1097/CCM.0000000000003011] [Medline: 29419557]
13. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924-933 [FREE Full text] [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]

XSL•FO
RenderX

14. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. EBioMedicine 2019;46:27-29 [FREE Full text] [doi: 10.1016/j.ebiom.2019.07.019] [Medline: 31303500]
15. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020;4(6):e16670 [FREE Full text] [doi: 10.2196/16670] [Medline: 32442148]
16. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]

Abbreviations

AI: artificial intelligence
CAQ: case analysis question
CQ: common question
MCQ: multichoice question
USMLE: United States Medical Licensing Examination

Edited by K Venkatesh, MN Kamel Boulos; submitted 26.04.23; peer-reviewed by A Arbabisarjou, N Mungoli, T Hou, W Nelson; comments to author 14.06.23; revised version received 04.10.23; accepted 11.12.23; published 09.02.24.

Please cite as:
Yu P, Fang C, Liu X, Fu W, Ling J, Yan Z, Jiang Y, Cao Z, Wu M, Chen Z, Zhu W, Zhang Y, Abudukeremu A, Wang Y, Liu X, Wang J
Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study
JMIR Med Educ 2024;10:e48514
URL: https://mededu.jmir.org/2024/1/e48514
doi:10.2196/48514
PMID:38335017

©Peng Yu, Changchang Fang, Xiaolin Liu, Wanying Fu, Jitao Ling, Zhiwei Yan, Yuan Jiang, Zhengyu Cao, Maoxiong Wu, Zhiteng Chen, Wengen Zhu, Yuling Zhang, Ayiguli Abudukeremu, Yue Wang, Xiao Liu, Jingfeng Wang. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Cocreating an Automated mHealth Apps Systematic Review Process With Generative AI: Design Science Research Approach

Guido Giunti[1,2,3], MD, PhD; Colin P Doherty[1,3,4], MD

[1]Academic Unit of Neurology, School of Medicine, Trinity College Dublin, Dublin, Ireland

[2]Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland

[3]FutureNeuro SFI Research Centre, Royal College of Surgeons in Ireland, Dublin, Ireland

[4]Department of Neurology, St James Hospital, Dublin, Ireland

Corresponding Author:
Guido Giunti, MD, PhD
Academic Unit of Neurology
School of Medicine
Trinity College Dublin
College Green
Dublin, D02
Ireland
Phone: 353 1 896 1000
Email: drguidogiunti@gmail.com

Abstract

Background: The use of mobile devices for delivering health-related services (mobile health [mHealth]) has rapidly increased, leading to a demand for summarizing the state of the art and practice through systematic reviews. However, the systematic review process is a resource-intensive and time-consuming process. Generative artificial intelligence (AI) has emerged as a potential solution to automate tedious tasks.

Objective: This study aimed to explore the feasibility of using generative AI tools to automate time-consuming and resource-intensive tasks in a systematic review process and assess the scope and limitations of using such tools.

Methods: We used the design science research methodology. The solution proposed is to use cocreation with a generative AI, such as ChatGPT, to produce software code that automates the process of conducting systematic reviews.

Results: A triggering prompt was generated, and assistance from the generative AI was used to guide the steps toward developing, executing, and debugging a Python script. Errors in code were solved through conversational exchange with ChatGPT, and a tentative script was created. The code pulled the mHealth solutions from the Google Play Store and searched their descriptions for keywords that hinted toward evidence base. The results were exported to a CSV file, which was compared to the initial outputs of other similar systematic review processes.

Conclusions: This study demonstrates the potential of using generative AI to automate the time-consuming process of conducting systematic reviews of mHealth apps. This approach could be particularly useful for researchers with limited coding skills. However, the study has limitations related to the design science research methodology, subjectivity bias, and the quality of the search results used to train the language model.

(JMIR Med Educ 2024;10:e48949) doi:10.2196/48949

KEYWORDS

generative artificial intelligence; mHealth; ChatGPT; evidence-base; apps; qualitative study; design science research; eHealth; mobile device; AI; language model; mHealth intervention; generative AI; AI tool; software code; systematic review; language model

Introduction

The delivery of health-related services through the use of mobile devices (mHealth) [1] has been growing at a tremendous pace. A decade ago, in the first "era of mHealth," the literature surrounding mHealth called for the generation of evidence demonstrating the impact of mHealth solutions on health system processes and patient outcomes [2]. In 2013, Labrique et al [2]

XSL•FO
RenderX

conducted a preliminary search on the US federal clinical trials database (ClinicalTrials.gov) and had to combine the keywords “mHealth,” “mobile,” and “cell AND phone” to obtain 1678 studies and their results. Today, that same number can be obtained using “mHealth” alone as a keyword. As the need for mHealth evidence has grown, so too has the necessity for summarizing both the state of the art and the practice.

Systematic reviews seek to collect and combine relevant evidence within the specific scope of a research question while also striving to minimize bias [3,4]. In PubMed alone, the number of systematic reviews published on digital health–related topics has increased a hundredfold in the last 10 years. In fact, the pace at which the mHealth field is developing for certain conditions like breast cancer is such that systematic reviews can be found every 2 or 3 years [5-9]. The systematic review process, however, is a time- and resource-intensive process, reportedly requiring a median of 5 researchers and approximately 40 weeks of work to reach submission [10-12].

The emergence of generative AI has been seen as a breakthrough in the field of automation. With the ability to generate content such as text, images, and even music, AI has been reported as a potential solution to tedious time-consuming and labor-intensive tasks [13]. For instance, generative AI can be used to automatically generate product descriptions, news articles, or even code [14]. By eliminating the need for human intervention, generative AI can free up valuable time and resources for more complex tasks, thereby improving efficiency and accuracy. ChatGPT, a natural language processing model with a capacity of 175 billion parameters, has been trained on extensive amounts of data and is designed to produce human-like responses to user inputs. Since its release in November 2022, ChatGPT has received significant attention from media and academia alike, provoking ethical discussions on scientific authorship [15,16], attempting to pass medical license and specialist examinations [17-19], and even designing medical education curricula [20].

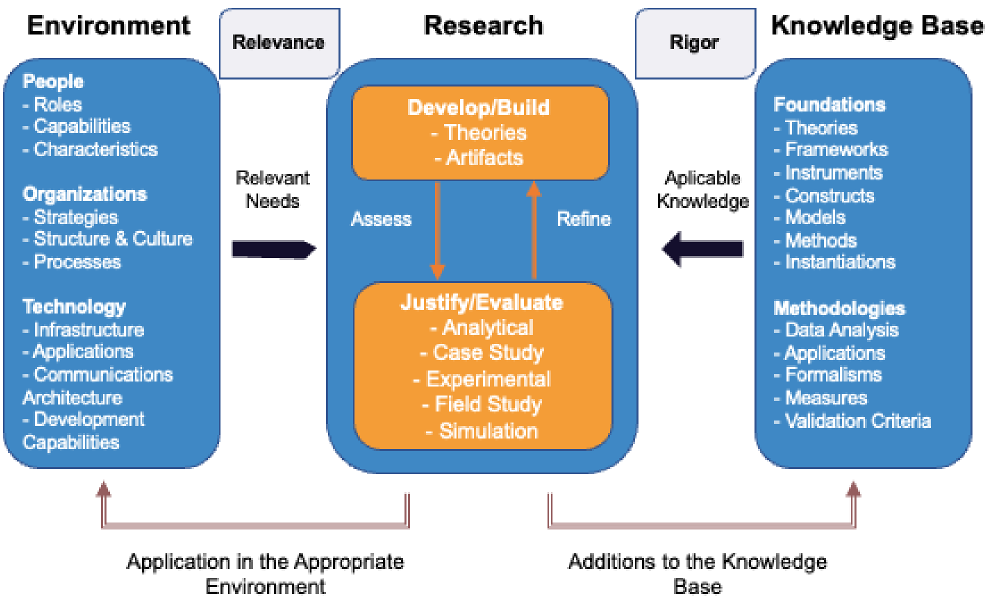
The objective of this study was to explore the feasibility of using generative AI tools to automate time-consuming and resource-intensive tasks in a systematic review process and assess the scope and limitations of using such tools.

Methods

Study Design

This study uses a design science research (DSR) methodology. DSR is a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts [21]. DSR commonly involves the identification of a problem or opportunity, followed by the development, implementation, and evaluation of a solution. In DSR, as well as in action research, the process happens within an organization that provides context and that would be changed as a result of the use of the artifact [21]. An overview of the process adapted from Hevner [22] can be seen in Figure 1.

Figure 1. Design science research overview, adapted from Hevner 2004 [22].



Problem Definition

The problem to which DSR was applied was the time-consuming and resource-intensive process of conducting systematic reviews of mHealth applications.

Organizational Context

The organizational context consisted of the More Stamina team of researchers, software developers, and health care professionals, working collaboratively within the host research institutions (ie, the University of Oulu and Trinity College Dublin).

The More Stamina project aims to create an evidence-driven gamified mHealth solution for people with multiple sclerosis (MS), where each step of the development follows a scientific process, as follows: MS needs as well as barriers and facilitators were explored through qualitative studies [23]; the state of the practice for MS apps was systematically reviewed [24,25]; user-centered design was used to create “MS personas” [23];

XSL•FO
RenderX

cocreation sessions took place to produce solution concepts [26]; the design, prototyping, and initial usability testing were described [27]; early health technology assessment was used to guide software development [28]; patient representatives were involved throughout the project [29]; and user testing and feasibility studies were ongoing in a multicenter study [30].

A script using the software application for audience targeting called 42matters [31] was used in the past to extract information from different app stores. The script is no longer functional, and person-hours from the software development team were not able to be dedicated to this task.

Background Studies

The research plans and outlines from previous studies, where systematic review methodologies were used to identify, select, collect, and analyze features and content of mHealth apps [6,24,25], served as models for our study. In those studies, a search strategy was defined, using relevant main keywords for each condition. App stores were searched, taking steps to ensure that no previous search history or cookies influenced results. Screening took place based on mHealth applications' titles, descriptions, and metadata.

Solution

The solution was to apply a cocreation process with a generative AI (ie, ChatGPT 3.5, as of June 2023) to produce software code that automated the process for conducting systematic reviews.

Cocreation Goal

The goal of the cocreation process was to use ChatGPT as a design and development partner for the automation process. The generative AI was to be interacted with as if it were a valid interloper who was more technologically skilled than the user and was guiding them through the process over text messages.

Development and Implementation

Development and implementation of the automated process happened through iterative and continuous conversations with the generative AI by one of the authors (GG). GG is a primary care physician with over a decade of experience leading digital health software design and development. Table 1 provides an overview of his digital skills background using the European Qualifications Framework and with a self-assessment score from 1 to 10 to describe his competency level. Regardless of the skill level, the development cycle was to be conducted as if no coding skill was present on the part of the user.

Table 1. Digital skills background.

Competency	Level	Experience	Self-assessment score (of 10)
Scrum master	Certified Scrum Master	Agile methodologies and team management	7
Product owner	Certified Scrum Product Owner	Product road mapping and stakeholder management	8
Game design	Intermediate	Game mechanics, storytelling, and level design	7
Web design	Advanced	User experience and user interface design and responsive design	8
JavaScript	Beginner	Front-end development	5
HTML5	Intermediate	Front-end development	7
PHP: Hypertext Preprocessor	Beginner	Front-end and back-end development	4

Evaluation

The results were evaluated for their effectiveness and efficiency in replicating the initial steps of the background studies. The measurements were considered with respect to the amount of time required to generate a spreadsheet containing the necessary information for human reviewers to start the systematic review process. The output was compared to the output generated by the previous script, which required further processing of the data.

Ethical Considerations

No ethics board review is needed as the work does not use patient data or involve human participants.

Results

The following section reports the overall design, development, and implementation process as well as the evaluation results and highlights of the experience. Prompts provided by GG and generative AI are tagged accordingly.

Cocreation Process

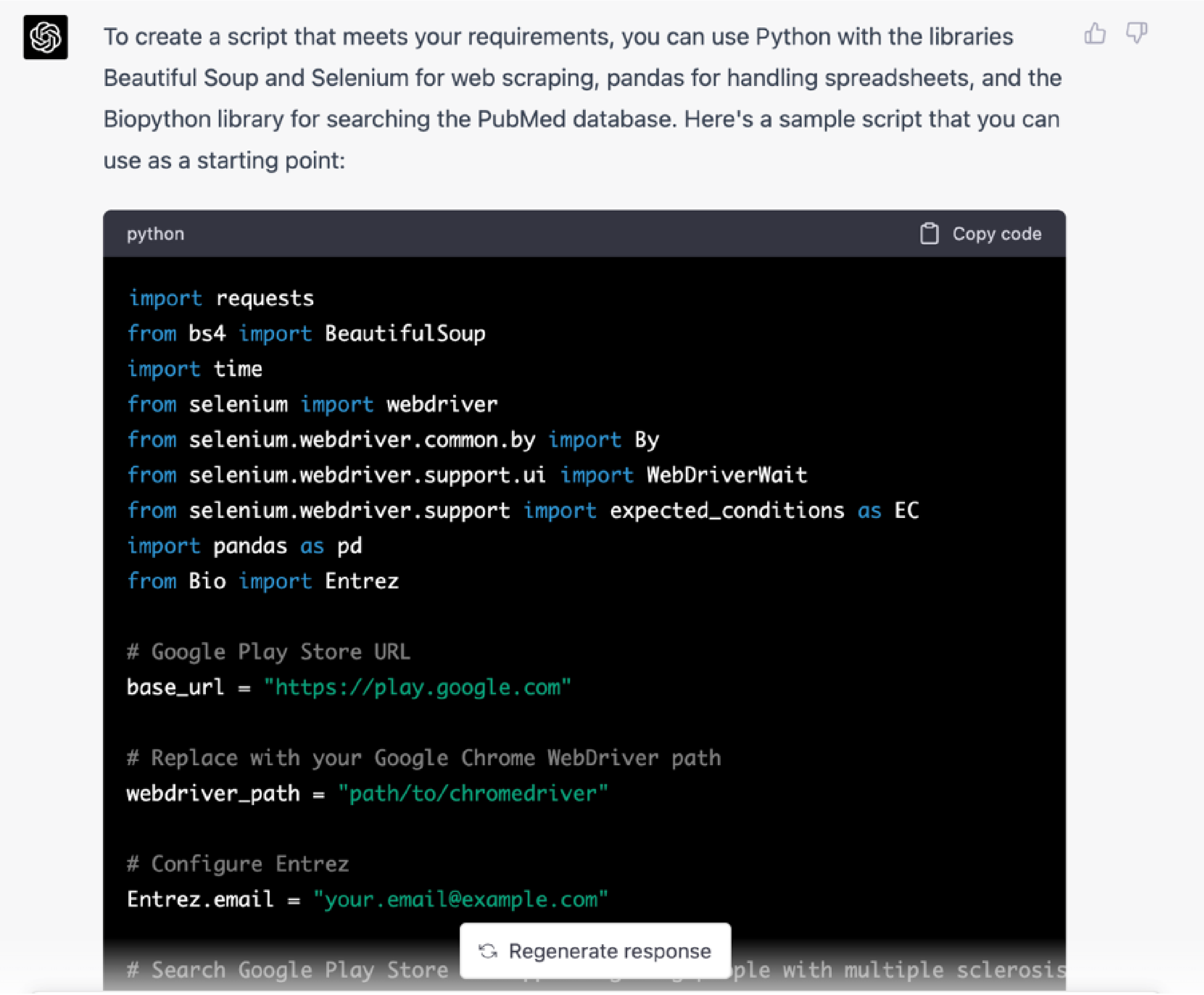
ChatGPT requires prompts to trigger interaction. For this purpose, a lay language description of the desired outcome was engineered by GG (triggering prompt), as follows:

> *I need a script that crawls the Google Play Store for apps that target people with multiple sclerosis. The script should generate a spreadsheet file that contains name, links, description and other metadata, as well as have a specific column to highlight if the description mentions any medical or clinical evidence. The script should then search Pubmed database looking for each app listed on the spreadsheet, and create a separate spreadsheet relating the apps with any scientific publication that mentions them in Pubmed.* [GG]

The generative AI's response was to provide a sample script in Python that was suggested as a starting point, as shown in Figure 2.

XSL•FO
RenderX

Figure 2. ChatGPT's first response to the initial prompt.



Following the premise of the cocreation goal, ChatGPT was asked (by GG) “I am new to Python, how do I run a Python script?” to which the response was the following:

> *To run a Python script, you need to follow these steps:*
>
> *1. Install Python on your computer, if it's not already installed. You can download Python from the official website: https://www.python.org/downloads/.*
>
> *2. Open a text editor (e.g. Notepad, Sublime Text, Atom) and create a new file.*
>
> *3. Write your Python code in the file.*
>
> *4. Save the file with a .py extension (e.g. “my_script.py”).*
>
> *5. Open a command prompt or terminal and navigate to the directory where your Python script is saved.*
>
> *6. Type “python” followed by the name of your script and press enter. For example, if your script is named “my_script.py”, you would type “python my_script.py” and press enter.*
>
> *7. Your Python script will now run, and you should see any output or results in the command prompt or terminal.*
>
> *Note: If you're using an Integrated Development Environment (IDE) like PyCharm or Visual Studio Code, you can run your script directly from the IDE by clicking on the “Run” button or pressing a keyboard shortcut.* [ChatGPT]

ChatGPT provided step-by-step guidance for the installation and setup of Visual Studio Code, along with the required libraries and dependencies. After the development environment was ready, the suggested Python script was copied and pasted to check its functionality. The code produced several errors, which were fed back to ChatGPT by GG, as follows:

> *I am getting “ModuleNotFoundError: No module named 'requests'”*
>
> *Now it says “ModuleNotFoundError: No module named 'bs4'”*

ChatGPT stated that the “requests” library was imported but not used, so it should be removed. Subsequent error messages were troubleshot with the AI. At times, the script would stop responding due to a time-out error or issues with the output. To make the debugging process easier to follow and understand, ChatGPT was prompted by GG to do the following:

> *add error messages that explain what went wrong.*

XSL•FO
RenderX

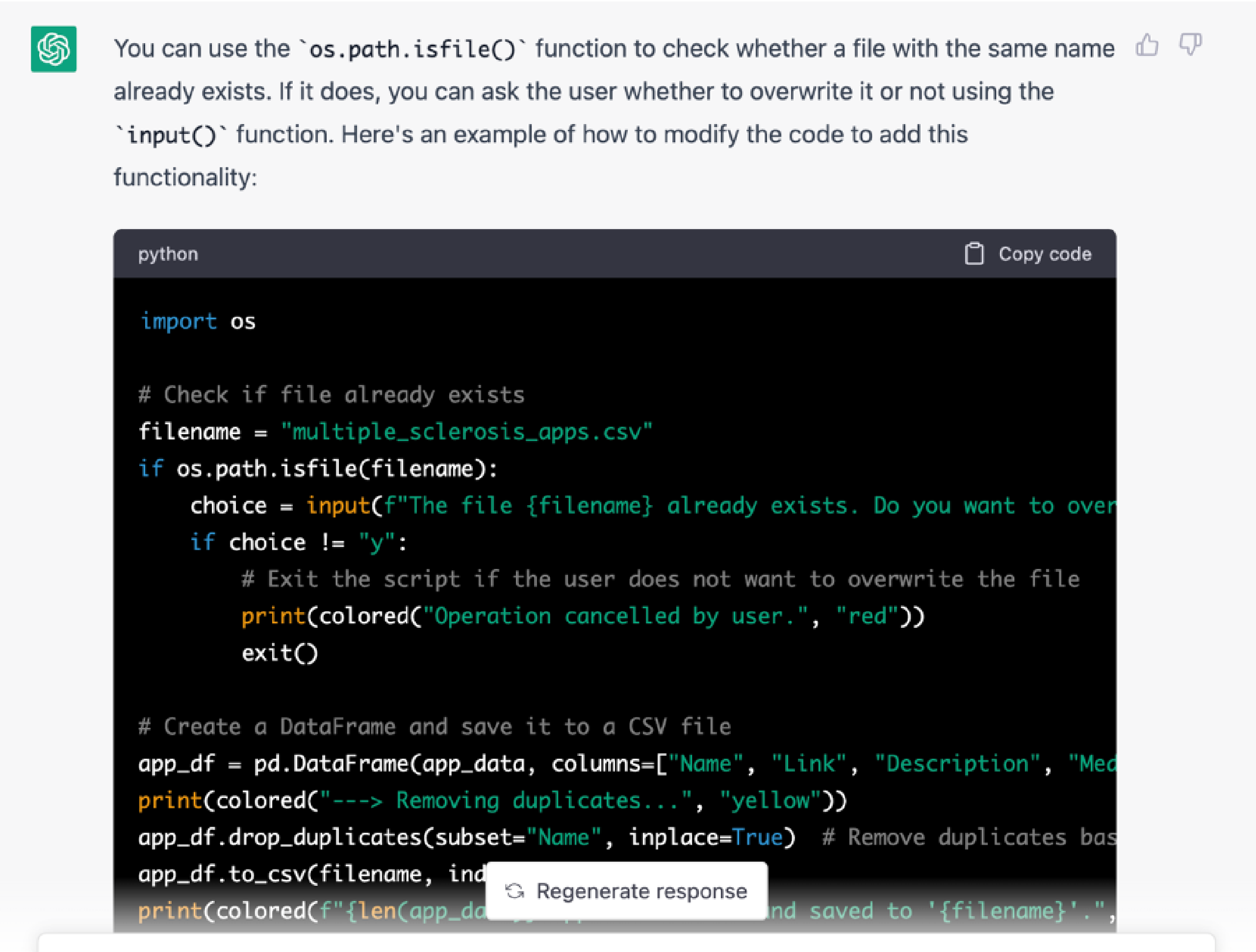
add indicators for each step that it is working.

could you add a progress bar indicator to each step?

I would like the responses to be colored.

Upon review, it was clear that some error caused the output files to be overwritten. The author requested the following: "how would you add a condition to check whether a file with the same name already exists and if it does ask whether to overwrite it?" The AI added the pertinent libraries to the script and produced an input prompt for the user (Figure 3).

Figure 3. Added libraries and code for checking existing files.



The triggering prompt expected the script to not only be searching app stores for matches with "multiple sclerosis" but also to cross-reference them with scientific publications found in PubMed. As described in the Background Studies section of this paper, this process was done by looking for keywords present in the app description. As it were, the script was designed to only search for "multiple sclerosis" and crashed when it reached the PubMed phase. ChatGPT diagnosed a time-out error due to the number of publications and provided a potential solution to improve the indicator bars:

> *The progress bar for PubMed search may not be moving because the total value for it is set to 3600, which is the number of seconds in an hour. This is the unit being used for the progress bar, not the total number of items to be searched. To fix this, you should update the total value for the progress bar to be the total number of apps being searched, which is the length of the app_df dataframe multiplied by the maximum number of PubMed IDs to be retrieved for each app (100 in this case). You can update the progress bar as follows…* [ChatGPT]

After these issues were sorted and the script could properly fetch PubMed results, more keywords were entered into the script by prompting ChatGPT with the following:

> *how would you make it so that the items in the medical_keywords list are taken from a csv file called "keywords.csv"?* [GG]

A screenshot of the final Python script running can be seen in Figure 4.

XSL•FO
RenderX

Figure 4. Screenshot of the final script running.

Evaluation

As explained in the Background Studies section of this paper, app data extraction from the Google Play Store resulted in a spreadsheet file that contained the mHealth app's name, store link, app description, developer's name, developers' URL, price, number of downloads, and app rating. During the screening phase of the studies, the research team read the apps' descriptions and flagged those that contained keywords or sentences suggestive of the evidence base for in-depth review.

The ChatGPT-generated code resulted in a CSV file that contained the app's name, store link, app description, and a column titled "Medical Evidence." There were no columns containing metadata, and the Medical Evidence column only contained "Yes" or "No," accordingly. Closer inspection revealed that the script was searching for a full match on the apps' titles in PubMed results. The resulting document was useful as an intermediate outcome but was deemed unsuitable as a final output. The overall cocreation process had a total duration of 4 hours and 39 minutes, providing a working script version available on GitHub [32].

Using the results from the ChatGPT-generated script to fully automate the process would likely require further work refining the script, either by using the steps of the background studies to base the script or by providing clearer starting prompts for the generative AI. However, leveraging this approach as a means to advance work when the software developing team was otherwise engaged was useful.

Highlights

Some highlights of this study are as follows:

- The overall cocreation process exercise had a total duration of 4 hours and 39 minutes.
- There were several misunderstandings during the interactions, not unlike the challenges one might encounter when messaging a more experienced coder.
- Structured thinking ahead of time reduced the number of misunderstandings.
- No knowledge of Python scripting was required by the author.
- The resulting output was useful to continue a systematic review but not sufficient to replace the final outputs.

Discussion

Principal Results

This study is the first to describe the cocreation process with a generative AI in developing an automated script for conducting a systematic review of mHealth apps. The study provides insights into the potential of using this kind of AI tools for researchers with little to no coding skills, and it identifies an innovative way of approaching a research problem and facilitating interdisciplinary collaborations. This study also makes a methodological contribution, expanding knowledge as it uses DSR, an approach that is not commonly used in health care and health informatics [33].

Comparison With Prior Work

The resource-intensive process and the burden that systematic reviews represent have been highlighted in the literature before. The use of multiple databases, such as MEDLINE, Embase, Cochrane Library, and Web of Science as well as clinical trial registries like ClinicalTrials.gov are common practices to increase results [34]. However, this tactic requires a lengthy deduplication process, involving long manual procedures, potentially introducing quality-affecting errors and biases [35-37]. In fact, automation attempts using AI models have been made in the past, with a focus on the deduplication problem, as seen in studies by Borissov et al [38] and Bramer et al [39].

Performing a systematic review is a common step in doctoral researchers' studies [40,41], as a means of introducing the candidate to the topic. The use of generative AI to cocreate scripts like the one presented in this study could help automate the time-consuming process, allowing researchers to focus on other aspects of the research process.

The ethical implications of using generative AI models, such as ChatGPT, to generate scientific authorship have sparked discussions [15,16]. AI's potential for assisting in academic research needs to be considered and weighed against the potential for its misuse. Although generative AI can assist in the development of a systematic review script, it is important to note that the final review still requires human oversight and input to not only assess the accuracy and relevance of the results but also ensure that the ethical principles have been followed.

XSL•FO
RenderX

Beyond research, there are wider implications for the use of generative AI in both medical education and the upskilling of the health care workforce. The need for more digital skills training for health care professionals is widely recognized [42], and other authors have further explored medical degree programs' curricula to examine how AI is included [43,44]. A recent publication explored the specific competencies needed for the effective and ethical use of AI in health care [45]. Understanding basic knowledge of AI and its applications as well as how to integrate AI into the general workflow of different tasks ranked among the top 6 key competency domains.

The role of generative AI in evolving health care education is pivotal, especially as universities adapt to its challenges. Generative AI has the potential to streamline processes like systematic reviews and clinical information retrieval, thereby allowing health care professionals to focus more on patient-centered, empathetic care and the co-design of effective treatment outcomes.

Limitations

The results of this study must be considered within its limitations. The DSR methodology was developed for this specific problem, which limits applicability in other contexts. In addition, subjectivity is a common bias present in DSR, which can make it difficult to establish the reliability and validity of the results. The main goal of DSR is to generate prescriptive knowledge, which provides guidelines on how to effectively design and implement solutions in the organizational context. However, as DSR focuses more on developing practical solutions rather than generating new theoretical insights, it was aligned with the goal of this study. DSR differs from traditional research paradigms by focusing more on creating and evaluating new solutions rather than on understanding existing phenomena. Further, while generative AI can assist in the development of a systematic review script, the result will be greatly affected by the training data used for the language model. Additionally, there may be limitations in the quality of the search results obtained from the previous studies, which only become apparent through automated processes.

Conclusions

This study outlined the cocreation process of an automated script for systematic reviews of mHealth apps, using generative AI. The study shed light on the potential of such AI tools for researchers with limited coding abilities and highlighted a novel approach for addressing research problems and promoting interdisciplinary collaborations.

Acknowledgments

GG would like to thank Prof Octavio Rivera-Romero, Dr Estefania Guisado-Fernandez, Dr Diego Giunta, Dr Analia Baum, and Prof Minna Isomursu for their collaboration and support.

This study has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement (101034252). The publication has also emanated from research supported (in part) by a research grant from Science Foundation Ireland (SFI) under grant number 16/RC/3948 and Business Finland's More Stamina Research to Business project.

The authors are grateful for ChatGPT, whose collaboration was essential for the completion and inception of this study.

Conflicts of Interest

None declared.

References

1. Ryu S. Book review: mHealth: new horizons for health through mobile technologies: based on the findings of the second global survey on eHealth (global observatory for eHealth series, volume 3). Healthc Inform Res 2012;18(3):231. [doi: 10.4258/hir.2012.18.3.231]
2. Labrique A, Vasudevan L, Chang LW, Mehl G. H_pe for mHealth: more "y" or "o" on the horizon? Int J Med Inform 2013 May;82(5):467-469 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.11.016] [Medline: 23279850]
3. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009 Jun;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
4. Nagendrababu V, Dilokthornsakul P, Jinatongthai P, Veettil SK, Pulikkotil SJ, Duncan HF, et al. Glossary for systematic reviews and meta-analyses. Int Endod J 2020 Mar 25;53(2):232-249. [doi: 10.1111/iej.13217] [Medline: 31520403]
5. Bender JL, Yue RYK, To MJ, Deacken L, Jadad AR. A lot of action, but not in the right direction: systematic review and content analysis of smartphone applications for the prevention, detection, and management of cancer. J Med Internet Res 2013;15(12):e287 [FREE Full text] [doi: 10.2196/jmir.2661] [Medline: 24366061]
6. Giunti G, Giunta DH, Guisado-Fernandez E, Bender JL, Fernandez-Luque L. A biopsy of Breast Cancer mobile applications: state of the practice review. Int J Med Inform 2018 Dec;110:1-9 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.10.022] [Medline: 29331247]
7. Rincon E, Monteiro-Guerra F, Rivera-Romero O, Dorronzoro-Zubiete E, Sanchez-Bocanegra CL, Gabarron E. Mobile phone apps for quality of life and well-being assessment in breast and prostate cancer patients: systematic review. JMIR Mhealth Uhealth 2017 Dec 04;5(12):e187 [FREE Full text] [doi: 10.2196/mhealth.8741] [Medline: 29203459]

XSL•FO
RenderX

8. Adam R, McMichael D, Powell D, Murchie P. Publicly available apps for cancer survivors: a scoping review. BMJ Open 2019 Oct 30;9(9):e032510 [FREE Full text] [doi: 10.1136/bmjopen-2019-032510] [Medline: 31575584]
9. Wanchai A, Anderson EA, Armer JM. A systematic review of m-health apps on managing side effects of breast cancer treatment. Support Care Cancer 2022 Dec 27;31(1):86. [doi: 10.1007/s00520-022-07464-x] [Medline: 36574048]
10. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
11. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017 Feb 27;7(2):e012545 [FREE Full text] [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
12. O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Wolfe MS. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev 2018 Jan 09;7(1):3 [FREE Full text] [doi: 10.1186/s13643-017-0667-4] [Medline: 29316980]
13. Benbya H, Davenport TH, Pachidi S. Artificial intelligence in organizations: current state and future opportunities. SSRN Journal 2020:1-15. [doi: 10.2139/ssrn.3741983]
14. Weisz J, Muller M, Ross S, Martinez F, Houde S, Agarwal M, et al. Better together? An evaluation of AI-supported code translation. 2022 Presented at: 27th International Conference on Intelligent User Interfaces; 22 - 25 March; New York, NY p. 369-391. [doi: 10.1145/3490099.3511157]
15. Thorp HH. ChatGPT is fun, but not an author. Science 2023 Jan 27;379(6630):313-313. [doi: 10.1126/science.adg7879] [Medline: 36701446]
16. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023 Jan 24;613(7945):612-612. [doi: 10.1038/d41586-023-00191-1] [Medline: 36694020]
17. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Mar 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
18. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Mar 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
19. Fuentes-Martín Á, Cilleruelo-Ramos ?, Segura-Méndez B, Mayol J. Can an artificial intelligence model pass an examination for medical specialists? Arch Bronconeumol 2023 Aug;59(8):534-536. [doi: 10.1016/j.arbres.2023.03.017] [Medline: 37055267]
20. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
21. vom BJ, Hevner A, Maedche A. Introduction to design science research. In: Design Science Research. Cases. Switzerland: Springer Cham; Sep 24, 2020:13.
22. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Quarterly 2004;28(1):75. [doi: 10.2307/25148625]
23. Giunti G, Kool J, Rivera Romero O, Dorronzoro Zubiete E. Exploring the specific needs of persons with multiple sclerosis for mhealth solutions for physical activity: mixed-methods study. JMIR Mhealth Uhealth 2018 Feb 09;6(2):e37 [FREE Full text] [doi: 10.2196/mhealth.8996] [Medline: 29426814]
24. Giunti G, Guisado-Fernandez E, Caulfield B. Connected health in multiple sclerosis: a mobile applications review. 2017 Presented at: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS); 22 - 24 June; Thessaloniki, Greece p. 660-665. [doi: 10.1109/cbms.2017.27]
25. Giunti G, Guisado FE, Dorronzoro ZE, Rivera RO. Supply and demand in mHealth apps for persons with multiple sclerosis: systematic search in app stores and scoping literature review. JMIR Mhealth Uhealth 2018 May 23;6(5):e10512 [FREE Full text] [doi: 10.2196/10512] [Medline: 29792295]
26. Giunti G. Gamified dDesign for health workshop. Stud Health Technol Inform 2016;225:605-606. [Medline: 27332273]
27. Giunti G, Mylonopoulou V, Rivera Romero O. More stamina, a gamified mHealth solution for persons with multiple sclerosis: research through design. JMIR Mhealth Uhealth 2018 Mar 02;6(3):e51 [FREE Full text] [doi: 10.2196/mhealth.9437] [Medline: 29500159]
28. Giunti G, Haverinen J, Reponen J. Informing the product development of an mHealth solution for people with multiple sclerosis through early health technology assessment. Stud Health Technol Inform 2022 Jul 06;290:1042-1043. [doi: 10.3233/SHTI220258] [Medline: 35673196]
29. Yrttiaho T, Isomursu M, Giunti G. Experiences using patient and public involvement in digital health research for multiple sclerosis. Stud Health Technol Inform 2022 May 25;294:735-739. [doi: 10.3233/SHTI220574] [Medline: 35612194]
30. Giunti G, Rivera-Romero O, Kool J, Bansi J, Sevillano JL, Granja-Dominguez A, et al. Evaluation of more stamina, a mobile app for fatigue management in persons with multiple sclerosis: protocol for a feasibility, acceptability, and usability study. JMIR Res Protoc 2020 Aug 04;9(8):e18196 [FREE Full text] [doi: 10.2196/18196] [Medline: 32749995]

XSL•FO
RenderX

31. Girardello A, Budde A, Wang B, Delchev I. 42matters. 2016. URL: https://www.42matters.com [accessed 2016-02-26]
32. GitHub. URL: https://github.com/guidogiunti/ChatGPT-SR-script [accessed 2024-02-07]
33. Hevner A, Wickramasinghe N. Design science research opportunities in health care. In: Theories to Inform Superior Health Informatics Research and Practice. Cham, Switzerland: Springer; 2018:3-18.
34. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017 Dec 06;6(1):245 [FREE Full text] [doi: 10.1186/s13643-017-0644-y] [Medline: 29208034]
35. Qi X. Duplicates in systematic reviews: a critical, but often neglected issue. WJMA 2013;1(3):97. [doi: 10.13105/wjma.v1.i3.97]
36. McKeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev 2021 Jan 23;10(1):38 [FREE Full text] [doi: 10.1186/s13643-021-01583-y] [Medline: 33485394]
37. Qi X, Yang M, Ren W, Jia J, Wang J, Han G, et al. Find duplicates among the PubMed, EMBASE, and Cochrane Library Databases in systematic review. PLoS One 2013 Aug 20;8(8):e71838 [FREE Full text] [doi: 10.1371/journal.pone.0071838] [Medline: 23977157]
38. Borissov N, Haas Q, Minder B, Kopp-Heim D, von Gernler M, Janka H, et al. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. Syst Rev 2022 Aug 17;11(1):172 [FREE Full text] [doi: 10.1186/s13643-022-02045-9] [Medline: 35978441]
39. Bramer WM, Giustini D, De Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc 2016 Sep 12;104(3):240-243. [doi: 10.5195/jmla.2016.24]
40. Riaz M, Sulayman M, Salleh N, Mendes E. Experiences conducting systematic reviews from novices' perspective. 2010 Presented at: 14th International Conference on Evaluation and Assessment in Software Engineering; 12 - 13 April; Keele, UK. [doi: 10.14236/ewic/ease2010.6]
41. Pickering C, Byrne J. The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. High Educ Res Dev 2013 Nov 11;33(3):534-548. [doi: 10.1080/07294360.2013.841651]
42. Giunti G, Guisado-Fernandez E, Belani H, Lacalle-Remigio JR. Mapping the access of future doctors to health information technologies training in the European union: cross-sectional descriptive study. J Med Internet Res 2019 Aug 12;21(8):e14086 [FREE Full text] [doi: 10.2196/14086] [Medline: 31407668]
43. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. Acad Med 2021 Nov 01;96(11S):S62-S70. [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
44. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
45. Russell R, Lovett Novak L, Patel M, Garvey KV, Craig KJT, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 01;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]

Abbreviations

AI: artificial intelligence
DSR: design science research
mHealth: mobile health
MS: multiple sclerosis

Edited by G Eysenbach, T de Azevedo Cardoso, K Venkatesh; submitted 12.05.23; peer-reviewed by D Carvalho, X Zhao; comments to author 14.06.23; revised version received 28.11.23; accepted 28.01.24; published 12.02.24.

Please cite as:
Giunti G, Doherty CP
Cocreating an Automated mHealth Apps Systematic Review Process With Generative AI: Design Science Research Approach
JMIR Med Educ 2024;10:e48949
URL: https://mededu.jmir.org/2024/1/e48949
doi:10.2196/48949
PMID:38345839

©Guido Giunti, Colin P Doherty. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,

XSL•FO
RenderX

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.

XSL•FO
RenderX

Original Paper

Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models

Tassallah Abdullahi[1], MSc; Ritambhara Singh[1,2], PhD; Carsten Eickhoff[3], PhD

[1]Department of Computer Science, Brown University, Providence, RI, United States
[2]Center for Computational Molecular Biology, Brown University, Providence, RI, United States
[3]School of Medicine, University of Tübingen, Tübingen, Germany

Corresponding Author:
Carsten Eickhoff, PhD
School of Medicine
University of Tübingen
Schaffhausenstr, 77
Tübingen, 72072
Germany
Phone: 49 7071 29 843
Email: carsten.eickhoff@uni-tuebingen.de

Abstract

Background: Patients with rare and complex diseases often experience delayed diagnoses and misdiagnoses because comprehensive knowledge about these diseases is limited to only a few medical experts. In this context, large language models (LLMs) have emerged as powerful knowledge aggregation tools with applications in clinical decision support and education domains.

Objective: This study aims to explore the potential of 3 popular LLMs, namely Bard (Google LLC), ChatGPT-3.5 (OpenAI), and GPT-4 (OpenAI), in medical education to enhance the diagnosis of rare and complex diseases while investigating the impact of prompt engineering on their performance.

Methods: We conducted experiments on publicly available complex and rare cases to achieve these objectives. We implemented various prompt strategies to evaluate the performance of these models using both open-ended and multiple-choice prompts. In addition, we used a majority voting strategy to leverage diverse reasoning paths within language models, aiming to enhance their reliability. Furthermore, we compared their performance with the performance of human respondents and MedAlpaca, a generative LLM specifically designed for medical tasks.

Results: Notably, all LLMs outperformed the average human consensus and MedAlpaca, with a minimum margin of 5% and 13%, respectively, across all 30 cases from the diagnostic case challenge collection. On the frequently misdiagnosed cases category, Bard tied with MedAlpaca but surpassed the human average consensus by 14%, whereas GPT-4 and ChatGPT-3.5 outperformed MedAlpaca and the human respondents on the moderately often misdiagnosed cases category with minimum accuracy scores of 28% and 11%, respectively. The majority voting strategy, particularly with GPT-4, demonstrated the highest overall score across all cases from the diagnostic complex case collection, surpassing that of other LLMs. On the Medical Information Mart for Intensive Care-III data sets, Bard and GPT-4 achieved the highest diagnostic accuracy scores, with multiple-choice prompts scoring 93%, whereas ChatGPT-3.5 and MedAlpaca scored 73% and 47%, respectively. Furthermore, our results demonstrate that there is no one-size-fits-all prompting approach for improving the performance of LLMs and that a single strategy does not universally apply to all LLMs.

Conclusions: Our findings shed light on the diagnostic capabilities of LLMs and the challenges associated with identifying an optimal prompting strategy that aligns with each language model's characteristics and specific task requirements. The significance of prompt engineering is highlighted, providing valuable insights for researchers and practitioners who use these language models for medical training. Furthermore, this study represents a crucial step toward understanding how LLMs can enhance diagnostic reasoning in rare and complex medical cases, paving the way for developing effective educational tools and accurate diagnostic aids to improve patient care and outcomes.

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e51391) doi:10.2196/51391

KEYWORDS

clinical decision support; rare diseases; complex diseases; prompt engineering; reliability; consistency; natural language processing; language model; Bard; ChatGPT 3.5; GPT-4; MedAlpaca; medical education; complex diagnosis; artificial intelligence; AI assistance; medical training; prediction model

Introduction

Background

Natural language processing has witnessed remarkable advances with the introduction of generative large language models (LLMs). In November 2022, OpenAI released ChatGPT-3.5 (OpenAI), a large natural language processing chatbot trained on a large corpus collected from the internet to generate humanlike text in response to user queries. ChatGPT-3.5 has seen massive popularity, and users have praised its creativity and language comprehension for several tasks, such as text summarization and writing computer programs [1]. In March 2023, OpenAI responded to the success of ChatGPT-3.5 by introducing an enhanced iteration called GPT-4, specifically designed to address intricate queries and nuanced directives more effectively. Shortly thereafter, Google released their comparable model, Bard (Google LLC), which joined the league of impressive LLMs. What sets Bard apart is its real-time access to and use of internet information, enriching its response generation with up-to-date information [2]. In contrast, GPT-4 possesses multimodal capabilities, including image inputs, albeit not publicly available during the study [3].

These LLMs were not originally designed for medical applications. However, several studies [4,5] have shown their extraordinary capabilities in excelling in various medical examinations, such as the Self-Assessment in Neurological Surgery examination and the USMLE (United States Medical Licensing Examination). Their results demonstrated the ability of these models to handle clinical information and complex counterfactuals. Furthermore, numerous investigations [6-8] have revealed the remarkable advantages of harnessing the power of LLMs in diverse medical scenarios. Notably, Lee et al [8] demonstrated using LLMs as a reliable conversational agent to collect patient information to assist in medical notetaking, whereas Patel and Lam [9] delved into using LLMs as a valuable tool for generating comprehensive patient discharge summaries. The ability of LLMs to process and generate medical text has unlocked new opportunities to enhance diagnostic reasoning, particularly in tackling rare and complex medical cases.

Rare diseases are characterized by their low prevalence in the general population, whereas complex diseases are conditions with overlapping factors and multiple comorbidities that are often difficult to diagnose [10,11]. Sometimes, a condition can be rare and complex if it is infrequent and challenging to diagnose accurately [11]. Rare and complex diagnoses present significant challenges across various medical levels and often require extensive medical knowledge or expertise for accurate diagnosis and management [10,11]. This may be because, during their education, physicians are trained to prioritize ruling out common diagnoses before considering rare ones during patient evaluation [12]. In addition, most medical education programs rarely cover some complex conditions, and guidance for practicing clinicians is often outdated and inappropriate [13,14]. As a result, most physicians perceive their knowledge of rare diseases as insufficient or very poor, and only a few feel adequately prepared to care for patients with these conditions [12,15]. This knowledge gap increases the risk of misdiagnosis among individuals with rare and complex conditions. Furthermore, the scarcity of available data and the relatively small number of affected individuals create a complicated diagnostic landscape, even for experienced and specialized clinicians [10]. Consequently, patients often endure a prolonged and arduous diagnostic process. Therefore, there is a pressing need for comprehensive educational tools and accurate diagnostic aids to fill the knowledge gap and address these challenges effectively.

This study aims to explore the potential of 3 LLMs, namely Bard, GPT-4, and ChatGPT-3.5, as continuing medical education (CME) systems to enhance the diagnoses of rare and complex conditions. Although these models have demonstrated impressive success in standardized medical examinations [4,5], it is important to acknowledge that most examinations reflect general clinical situations, which may not fully capture the intricacies encountered in real-world diagnostic scenarios. Furthermore, these standardized tests often feature questions that can be answered through memorization [16]. In contrast, real-world complex diagnostic scenarios that physicians face involve dynamic, multifaceted patient cases with numerous variables and uncertainties. Although previous studies by Liu et al [17] and Cascella et al [18] have highlighted the ability of LLMs to support health care professionals in real-world scenarios, their effectiveness in diagnosing rare and complex conditions remains an area of exploration. Despite the promising use of LLMs in medical applications, studies have reported that their responses to user queries are often nondeterministic (ie, depending on the query format) and exhibit significant variance [17,19]. This attribute may pose challenges in clinical decision support scenarios because the dependability of a system is uncertain when its behavior cannot be accurately predicted. However, no investigation has been conducted to show how different input formats (prompts) affect LLM responses in the medical context.

Prompt engineering is a technique for carefully designing queries (inputs) to improve the performance of generative language models [20,21]. We can guide LLMs to generate more accurate and reliable responses by carefully crafting effective prompts. Our study investigated effective prompting strategies to improve the accuracy and reliability of LLMs in diagnosing rare and complex conditions within an educational context. We evaluated the performance of LLMs by comparing their responses to those of human respondents and the responses of

XSL•FO
RenderX

MedAlpaca [22], an open-source generative LLM designed for medical tasks. Given the documented advantages of using LLMs as a complementary tool rather than a substitute for clinicians [17,18], our study incorporated LLMs with the understanding that clinicians may use them beyond real-time diagnostic scenarios. Although our premise is based on a clinician having established an initial diagnostic hypothesis and seeking further assistance to refine the precise diagnosis, we acknowledge the broader utility of LLMs. They can be valuable in real-time decision support and retrospective use during leisure or documentation, allowing physicians to experiment with and enhance their understanding of rare and complex diseases. This approach recognizes the inherent uncertainty in diagnosis and harnesses the capabilities of LLMs to assist clinicians in various aspects of their diagnostic processes. In the context of CME, our study highlights the possibility of integrating LLMs as a valuable addition. By providing further assistance in refining complex and rare diagnoses, these LLMs could support evidence-based decision-making among health care professionals for improved patient outcomes.

Objectives

Our study has 2 main objectives: first, to examine the potential of LLMs as a CME tool for diagnosing rare and complex conditions, and second, to highlight the impact of prompt formatting on the performance of LLMs. Understanding these aspects could significantly contribute to advancing diagnostic practices and effectively using LLMs to improve patient care.

Methods

Data Sets

We used 2 data sets to examine the capacity of LLMs to diagnose rare and complex conditions as follows:

1. Diagnostic case challenge collection (DC3) [11] comprises 30 complex diagnostic cases curated by medical experts in the *New England Journal of Medicine* web-based case challenges. The original cases contained text and image descriptions of patients' medical history, diagnostic imaging, and laboratory results; however, we used only textual information to form prompts (queries). The web-based polls recorded an average of 5850 (SD 2522.84) respondents per case, many of whom were health care professionals. The participants were required to identify the correct diagnosis from a list of differential diagnoses. Case difficulty was categorized based on the percentage of correct responses received from the respondents on the web-based survey. The case categories were: "rarely misdiagnosed cases" (with ≥21/30, 70% correct responses), "moderately misdiagnosed cases" (with >9/30, 30% and <21/30, 70% correct responses), and "frequently misdiagnosed cases" (with ≤9/30, 30% correct responses). Furthermore, the final diagnoses determined by the treating physicians of the cases were provided alongside the poll results, enabling the comparison of the performance of human respondents with that of the targeted LLMs.
2. Medical Information Mart for Intensive Care-III (MIMIC-III) [23] comprises deidentified electronic health record data from approximately 50,000 Boston Beth Israel Deaconess Medical Center intensive care unit patients. We focused on discharge summaries containing the accumulated patient information from admission to discharge. Similar to previous work on clinical outcome prediction by van Aken et al [24] and Abdullahi et al [25], we filtered document sections unrelated to admissions, such as discharge information or hospital course and retained sections related to admissions, such as chief complaint, history of illness or present illness, medical history, admission medications, allergies, physical examination, family history, and social history. Each discharge summary had a discharge diagnosis section that indicated the patient's final diagnosis for that admission. We reviewed the discharge summaries to identify rare diseases and referred to the Orphanet website [26]. In this study, we randomly selected 15 unique, rare conditions as our target. These cases were selected as pilot studies for a focused and in-depth analysis.

Models

In this study, we conducted experiments using LLMs designed for conversational context. Specifically, we used the July 6, 2023, version of Bard; the July 4, 2023, versions of GPT-4 and ChatGPT-3.5; and the publicly available version of MedAlpaca 7b [22]. We entered prompts individually through the chat interface to evaluate Bard, GPT-4, and ChatGPT-3.5, treating each prompt as a distinct conversation. MedAlpaca differs from Bard, ChatGPT-3.5, and GPT-4 in that it requires users to submit queries or prompts through a Python (Python Software Foundation) script. Consequently, we used a single Python script for each prompt strategy to submit queries for each data set. It is worth noting that Bard has certain limitations compared with ChatGPT-3.5 and GPT-4. Bard has a restricted capacity to handle lengthy queries. Moreover, Bard is more sensitive to noisy input and specific characters. For example, the MIMIC-III data set contained deidentified patients' notes filled with special characters such as "[**Hospital 18654**]" and laboratory results written in shorthand, for example, ** Hgb-9.6* Hct-29.7* MCV-77* MCH-24.9**. Consequently, to work effectively with Bard, we preprocessed the text by removing special characters and retaining only alphanumeric characters.

Prompting Strategies

Direct (standard prompting) and iterative prompting (chain of thought prompting) [27] are the 2 major prompting methods. Iterative prompting is a promising method for improving LLM performance on specialized tasks; however, it requires a predefined set of manually annotated reasoning steps, which can be time consuming and difficult to create, especially for specialized domains. Most users opt for a direct prompt method to save time and obtain an immediate response. Therefore, to analyze the effect of prompt formats on LLM performance, we assessed each model's performance for every case using the 3 distinct direct prompt strategies outlined in Table 1. These strategies varied from open-ended to multiple-choice formats.

XSL•FO
RenderX

Table 1. Prompt strategies.

Approach	Prompt strategy description	Prompt sample
Approach 1 (open-ended prompt)	In this approach, prompts were formatted in an open-ended fashion. Formatting a prompt using this method allows the model to formulate a hypothesis for the case and explain why and what it thinks is the diagnosis. Here, we scored a model based on its ability to provide the correct diagnosis without additional assistance.	"What is the diagnosis? The case is: A 32-year-old man was evaluated in the emergency department of this hospital for the abrupt onset of postprandial chest pain..."
Approach 2 (multiple-choice prompt)	We formatted prompts as multiple-choice questions, and the LLMs[a] were expected to select a single diagnosis from a list of options. The models were assigned a positive score in this task if they selected the correct diagnosis from the options.	"Choose the most likely diagnosis from the following: Option I: Cholecystitis, Option II: Acute coronary syndrome, Option III: Pericarditis, Option IV: Budd-Chiari syndrome. The case is: A 32-year-old man was evaluated in the emergency department of this hospital for the abrupt onset of postprandial chest pain..."
Approach 3 (ranking prompt)	The prompts were presented as a case and a list of diagnoses to be ranked by the LLMs. Models were assigned a positive score if the correct diagnosis was ranked first in this format.	"Rank the following diagnoses according to the most likely. Option I: Cholecystitis, Option II: Acute coronary syndrome, Option III: Pericarditis, Option IV: Budd-Chiari syndrome. The case is: A 32-year-old man was evaluated in the emergency department of this hospital for the abrupt onset of postprandial chest pain..."

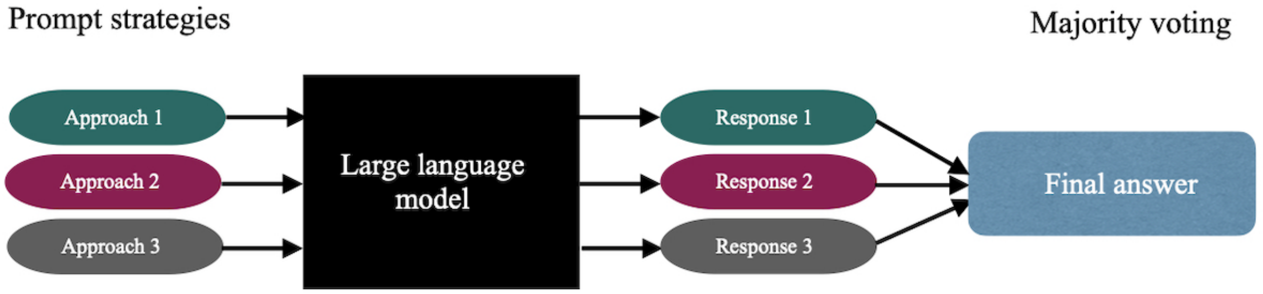
[a]LLM: large language model.

Building upon prior research by Wang et al [28] and Li et al [29], we hypothesized that using a diverse range of prompts can reveal distinct reasoning paths while maintaining consistency in the correct responses regardless of the variations. When using multiple-choice prompts for the DC3 cases, we presented the same options available in the original web-based polls to the models, but on the MIMIC-III data set, we generated random wrong answers that were closely related to the correct diagnosis. We evaluated each LLM by assigning a positive or negative score (binary score) based on their responses. A positive score was assigned only if the models correctly selected the diagnosis for either data set. Conversely, we omitted the options for open-ended prompts, expecting the models to generate the correct diagnosis independently. Positive scores were awarded only if the models accurately provided the correct diagnosis.

Prompt Ensemble: Majority Voting

To safely use imperfect language models, users must determine when to trust their predictions, particularly in critical situations, such as clinical decision support. Therefore, we used a majority voting (prompt ensembling) strategy to enhance the reliability of LLMs' responses. The majority voting approach involves aggregating multiple responses and selecting the most common answer. By applying this approach to responses generated by different LLMs, we can observe the level of agreement and infer the consistency in their outputs for a given prompt. Specifically, we hypothesized that using a majority voting approach from the ensemble of prompt responses would boost the reliability of language models, minimizing potential errors, variations, and biases associated with individual prompting approaches. To achieve this, in independent chats, we prompted the LLM with 3 distinct prompt formats per case, as presented in Table 1. Subsequently, we collected the responses of each model and applied majority voting to aggregate its predictions, as presented in Figure 1. In majority voting, each prompt produced a response from the language model, and the majority response was chosen as the final response. In a scenario where all prompt strategies resulted in different responses, we assumed that the model was unsure of that question and scored the final response as a failure case. We limited the number of prompts in the ensemble to 3 because studies by Wang et al [28] and Li et al [29] have shown that we obtain diminishing returns as we increase the overall number of prompts in an ensemble.

Figure 1. Our proposed method contains the following steps: (1) prompt a language model using a distinct set of prompts, (2) obtain diverse responses, and (3) choose the most consistent response as the final answer (majority voting).



Ethical Considerations

No ethics approval was pursued for this research, given that the data was publicly accessible and deidentified. This aligns with the guidelines outlined in the National Institutes of Health investigator manual for human subjects research [30].

XSL•FO
RenderX

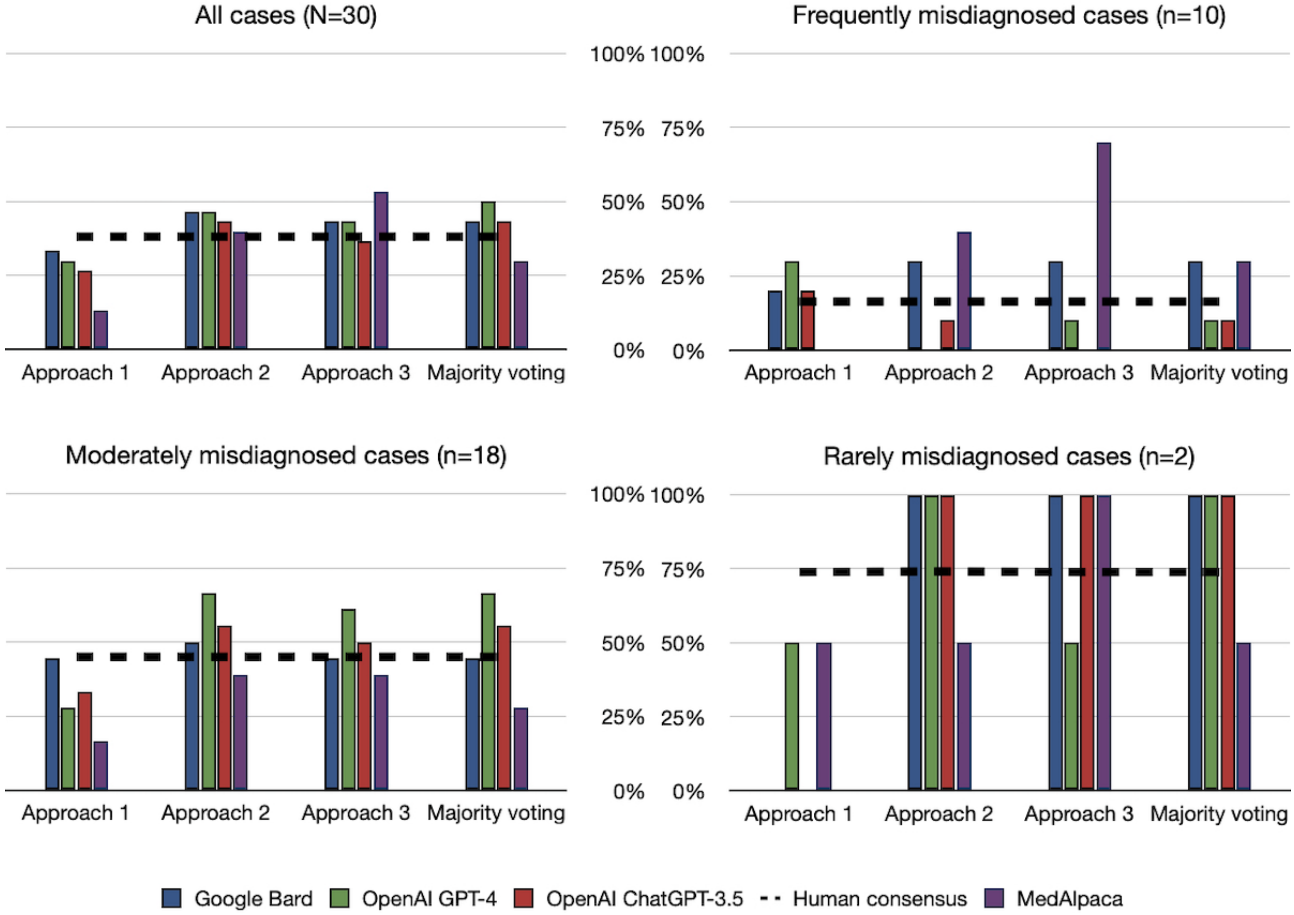
Results

Performance Across Prompt Strategies

Figure 2 reveals the performance of LLMs across different prompts on the DC3 data set. Overall, approach 2 (multiple-choice prompt) yielded the highest score for all 30 cases, with GPT-4 and Bard achieving an accuracy score of 47% (14/30) and ChatGPT-3.5 obtaining a score of 43% (13/30). However, when considering case difficulty, the results varied. On the frequently misdiagnosed cases category, GPT-4 and ChatGPT-3.5 performed better with open-ended prompts (approach 1), scoring 30% (3/10) and 20% (2/10), respectively. In contrast, Bard demonstrated superior performance with multiple-choice prompts for selection and ranking (approaches 2 and 3), achieving a score of 30% (3/10). ChatGPT-3.5 and Bard performed equally well on the rarely misdiagnosed cases category using approaches 2 and 3, achieving a perfect score of 100% (2/2). Furthermore, GPT-4 attained a score of 100% (2/2) but only with approach 2. For the moderately misdiagnosed cases category, all LLMs achieved their best performance with approach 2, scoring 67% (12/18), 56% (10/18), and 50% (9/18) for GPT-4, ChatGPT-3.5, and Bard, respectively. Table S1 in the Multimedia Appendix 1 presents the inconsistencies in the correct responses across the approaches for different cases. For example, Bard could only diagnose milk alkali syndrome using approach 1 but failed to use other prompt approaches. ChatGPT-3.5 correctly diagnosed primary adrenal insufficiency (Addison disease) with only approach 2, whereas GPT-4 was able to diagnose acute hepatitis E virus infection with only approach 1. These results indicate that no universal prompt approach is optimal for all LLMs when dealing with complex cases.

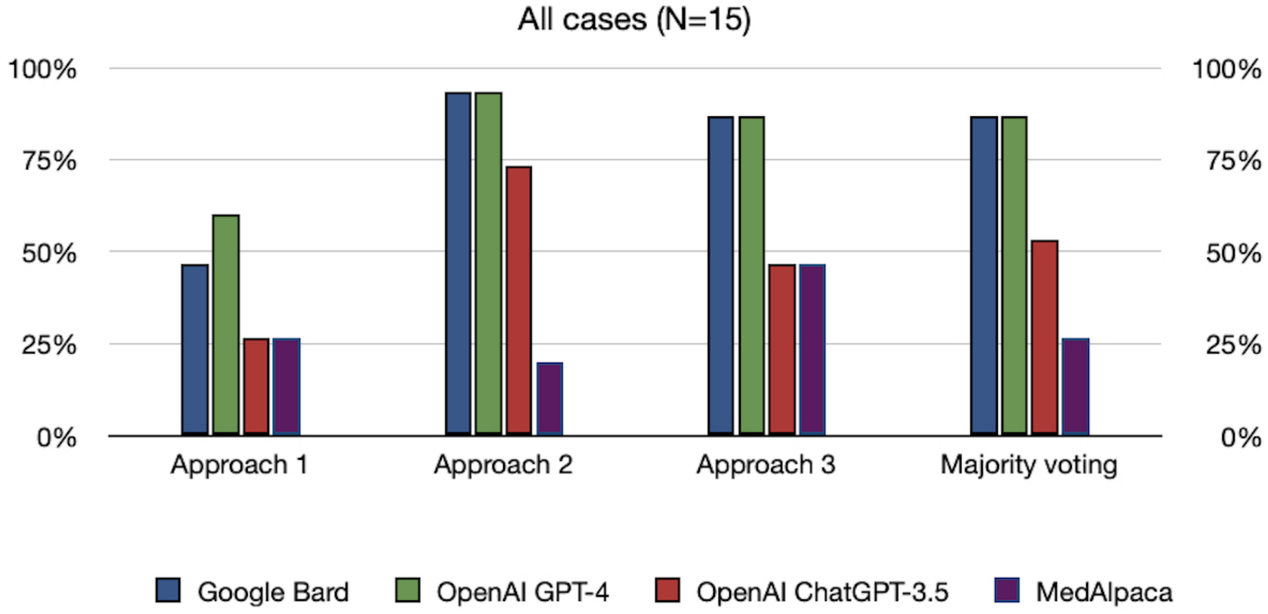
Results on the MIMIC-III data set in Figure 3 showed that the LLMs also performed best using approach 2 (multiple-choice prompt), with Bard and GPT-4 obtaining scores of 93% (14/15) each and ChatGPT-3.5 obtaining 73% (11/15). Using approach 3 (ranking prompt) resulted in a slight drop in performance for GPT-4 and Bard, with a 6% decrease, whereas the performance of ChatGPT-3.5 dropped by 26%. Approach 1 (open-ended prompt) proved challenging for the LLMs, with scores of 47% (7/15), 60% (9/15), and 27% (4/15) for Bard, GPT-4, and ChatGPT-3.5, respectively. Table S2 in the Multimedia Appendix 1 illustrates that approach 1 was only beneficial to GPT-4 in diagnosing amyloidosis, whereas it was consistently never the sole correct approach for Bard and ChatGPT-3.5. These results aligned with the findings from the DC3 data set and emphasized the varying performances of different models and prompt approaches across tasks.

Figure 2. Results of the diagnostic case challenge collection data set comparing prompt strategies. OpenAI GPT-4 outperformed all other models, achieving the highest score in all 30 cases using the majority voting approach. Furthermore, all large language models except MedAlpaca outperformed the human consensus (denoted by a black dashed line) across all cases, regardless of the difficulty, using at least 1 prompt approach. GPT-4: generative pretrained transformer-4.



XSL•FO
RenderX

Figure 3. Results of the Medical Information Mart for Intensive Care-III data set across prompt strategies. Approach 1 (open-ended prompt) proved challenging for all the large language models compared with approach 2 (multiple-choice prompt) and approach 3 (ranking prompt).



Performance With Majority Voting

Previous experiments have demonstrated that there is no perfect prompting strategy because LLM users may not know beforehand which prompt will produce a correct response. We used the majority voting approach to estimate consistency, maximize the benefits of different prompt strategies, and enhance the reliability of the LLMs' responses. Figure 2 illustrates the results for all DC3 cases. Majority voting improved the overall performance of GPT-4 from 47% to 50%, whereas the performance of ChatGPT-3.5 remained at 43% because majority voting did not decrease its performance compared with that of approach 2. In contrast, the performance of Bard decreased from 47% to 43% compared with that of approach 2. Summarizing the overall performance based on query difficulty, majority voting resulted in a perfect score of 100% for the rarely misdiagnosed cases category across all the LLMs. For the frequently misdiagnosed cases category in DC3, Bard achieved the highest score with majority voting and multiple-choice prompts, whereas GPT-4 performed best for the moderately misdiagnosed cases category with majority voting and approach 2. In addition, GPT-4 outperformed all other LLMs across all DC3 cases using the majority voting approach, regardless of the case difficulty. This score surpassed the performance of the individual prompt approaches in all cases.

Results on the MIMIC-III data set in Figure 3 showed that, the scores with majority voting were 87% (13/15) for GPT-4 and Bard each and 53% (8/15) for ChatGPT-3.5. These results indicate that the ensemble method did not substantially improve their performance compared with their best individual approach. It is worth noting that although the majority voting approach did not consistently outperform individual approaches in terms of the highest number of correct responses, it did provide a means to consolidate predictions and mitigate potential errors and biases from single approaches.

Comparison With Human Respondents

In the DC3 cases, although the human respondents had the advantage of accessing supporting patient information such as image scans and magnetic resonance imaging, the LLMs consistently outperformed the average human consensus. As shown in Figure 2, using the majority voting approach, all LLMs achieved a higher performance than the human consensus (denoted by a black dashed line), with a minimum margin of 5% across all 30 cases. Specifically, when considering query difficulty, the LLMs demonstrated even greater superiority. In the rarely misdiagnosed cases category, all LLMs surpassed the average human consensus by a substantial margin of 26%. For the moderately misdiagnosed cases category, GPT-4 and ChatGPT-3.5 maintained their advantage over human respondents, achieving a minimum margin of 11% with the majority voting approach. In contrast, only Bard outperformed the human average consensus on the frequently misdiagnosed cases category, with a margin of 14%.

We conducted a Spearman rank correlation test to analyze the pattern in the responses between each LLM and the human respondents. This involved correlating the average percentage of correct responses for each LLM across the prompt strategies with that of correct human responses. The results of the Spearman correlation test revealed that Bard had a relatively weak correlation coefficient of 0.30, whereas GPT-4 and ChatGPT-3.5 exhibited moderate positive correlations of 0.51 and 0.50, respectively. This suggested that the diagnostic performance patterns of GPT-4 and ChatGPT-3.5 aligned moderately with those of the human respondents. The observed correlation in answering patterns between human respondents and LLMs may stem from the inherent data bias present in the training data sets. The LLMs learn from vast amounts of data, and if the training data are biased toward certain diagnostic or decision-making patterns commonly expressed by human physicians, the model is likely to replicate those patterns. Although the correlation suggested that the LLMs have the

XSL•FO
RenderX

potential to be valuable tools in medical education, it is important to note their correlation with human physicians and that the performance of LLMs does not necessarily mean that they are as good as human physicians in diagnosing and treating diseases.

We could not directly compare the performance of human respondents on the MIMIC-III data sets because of the unavailability of data. Overall, the results indicated that the LLMs consistently outperformed the average human consensus in diagnosing medical cases, showcasing their potential as a tool to complement and enhance care quality and education for complex diagnostic cases.

Comparison With MedAlpaca

On the DC3 data sets, Bard, GPT-4, and ChatGPT-3.5 outperformed MedAlpaca across all cases using the majority voting approach by a minimum margin of 13%. MedAlpaca also displayed the worst performance in the open-ended prompts, irrespective of query difficulty. However, when multiple-choice options were provided, MedAlpaca outperformed the other LLMs in the frequently misdiagnosed cases category. Similar to the DC3 data set, MedAlpaca consistently demonstrated its best performance using the ranking prompt on the MIMIC-III data sets. However, its overall performance was significantly poorer than the other LLMs, with each LLM outperforming the model by at least 26% using the majority voting approach. In contrast to the general-purpose LLMs (eg, Bard, GPT-4, and ChatGPT-3.5), investigating the MedAlpaca model was finetuned using diverse medical tasks and assessed using multiple-choice medical examinations. This tailored training approach likely contributed to its notable performance, particularly excelling in DC3 cases (frequently misdiagnosed instances) and demonstrating optimal results in multiple-choice queries.

Qualitative Analysis

In our experiments, we manually observed the responses of each LLM to all our prompts and noted that each LLM consistently justified its diagnosis choice except for MedAlpaca. Specifically, each LLM offered a logical explanation for its chosen response regardless of the prompting strategy. For further investigation, we analyzed each LLM’s responses in 3 scenarios: (1) when presented with multiple-choice options containing the true diagnosis and they responded accurately, (2) when their response was incorrect, and (3) when given only incorrect multiple-choice options to pick from. In the first scenario, as presented in Multimedia Appendix 1, all LLMs (eg, Bard, GPT-4, and ChatGPT-3.5) mentioned that their rationale for diagnosing *miliary tuberculosis* was owing to relevant symptoms presented in the case, such as a *history of respiratory illness and the presence of mesenteric lymph nodes and numerous tiny nodules throughout both lungs distributed in a miliary pattern.* This pattern of offering insightful reasons for the likelihood of a diagnosis and explaining why other diagnostic options are less probable is valuable for educational purposes. In the second scenario, we observed that there was a notable disparity in the accuracy of human respondents. Only 6% (217/3624) of the human participants provided the correct response, with most votes (1232/3624, 34%) favoring *ulcerative colitis*, whereas 23% (833/3624) of the human responses opted for *salmonellosis*. Notably, Bard and GPT-4 displayed similar behavior by selecting salmonellosis, whereas ChatGPT-3.5 and MedAlpaca chose *ulcerative colitis*.

Another notable finding occurred in the responses of GPT-4 and ChatGPT-3.5. Regardless of the correctness of their chosen diagnoses, these models consistently recommended further tests to confirm their responses. This behavior suggested a general tendency toward advocating additional examinations to validate their diagnoses, potentially reflecting a cautious approach. In contrast, Bard adopted a different approach. Instead of recommending further tests, Bard highlighted that the provided query information supported the diagnosis without suggesting additional confirmatory measures. In the scenario where only incorrect options were given, Bard, ChatGPT-3.5, and MedAlpaca made choices and justified their responses. In contrast, GPT-4 explicitly mentioned that none of the provided options matched the case presentation. Furthermore, GPT-4 suggested a more probable diagnosis and recommended additional testing to explore its feasibility.

Discussion

Principal Findings

Previous studies [4,5] have presented the impressive success of LLMs in standardized medical examinations. We conducted experiments to assess the potential of LLMs as a CME system for rare and complex diagnoses, and our findings demonstrated that LLMs have the potential to be a valuable tool for rare disease education and differential diagnosis. Although LLMs demonstrated superior performance compared with the average human consensus in diagnosing complex diseases, it is essential to note that this does not imply their superiority over physicians. Numerous unknown factors, including the level of respondents’ expertise, may influence the outcome of web-based polls. Furthermore, we examined the knowledge capacity of LLMs through open-ended and multiple-choice prompts and found that LLMs, including MedAlpaca, performed better with multiple-choice prompts. This improvement can be attributed to the options provided, which narrowed the search space for potential diagnoses from thousands to a few likely possibilities. Consequently, we surmise that LLMs are not yet ready to be used as stand-alone tools, which aligns with the findings of previous studies [5,17,18]. Our observations revealed the consistent outperformance of general-purpose LLMs over MedAlpaca in various experiments. Their superior ability to provide valuable justifications for making diagnoses was particularly noteworthy, a strength not matched by MedAlpaca. This difference may stem from MedAlpaca’s exclusive finetuning and assessment for multiple-choice medical examinations, which slightly differ in format from the clinical cases in our experiments.

A notable finding in the response of LLMs to queries was their consistent provision of coherent and reasoned explanations, regardless of the query format. For instance, when diagnosing *miliary tuberculosis*, all 3 LLMs emphasized that the patient’s systemic symptoms, exposure risks, chest radiograph, computed tomography scan findings, and the suspected compromised

XSL•FO
RenderX

immune state collectively support the diagnosis of *miliary tuberculosis*. Furthermore, Bard and GPT-4 ruled out other diagnoses presented in the multiple-choice prompt by highlighting their less typical presentations and lack of certain associated symptoms or risk factors. In addition, the conversational nature of LLMs allows users to ask follow-up questions for further context. These attributes hold great potential for educating users and offering them insights. However, we observed that LLMs provided logical explanations, even when their diagnoses were incorrect. ChatGPT-3.5 and GPT-4 may suggest additional testing to validate their selected diagnosis or use cautious terms like "potential diagnosis." However, it remains unclear whether these recommendations stem from the models' internal confidence or whether there are features intentionally designed by the developers for cautious use. The absence of explicit information regarding the level of uncertainty of LLMs for a specific case is concerning as it could potentially mislead clinicians. The ability to quantify uncertainty is crucial in medical decision-making, in which accurate diagnoses and treatment recommendations are paramount. Clinicians heavily rely on confidence levels and probability assessments to make informed judgments [29]. Without an indication of uncertainty, there is a risk that clinicians may trust the logical explanations provided by the LLMs even when they are incorrect, leading to misdiagnoses or inappropriate treatment plans.

Considering the delicate role of clinical decision support, it is essential to address validity and reliability as crucial aspects of uncertainty. Moreover, a reliable system is of paramount importance for medical education. However, the stochastic nature of LLMs introduces doubts among clinicians regarding their reliability. Although a specific metric to quantitatively assess the reliability of the LLMs used in this study is currently lacking, we acknowledge the significance of consistency in achieving reliability. To address this, we used different prompting strategies and implemented a majority voting approach to select the most consistent response from each LLM. After examining the individual prompt strategies, we anticipated consistent responses across strategies for a specific case. However, our findings revealed that the responses of LLMs were sensitive to concrete prompt formats, particularly in complex diagnoses. For instance, ChatGPT-3.5 and GPT-4 performed better with the open-ended prompt (approach 1) in the frequently misdiagnosed cases category of DC3 cases but struggled with similar cases using multiple-choice and ranking prompts (approaches 2 and 3). In contrast, Bard performed better with multiple-choice prompts. These results highlighted that there is no one-size-fits-all prompting approach nor does a single strategy apply universally to all LLMs. Although the majority voting strategy did not yield optimal results for all models across data sets, it served as a means to consolidate responses from multiple prompts and provided a starting point for incorporating reliability.

Several studies [10-12,14,15] have emphasized the significance of enhancing the education of clinicians at all levels to provide better support for rare and complex diagnoses. In this pursuit, the studies by Lee et al [8] and Decherchi et al [31] have highlighted the potential advantages of artificial intelligence (AI) systems, whereas the studies by Abdullahi et al [25] and Sutton et al [32] have reported a lack of acceptance of AI tools among clinicians. For instance, younger medical students and residents appeared more receptive to integrating technology [33]. One notable reason for this lack of acceptance is that conventional AI systems typically require training before clinicians can effectively use them, which can be burdensome and time consuming [32]. In contrast, conversational LLMs, such as ChatGPT-3.5, Bard, and GPT-4, offer a distinct advantage with their simple interface and dialogue-based nature. These conversational LLMs eliminate the need for extensive training, increasing their potential for high acceptance across all levels of medical practice. Although the exciting ease of use, conversational nature, impressive display of knowledge, and logical explanations of LLMs have the potential for user education and insights, their current limitations in reliability and expressing uncertainty must be addressed to ensure their effective and responsible use in critical domains, such as health care.

Limitations

First, the limitations of the knowledge of ChatGPT-3.5 and GPT-4 to the latest trends and updates in health care (or medical) data till 2021 pose the risk of potentially incomplete information and hamper the effectiveness of the models as a CME tool, especially when addressing emerging diseases. In contrast, although continuous updates to Bard are advantageous for keeping the model up-to-date, this attribute may impact the reproducibility of our study. Second, it is notable that our experiments had a limited scope owing to a small sample size consisting of only 30 diseases from the DC3 data set and 15 cases from the MIMIC-III data set. In addition, although we took precautions to preprocess the MIMIC-III notes to prevent leakage of the final diagnosis, the discharge summaries may still contain nuanced information that could make the diagnosis obvious. Furthermore, the closed nature of the LLMs used in this study restricted our technique for measuring reliability to a majority voting approach, which consolidated responses from diverse prompts. Although majority voting can help to mitigate the variability of LLM output, it is notable that LLMs may still generate different responses for the same prompt. This variability should be considered when interpreting the results of this study. However, when these LLMs are released with an enhanced iteration that allows for finetuning and calibration, future work should incorporate more effective mechanisms to estimate and communicate uncertainty. An example of such an approach could involve assigning a confidence score to the probability score of their responses. This methodology could allow clinicians to make informed decisions regarding whether to accept or reject responses that fall within a desired threshold.

Conclusions

In this study, we conducted experiments to assess the potential of LLMs, including ChatGPT-3.5, GPT-4, and Bard, as a CME system for rare and complex diagnoses. First, we evaluated their diagnostic capability specifically for rare and complex cases. Subsequently, we explored the impact of prompt formatting on their performance. Our results revealed that these LLMs possessed potential diagnostic capacities for rare and complex

XSL•FO
RenderX

medical cases, surpassing the average crowd consensus on the DC3 cases. For selected rare cases from the MIMIC-III data set, Bard and GPT-4 achieved a diagnostic accuracy of 93%, whereas ChatGPT-3.5 achieved an accuracy of 73%. Our findings highlighted that users might discover an approach that yields favorable results for various queries by exploring different prompt formats. In contrast, using majority voting of responses from multiple prompt strategies offers the benefit of a robust and reliable model, instilling confidence in the generated responses. However, determining the best prompt strategy versus relying on the majority voting approach involves a tradeoff between exploration and exploitation. Although prompt engineering research is continuing, we hope that future studies will yield better solutions to enhance the reliability and consistency of the responses of LLMs. Overall, our study's results and conclusions provide a benchmark for the performance of LLMs and shed light on their strengths and limitations in generating responses, expressing uncertainty, and providing diagnostic recommendations. The insights gained from this study can serve as a foundation for further exploration and research on using LLMs as medical education tools to enhance their performance and capabilities as conversational language models.

Acknowledgments

We acknowledge support from the Open Access Publication Fund of the University of Tübingen.

Data Availability

The URLs for the diagnostic case challenge collection data set can be obtained via A Diagnostic Case Challenge Collection [34]. The Medical Information Mart for Intensive Care data sets can be accessed via the database, Medical Information Mart for Intensive Care-III Clinical Database v1.4 [35], after obtaining permission from Physionet.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comprehensive tables detailing the performance of each model across data sets, with included examples of prompts and responses for each model.

[DOCX File , 46 KB - mededu_v10i1e51391_app1.docx]

References

1. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-03-23]
2. Manyika J, Hsiao S. An overview of Bard: an early experiment with generative AI. Google. URL: https://ai.google/static/documents/google-about-bard.pdf [accessed 2024-01-26]
3. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv Preprint posted online March 15, 2023. [FREE Full text]
4. Resnick DK. Commentary: performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023 Jul 19 (forthcoming). [doi: 10.1227/neu.0000000000002618] [Medline: 37466324]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75 [FREE Full text] [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
7. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/nejmsr2214184]
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/s2589-7500(23)00021-3]
10. Mitani AA, Haneuse S. Small data challenges of studying rare diseases. JAMA Netw Open 2020 Mar 02;3(3):e201965 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.1965] [Medline: 32202640]
11. Eickhoff C, Gmehlin F, Patel AV, Boullier J, Fraser H. DC3 -- a diagnostic case challenge collection for clinical decision support. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. 2019 Presented at: ICTIR '19; October 2-5, 2019; Santa Clara, CA. [doi: 10.1145/3341981.3344239]

XSL•FO
RenderX

12. Walkowiak D, Domaradzki J. Are rare diseases overlooked by medical education? Awareness of rare diseases among physicians in Poland: an explanatory study. Orphanet J Rare Dis 2021 Sep 28;16(1):400 [FREE Full text] [doi: 10.1186/s13023-021-02023-9] [Medline: 34583737]
13. Sartorious N. Comorbidity of mental and physical diseases: a main challenge for medicine of the 21st century. Shanghai Arch Psychiatry 2013 Apr;25(2):68-69 [FREE Full text] [doi: 10.3969/j.issn.1002-0829.2013.02.002] [Medline: 24991137]
14. Bateman L, Bested AC, Bonilla HF, Chheda BV, Chu L, Curtin JM, et al. Myalgic encephalomyelitis/chronic fatigue syndrome: essentials of diagnosis and management. Mayo Clin Proc 2021 Nov;96(11):2861-2878 [FREE Full text] [doi: 10.1016/j.mayocp.2021.07.004] [Medline: 34454716]
15. Faviez C, Chen X, Garcelon N, Neuraz A, Knebelmann B, Salomon R, et al. Diagnosis support systems for rare diseases: a scoping review. Orphanet J Rare Dis 2020 Apr 16;15(1):94 [FREE Full text] [doi: 10.1186/s13023-020-01374-z] [Medline: 32299466]
16. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb 9;2(2):e0000205 [FREE Full text] [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]
17. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
18. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
19. Qin G, Eisner J. Learning how to ask: querying LMs with mixtures of soft prompts. arXiv Preprint posted online April 14, 2021. [FREE Full text] [doi: 10.18653/v1/2021.naacl-main.410]
20. Si C, Gan Z, Yang Z, Wang S, Wang J, Boyd-Graber J, et al. Prompting GPT-3 to be reliable. arXiv Preprint posted online October 17, 2022. [FREE Full text]
21. Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, et al. Large language models are human-level prompt engineers. arXiv Preprint posted online November 3, 2022. [FREE Full text]
22. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser, T, Löser A, et al. MedAlpaca -- an open-source collection of medical conversational AI models and training data. arXiv Preprint posted online April 14, 2023. [FREE Full text]
23. Johnson AE, Pollard TJ, Shen LW, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
24. van Aken B, Papaioannou JM, Mayrdorfer M, Budde K, Gers F, Loeser A. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021 Presented at: 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; April 21-23, 2021; Online. [doi: 10.18653/v1/2021.eacl-main.75]
25. Abdullahi TA, Mercurio L, Singh R, Eickhoff C. Retrieval-based diagnostic decision support. JMIR Preprints Preprint posted online June 25, 2023. [FREE Full text] [doi: 10.2196/preprints.50209]
26. Orphanet: about rare diseases. Orphanet. URL: https://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN [accessed 2023-07-03]
27. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv Preprint posted online January 28, 2022. [FREE Full text]
28. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv Preprint posted online March 21, 2022. [FREE Full text]
29. Li Y, Lin Z, Zhang S, Fu Q, Chen B, Lou JG, et al. Making language models better reasoners with step-aware verifier. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023 Presented at: 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 9-14, 2023; Toronto, ON. [doi: 10.18653/v1/2023.acl-long.291]
30. NIH investigator manual for human subjects research. Office of Intramural Research. Office of Human Subjects Research Protections. URL: https://ohsrp.nih.gov/confluence/display/ohsrp/Chapter+1+-+Types+of+Research_+Human+Subjects+Research+Vs.+Not+Human+Subjects+Research [accessed 2024-01-31]
31. Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi L. Opportunities and challenges for machine learning in rare diseases. Front Med (Lausanne) 2021 Oct 5;8:747612 [FREE Full text] [doi: 10.3389/fmed.2021.747612] [Medline: 34676229]
32. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 06;3(1):17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
33. Eckleberry-Hunt J, Lick D, Hunt R. Is medical education ready for generation Z? J Grad Med Educ 2018 Aug;10(4):378-381 [FREE Full text] [doi: 10.4300/JGME-D-18-00466.1] [Medline: 30154963]
34. codiag-public / dc3. GitHub. URL: https://github.com/codiag-public/dc3/blob/master/cases.url [accessed 2024-01-31]

XSL•FO
RenderX

35. Johnson A, Pollard T, Mark R. MIMIC-III clinical database (version 1.4). PhysioNet. 2016. URL: https://physionet.org/content/mimiciii/1.4/ [accessed 2024-01-31]

Abbreviations

AI: artificial intelligence
CME: continuing medical education
DC3: diagnostic case challenge collection
LLM: large language model
MIMIC-III: Medical Information Mart for Intensive Care-III
USMLE: United States Medical Licensing Examination

Edited by G Eysenbach, K Venkatesh, MN Kamel Boulos; submitted 30.07.23; peer-reviewed by L Modersohn, S Ghanvatkar; comments to author 20.10.23; revised version received 07.11.23; accepted 11.12.23; published 13.02.24.

Please cite as:
Abdullahi T, Singh R, Eickhoff C
Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models
JMIR Med Educ 2024;10:e51391
URL: https://mededu.jmir.org/2024/1/e51391
doi:10.2196/51391
PMID:38349725

©Tassallah Abdullahi, Ritambhara Singh, Carsten Eickhoff. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Viewpoint

Using ChatGPT-Like Solutions to Bridge the Communication Gap Between Patients With Rheumatoid Arthritis and Health Care Professionals

Chih-Wei Chen[1,2,3,4*], MPhil; Paul Walter[1,5,6], MSc; James Cheng-Chung Wei[3,7*], MD, PhD

[1]National Applied Research Laboratories, Taipei, Taiwan
[2]National Council for Sustainable Development, Taipei, Taiwan
[3]Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan
[4]Faculty of Engineering Sciences, University College London (UCL), London, United Kingdom
[5]Faculty of Pharmacy, Paris-Saclay University, Orsay, France
[6]Mines Saint-Etienne, Saint-Etienne, France
[7]Department of Allergy, Immunology & Rheumatology, Chung Shan Medical University Hospital, Taichung, Taiwan
[*]these authors contributed equally

Corresponding Author:
Chih-Wei Chen, MPhil
National Applied Research Laboratories
3F, No 106, Sector 2
Heping East Road
Taipei, 106214
Taiwan
Phone: 886 975303092
Email: chihwei.chen@udm.global

Abstract

The communication gap between patients and health care professionals has led to increased disputes and resource waste in the medical domain. The development of artificial intelligence and other technologies brings new possibilities to solve this problem. This viewpoint paper proposes a new relationship between patients and health care professionals—"shared decision-making"—allowing both sides to obtain a deeper understanding of the disease and reach a consensus during diagnosis and treatment. Then, this paper discusses the important impact of ChatGPT-like solutions in treating rheumatoid arthritis using methotrexate from clinical and patient perspectives. For clinical professionals, ChatGPT-like solutions could provide support in disease diagnosis, treatment, and clinical trials, but attention should be paid to privacy, confidentiality, and regulatory norms. For patients, ChatGPT-like solutions allow easy access to massive amounts of information; however, the information should be carefully managed to ensure safe and effective care. To ensure the effective application of ChatGPT-like solutions in improving the relationship between patients and health care professionals, it is essential to establish a comprehensive database and provide legal, ethical, and other support. Above all, ChatGPT-like solutions could benefit patients and health care professionals if they ensure evidence-based solutions and data protection and collaborate with regulatory authorities and regulatory evolution.

(JMIR Med Educ 2024;10:e48989) doi:10.2196/48989

KEYWORDS
rheumatoid arthritis; ChatGPT; artificial intelligence; communication gap; privacy; data management

Introduction

In recent years, the communication gap has led to intense relationships between patients and health care professionals. The use of ChatGPT-like solutions in health care has enormous potential to improve the patient-provider relationship, such as patient clinic letter writing [1], medical note-taking and consultation [2], and rheumatoid arthritis treatment. Although ChatGPT (OpenAI) [3] is not the only solution available, the technology has generated a lot of traction due to its advanced features, such as the ability to enhance rule-based chatbots.

However, it is important to note that ChatGPT-like solutions should not be viewed as a stand-alone solution but as an integrated interface in a larger ecosystem that allows access to

XSL•FO
RenderX

multiple data sources. In terms of the patient-provider relationship, the use of ChatGPT can enable more fluid and effective communication between the 2 parties, which can improve the quality of care. In particular, the use of ChatGPT-like solutions in the context of methotrexate treatment for rheumatoid arthritis could have a significant impact. This viewpoint paper proposes a new relationship between patients and providers—"shared decision-making"; explains the potential of ChatGPT-like solutions in improving the patient–health care professional relationship from the clinical and patient perspectives; and suggests the importance of establishing a comprehensive database to promote the implementation of "shared decision-making" between patients and health care professionals.

Toward Shared Decision-Making

In conventional medical settings, the relationship between patients and health care professionals was not equal, mainly because of the huge information gap between them, since patients lacked medical knowledge and decision-making capacity. In recent years, the rapid development of ChatGPT possesses enormous potential to bridge the information gap and improve the relationship between patients and health care professionals. For instance, ChatGPT could help provide the risk-benefit analysis of different treatment options, assisting health care professionals and patients to understand the advantages and disadvantages of each option and then make informed decisions together. It could also assist patients in understanding the complex medical jargon and technical details and provide information about the disease, treatment options, potential risks, and expected outcomes, allowing patients to participate in making informed decisions with health care professionals together. ChatGPT-like solutions allow bilateral communications between patients and health care professionals toward shared decision-making.

Clinical Perspective

From the clinical standpoint, early diagnosis of rheumatoid arthritis is crucial [4] for health care professionals and should be based on clinical examinations and biological results such as serological tests [5]. However, the differential diagnosis of complex diseases such as rheumatoid arthritis–associated interstitial lung disease [6] remains a major concern [7], as it is responsible for a significant increase in mortality [8]. ChatGPT-like solutions could bring complementary support to diagnose the disease and predict its evolution. Thus, to the query "What could be the reason for cough and dyspnea in a patient with rheumatoid arthritis?" ChatGPT suggests interstitial lung disease in the first place. By integrating external data on risk factors (age, sex, and smoking), biological results (pulmonary function testing, autoantibodies, and biopsy), and imaging (high-resolution computed tomography and ultrasound) [6], ChatGPT-like solutions can assist in suggesting additional tests and confirming the diagnosis.

The initiation of treatment for rheumatoid arthritis should be in accordance with the latest official recommendations, such as those from the European League Against Rheumatism [9] and the American College of Rheumatology (ACR) [4]. An advanced tool such as ChatGPT provides clinicians with exhaustive information on the latest guidelines for the management of rheumatoid arthritis. For instance, if a clinician asks "What are the current guidelines for treating rheumatoid arthritis according to the ACR?" ChatGPT can retrieve the key points of rheumatoid arthritis management in accordance with the official ACR guidelines [4]. However, in the specific case of a request regarding recommendations for treating rheumatoid arthritis–associated interstitial lung disease, ChatGPT erroneously refers to nonexistent ACR guidance [10]. Currently, the tool has limitations, such as data exclusion after 2021 and response size limits.

The determination of a patient's drug dose by the clinician is based on a comprehensive evaluation of the results of the biological tests and clinical examination. However, dose adjustment may not always be performed according to a standardized procedure and evidence-based solution, although this is crucial to ensure the effectiveness and tolerability of the treatment for the patient. Methotrexate is the most common treatment for rheumatoid arthritis, and an initial dose of 7.5-15 mg once a week is recommended, followed by a gradual increase in dose. However, poor patient adherence and nonpersistence to methotrexate therapy have been reported [11] mainly due to low dose tolerance. Optimization of methotrexate dose is therefore essential for treating rheumatoid arthritis [12]. The use of methotrexate monotherapy has shown similar efficacy to the combined use of methotrexate monotherapy with biologic disease–modifying antirheumatic drugs [13]. Process automation and integration of complementary data, based on solutions such as ChatGPT, could improve outcome prediction, contribute to drug dose optimization, and thus reduce costs to the health care system.

Access to information on ongoing clinical trials and their results would enable clinicians to propose treatments for people with rare conditions in rheumatoid arthritis. Compiling data on clinical trials and patient characteristics would allow clinicians to propose alternatives, for example, for patients who have failed current therapies. Identifying subpopulations would facilitate patient recruitment and bring more effective and safer drugs to market. However, one challenge is to deidentify data to comply with the US Health Insurance Portability and Accountability Act (HIPAA) [14]. As such, it is important for clinicians to prioritize patient privacy and confidentiality when accessing and using such data. In addition, it is necessary for further interdisciplinary research to improve the accuracy and persuasiveness of artificial intelligence (AI) chatbots to influence patients' behaviors [15]. Moreover, the application of AI and machine learning in health care should still be regulated by establishing norms to reduce bias and reflect the real problems [16].

Patient Perspective

From the patient's perspective, it allows easy access to a large volume of information with a certain degree of scientific evidence, which improves the patient's knowledge of rheumatoid arthritis and their health literacy. ChatGPT-like

XSL•FO
RenderX

solutions thus contribute to dealing with the proliferation of unreliable sources of emerging information and widespread disinformation [17]. It is also a tool that could not only enable empowerment by acting interactively throughout the care pathway but also promote patient adherence to treatment. However, some concerns persist regarding the lack of supervision of this type of solution and the liability involved [18]. For example, in the case of methotrexate side effects, to the query "I have gastrointestinal problems and fatigue, is this related to my methotrexate intake?" ChatGPT suggests that the doctor can adjust the dosage. It does not provide suggestions to state that concomitant folate or folic acid changes would reduce toxicity. It also raises questions about the risk of patients adjusting their own dosage. ChatGPT-like solutions can strengthen expert patients' collaboration, allowing the cocreation of care pathways; however, it can also be a source of conflict by pitting the tool's and the caregiver's advice against each other. Therefore, it is crucial to better supervise this tool from the beginning of its development, in order to clearly distinguish between its general public and medical use and to define the responsibilities of each. The use of ChatGPT-like solutions can improve communication and access to information for patients with rheumatoid arthritis but must be carefully managed to ensure safe and effective care.

A ChatGPT-like solution allows the patient to have continuous access to information in an interactive way that promotes understanding outside the clinical setting. This solution can play an important role in therapeutic education by providing information on the self-management of rheumatoid arthritis, on a drug such as methotrexate, or on the administration methods (oral and subcutaneous). Therefore, the query of "What precautions should be taken when taking methotrexate?" could instantly provide basic and exhaustive information (taking it with food, avoiding alcohol, staying hydrated, using contraceptives, etc) and could contribute to therapeutic education [19]. In addition, a ChatGPT-like solution could be used to communicate medical information on potential benefits and assist in administration [20], for example, when modifying the route of administration of methotrexate. This would have an impact on facilitating the acceptability of subcutaneous methotrexate, allowing better bioavailability and clinical efficacy. It would also reduce the time required to initiate treatment and avoid the use of biologics, thus having a significant impact on health care costs [21].

Further integration and analysis of patient requests would also accelerate the transition to more personalized medicine. ChatGPT-like solutions could identify patient profiles and adapt communication strategies to overcome resistance and nudge behavior. These solutions will have to be adapted to each country in terms of public health systems and beliefs.

Establishment of a Comprehensive Database

The database is one of the critical elements of digital infrastructure for digital technology applications [22], especially AI-based solutions that require huge amounts of data to achieve more accurate results. However, using AI-based technology can be limited by the nontransparent learning process, difficulties in explanation and validation, and the influence of improper data [23]. Hence, the establishment of a comprehensive database, which is sourced from real-world data and updated on time and precisely, could contribute to overcoming limitations caused by insufficient data and support evidence-based clinical applications.

In recent decades, the Taiwan government launched the National Health Insurance (NHI) system that collected health-related data of health care providers, citizens, and legal residents. Since its establishment, the NHI database has been continuously improved by using the latest technologies to accommodate the increasing needs. During the COVID-19 pandemic, the NHI database successfully supported the Taiwan government in tracking patients, distributing face masks, and containing the infections [24,25].

On the other hand, using mobile health tools also contributes to the establishment of a comprehensive database. In recent years, tools such as the Apple Watch have been widely used to collect data about health conditions and identify possible illnesses of people. Mobile health tools allow the collection active data and passive data, which could better inform the health condition of the people [26].

Above all, the establishment of a comprehensive database is fundamental to applying AI-based solutions in the digital governance of health care. Moreover, applying AI-based solutions and other digital technologies should also be accompanied with comprehensive planning and flexible strategies to achieve effective digital governance in health care [22].

Conclusions

In conclusion, ChatGPT-like solutions have the potential to improve the patient-provider relationship through "shared decision-making." ChatGPT solutions should optimize the patient's care pathway while improving the patient's experience of using methotrexate in rheumatoid arthritis. However, there is a need to ensure evidence-based solutions and quantify these benefits. In the future, we may question the compatibility of the business model of mass-market solutions with health care system purposes, particularly concerning data protection. Using federated learning might be a way for developers to overcome this limitation. The implementation in a specific health care context should increase in the coming years with the development of solutions in specific domains such as Bio-Generative Pre-Trained Transformer. A deployment in clinical settings will require collaboration with regulatory authorities and potentially an evolution of the software as a medical device regulatory framework [27].

The need to include individuals in the design of these solutions is also crucial to consider from an efficiency point of view to avoid certain biases and from an ethical point of view. This solution also facilitates access to health care information for the entire world population in pursuit of the sustainable development goals set by the United Nations.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181 [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
3. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-02-20]
4. Fraenkel L, Bathon JM, England BR, St Clair EW, Arayssi T, Carandang K, et al. 2021 American College of Rheumatology guideline for the treatment of rheumatoid arthritis. Arthritis Care Res (Hoboken) 2021;73(7):924-939 [FREE Full text] [doi: 10.1002/acr.24596] [Medline: 34101387]
5. Cush JJ. Rheumatoid arthritis: early diagnosis and treatment. Med Clin North Am 2021;105(2):355-365. [doi: 10.1016/j.mcna.2020.10.006] [Medline: 33589108]
6. Dai Y, Wang W, Yu Y, Hu S. Rheumatoid arthritis-associated interstitial lung disease: an overview of epidemiology, pathogenesis and management. Clin Rheumatol 2021;40(4):1211-1220. [doi: 10.1007/s10067-020-05320-z] [Medline: 32794076]
7. Bendstrup E, Møller J, Kronborg-White S, Prior TS, Hyldgaard C. Interstitial lung disease in rheumatoid arthritis remains a challenge for clinicians. J Clin Med 2019;8(12):2038 [FREE Full text] [doi: 10.3390/jcm8122038] [Medline: 31766446]
8. Hyldgaard C, Hilberg O, Pedersen AB, Ulrichsen SP, Løkke A, Bendstrup E, et al. A population-based cohort study of rheumatoid arthritis-associated interstitial lung disease: comorbidity and mortality. Ann Rheum Dis 2017;76(10):1700-1706. [doi: 10.1136/annrheumdis-2017-211138] [Medline: 28611082]
9. Smolen JS, Landewé RBM, Bergstra SA, Kerschbaumer A, Sepriano A, Aletaha D, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2022 update. Ann Rheum Dis 2023;82(1):3-18 [FREE Full text] [doi: 10.1136/ard-2022-223356] [Medline: 36357155]
10. Cassone G, Manfredi A, Vacchi C, Luppi F, Coppi F, Salvarani C, et al. Treatment of rheumatoid arthritis-associated interstitial lung disease: lights and shadows. J Clin Med 2020;9(4):1082 [FREE Full text] [doi: 10.3390/jcm9041082] [Medline: 32290218]
11. Curtis JR, Bykerk VP, Aassi M, Schiff M. Adherence and persistence with methotrexate in rheumatoid arthritis: a systematic review. J Rheumatol 2016;43(11):1997-2009 [FREE Full text] [doi: 10.3899/jrheum.151212] [Medline: 27803341]
12. Bello AE, Perkins EL, Jay R, Efthimiou P. Recommendations for optimizing methotrexate treatment for patients with rheumatoid arthritis. Open Access Rheumatol 2017;9:67-79 [FREE Full text] [doi: 10.2147/OARRR.S131668] [Medline: 28435338]
13. Gaujoux-Viala C, Hudry C, Zinovieva E, Herman-Demars H, Flipo RM. MTX optimization or adding bDMARD equally improve disease activity in rheumatoid arthritis: results from the prospective study STRATEGE. Rheumatology (Oxford) 2021;61(1):270-280 [FREE Full text] [doi: 10.1093/rheumatology/keab274] [Medline: 33774669]
14. Office for Civil Rights, Department of Health and Human Services. Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; other modifications to the HIPAA rules. Fed Regist 2013;78(17):5565-5702 [FREE Full text] [Medline: 23476971]
15. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. J Med Internet Res 2020;22(9):e22845 [FREE Full text] [doi: 10.2196/22845] [Medline: 32996892]
16. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
17. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. Cureus 2023;15(2):e35029 [FREE Full text] [doi: 10.7759/cureus.35029] [Medline: 36819954]
18. Sanderson K. GPT-4 is here: what scientists think. Nature 2023;615(7954):773. [doi: 10.1038/d41586-023-00816-5] [Medline: 36928404]
19. Fayet F, Pereira B, Fan A, Rodere M, Savel C, Berland P, et al. Therapeutic education improves rheumatoid arthritis patients' knowledge about methotrexate: a single center retrospective study. Rheumatol Int 2021;41(11):2025-2030. [doi: 10.1007/s00296-021-04893-5] [Medline: 34050794]
20. Molina JT, Robledillo JCL, Ruiz NC. Potential benefits of the self-administration of subcutaneous methotrexate with autoinjector devices for patients: a review. Drug Healthc Patient Saf 2021;13:81-94 [FREE Full text] [doi: 10.2147/DHPS.S290771] [Medline: 33824602]

XSL•FO
RenderX

21. Sharma P, Scott DGI. Optimizing methotrexate treatment in rheumatoid arthritis: the case for subcutaneous methotrexate prior to biologics. Drugs 2015;75(17):1953-1956. [doi: 10.1007/s40265-015-0486-7] [Medline: 26474779]
22. Chen CW, Wei JCC. Employing digital technologies for effective governance: Taiwan's experience in COVID-19 prevention. Health Policy Technol 2023;12(2):100755 [FREE Full text] [doi: 10.1016/j.hlpt.2023.100755] [Medline: 37287501]
23. Hunter DJ, Holmes C. Where medical statistics meets artificial intelligence. N Engl J Med 2023;389(13):1211-1219. [doi: 10.1056/NEJMra2212850] [Medline: 37754286]
24. Chen CM, Jyan HW, Chien SC, Jen HH, Hsu CY, Lee PC, et al. Containing COVID-19 among 627,386 persons in contact with the diamond princess cruise ship passengers who disembarked in Taiwan: big data analytics. J Med Internet Res 2020;22(5):e19540 [FREE Full text] [doi: 10.2196/19540] [Medline: 32353827]
25. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. JAMA 2020;323(14):1341-1342. [doi: 10.1001/jama.2020.3151] [Medline: 32125371]
26. Spadaro B, Martin-Key NA, Bahn S. Building the digital mental health ecosystem: opportunities and challenges for mobile health innovators. J Med Internet Res 2021;23(10):e27507 [FREE Full text] [doi: 10.2196/27507] [Medline: 34643537]
27. Artificial intelligence and machine learning in software as a medical device. US Food and Drug Administration (FDA). URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [accessed 2024-02-09]

Abbreviations

ACR: American College of Rheumatology
AI: artificial intelligence
HIPAA: Health Insurance Portability and Accountability Act
NHI: National Health Insurance

Edited by K Venkatesh, MN Kamel Boulos; submitted 14.05.23; peer-reviewed by Y Zhuang, A Hidki, P Iyer; comments to author 10.08.23; revised version received 09.10.23; accepted 05.02.24; published 27.02.24.

Please cite as:
Chen CW, Walter P, Wei JCC
Using ChatGPT-Like Solutions to Bridge the Communication Gap Between Patients With Rheumatoid Arthritis and Health Care Professionals
JMIR Med Educ 2024;10:e48989
URL: https://mededu.jmir.org/2024/1/e48989
doi:10.2196/48989
PMID:38412022

©Chih-Wei Chen, Paul Walter, James Cheng-Chung Wei. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Exploring the Feasibility of Using ChatGPT to Create Just-in-Time Adaptive Physical Activity mHealth Intervention Content: Case Study

Amanda Willms[1*], MSc; Sam Liu[1*], PhD

School of Exercise Science, Physical and Health Education, University of Victoria, Victoria, BC, Canada

[*]all authors contributed equally

Corresponding Author:
Amanda Willms, MSc
School of Exercise Science, Physical and Health Education
University of Victoria
PO Box 3010 STN CSC
Victoria, BC, V8W 2Y2
Canada
Phone: 1 250 721 8392
Email: awillms@uvic.ca

Abstract

Background: Achieving physical activity (PA) guidelines' recommendation of 150 minutes of moderate-to-vigorous PA per week has been shown to reduce the risk of many chronic conditions. Despite the overwhelming evidence in this field, PA levels remain low globally. By creating engaging mobile health (mHealth) interventions through strategies such as just-in-time adaptive interventions (JITAIs) that are tailored to an individual's dynamic state, there is potential to increase PA levels. However, generating personalized content can take a long time due to various versions of content required for the personalization algorithms. ChatGPT presents an incredible opportunity to rapidly produce tailored content; however, there is a lack of studies exploring its feasibility.

Objective: This study aimed to (1) explore the feasibility of using ChatGPT to create content for a PA JITAI mobile app and (2) describe lessons learned and future recommendations for using ChatGPT in the development of mHealth JITAI content.

Methods: During phase 1, we used Pathverse, a no-code app builder, and ChatGPT to develop a JITAI app to help parents support their child's PA levels. The intervention was developed based on the Multi-Process Action Control (M-PAC) framework, and the necessary behavior change techniques targeting the M-PAC constructs were implemented in the app design to help parents support their child's PA. The acceptability of using ChatGPT for this purpose was discussed to determine its feasibility. In phase 2, we summarized the lessons we learned during the JITAI content development process using ChatGPT and generated recommendations to inform future similar use cases.

Results: In phase 1, by using specific prompts, we efficiently generated content for 13 lessons relating to increasing parental support for their child's PA following the M-PAC framework. It was determined that using ChatGPT for this case study to develop PA content for a JITAI was acceptable. In phase 2, we summarized our recommendations into the following six steps when using ChatGPT to create content for mHealth behavior interventions: (1) determine target behavior, (2) ground the intervention in behavior change theory, (3) design the intervention structure, (4) input intervention structure and behavior change constructs into ChatGPT, (5) revise the ChatGPT response, and (6) customize the response to be used in the intervention.

Conclusions: ChatGPT offers a remarkable opportunity for rapid content creation in the context of an mHealth JITAI. Although our case study demonstrated that ChatGPT was acceptable, it is essential to approach its use, along with other language models, with caution. Before delivering content to population groups, expert review is crucial to ensure accuracy and relevancy. Future research and application of these guidelines are imperative as we deepen our understanding of ChatGPT and its interactions with human input.

(JMIR Med Educ 2024;10:e51426) doi:10.2196/51426

XSL•FO
RenderX

KEYWORDS

ChatGPT; digital health; mobile health; mHealth; physical activity; application; mobile app; mobile apps; content creation; behavior change; app design

Introduction

Physical inactivity is a key modifiable risk factor for many chronic conditions, including cardiovascular disease, type 2 diabetes, and cancers, throughout the lifespan [1]. Despite this evidence, adults and adolescents alike are not consistently meeting the recommended guidelines to prevent developing these chronic conditions [2]. Previous studies have shown that 150 minutes of moderate-to-vigorous physical activity (MVPA) can reduce the risk of all-cause mortality by at least 30%, along with reducing the risk for chronic conditions such as cardiovascular disease (30%), colon cancer (20%), and breast cancer (14%) [3]. Although many chronic diseases affect adults, healthy lifestyle habits need to be developed early from childhood. Children aged 8 to 12 years are more flexible than adults in their ability to change behaviors because they are just beginning to develop self-regulation skills, habits, and identities for healthy living [4,5]. Thus, many countries such as Canada [6], the United States [7], and the United Kingdom [8] have set guidelines recommending 60 minutes of MVPA per day for children 17 years and younger [2]. However, despite these recommendations, physical inactivity is prevalent among children, with less than one-quarter of children meeting the guidelines in countries such as Canada [9] and the United States [9]. Consequently, promoting regular PA to prevent chronic diseases and maintain lifelong health has been a key priority for governments worldwide.

Recent studies suggest that family-based PA programs can be an effective strategy to improve PA levels in children [10,11]. These programs focus on providing guidance for parents to support their child's PA (eg, encouragement, providing opportunity, and logistic support) [12]. With advancements in mobile health (mHealth) technologies and improved access to smartphones, emerging evidence indicates that PA interventions delivered through mHealth technology can be effective while improving scalability and personalization. However, the effectiveness and engagement of interventions vary depending on the intervention design and the degree of tailoring [13,14]. Studies have demonstrated that tailored mHealth interventions are more effective in improving behavior and health outcomes compared with nontailored interventions [15]. A recent advancement in tailored mHealth interventions is the development of just-in-time adaptive interventions (JITAIs), which use mHealth technology to assess the dynamically changing needs of individuals and deliver tailored support in real time [14,16]. Thus far, JITAIs have shown great promise in promoting PA among adults [17], university students [18], and chronic disease populations [19]. Further, innovative mHealth "no-code" development platforms, such as Pathverse, have made the development and implementation of JITAIs much easier and cost-effective [20,21]. However, the development of content for JITAIs can be extremely labor-intensive due to the need to create various versions of health-related content for different tailored algorithms. Although the documentation of content creation timelines for PA JITAIs is in its infancy, a typical timeline for PA content creation from the formative phase to pilot testing reportedly ranges from 12 [22] to 15 months [23,24].

Specific to JITAIs, the typical process of creating evidence-based and engaging content for these mHealth interventions typically involves the following steps [21,25]: (1) defining the behavior change theories and behavior change techniques (BCTs) required for the intervention [26]; (2) gathering evidence from various sources, such as previous literature, public health sources, gray literature, and blogs, and then adapting it to suit the needs of the intervention and deliver it through the chosen medium; and (3) writing content that is engaging and matches the literacy level of the target population for the app. These steps can often be time-consuming, with the need for researchers to follow these steps iteratively and repetitively for the duration of the design of the intervention. Further, despite the consideration of these steps, several challenges still arise in the development of JITAI content. Existing studies have identified limitations, such as the need for more extensive content within interventions, struggles in creating novel and meaningful messages, and challenges in tailoring messages to diverse user preferences [27,28]. These studies have also recognized the resource constraints in developing content to meet these needs and the complex, multidimensional nature of creating tailored and engaging content for their sample. Therefore, an artificial intelligence (AI) tool such as ChatGPT (OpenAI) [29] can be extremely useful in making the process of generating JITAI content for mHealth interventions faster and more cost-effective. ChatGPT offers a solution to the need for more content within interventions by leveraging its vast training data and the ability to generate a diverse set of messages efficiently. Further, its generative capabilities and the ability for users to continually prompt new rules address the challenge of creating novel content, reducing the risk of messages being perceived as overly simplistic.

ChatGPT was first launched by OpenAI in November 2022 and is an open AI language model that generates human-like responses to text-based prompts [30]. It can understand and generate responses in various languages, as well as debug code, write stories in different genres and lengths, summarize information from complex texts, offer explanations on various topics, and even reject answering inappropriate prompts [31]. Unlike other generative large language models (LLMs), ChatGPT stands out as the inaugural member of a series of highly scaled LLMs that attain state-of-the-art performance with minimal need for fine-tuning [32]. Further, ChatGPT is highly sophisticated in that it is able to provide continuous dialog by remembering what the user has said earlier in the conversation thread [33].

Although ChatGPT hosts an impressive suite of features and capabilities, there are also several ethical and privacy concerns

XSL•FO
RenderX

to keep in mind while using this service. First, it is important to note that ChatGPT "learns" its information from human input. This is subject to error and is limited based on what others have input into its system. Further, when generating health information content, in particular, this LLM has been extensively trained with data up to 2021, thus limiting some of the relevance and accuracy of current practices [34]. Second, ChatGPT stores its data in the United States, which, depending on the type of information being input into the United States, may be subjected to privacy concerns based on US freedom and privacy laws. To build on this consideration of data storage, it is crucial not to input any personal health information or other sensitive data into ChatGPT, as this LLM continues to learn from text prompts.

Since its inception, ChatGPT has been widely cited in various bodies of behavioral science literature as a virtual assistant, chatbot, and language translation tool [35]. To generate output from the program, a concept called prompt engineering is one method that explains how ChatGPT generates output [36]. In LLMs, a prompt is defined as an instruction to the model that customizes, enhances, or refines the output [37]. However, there is currently a lack of studies examining the feasibility of using ChatGPT to help develop intervention content for JITAIs aimed to promote PA when given a behavior change theory and a behavior target outcome.

Thus, the primary objective of this paper was to present an autoethnographic case study that explored the feasibility, including the acceptability and ease of use, of using ChatGPT to create content for a family-based PA JITAI mobile app. The secondary objective was to describe lessons learned and future recommendations for using ChatGPT in developing mHealth intervention content.

Methods

Study Design

This case study consisted of 2 phases, which took place from March 1, 2023, to April 30, 2023. In phase 1 (0-2 months), we used ChatGPT-3 to develop a 10-week family-based PA JITAI. In phase 2 (3-4 months), we described lessons learned based on our experience of using ChatGPT in phase 1 and provided future recommendations for using ChatGPT in the development of mHealth interventions.

Ethical Considerations

This paper outlines the procedural aspects of using ChatGPT for content generation for a subsequent study. Given that it operates independently without involvement of human participants or sensitive data, formal ethics approval from our institution was deemed unnecessary.

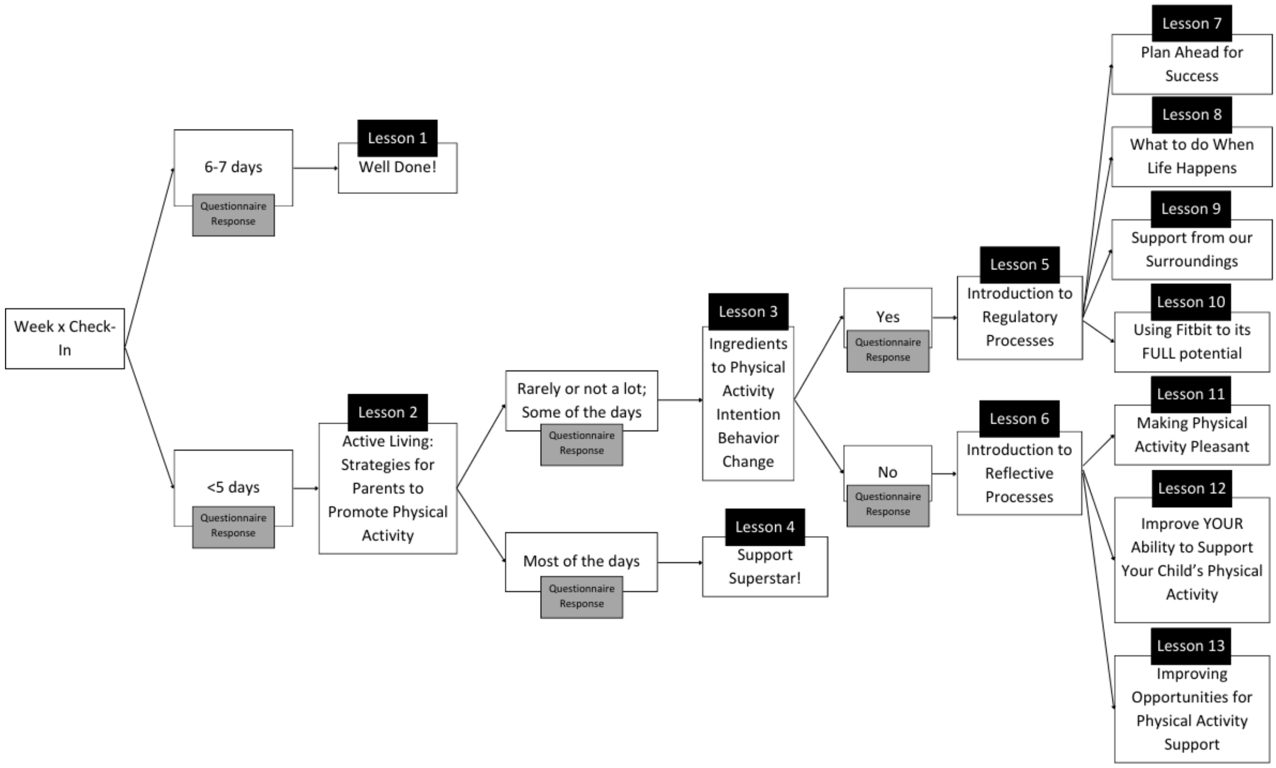
Phase 1

We explored the feasibility of using ChatGPT to create content for the PA JITAI mobile app. To determine the feasibility of using ChatGPT to rapidly create JITAI content, we used an autoethnographic case study approach [38]. This method enabled the researchers (AW and SL) to reflect on their experience of using ChatGPT. While using ChatGPT, the researchers created field notes and had a meeting to discuss their independent experiences with using ChatGPT-3. Specifically, we reflected on the acceptability and ease of use as key areas of focus for feasibility [39]. Results of the meeting were themed into acceptability and ease of use of using ChatGPT. Assessing acceptability metrics involved reflecting on the satisfaction of the response generated by ChatGPT. The ease-of-use assessment involved reflecting on ChatGPT usability [39]. In this phase, we used 2 tools, Pathverse and ChatGPT. Pathverse is a no-code app builder platform that supports mHealth research [20,40]. It consists of a web portal for researchers to create engaging mobile app interventions with "drag and drop" features instead of coding. The content is then instantly displayed on the Pathverse mobile app. We used ChatGPT-3 to generate the content needed to be added to Pathverse. To gather feasibility data, we generated intervention content to support parents to help their child (8-12 years of age) to be physically active.

The content generated for this app was developed based on the Multi-Process Action Control (M-PAC) framework [41,42]. The M-PAC framework addresses the intention-behavior gap through the understanding that ongoing reflective processes (ie, affective attitude and perceived opportunity) and regulation processes (ie, behavioral and cognitive tactics to maintain intention focus) are necessary for the intention to become an action [41]. Specific to a JITAI, the M-PAC framework was selected as the framework for this intervention to dynamically and contextually address users' failed intentions to be physically active. Thus, the just-in-time intervention options can be tailored to the specific circumstances of the individual, aligning with either the reflective, regulatory, or reflexive process [41,42]. The M-PAC framework was additionally chosen as we have seen success with this framework and its associated BCTs (ie, action planning, repetition, and habit formation) in previous family-based PA programs [43]. To address these circumstances, our research team created decision tree algorithms to tailor the family lessons and challenges recommended throughout the weeks. The algorithms were designed using the M-PAC framework and take into consideration (1) child MVPA minutes, (2) parent support behavior, and (3) parent self-efficacy and motivation for supporting their child's PA (Figure 1). Based on the decision tree, weekly tailored lessons needed to be created to target each M-PAC construct. Topics included parental support, affective attitudes toward supporting their children's PA, capability, opportunity, self-monitoring of PA, and restructuring the environment for PA. These topics stemmed from previous research for family-based PA interventions using the M-PAC framework [43]. With these considerations, a variety of prompts were created based on these topics.

XSL•FO
RenderX

Figure 1. Names of modules in the decision tree algorithm for personalized lessons.



There are various components to consider when generating a prompt for ChatGPT. Specific to academic uses of ChatGPT, the elements to be included in a prompt include an instruction (ie, an overview of the output you would like to receive), context (ie, other background information to help tailor the output), input data (ie, additional specifications for the output that may include its strengths or limitations), and output indicator (ie, how you would like the output to be presented, including word count and paragraph format) [44]. When creating a prompt for this case study, we included the target behavior and the M-PAC framework, with each output to be delivered in bullet point form. Once the content was created, we then used the Pathverse mHealth no-code app design tool to develop the JITAI app [20,38-40]

Phase 2

We summarized lessons learned and future recommendations for using ChatGPT in the development of mHealth interventions. Our team identified common themes and patterns emerging from the process of creating the JITAI content using ChatGPT. We then compared our data with previous literature to develop recommendations for future use. This involved a literature search to identify relevant studies and lessons learned from using ChatGPT in mHealth interventions. The primary aim of the literature search was to gather a wide range of insights into the acceptability, including the application of ChatGPT and its effectiveness in this context and challenges associated with integrating ChatGPT into mHealth interventions to refine our recommendations.

Results

Phase 1: Exploring the Feasibility of Using ChatGPT to Create JITAI Content

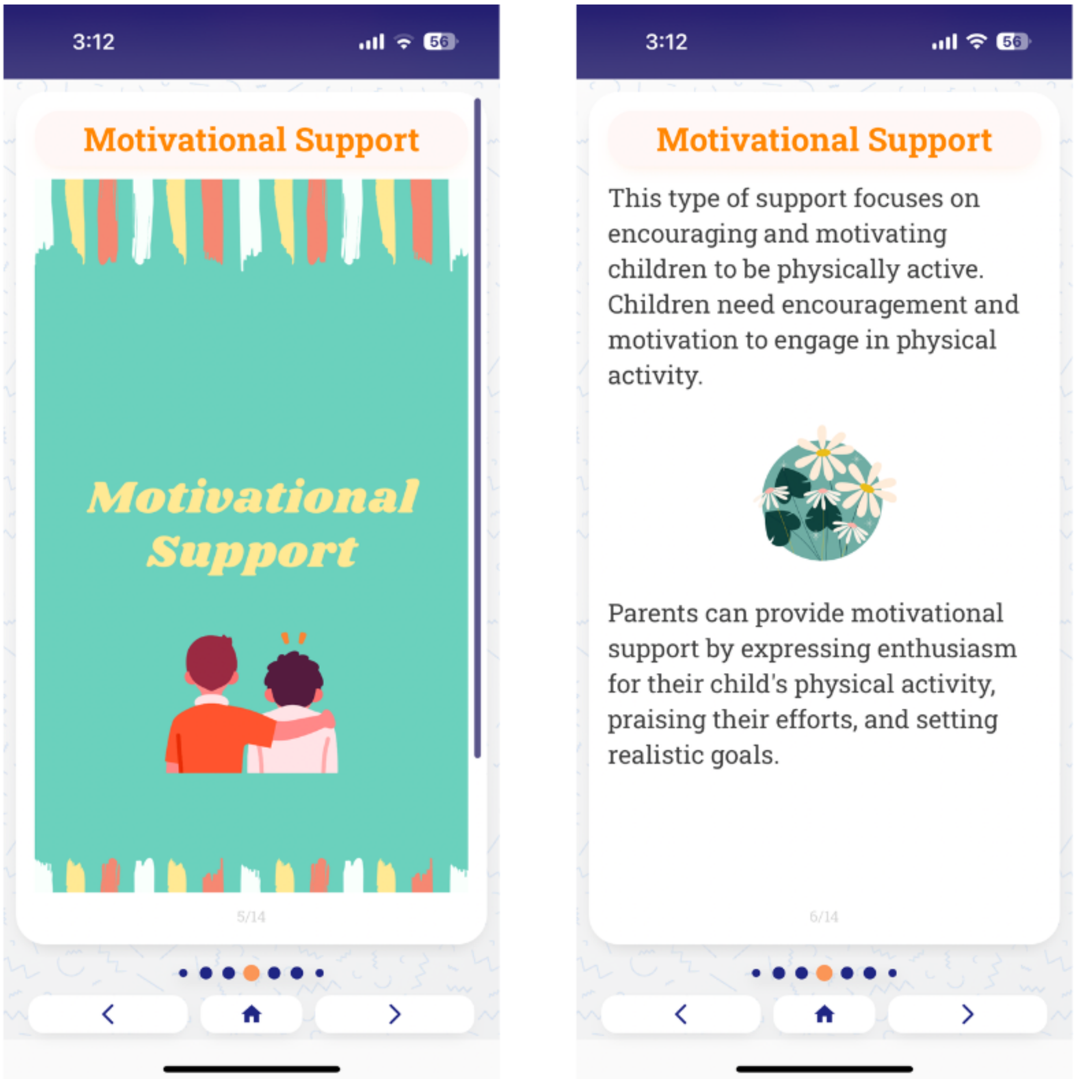
The results of phase 1 are first reported on how the researchers (AW and SL) used ChatGPT to generate content, followed by an analysis of the feasibility of the use of ChatGPT in this context. Overall, we created 13 lessons with the help of ChatGPT in phase 1. Figure 2 displays an example of how this content was displayed in the mobile app. We provided specific prompts about the length of the content generated, the target constructs of the M-PAC framework, the tone of the lesson, and the literacy levels needed. We used multiple question prompts to optimize text output. Table 1 provides examples of prompts used for different lessons. We started with broad prompts (eg, explain the various constructs in the M-PAC framework) and then used specific prompts based on the output (eg, provide specific fun examples to help parents improve opportunities to support child PA; Table 1). After the prompts were input into ChatGPT, the output was copied into a separate document for review by the researchers (AW and SL). If more or alternate content was needed, prompts such as "provide additional information about [this topic]" were used. To ensure that the output given by ChatGPT was relevant and accurate, we referred to previous literature and previous content examples following the M-PAC framework [21,45,46]. Once the content was deemed acceptable and accurate by the researchers, it was uploaded to the Pathverse platform. This step additionally involved creating graphics to include along with the text responses and formatting the content into different app "pages" with fewer than 400 characters per page of the mobile app.

XSL•FO
RenderX

We evaluated the acceptability of ChatGPT for creating mHealth content by reflecting on content accuracy, relevance, and tone. Both researchers found that ChatGPT demonstrated an acceptable level of accuracy and relevance and provided relevant responses to the prompts. However, on some occasions, ChatGPT provided false academic references. This is a serious issue that needs to be addressed to prevent misinformation. Thus, both authors reflected the need to place a filtering mechanism to ensure that the content generated was appropriate. Furthermore, some of the answers lacked specificity (eg, provide examples of PA programs in my area). This may be due to the fact that ChatGPT-3 was trained using data up to September 2021. Finally, we found the tone of ChatGPT responses to be acceptable for research purposes. The overall tone matched the prompt given (eg, write in a fun and positive voice). Overall, ChatGPT did not generate any inappropriate content. There is an evident need to provide clear prompts in order for ChatGPT to provide optimal responses. Additionally, multiple questions are often needed to optimize ChatGPT responses. The researchers additionally agreed that providing a role to ChatGPT, for example, telling the LLM that it is a health researcher delivering a family-based PA intervention, may have further refined the tone and quality of the response given.

When reflecting on the feasibility of implementing ChatGPT for this case study, we (AW and SL) found ChatGPT to be easy to use. Both researchers (AW and SL), with varying levels of technical expertise, found the user interface to be intuitive. The ease of use also allowed us to test various prompts to help optimize the ChatGPT responses. Overall, we found that minimal training or prior experience is needed to use this tool, and it has the potential to make it widely accessible for researchers.

Figure 2. Screenshots of physical activity content generated by ChatGPT in the mobile app Pathverse.



XSL•FO
RenderX

Table 1. Prompts and responses from a conversation with ChatGPT in generating content for a 10-week family-based physical activity mobile app.

Prompt entered	ChatGPT response
Write a fun to read article about the role of parental capabilities to support for child physical activity based on the M-PAC[a] framework	• Are you struggling to find ways to support your child's physical activity? Do you want to improve your parental capabilities to make physical activity a fun and regular part of your child's life? Look no further than the M-PAC framework, which provides a roadmap for parents to support their child's physical activity. The M-PAC framework stands for Multi-Process Action Control, and it emphasizes the importance of three key components for parents to successfully support their child's physical activity: instrumental support, regulatory support, and motivational support. Instrumental support refers to the tangible resources and skills that parents provide to support their child's physical activity. This could include things like providing transportation to and from activities, purchasing sports equipment, or enrolling your child in a physical activity program. Regulatory support focuses on creating an environment that supports physical activity. This could include setting rules and boundaries around screen time, creating a routine for physical activity, or providing healthy food options to fuel physical activity. Motivational support is all about fostering a positive attitude towards physical activity. This could include providing praise and encouragement, setting goals, and finding ways to make physical activity fun and enjoyable.
Provide specific fun examples to help parents improve their capabilities for supporting child physical activity in bullet points	• Certainly! Here are some specific fun examples to help parents improve their capabilities for supporting child physical activity: 1. Instrumental Support: • Have a weekly family game night that involves physical activity such as charades, Pictionary or dance party. • Organize a scavenger hunt with physical activity challenges for your child and their friends. • Plan a family hike or outdoor adventure at a nearby park or nature reserve. 2. Regulatory Support: • Start a vegetable garden together and teach your child the importance of physical activity and nutrition for a healthy lifestyle. • Use a fitness tracker or pedometer with your child to encourage them to walk more and set goals for daily steps. • Plan a family bike ride or roller skating adventure on a weekend afternoon. 3. Motivational Support: • Make a fun reward system for reaching physical activity goals, such as a family outing to a local amusement park or water park. • Watch sports games or videos of active events with your child and talk about the benefits of physical activity for overall health. • Encourage your child to invite friends over for a game of soccer, basketball or other team sport to promote socialization and physical activity. By incorporating these fun examples into your family's routine, you can improve your capabilities for supporting your child's physical activity and make healthy habits a fun and rewarding part of your family's lifestyle.

[a]M-PAC: Multi-Process Action Control.

Phase 2: Recommendations for the Future Use of ChatGPT

Overview

On the basis of our experience in developing the above content and based on previous research [21,25,47], we have compiled the following recommendations for using ChatGPT for similar purposes. First, a 6-step recommendation was proposed to help make the process of using ChatGPT more efficient for future research. These steps included defining the target behavior for the intervention, applying behavior change theory, creating the intervention structure, inputting this information into ChatGPT, refining the output, and customizing the content to be delivered to the target population. Second, we discussed ethical considerations for the use of ChatGPT in this research space. Lastly, we discussed transparency with using ChatGPT in academic research.

Step 1: Determine Target Behavior

The first step of these guidelines involves determining a target behavior or behaviors for the target population of your intervention. This could be based on previous research for certain demographic groups or based on a gap in the current literature. The selected target behavior defines the purpose of the intervention and the outcomes on which the intervention will be assessed [25]. This is considered an essential first step as it will guide the remaining steps of these recommendations.

Step 2: Ground in Behavior Change Theory

The second step recognizes the need to deliver digital health content grounded in behavior change theory. Based on previous literature and considering the target behavior selected in step 1, it is advised to select a health behavior theory to guide the intervention. Thus, constructs of the behavior change theory must be considered when searching for and developing digital health intervention content. Further, other elements of the intervention, such as BCTs, to strengthen the behavior change theory [26] should be considered during this step.

XSL•FO
RenderX

Step 3: Design Intervention Structure

Step 3 involves designing the intervention structure. In this step, the length of the intervention and the length and amount of content to be delivered should be considered first. After this information has been determined, it is recommended to consider the medium of delivery of the digital health intervention content. Previous research has shown varying success for both web-based interventions and mobile-based interventions [48,49]. Additionally, there are important considerations for best practices with delivering content through these different mediums, which are explored later in this development process.

It is important to note that this step may involve an additional agenda. Examining previous literature, using participatory action research or co-design principles, or other methods may be necessary to ensure that you are gathering content that will be both engaging to the participants and promoting adherence to the target behavior.

Step 4: Input Intervention Structure and Behavior Change Constructs Into ChatGPT

The next step is to input the information gathered from steps 1 to 3 and create varying prompts into ChatGPT. If this is your first time logging into ChatGPT through OpenAI, you will need to create a free account. Once your account has been created, you may type your prompt into the text box at the bottom of the screen. Determining an optimal prompt to input includes considering the target behavior, the proposed structure of the intervention, the behavior change theory and its constructs, and BCTs. Further, it is important to consider the rules in which ChatGPT delivers its output, for instance, whether you would prefer the response to be in paragraph form or bullet points. This step is iterative as you receive responses and continue to modify your prompt until you receive the desired output. Additionally, it has been previously recommended to consider assigning a role and tone for ChatGPT to embody in its response or to provide a similar example, when available [50].

Step 5: Revise the Output of ChatGPT

This step involves revising the response received from ChatGPT. There is a possibility that the language model has created errors or has provided incorrect references with their output. We compared the results with previous literature and revise and adapted as necessary to ensure that the most accurate information is being provided. Including information from the previous literature in the next prompt may continue to provide more refined ChatGPT responses.

Step 6: Customize the Content to be Delivered

The final step of this framework is to customize the content to meet the needs of your intervention. This involves considering the layout and design of how you will deliver the content on your selected medium from step 3, as well as any images or graphics used to supplement the given content. This step may involve working with an additional team to develop a web-based or mobile-based platform to support the health behavior change intervention. Further, user experience and design should be considered to improve usability and satisfaction of the content [51-54]. Table 2 summarizes the steps of these guidelines and considerations to meet the needs of each step.

By following these guidelines and using ChatGPT to assist in the rapid creation of digital health content, many ethical considerations arise. The first consideration, as highlighted above, is ensuring that the responses from ChatGPT are accurate and validated to be used as health information in a research study. This can be done by referencing previous literature or creating a panel of experts in the field to review the output created by ChatGPT. Further, it is vital to ensure that users engaging with AI-generated content through ChatGPT or other LLMs are adequately informed about its limitations, decision-making capabilities, and the crucial nature of their involvement. Transparent communication and obtaining informed consent are pivotal to respect user autonomy and comprehension. Although ChatGPT demonstrates remarkable efficiency in generating responses to prompts, evaluating its applicability within the intervention's context remains crucial to ensure substantial value to using ChatGPT.

As ChatGPT inevitably continues to support academic research across disciplines, it is also important to consider how ChatGPT is being cited by those who use it. There has been a variety of techniques used so far, with some authors including ChatGPT as an author [55] and others acknowledging the use of ChatGPT [34] to assist with their manuscript.

XSL•FO
RenderX

Table 2. Proposed recommendations for developing digital health content using ChatGPT and a summary of considerations for using this tool.

Step	Task	Consideration
1	Determine target behavior	• Previous research • Needs of the target population
2	Ground in behavior change theory	• Stage of readiness of participants • Needs of the population group
3	Design the intervention structure	• Web or mobile based • Length of the intervention • Amount of content to be delivered in each bout (ie, how many words, characters, or pages of content to be delivered) • Use co-design or other frameworks to ensure that the intervention aligns with the needs of the target population
4	Input intervention structure and behavior change constructs into ChatGPT	• Structure prompt to input into ChatGPT (considering instruction, context, input data, and output indicator) • Iteratively adapt prompts based on desired output • Order in which relevant information relating to each construct is delivered, if not predefined by the literature
5	Revise the output of ChatGPT	• Refer outputs to previous literature to ensure accuracy • Confirm whether references used by ChatGPT are accurate
6	Customize the content to be delivered	• Layout and design of content • Images or graphics to supplement text output • User experience and design of the intervention platform

Discussion

Principal Findings

The primary objective of this study was to explore the feasibility of using ChatGPT to develop content for a mobile-based JITAI to promote parental support for their children's PA. The secondary objective was to propose recommendations for using ChatGPT for future work in this area. To our knowledge, the process of using ChatGPT to develop health intervention content has not yet been documented, so we considered the key components required to develop effective behavior change interventions. We found that using ChatGPT was overall acceptable for this case study. However, a human check by researchers in the field is imperative to ensure the relevance and accuracy of the output provided. The use of ChatGPT and similar LLMs is rapidly evolving, and as such, these proposed recommendations are highly dynamic to the developing nature of these technologies.

This study has several implications for researchers using ChatGPT when developing mHealth app content. First, ChatGPT can help researchers improve the efficiency of creating digital health content for various tailored lessons. Previously, it was determined that ChatGPT can expedite the research process by allowing researchers to focus on steps of the research design process that require more human input, for example, focusing on the experimental design [56,57]. The improvement and versatility of text generation, knowledge translation, and literature review have been documented in various studies that have used ChatGPT in health care education [58]. As seen in this study, ChatGPT can help create various versions of content (varying in writing styles and tones) using a series of different prompts. Further, coupled with the efficiency of developing intervention content, this study has highlighted the ability to efficiently create a variety of tailored content specific to PA messaging. The need for more variety and content options has been previously stated as a limitation in previous studies that did not use ChatGPT for the creation of content [28]. Overall, this study highlights one use case that benefited from the use of ChatGPT to rapidly create digital health content. As ChatGPT is in its infancy, we expect it to evolve quickly [58].

Second, this study highlights the current limitations of using ChatGPT for creating mHealth behavior interventions. Although ChatGPT has great potential to improve the efficiency with which digital health content creation can occur, it is not possible to replicate responses by ChatGPT while using the same prompt [58,59]. This unpredictability poses a significant challenge for health researchers and developers who may require stable and reliable outputs [58]. Because of the probabilistic nature of ChatGPT and similar LLMs, the responses generated from ChatGPT are generated based on a probability distribution, meaning the same response will not be generated [60]. Further, a significant concern is the generation of references by ChatGPT that do not exist or are inaccurate. This lack of interpretability hampers the transparency of mHealth content development, making it difficult for researchers to have a clear understanding of the AI's decision-making process. Other limitations have been recognized by previous work around ChatGPT. These include limited accuracy, bias and limitations of data, lack of context, and the potential of limited engagement with the content [34]. To mitigate these challenges, we highly recommend a rigorous human fact-checking process, as indicated in our recommendations for mHealth intervention content development

XSL•FO
RenderX

using ChatGPT, and fine-tuning specific prompts to ensure that the information given by ChatGPT is relevant.

Finally, the integration of ChatGPT with existing mHealth app development tools, such as Pathverse, holds the potential to significantly enhance the efficiency and effectiveness of developing and evaluating JITAI apps. By incorporating ChatGPT's language generation capabilities into Pathverse, developers can expedite the creation of content-rich JITAIs. Additionally, reinforcement learning algorithms can play a crucial role in JITAIs by dynamically adapting the intervention based on real-time data and user feedback [61]. Developers can leverage ChatGPT's language generation capabilities using its application programming interface to assist with content creation [61]. With the integration of ChatGPT, these algorithms can benefit from the AI-generated content to offer more tailored and contextually relevant interventions. By combining the strengths of reinforcement learning and ChatGPT, JITAI apps can become more adaptive and responsive to individual user's needs, thereby increasing their effectiveness in promoting behavior change and improving health outcomes.

There are several limitations to this study. First, we used ChatGPT to create content for only 1 JITAI, potentially restricting the generalizability of the study findings. Second, because of ChatGPT's tendency to provide different responses for the same prompt, it was challenging to accurately characterize the content's reproducibility and consistency. Lastly, as ChatGPT is rapidly evolving, the use case described in this study may have limited applicability a few years from now. We also want to add that although ChatGPT-3 is currently free to use, it is likely that as it improves, it is likely to come with an associated cost.

Conclusions

By using ChatGPT, we were able to expedite the process of creating 13 lessons that were guided by the M-PAC framework, thus highlighting the incredible opportunity ChatGPT presents to rapidly create content for various mHealth JITAIs. Although we found that ChatGPT was acceptable for this case study, we still encourage the cautious use of ChatGPT and other LLMs in similar contexts. The use of ChatGPT expedited the process of content development to 2 months, the bulk of which was spent on reviewing the content by experts in the field before delivering to population groups. This process was imperative to ensure that accurate and relevant content was being created to be delivered. The results from this study found implications in 3 areas. The first is efficiency in generating a variety of content based on different prompts. Second, this study highlighted the potential limitations of ChatGPT, including the inability to replicate responses from the same prompts and the need for human input to ensure that the output from ChatGPT is accurate. Finally, this case study has highlighted the efficiency of using no-code app builders, such as Pathverse, to disseminate information generated by ChatGPT. It is without a doubt that as ChatGPT and other LLMs continue to improve in sophistication and accuracy, they will continue to integrate into intervention design and other various contexts for researchers. Further research and applications of ChatGPT and the guidelines proposed in this study are imminent in this field as we continue to understand ChatGPT.

Acknowledgments

The authors acknowledge that ChatGPT was used to generate results for this study. For a summary of the ChatGPT conversations, see Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ChatGPT Transcript.
[PDF File (Adobe PDF File), 172 KB - mededu_v10i1e51426_app1.pdf]

References

1. Anderson E, Durstine JL. Physical activity, exercise, and chronic diseases: a brief review. Sports Med Health Sci 2019;1(1):3-10 [FREE Full text] [doi: 10.1016/j.smhs.2019.08.006] [Medline: 35782456]
2. Global status report on physical activity 2022. World Health Organization. Geneva; 2022. URL: https://www.who.int/teams/health-promotion/physical-activity/global-status-report-on-physical-activity-2022 [accessed 2024-01-27]
3. Bryan SN, Katzmarzyk PT. The association between meeting physical activity guidelines and chronic diseases among Canadian adults. J Phys Act Health 2011;8(1):10-17. [doi: 10.1123/jpah.8.1.10] [Medline: 21297180]
4. Rhodes RE, Perdew M, Malli S. Correlates of parental support of child and youth physical activity: a systematic review. Int J Behav Med 2020;27(6):636-646. [doi: 10.1007/s12529-020-09909-1] [Medline: 32529629]
5. Lithopoulos A, Liu S, Rhodes RE, Naylor PJ. The role of identity in parental support for physical activity and healthy eating among overweight and obese children. Health Psychol Behav Med 2020;8(1):185-201 [FREE Full text] [doi: 10.1080/21642850.2020.1750959] [Medline: 34040867]
6. Tremblay MS, Warburton DER, Janssen I, Paterson DH, Latimer AE, Rhodes RE, et al. New Canadian physical activity guidelines. Appl Physiol Nutr Metab 2011;36(1):36-46; 47-58 [FREE Full text] [doi: 10.1139/H11-009] [Medline: 21326376]

XSL•FO
RenderX

7. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. JAMA 2018;320(19):2020-2028 [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
8. Physical activity guidelines: UK chief medical officers' report. Department of Health and Social Care. 2019. URL: https://www.gov.uk/government/publications/physical-activity-guidelines-uk-chief-medical-officers-report [accessed 2024-01-27]
9. Merlo CL, Jones SE, Michael SL, Chen TJ, Sliwa SA, Lee SH, et al. Dietary and physical activity behaviors among high school students—youth risk behavior survey, United States, 2019. MMWR Suppl 2020;69(1):64-76 [FREE Full text] [doi: 10.15585/mmwr.su6901a8] [Medline: 32817612]
10. Liu S, Weismiller J, Strange K, Forster-Coull L, Bradbury J, Warshawski T, et al. Evaluation of the scale-up and implementation of mind, exercise, nutrition … do it! (MEND) in British Columbia: a hybrid trial type 3 evaluation. BMC Pediatr 2020;20(1):392 [FREE Full text] [doi: 10.1186/s12887-020-02297-1] [Medline: 32819325]
11. Perdew M, Liu S, Rhodes R, Ball GDC, Mâsse LC, Hartrick T, et al. The effectiveness of a blended in-person and online family-based childhood obesity management program. Child Obes 2021;17(1):58-67. [doi: 10.1089/chi.2020.0236] [Medline: 33370164]
12. Liu S, Marques IG, Perdew MA, Strange K, Hartrick T, Weismiller J, et al. Family-based, healthy living intervention for children with overweight and obesity and their families: a 'real world' trial protocol using a randomised wait list control design. BMJ Open 2019;9(10):e027183 [FREE Full text] [doi: 10.1136/bmjopen-2018-027183] [Medline: 31676642]
13. Smith N, Liu S. A systematic review of the dose-response relationship between usage and outcomes of online physical activity weight-loss interventions. Internet Interv 2020;22:100344 [FREE Full text] [doi: 10.1016/j.invent.2020.100344] [Medline: 32995302]
14. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
15. Davis A, Sweigart R, Ellis R. A systematic review of tailored mHealth interventions for physical activity promotion among adults. Transl Behav Med 2020;10(5):1221-1232. [doi: 10.1093/tbm/ibz190] [Medline: 33044542]
16. Wang L, Miller LC. Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. Health Commun 2020;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
17. Bond DS, Thomas JG, Raynor HA, Moon J, Sieling J, Trautvetter J, et al. B-MOBILE—a smartphone-based intervention to reduce sedentary time in overweight/obese individuals: a within-subjects experimental trial. PLoS One 2014;9(6):e100821 [FREE Full text] [doi: 10.1371/journal.pone.0100821] [Medline: 24964010]
18. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. Med Sci Sports Exerc 2010;42(11):2134-2140 [FREE Full text] [doi: 10.1249/MSS.0b013e3181e0b3ff] [Medline: 20386334]
19. Hermens H, op den Akker H, Tabak M, Wijsman J, Vollenbroek M. Personalized coaching systems to support healthy behavior in people with chronic conditions. J Electromyogr Kinesiol 2014;24(6):815-826. [doi: 10.1016/j.jelekin.2014.10.003] [Medline: 25455254]
20. Liu S, La H, Willms A, Rhodes RE. A "No-Code" app design platform for mobile health research: development and usability study. JMIR Form Res 2022;6(8):e38737 [FREE Full text] [doi: 10.2196/38737] [Medline: 35980740]
21. Willms A, Rhodes RE, Liu S. The development of a hypertension prevention and financial-incentive mHealth program using a "no-code" mobile app builder: development and usability study. JMIR Form Res 2023;7:e43823 [FREE Full text] [doi: 10.2196/43823] [Medline: 37018038]
22. Hingle M, Nichter M, Medeiros M, Grace S. Texting for health: the use of participatory methods to develop healthy lifestyle messages for teens. J Nutr Educ Behav 2013;45(1):12-19 [FREE Full text] [doi: 10.1016/j.jneb.2012.05.001] [Medline: 23103255]
23. Berg M, Adolfsson A, Ranerup A, Sparud-Lundin C, University of Gothenburg Centre for Person-Centred Care. Person-centered web support to women with type 1 diabetes in pregnancy and early motherhood—the development process. Diabetes Technol Ther 2013;15(1):20-25. [doi: 10.1089/dia.2012.0217] [Medline: 23297670]
24. Hanson E, Magnusson L, Arvidsson H, Claesson A, Keady J, Nolan M. Working together with persons with early stage dementia and their family members to design a user-friendly technology-based support service. Dementia 2016;6(3):411-434. [doi: 10.1177/1471301207081572]
25. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
26. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
27. Coughlin LN, Nahum-Shani I, Philyaw-Kotov ML, Bonar EE, Rabbi M, Klasnja P, et al. Developing an adaptive mobile intervention to address risky substance use among adolescents and emerging adults: usability study. JMIR Mhealth Uhealth 2021;9(1):e24424 [FREE Full text] [doi: 10.2196/24424] [Medline: 33448931]

XSL•FO
RenderX

28. Mair JL, Hayes LD, Campbell AK, Buchan DS, Easton C, Sculthorpe N. A personalized smartphone-delivered Just-in-time adaptive intervention (JitaBug) to increase physical activity in older adults: mixed methods feasibility study. JMIR Form Res 2022;6(4):e34662 [FREE Full text] [doi: 10.2196/34662] [Medline: 35389348]
29. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-02-20]
30. Pocock K. What is ChatGPT? why you need to care about GPT-4. PC Guide. 2024. URL: https://www.pcguide.com/apps/what-is-chat-gpt/ [accessed 2024-01-27]
31. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: jack of all trades, master of none. Inf Fusion 2023;99:101861 [FREE Full text] [doi: 10.1016/j.inffus.2023.101861]
32. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901 [FREE Full text]
33. Jiao W, Wang W, Huang J, Wang X, Shi S, Tu Z. Is ChatGPT a good translator? yes with GPT-4 as the engine. ArXiv Preprint posted online on November 2, 2023. [doi: 10.48550/arXiv.2301.08745]
34. Biswas SS. Role of ChatGPT in public health. Ann Biomed Eng 2023;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
35. Arslan S. Exploring the potential of ChatGPT in personalized obesity treatment. Ann Biomed Eng 2023;51(9):1887-1888. [doi: 10.1007/s10439-023-03227-9] [Medline: 37145177]
36. Liu Y, Deng G, Xu Z, Li Y, Zheng Y, Zhang Y, et al. Jailbreaking ChatGPT via prompt engineering: an empirical study. ArXiv Preprint posted online on May 23, 2023. [FREE Full text]
37. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023;55(9):1-35 [FREE Full text] [doi: 10.1145/3560815]
38. Muncey T. Doing autoethnography. Int J Qual Methods 2016;4(1):69-86 [FREE Full text] [doi: 10.1177/160940690500400105]
39. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
40. Pathverse. 2021. URL: https://pathverse.ca/en/ [accessed 2024-01-27]
41. Rhodes RE. Chapter five—the evolving understanding of physical activity behavior: a multi-process action control approach. Adv Motiv Sci 2017;4:171-205. [doi: 10.1016/bs.adms.2016.11.001]
42. Rhodes RE, Yao CA. Models accounting for intention-behavior discordance in the physical activity domain: a user's guide, content overview, and review of current evidence. Int J Behav Nutr Phys Act 2015;12:9 [FREE Full text] [doi: 10.1186/s12966-015-0168-6] [Medline: 25890238]
43. Nuss K, Coulter R, DeSilva B, Buenafe J, Sheikhi R, Naylor PJ, et al. Evaluating the effectiveness of a family-based virtual childhood obesity management program delivered during the COVID-19 pandemic in Canada: prospective study. JMIR Pediatr Parent 2022;5(4):e40431 [FREE Full text] [doi: 10.2196/40431] [Medline: 36054663]
44. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 2023;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
45. Hartson KR, Della LJ, King KM, Liu S, Newquist PN, Rhodes RE. Application of the IDEAS framework in adapting a web-based physical activity intervention for young adult college students. Healthcare (Basel) 2022;10(4):700 [FREE Full text] [doi: 10.3390/healthcare10040700] [Medline: 35455877]
46. Liu S, Husband C, La H, Juba M, Loucks R, Harrison A, et al. Development of a self-guided web-based intervention to promote physical activity using the multi-process action control framework. Internet Interv 2019;15:35-42 [FREE Full text] [doi: 10.1016/j.invent.2018.11.003] [Medline: 30568879]
47. Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol 2015;34(10):971-982 [FREE Full text] [doi: 10.1037/hea0000161] [Medline: 25642841]
48. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021;9(1):e21563 [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
49. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
50. Cook J. How to write effective prompts for ChatGPT: 7 essential steps for best results. Forbes. 2023. URL: http://tinyurl.com/24zsaaph [accessed 2023-07-30]
51. Saparamadu AADNS, Fernando P, Zeng P, Teo H, Goh A, Lee JMY, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021;9(3):e18079 [FREE Full text] [doi: 10.2196/18079] [Medline: 33769297]
52. Walden A, Garvin L, Smerek M, Johnson C. User-centered design principles in the development of clinical research tools. Clin Trials 2020;17(6):703-711. [doi: 10.1177/1740774520946314] [Medline: 32815381]

XSL•FO
RenderX

53. Schnall R, Rojas M, Bakken S, Brown W, Carballo-Dieguez A, Carry M, et al. A user-centered model for designing consumer mobile health (mHealth) applications (apps). J Biomed Inform 2016;60:243-251 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.002] [Medline: 26903153]
54. Hentati A, Forsell E, Ljótsson B, Kaldo V, Lindefors N, Kraepelien M. The effect of user interface on treatment engagement in a self-guided digital problem-solving intervention: a randomized controlled trial. Internet Interv 2021;26:100448 [FREE Full text] [doi: 10.1016/j.invent.2021.100448] [Medline: 34471610]
55. King MR, ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cell Mol Bioeng 2023;16(1):1-2 [FREE Full text] [doi: 10.1007/s12195-022-00754-8] [Medline: 36660590]
56. No authors listed. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023;613(7945):612 [FREE Full text] [doi: 10.1038/d41586-023-00191-1] [Medline: 36694020]
57. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]
58. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
59. Holzinger A, Keiblinger K, Holub P, Zatloukal K, Müller H. AI for life: trends in artificial intelligence for biotechnology. N Biotechnol 2023;74:16-24 [FREE Full text] [doi: 10.1016/j.nbt.2023.02.001] [Medline: 36754147]
60. Why ChatGPT and other LLMs generate different answers to same questions. Ayfie. 2023. URL: https://blog.ayfie.com/why-chatgpt-and-other-llms-generate-different-answers-to-same-questions [accessed 2023-12-14]
61. Gönül S, Namlı T, Coşar A, Toroslu İ. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. Artif Intell Med 2021;115:102062. [doi: 10.1016/j.artmed.2021.102062] [Medline: 34001322]

Abbreviations

AI: artificial intelligence
BCT: behavior change technique
JITAI: just-in-time adaptive intervention
LLM: large language model
mHealth: mobile health
M-PAC: Multi-Process Action Control
MVPA: moderate-to-vigorous physical activity
PA: physical activity

Edited by K Venkatesh; submitted 31.07.23; peer-reviewed by S Biswas, J Mair; comments to author 08.11.23; revised version received 15.12.23; accepted 27.12.23; published 29.02.24.

Please cite as:
Willms A, Liu S
Exploring the Feasibility of Using ChatGPT to Create Just-in-Time Adaptive Physical Activity mHealth Intervention Content: Case Study
JMIR Med Educ 2024;10:e51426
URL: https://mededu.jmir.org/2024/1/e51426
doi:10.2196/51426
PMID:38421689

©Amanda Willms, Sam Liu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Incorporating ChatGPT in Medical Informatics Education: Mixed Methods Study on Student Perceptions and Experiential Integration Proposals

Sabrina Magalhães Araujo[1*], RN, MSc; Ricardo Cruz-Correia[1,2,3*], PhD

[1]Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Porto, Portugal

[2]Department of Community Medicine, Information and Decision Sciences, Faculty of Medicine, University of Porto, Porto, Portugal

[3]Working Group Education, European Federation for Medical Informatics, Le Mont-sur-Lausanne, Switzerland

[*]all authors contributed equally

Corresponding Author:
Sabrina Magalhães Araujo, RN, MSc
Center for Health Technology and Services Research
Faculty of Medicine
University of Porto
Rua Dr Plácido da Costa, s/n
Porto, 4200-450
Portugal
Phone: 351 220 426 91 ext 26911
Email: saraujo@med.up.pt

Abstract

Background: The integration of artificial intelligence (AI) technologies, such as ChatGPT, in the educational landscape has the potential to enhance the learning experience of medical informatics students and prepare them for using AI in professional settings. The incorporation of AI in classes aims to develop critical thinking by encouraging students to interact with ChatGPT and critically analyze the responses generated by the chatbot. This approach also helps students develop important skills in the field of biomedical and health informatics to enhance their interaction with AI tools.

Objective: The aim of the study is to explore the perceptions of students regarding the use of ChatGPT as a learning tool in their educational context and provide professors with examples of prompts for incorporating ChatGPT into their teaching and learning activities, thereby enhancing the educational experience for students in medical informatics courses.

Methods: This study used a mixed methods approach to gain insights from students regarding the use of ChatGPT in education. To accomplish this, a structured questionnaire was applied to evaluate students' familiarity with ChatGPT, gauge their perceptions of its use, and understand their attitudes toward its use in academic and learning tasks. Learning outcomes of 2 courses were analyzed to propose ChatGPT's incorporation in master's programs in medicine and medical informatics.

Results: The majority of students expressed satisfaction with the use of ChatGPT in education, finding it beneficial for various purposes, including generating academic content, brainstorming ideas, and rewriting text. While some participants raised concerns about potential biases and the need for informed use, the overall perception was positive. Additionally, the study proposed integrating ChatGPT into 2 specific courses in the master's programs in medicine and medical informatics. The incorporation of ChatGPT was envisioned to enhance student learning experiences and assist in project planning, programming code generation, examination preparation, workflow exploration, and technical interview preparation, thus advancing medical informatics education. In medical teaching, it will be used as an assistant for simplifying the explanation of concepts and solving complex problems, as well as for generating clinical narratives and patient simulators.

Conclusions: The study's valuable insights into medical faculty students' perspectives and integration proposals for ChatGPT serve as an informative guide for professors aiming to enhance medical informatics education. The research delves into the potential of ChatGPT, emphasizes the necessity of collaboration in academic environments, identifies subject areas with discernible benefits, and underscores its transformative role in fostering innovative and engaging learning experiences. The envisaged proposals hold promise in empowering future health care professionals to work in the rapidly evolving era of digital health care.

(JMIR Med Educ 2024;10:e51151) doi:10.2196/51151

XSL•FO
RenderX

KEYWORDS

education; medical informatics; artificial intelligence; AI; generative language model; ChatGPT

Introduction

Generative pre-trained transformers have evolved into potent language models with diverse education applications, including personalized and problem-based learning that emphasizes critical thinking [1]. They offer a chat interface for natural interactions, which can be valuable in engaging students in educational discussions. Additionally, these models can be adjusted to align with specific educational objectives and generate text embeddings, enabling tasks such as classification, recommendations, and similarity analysis. Furthermore, their accessibility through application programming interfaces (APIs) opens the door to integrate them into various educational applications beyond chat interfaces.

The integration of these artificial intelligence (AI) technologies, including ChatGPT, into the educational environment has the potential to improve the student learning experience [1-6]. By incorporating ChatGPT into the teaching and learning process in higher education, students can be supported throughout their educational journey and develop the necessary skills to effectively use AI in professional settings [4,7,8].

Integrating ChatGPT into the field of medical informatics education could not only enrich the learning experience but also empower students to apply AI skills, preparing them to tackle the complex challenges they will encounter in their future careers in health care [9]. For instance, professionals in medical informatics can use AI in developing decision support systems for diagnosing medical images and predicting patient outcomes based on data analysis. These applications demonstrate how a strong foundation in AI, facilitated by ChatGPT, can enhance the capabilities of future medical informatics professionals in delivering quality health care services.

Medical informatics, commonly referred to as biomedical or health informatics, is an umbrella term [10,11] that encompasses the use of information and communication technologies in health care. It is a fundamental field of study that caters to a diverse range of disciplines [12]. The field concentrates on using biomedical data, information, and knowledge for scientific inquiry, problem-solving, and decision-making endeavors that aim to advance care quality and delivery [13]. There is a growing interest in biomedical and health informatics (BMHI) education [14] due to the increasing demand for professionals who can address BMHI issues through the development, implementation, and evaluation of innovative technological solutions [15].

The educational requirements for BMHI vary depending on the level of education and career progression, with different pedagogical approaches needed to provide theoretical knowledge, practical skills, and a mature attitude [16]. Although medical informatics knowledge is globally applicable and requires international standards, education in this field is typically localized, with competencies being tailored to the specific environment in which they will be used [17]. Variations in educational and health care systems result in differences in BMHI education across countries. Nevertheless, despite this variability, fundamental similarities can be identified and used as a framework for recommendations [16,18].

The International Medical Informatics Association (IMIA), through its educational recommendations, has outlined essential knowledge domains for teaching in the field of BMHI, including the domain of computer science, data, and information [19]. Within this domain, there is a particular emphasis on imparting students with a deep understanding of the fundamental principles underlying emerging technologies, such as AI [19]. Notably, the most recent IMIA recommendation signifies the first explicit inclusion of AI as a topic within a BMHI knowledge domain. These IMIA recommendations offer a valuable framework for the development of educational programs, enabling the integration of essential competencies in medical informatics into the curricula of undergraduate medicine programs and master programs in the field, for instance. However, there is currently a dearth of specific recommendations regarding the inclusion of AI skills in the curriculum [9].

Since its launch, ChatGPT has ignited discussions surrounding its application in education. Rather than outright banning its use in universities, this presents an opportune moment to reassess teaching methodologies and examination practices in higher education, with the goal of preparing students for the digital world [2,3,5,7,20,21]. ChatGPT represents the initial step of a broader trend, and we must adapt to collaborate with it instead of opposing its presence [22]. In the education domain, ChatGPT will be able to offer interactive and personalized learning experiences, accessible on various devices. It can speed up routine tasks like assessments, allowing professors more time for personalized teaching. Furthermore, it can generate diverse educational content, ensure round-the-clock availability, assist in language learning, and promote innovative teaching methods.

It is crucial for faculty, professors, and students to be cognizant of both the potential and limitations of ChatGPT while also addressing ethical concerns [1,4,7,23,24] and ensuring accessibility in its implementation within academic settings. Encouraging the integration of ChatGPT in education requires the formulation of policies that promote best practices, nurturing students' critical thinking and equipping them with the necessary skills to effectively use AI tools [4,7,25]. To foster the development of critical thinking, engaging students in activities that prompt them to verify the accuracy, veracity, and potential biases of the text generated by ChatGPT is essential [2,26,27].

Building upon the aforementioned discussions and considerations, this study aims to contribute meaningfully to the broader objective of integrating AI education within the field of medical informatics. Recognizing the significant relevance of ChatGPT as an AI tool to the medical informatics courses offered in the master's programs in medicine and medical informatics at the Faculty of Medicine of the University of Porto (FMUP) in Portugal, this research seeks to address a proposal for integration of ChatGPT in the educational process.

XSL•FO
RenderX

The first objective is to compile the opinions of students enrolled in all levels of the medical faculty's programs with the aim of obtaining a general perception of their perspectives and experiences regarding the use of ChatGPT as an educational tool. Furthermore, this study endeavors to provide practical proposals for professors, offering examples of prompts for incorporating ChatGPT into their teaching activities, in order to enhance the educational benefits for students in medical informatics courses.

Methods

Ethical Considerations

A structured questionnaire was designed and submitted to the ethics committee of Faculty of Medicine of the University of Porto (105/CEFMUP/2023) to ensure ethical considerations in conducting this research and obtain approval for data collection. Students who participated in the questionnaire were explicitly informed that their participation in the research was entirely voluntary and were assured of confidentiality and anonymity regarding their responses.

Participants and Questionnaire

This study used a mixed methods approach involving students enrolled at all levels of the medical faculty's programs. The survey aimed to provide initial insight into medical informatics students' perspectives regarding the use of ChatGPT in teaching. The exploratory survey served as a preliminary assessment to outline proposals for incorporating the tool classes.

The questionnaire consisted of a total of 25 questions, comprising both closed-ended and open-ended formats, which were electronically distributed to 105 students enrolled in programs at the FMUP that offer medical informatics courses. The closed-ended questions aimed to assess the participants' familiarity with ChatGPT, their perception of the technology's use in educational contexts, and their attitudes toward using the application in academic and learning tasks. Participants were asked to indicate their level of agreement on a Likert scale, ranging from "strongly agree" to "strongly disagree," enabling nuanced responses.

In parallel, the open-ended questions encouraged participants to provide comprehensive and detailed feedback, sharing specific instances of their interactions and experiences with ChatGPT.

Data Analysis

The data collected from the questionnaire were analyzed using descriptive statistical techniques to summarize the quantitative responses. Thematic analysis was used to identify recurring themes and patterns in the qualitative responses, providing deeper insights into the students' perceptions and suggestions.

Literature Review and Description of Course Learning Outcomes

The methodology of this study also involved conducting a comprehensive literature review to explore the current publications pertaining to the implementation of AI in higher education settings. Specifically, the focus was on examining the integration of ChatGPT within the context of teaching medical informatics and assessing its alignment with international recommendations for effective pedagogy in this domain.

To assess the potential benefits of incorporating ChatGPT into educational practices, the study describes the learning outcomes of 2 proposed courses, carefully designed based on the competencies and skills expected of medical informatics students. The authors of this research, along with other esteemed members of the faculty, collaboratively deliberated on the proposals for using ChatGPT, aiming to optimize its functionalities both in master's programs in medicine and medical informatics.

In addition, it is expected to exemplify specific prompts to be used by students and professors to maximize the tool's potential to facilitate learning experiences. These prompts are carefully designed to engage students in critical thinking, problem-solving, and knowledge exploration while also aiding professors in delivering exemplary instruction.

Results

Questionnaire

In July 2023, the questionnaire was distributed to the students through their institutional email addresses. Out of the recipients, a noteworthy 25 university students actively participated by responding to the survey, resulting in a response rate of approximately 24%. The majority of respondents identified as male, accounting for 56% (n=14) of the total sample, with an average age of 35.2 (SD 8.6) years. Table 1 provides a concise summary of the key demographic characteristics of the participating students.

XSL•FO
RenderX

Table 1. Characteristics of the participants (N=25).

Characteristics	Values
Sex, n (%)	
Female	11 (44)
Male	14 (56)
Age (years)	
20-25, n (%)	2 (8)
26-30, n (%)	6 (24)
31-35, n (%)	8 (32)
>35, n (%)	9 (36)
Mean (SD)	35.2 (8.6)

Regarding the use of ChatGPT (Table 2), 52% (n=13) of the students indicated that they had their initial encounter with the chatbot during the second semester of 2022. Among them, a considerable proportion (n=4, 16%) reported using it on a daily basis, while the majority (several times during the week) found it to be a frequent resource. Impressively, 92% (n=23) of the students conveyed their satisfaction with the responses generated by ChatGPT, with 48% (n=12) expressing a high level of reliance on its answers and offering strong endorsements of its implementation among their peers. Nevertheless, a subset of participants (n=5, 20%) disclosed that they rarely place trust in the responses provided by the system. Furthermore, 96% (n=24) of the participants asserted that the tool comprehends the contextual intricacies of questions well. However, they noted that occasionally, to obtain the desired response, it becomes necessary to reformulate the query.

XSL•FO
RenderX

Table 2. Answers to questionnaire questions about the use of ChatGPT by medical faculty students (N=25).

Question and answers	Respondents, n (%)
Do you usually talk to your colleagues about ChatGPT?	
Ever	11 (44)
Occasionally	13 (52)
Often	1 (4)
When did you first use ChatGPT?	
Between March and April 2023	4 (16)
Between January and February 2023	7 (28)
Between July and December 2022	13 (52)
Between January and June 2022	0 (0)
In 2021	1 (4)
Do you use ChatGPT regularly?	
Yes every day	4 (16)
Yes, several times a week	13 (52)
I use it from time to time	8 (32)
How satisfied are you with ChatGPT's responses?	
Very satisfied	5 (20)
Satisfied	18 (72)
I have a neutral position on this	2 (8)
Do you trust the information provided by ChatGPT?	
Most of the time	12 (48)
Sometimes	8 (32)
Rarely	5 (20)
Does ChatGPT understand the context of your questions well?	
Very good	7 (28)
Good	17 (68)
No opinion	1 (4)
When using ChatGPT, do you have to rephrase questions to get the answers you want?	
Rarely	7 (28)
Sometimes	17 (68)
Often	1 (4)
Would you recommend ChatGPT to your colleagues?	
Definitely	18 (72)
Probably	5 (20)
I am not sure	1 (4)
Probably not	1 (4)

Concerning attitudes toward the use of ChatGPT for learning and academic purposes (Table 3), a majority of students demonstrated openness to adopting this form of chatbot and express intentions to incorporate it regularly into their educational endeavors. Nevertheless, it is noteworthy that 8% (n=2) of the participants held a dissenting perspective and are resolutely against its use in academic activities.

XSL•FO
RenderX

Table 3. Attitudes toward using ChatGPT for learning and academic tasks.

Statement and Likert scale	Respondents, n (%)
I think using a tool like ChatGPT would be a good idea to support learning	
Strongly agree	13 (52)
Agree	8 (32)
Neither agree nor disagree	4 (16)
I will start using ChatGPT to support learning and completing academic tasks	
Strongly agree	8 (32)
Agree	9 (36)
Neither agree nor disagree	7 (28)
Disagree	1 (4)
I will ask my colleagues about ChatGPT and how they use it	
Strongly agree	7 (28)
Agree	8 (32)
Neither agree nor disagree	10 (40)
I intend to create the habit of using ChatGPT to support learning and carry out my academic work	
Strongly agree	6 (24)
Agree	12 (48)
Neither agree nor disagree	5 (20)
Disagree	2 (8)
I will use ChatGPT or other similar chatbots whenever the opportunity arises	
Strongly agree	9 (36)
Agree	11 (44)
Neither agree nor disagree	3 (12)
Disagree	2 (8)
I have a bad feeling about ChatGPT and artificial intelligence in general	
Strongly agree	2 (8)
Agree	1 (4)
Neither agree nor disagree	8 (32)
Disagree	10 (40)
Strongly disagree	4 (16)
In my opinion, the use of ChatGPT or similar chatbots for academic tasks should not be allowed	
Strongly agree	1 (4)
Agree	1 (4)
Neither agree nor disagree	5 (20)
Disagree	6 (24)
Strongly disagree	12 (48)

Regarding the perceptions of ChatGPT as an academic support tool (Table 4), a significant proportion (n=18, 72%) of students concur with the notion that the ChatGPT tool has the potential to enhance and facilitate learning experiences. Furthermore, an overwhelming majority strongly agree that its implementation streamlines the execution of tasks, promoting efficiency within the academic context.

XSL•FO
RenderX

Table 4. Perceptions of ChatGPT as an academic support tool.

Statement and Likert scale	Respondents, n (%)
Using ChatGPT improves learning	
Strongly agree	12 (48)
Agree	6 (24)
Neither agree nor disagree	6 (24)
Disagree	1 (4)
Using the ChatGPT can make learning tasks easier to complete	
Strongly agree	12 (48)
Agree	10 (40)
Neither agree nor disagree	3 (12)
I find ChatGPT a very useful tool to support learning	
Strongly agree	15 (60)
Agree	7 (28)
Neither agree nor disagree	3 (12)
Using ChatGPT can increase my productivity as a student	
Strongly agree	16 (64)
Agree	6 (24)
Neither agree nor disagree	3 (12)
Using ChatGPT allows me to complete tasks faster	
Strongly agree	19 (76)
Agree	4 (16)
Neither agree nor disagree	2 (8)

In terms of the use of ChatGPT or other AI bots in the future, the majority of responses indicate that participants find it extremely useful, especially for medical writers who are not proficient in English, as it aids in restructuring and correcting texts. Some believe that the adoption of these tools will be inevitable and increasingly common, both in academic and professional contexts, resulting in enhanced process efficiency. However, there are also ethical concerns and apprehensions regarding the potential impact on the employability of programming professionals. Moreover, participants emphasize the importance of informed use, understand the limitations, and use these tools intelligently and as complementary to specific objectives. Some express caution, recognizing that although the tools have advantages, they do not create anything new but rather help organize thoughts and facilitate a better understanding of concepts.

The responses reveal diverse opinions on the benefits and drawbacks of using AI bots like ChatGPT in education. Some students highlight the capacity to customize responses to specific questions without worrying about boring the instructor and speeding up repetitive tasks, envisioning its potential to revolutionize the educational landscape. However, there are concerns about the long-term effects and the need for caution. The AI's biases and limitations are seen as potentially harmful to knowledge, and there are worries about the credibility of sources used and proper attribution of credits and bibliographic referencing.

While some view it as a valuable tutor or assistant that is always available, others caution against its potential to promote laziness, particularly in written work, which may discourage the development of quality writing skills. Nevertheless, the quick response time for academic tasks is regarded as an advantage by some, with no apparent disadvantages seen. The major concern is the risk of a bias in thinking, whereby AI-generated ideas or responses could influence one's own thought process, potentially inhibiting critical thinking. Despite this, many believe that ChatGPT can be a useful aid in executing certain tasks, as it does not create content but rather assists in the development and enhancement of ideas. The list of suggestions from medical informatics students on how to use ChatGPT or other AI bots in education can be seen in Figure 1.

XSL•FO
RenderX

Figure 1. List of suggestions from medical informatics course students on how to use ChatGPT in education generated with the assistance of artificial intelligence (AI).

Integration of ChatGPT

Overview

The analysis of student responses to the questionnaire has revealed a positive receptiveness and interest in the use of ChatGPT as an educational tool. Building on these findings, the next phase of our investigation focused on the proposed integration of ChatGPT into classroom settings. This section explores the proposal to incorporate ChatGPT into 2 specific courses, outlining how prompts developed by professors can be applied to enhance medical informatics students' learning experiences. Additionally, we will address the events held at the faculty to discuss and guide the implementation of ChatGPT in the context of education, emphasizing the collaboration among professors to foster educational innovation.

Master's Program in Medicine

The master's program in medicine at FMUP comprises an integrated cycle of studies totaling 360 credits, in accordance with the European Credit Transfer and Accumulation System (ECTS).

The course chosen for this study's proposal of implementing ChatGPT in the master's program in medicine is "DECIDES III: Decision, Data and Digital Health" (4 ECTS), which has been taught in the fourth year. By the end of this course unit, medical students are expected (1) to have knowledge and be able to discuss key topics related to health information systems and the integration of scientific evidence in health decision-making; (2) to proficiently and safely use health information systems; (3) to critically evaluate health scientific literature, particularly regarding health information systems, health technology assessment, and health decision analysis; and (4) to plan and interpret studies on health economic evaluation and decision analysis.

The course offered in the master's program in medicine proposes the use of ChatGPT in 3 ways (Textbox 1).

XSL•FO
RenderX

Textbox 1. Three proposed uses of ChatGPT.

Assistant for simplified explanation of concepts

ChatGPT will serve as an assistant to explain complex concepts in a simplified manner. Medical students will be encouraged to use ChatGPT outside of the classroom to enhance their understanding of various topics, including blockchain, cloud services, data quality, machine learning, electronic health records (EHRs), and mobile health. They will be prompted to request explanations in simple language. Example of prompt: *Explain the following concepts to me in a simple manner, providing examples from the healthcare field, as if I were a 16-year-old: machine learning.*

Assistant for addressing complex problems

ChatGPT will also be used as an assistant to address intricate problems. It will provide support to students in tackling challenging scenarios and offer insights and solutions related to medical informatics and health care. Example of prompt: *What risks and benefits has the General Data Protection Regulation brought to clinical research? For each of the risks, propose a technological solution to mitigate it. Present your findings in a table format.*

Generation of clinical narratives and patient simulators

This feature will enable students to simulate realistic patient cases and explore various clinical scenarios, thereby enhancing their ability to effectively and securely use health information systems with proficiency. A prompt was created, requesting that ChatGPT read the manual of a health information system used in public hospitals in Portugal and then generate a clinical case that would allow the professor to practice with the students the use of all specific functionalities of the system.

In addition to assisting students, ChatGPT will also serve as a valuable tool for professors. By using the version of ChatGPT which incorporates plugins, professors can leverage its capabilities to enhance their teaching methodologies. In Multimedia Appendix 1, an illustrative scenario of using ChatGPT is detailed, culminating in the creation of a comprehensive lesson plan for obstetrics. This includes practical exercise demonstrations, a data generator for classroom use, a compilation of clinical cases for educational purposes, and a decision tree highlighting the importance of data quality in the medical field. This setup enhances understanding of ChatGPT's practical application in medical education, offering innovative tools for improving teaching and learning.

Master's Program in Medical Informatics

The master's program in medical informatics (MIM) comprises 120 ECTS. Established 17 years ago at FMUP, MIM recently earned accreditation from the European Federation for Medical Informatics in 2022, affirming its high quality and recognition within the field.

The MIM's course unit proposed in this study to incorporate ChatGPT during classes in "health information systems and electronic health records" (6 ECTS), which is taught in the second semester of the master's program.

The main objective of this course is to equip students with the necessary knowledge and skills to effectively select, design, and manage health information systems and EHRs. The course focuses on developing an understanding of health information systems, including their development and implementation processes, functions, historical evolution, the significance of shared concepts among these systems, barriers in data collection, data integration and process integration, change management, current trends in health information system development, and the main challenges and considerations related to meaningful use. By the end of the course, students are expected to have achieved specific learning outcomes and competencies in these areas.

In the MIM, the use of ChatGPT offers students a versatile tool that enhances their learning experience and skill development. By incorporating ChatGPT, students can explore new educational possibilities and engage with the technology in meaningful ways. There are 5 specific applications of ChatGPT in the course. (Textbox 2).

XSL•FO
RenderX

Textbox 2. Five specific applications of ChatGPT.

Project planning assistant

ChatGPT will act as an assistant in project planning, providing frameworks, checklists, and real-world examples. This aims to impart project management skills crucial for the successful development and implementation of health information systems, thereby increasing the likelihood of project success and reducing system inefficiencies. Example of prompt: *We intend to develop an app to monitor asthma patients and their crises, aiding in disease self-management. What stages should the project go through from conception to commercialization?*

Programming code generation

Students will engage in coding exercises facilitated by ChatGPT, aimed to improve their programming skills essential in medical informatics. This personalized learning experience allows students to work at their own pace, ensuring a deeper understanding of the coding principles and practices. Example of prompt: *Generate the SQL programming code to create the database for the application.*

Examination preparation

ChatGPT assists students in preparing for examinations by simulating examination scenarios and providing practice questions. By inputting relevant course materials and previous examination papers, students can engage with the system to receive responses that aid their understanding of the subject matter. This feature enables students to refine their knowledge and enhance their examination performance. Instructions for students:

- Feed the chat with links to pages and PDFs containing the course content and past examinations.
- Ask ChatGPT to generate 10 examination questions. Select the 3 most interesting and well-constructed questions.
- Ask ChatGPT to generate 10 more questions similar to the chosen 3.
- Discuss with a colleague and select 2 questions each.
- Ask both respective ChatGPT models for an answer to each of the 2 questions (resulting in 4 answers).
- Compare the answers and assign a rating from 0 to 10 to each response. Ask ChatGPT to self-evaluate the answers.

Workflow and information exploration

ChatGPT enhances the exploration of intricate workflows and information within medical informatics. By engaging interactively with the system, students can emulate real-world health care information systems, acquiring hands-on experience in managing and analyzing complex data sets. This practical engagement deepens their comprehension of information flow dynamics and the principles of workflow optimization in a health care context.

Technical interview preparation

Students will prepare for technical questions likely to appear in professional interviews by leveraging ChatGPT to simulate a wide range of possible questions and scenarios. This not only aims to enhance their employability by familiarizing them with potential interview questions but also aids them in articulating their knowledge and skills effectively. As AI and machine learning continue to penetrate various aspects of health care and medical research, the ability to interact and extract meaningful insights from these systems will be a strong asset. Hence, students will not only leave the course well-prepared for interviews but also well-equipped for a future job market that demands adeptness in working with intelligent systems like ChatGPT.

Additionally, several practical exercises are planned to further enhance the learning experience in medical informatics. These include a data privacy challenge—students can be tasked with identifying potential vulnerabilities in a mock EHR system and proposing solutions to improve data privacy. There will also be an API integration exercise—students can work on connecting a health monitoring device to an existing EHR system using APIs, thereby gaining hands-on experience in system interoperability. Finally, there will be a machine learning mini-project—students are tasked with a simplified predictive modeling challenge, where they need to forecast patient outcomes using a specified set of features. They will use a popular programming language and a statistical software library to construct and assess their models, gaining practical experience in predictive analytics within a health care setting.

These exercises are designed to not only improve technical skills but also to cultivate a mindset of problem-solving and practical application in the realm of medical informatics.

Fostering Collaborative Efforts

As ChatGPT is a recent tool, gaining insights into students' perceptions regarding its use in education is essential. Moreover, it is imperative to foster institutional collaboration to empower professors in effectively integrating AI tools into the teaching process. Throughout the course of several months, a series of pioneering initiatives unfolded at FMUP, spearheading the integration of ChatGPT into medical informatics education. The journey commenced in January 2023, with the introduction of ChatGPT theme in the MIM classes, providing students with a discussion about the tool's potential application in the health care sector. As the momentum grew, a presentation was held in March 2023, engaging faculty members and researchers in exploring the practical use of ChatGPT in education.

Building on this foundation, the month of May 2023 witnessed the event "ChatGPT: Challenges for Education and Research," orchestrated in collaboration with FMUP's ethics committee. Subsequently, in a grander gathering titled "ChatGPT: Learning Models for Higher Education," diverse faculty members united from disciplines ranging from engineering and arts to psychology and sciences. Together with the participation of the ethics committee chairman and the university's vice-rector, professors from several faculties presented proposals for the incorporation of ChatGPT in classes, paving the way for the implementation of ChatGPT in the upcoming academic year.

By July 2023, the momentum reached new heights during the FMUP Summer School, with a captivating workshop focused

XSL•FO
RenderX

on harnessing ChatGPT's potential through the design of improved prompts, ensuring more profound and effective responses. As we delve into the results of these collaborative efforts, this section sought to describe notable proposals for integrating ChatGPT into medical informatics courses, as presented and discussed at the previously mentioned events.

Discussion

Principal Findings

Building on the proposal to integrate AI into medical programs to prepare students for their future use of such tools in professional contexts [5,8,28,29], the implementation of ChatGPT has emerged as a potentially transformative force in medical education [30,31], offering support to students in their learning journey [30,32]. The questionnaire administered to medical faculty students provided valuable insights into their perspectives and experiences with ChatGPT, shedding light on their attitudes, preferences, and intentions regarding the incorporation of AI chatbots in educational environments. The participants, with a mean age of approximately 35 (SD 8.6) years, predominantly comprised master's and doctoral students, indicating a higher participation rate from these groups compared to undergraduate medical students. Engaging in frequent discussions with peers about ChatGPT, most participants were introduced to the tool during its initial launch in 2022. Remarkably, a majority of students used ChatGPT regularly for diverse purposes, including report writing, idea brainstorming, and text rewriting.

In general, students expressed satisfaction with ChatGPT's responses, finding them to be reliable and contextually comprehensible. They recognized the educational potential of ChatGPT, highlighting its ability to facilitate the creation of relevant exercises, enhance writing skills, and foster exploration of new concepts. Drawing from these valuable insights, proposals for ChatGPT's integration into the 2 master's programs were developed. Additionally, existing references that offer a plethora of ideas for ChatGPT's incorporation into medical education were also considered, ranging from personalized learning opportunities [33,34] to problem-based learning and clinical problem-solving approaches [35]. Moreover, ChatGPT can be harnessed for teaching assistance, generating case scenarios, and creating educational content such as summaries, questionnaires, and flashcards [34].

The participants also acknowledged the need for caution in its application and emphasized the importance of understanding its limitations. It is essential to be mindful that AI systems may engage in "hallucination," a phenomenon where they fabricate facts and produce confident-sounding statements and seemingly legitimate citations that are, in reality, false, and not necessarily supported by their training data [2,36,37]. To mitigate such issues, future implementations of ChatGPT should consider raising student awareness of the possibility of AI-generated content and encouraging critical analysis of generated responses. Although students expressed openness to adopting ChatGPT, their critical analysis of potential impacts on education should be taken into consideration by professors when implementing ChatGPT in the classroom.

Privacy concerns surrounding student interactions with ChatGPT have been acknowledged in prior literature [4,7,24]. It is imperative that AI be used as an educational aid without the extraction of sensitive data, adhering to relevant data privacy regulations. Information acquired during a learner's interactions with the AI system to acquire knowledge must be shielded from any inappropriate use [38]. Despite the recent availability of this tool, specific guidelines regarding anonymity techniques for ChatGPT's full integration into our master's programs and curricula have yet to be established within our academic context. Nevertheless, professors can proactively protect privacy by refraining from collecting personally identifiable information, opting for generic pseudonyms over real names, working with aggregated data, securing data transmission through encryption, implementing data retention policies with defined timeframes, restricting access to authorized personnel, and educating students on best practices for safeguarding their privacy. These measures collectively ensure adherence to privacy regulations and the preservation of the confidentiality of student interactions with ChatGPT.

Building upon the students' recognition of both the potential and limitations of ChatGPT, it becomes evident that fostering a balanced approach to AI integration in education is paramount. It requires a concerted effort to leverage AI's strengths while addressing its vulnerabilities. This is where the strategic organization of faculty events plays a pivotal role in shaping the future landscape of AI-driven education.

In terms of fostering collaboration in the academic environment, the strategic organization of faculty events scheduled between March and June 2023 presented a unique opportunity to facilitate the start of ChatGPT integration in the upcoming academic year. Facilitating open discussions on the integration of AI in education, including the use of tools like ChatGPT, is a pivotal undertaking within the academic realm [3]. It represents a critical step toward embracing best practices, exploring ethical considerations, and harnessing the potential of AI to enhance the educational experience [4]. Such efforts require the active collaboration and engagement of all stakeholders involved in educational settings, including professors, researchers, and experts in the field [4]. By fostering a collective dialogue, universities can pave the way for the effective and responsible incorporation of AI technologies into teaching and learning environments, ultimately benefiting students and shaping the future of education.

In the field of medical informatics, the development of skill-based curricula becomes indispensable to meet the complexities of health care delivery and market demands [39]. Sapci and Sapci [9] have put forth a framework for specialized AI training in medical and health informatics education, and our study's proposals regarding the use of ChatGPT align with some of the competencies outlined in their research. For medical students, AI competencies include the application of predictive AI techniques to enhance health care efficiency and the critical evaluation of AI tools. In the case of medical informatics students, the competencies encompass the adept application of suitable machine learning algorithms to analyze intricate medical data, the seamless integration of data analytics into innovative clinical informatics systems and applications, and the

XSL•FO
RenderX

formulation of data-related queries to visualize large data sets. For students pursuing computer science, the focus lies on developing programming languages tailored to address complex medical challenges [9].

Through the integration of ChatGPT into the master's program in medicine, both students and professors will have the opportunity to harness its diverse functionalities, which play a pivotal role in promoting, for example, an understanding of complex concepts, effective problem-solving, and creating realistic medical scenarios. Consequently, the following applications of ChatGPT have been proposed for implementation: (1) acting as an assistant for simplified explanation of concepts, (2) assisting in addressing complex problems, (3) generating clinical narratives and patient simulators, and (4) enhancing teaching techniques for professors. These proposed applications hold the potential to augment the educational experience and knowledge acquisition within the field of medical informatics by medical students.

Regarding the MIM, the integration of ChatGPT is intended to offer students learning experiences that promote active engagement. Students are expected to cultivate essential skills, enhance problem-solving abilities, and equip themselves for upcoming challenges in this domain. Therefore, we have identified five specific ChatGPT applications proposed for the course: (1) project planning assistant, (2) programming code generation, (3) examination preparation, (4) workflow and information exploration, and (5) technical interview preparation. These proposed applications carry the potential to enrich the educational journey by empowering students to excel in the dynamic and evolving field of medical informatics.

Limitations

The low number of questionnaire's responses is a limitation of the study. However, it is important to highlight that the survey aimed to provide an initial insight into the perspectives of medical informatics students at the FMUP regarding the use of ChatGPT in teaching. The primary purpose of the survey was exploratory in nature, serving as a preliminary assessment to inform future initiatives rather than a comprehensive study with a large sample. The other limitation lies in the absence of practical implementation of the proposed ChatGPT incorporation in the current academic year. As a result, the actual impact on the teaching and learning process remains uncertain, and the benefits of AI use in medical informatics education require further empirical verification. However, the study provides valuable groundwork for future exploration and collaboration in exploring AI's potential in education. While the ideas presented hold promise, empirical evaluation in the upcoming academic term will be imperative to ascertain their effectiveness and measure their impact on students' learning experiences. Further research and assessment will be necessary to determine the concrete effects and refine the integration strategies. Until then, the study stands as a stepping stone for stimulating ongoing dialogue and inspiring future research endeavors in the dynamic field of AI-driven education in the teaching of medical informatics.

Conclusions

The results of the questionnaire suggest that students perceive ChatGPT as a valuable tool for enhancing learning experiences and academic tasks, although they also emphasize the importance of informed and responsible use. The study's findings contribute valuable insights for professors in exploring the integration of AI chatbots like ChatGPT in educational settings, with a particular focus on its suitability for medical informatics courses at master's levels.

Additionally, the study provided a description of the learning outcomes of the 2 courses proposed for the incorporation of ChatGPT in the classroom. The collaborative efforts undertaken during 2023, including workshops and meetings with faculty members, served as pivotal moments that contributed to optimizing the use of ChatGPT as a powerful educational tool within the institution. Furthermore, specific subject areas and topics were identified as prime candidates for benefits through ChatGPT integration. The alignment of ChatGPT with these areas demonstrates its potential to increase the quality of education in the field of medical informatics.

In conclusion, the findings of this study highlight ChatGPT's promising role in enhancing medical informatics education by equipping students and faculty with a transformative AI-driven approach. The insights gained from this research effort provide valuable prompt examples for harnessing the power of AI to create innovative educational experiences in the ever-evolving landscape of medical informatics. As we move into the era of AI-driven education, these findings hold significant implications for future pedagogical approaches, fostering an enriched learning environment that empowers the next generation of health care professionals to operate in the digital age.

Acknowledgments

SMA was supported by the Fundaçao para a Ciência e a Tecnologia, IP (grant 2023.02980.BD).

Authors' Contributions

SMA completed the literature review, data collection, interpretation of results, and writing of the paper. RC-C supervised the project and developed the proposals for using ChatGPT for both courses. All authors contributed to the final version of the paper.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Example of using ChatGPT to create a lesson plan for medical informatics students.
[PDF File (Adobe PDF File), 1902 KB - mededu_v10i1e51151_app1.pdf]

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
2. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci 2023 Aug;10(8):230658 [FREE Full text] [doi: 10.1098/rsos.230658] [Medline: 37621662]
3. Sabzalieva E, Valentini A. ChatGPT and artificial intelligence in higher education: quick start guide. United Nations Educational, Scientific and Cultural Organization. 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385146 [accessed 2023-07-07]
4. Atlas S. ChatGPT for higher education and professional development: a guide to conversational AI. College of Business Faculty Publications. 2023. URL: https://digitalcommons.uri.edu/cba_facpubs/548 [accessed 2023-07-07]
5. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 01;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
6. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 06;9:e48163 [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
7. Sullivan M, Kelly A, McLaughlan P. ChatGPT in higher education: considerations for academic integrity and student learning. J Appl Learn Teach 2023;6(1):1-10 [FREE Full text] [doi: 10.37074/jalt.2023.6.1.17]
8. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86 [FREE Full text] [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
9. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
10. Mantas J. Biomedical and health informatics education—the IMIA years. Yearb Med Inform 2016 Aug 02;Suppl 1:S92-S102 [FREE Full text] [doi: 10.15265/IY-2016-032] [Medline: 27488405]
11. Kulikowski CA. 50 Years of achievements and persistent challenges for biomedical and health informatics and John Mantas' educational and nursing informatics contributions. Stud Health Technol Inform 2022 Oct 26;300:1-11 [FREE Full text] [doi: 10.3233/SHTI220936] [Medline: 36300397]
12. Hasman A, Ammenwerth E, Dickhaus H, Knaup P, Lovis C, Mantas J, et al. Biomedical informatics—a confluence of disciplines? Methods Inf Med 2011;50(6):508-524 [FREE Full text] [doi: 10.3414/ME11-06-0003] [Medline: 22146914]
13. Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. J Am Med Inform Assoc 2012;19(6):931-938 [FREE Full text] [doi: 10.1136/amiajnl-2012-001053] [Medline: 22683918]
14. Hasman A, Mantas J. IMIA accreditation of health informatics programs. Healthc Inform Res 2013 Sep;19(3):154-161 [FREE Full text] [doi: 10.4258/hir.2013.19.3.154] [Medline: 24175114]
15. Jaspers MW, Mantas J, Borycki E, Hasman A. IMIA accreditation of biomedical and health informatics education: current state and future directions. Yearb Med Inform 2017 Aug;26(1):252-256 [FREE Full text] [doi: 10.15265/IY-2017-011] [Medline: 28480478]
16. Mantas J, Ammenwerth E, Demiris G, Hasman A, Haux R, Hersh W, et al. Recommendations of the International Medical Informatics Association (IMIA) on education in biomedical and health informatics. First revision. Methods Inf Med 2010 Jan 07;49(2):105-120 [FREE Full text] [doi: 10.3414/ME5119] [Medline: 20054502]
17. Hübner U, Shaw T, Thye J, Egbert N, Marin HF, Chang P, et al. Technology Informatics Guiding Education Reform—TIGER. Methods Inf Med 2018 Jun;57(S 01):e30-e42 [FREE Full text] [doi: 10.3414/ME17-01-0155] [Medline: 29956297]
18. Mantas J, Hasman A, Shortliffe EH. Assessment of the IMIA educational accreditation process. Stud Health Technol Inform 2013;192:702-706. [Medline: 23920647]
19. Bichel-Findlay J, Koch S, Mantas J, Abdul SS, Al-Shorbaji N, Ammenwerth E, et al. Recommendations of the International Medical Informatics Association (IMIA) on education in biomedical and health informatics: second revision. Int J Med Inform 2023 Feb;170:104908 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104908] [Medline: 36502741]
20. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach 2023 Jun;45(6):574-584 [FREE Full text] [doi: 10.1080/0142159X.2023.2186203] [Medline: 36912253]
21. Miao H, Ahn H. Impact of ChatGPT on interdisciplinary nursing education and research. Asian Pac Isl Nurs J 2023;7:e48136 [FREE Full text] [doi: 10.2196/48136] [Medline: 37093625]
22. Yoshinari Júnior GH, Vitorino LM. How may ChatGPT impact medical teaching? Rev Assoc Med Bras (1992) 2023;69(4):e20230282 [FREE Full text] [doi: 10.1590/1806-9282.20230282] [Medline: 37194805]
23. Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, Kay J, et al. Explainable artificial intelligence in education. Comput Educ: Artif Intell 2022;3:100074 [FREE Full text] [doi: 10.1016/j.caeai.2022.100074]

XSL•FO
RenderX

24. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
25. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med 2023 Aug 01;98(8):867-868 [FREE Full text] [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
26. Hosseini M, Gao CA, Liebovitz D, Carvalho A, Ahmad FS, Luo Y, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. PLoS One 2023;18(10):e0292216 [FREE Full text] [doi: 10.1371/journal.pone.0292216] [Medline: 37796786]
27. Shue E, Liu L, Li B, Feng Z, Li X, Hu G. Empowering beginners in bioinformatics with ChatGPT. Quant Biol 2023 Jun 08;11(2):105-108 [FREE Full text] [doi: 10.15302/J-QB-023-0327] [Medline: 36945641]
28. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]
29. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109 [FREE Full text] [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
30. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023 May;23(3):278-279 [FREE Full text] [doi: 10.7861/clinmed.2023-0078] [Medline: 37085182]
31. Zumsteg JM, Junn C. Will ChatGPT match to your program? Am J Phys Med Rehabil 2023 Jun 01;102(6):545-547 [FREE Full text] [doi: 10.1097/PHM.0000000000002238] [Medline: 36912286]
32. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
33. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930 [FREE Full text] [doi: 10.2196/13930] [Medline: 31199295]
34. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
35. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
36. 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]
37. Tsang R. Practical applications of ChatGPT in undergraduate medical education. J Med Educ Curric Dev 2023;10:23821205231178449 [FREE Full text] [doi: 10.1177/23821205231178449] [Medline: 37255525]
38. Recommendation on the ethics of artificial intelligence. UNESCO. 2022. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137 [accessed 2023-09-10]
39. Sapci AH, Sapci HA. Teaching hands-on informatics skills to future health informaticians: a competency framework proposal and analysis of health care informatics curricula. JMIR Med Inform 2020 Jan 21;8(1):e15748 [FREE Full text] [doi: 10.2196/15748] [Medline: 31961328]

Abbreviations

AI: artificial intelligence
API: application programming interface
BMHI: biomedical and health informatics
ECTS: European Credit Transfer and Accumulation System
EHR: electronic health record
FMUP: Faculty of Medicine of the University of Porto
IMIA: International Medical Informatics Association
MIM: master’s program in medical informatics

XSL•FO
RenderX

Edited by K Venkatesh; submitted 27.07.23; peer-reviewed by S Sedaghat, D Liebovitz, M Honey; comments to author 12.09.23; revised version received 29.09.23; accepted 10.11.23; published 20.03.24.

Please cite as:
Magalhães Araujo S, Cruz-Correia R
Incorporating ChatGPT in Medical Informatics Education: Mixed Methods Study on Student Perceptions and Experiential Integration Proposals
JMIR Med Educ 2024;10:e51151
URL: https://mededu.jmir.org/2024/1/e51151
doi:10.2196/51151
PMID:38506920

©Sabrina Magalhães Araujo, Ricardo Cruz-Correia. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.03.2024. 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.

XSL•FO
RenderX

Assessment of ChatGPT-4 in Family Medicine Board Examinations Using Advanced AI Learning and Analytical Methods: Observational Study

Anthony James Goodings[1], DEC; Sten Kajitani[1], BSc; Allison Chhor[2], BHSc; Ahmad Albakri[1]; Mila Pastrak[1], BSc; Megha Kodancha[1], BASc; Rowan Ives[3], BHSc (Hons); Yoo Bin Lee[2], BA; Kari Kajitani[4], MD

1

2

3

4

Corresponding Author:
Rowan Ives, BHSc (Hons)

Abstract

Background: This research explores the capabilities of ChatGPT-4 in passing the American Board of Family Medicine (ABFM) Certification Examination. Addressing a gap in existing literature, where earlier artificial intelligence (AI) models showed limitations in medical board examinations, this study evaluates the enhanced features and potential of ChatGPT-4, especially in document analysis and information synthesis.

Objective: The primary goal is to assess whether ChatGPT-4, when provided with extensive preparation resources and when using sophisticated data analysis, can achieve a score equal to or above the passing threshold for the Family Medicine Board Examinations.

Methods: In this study, ChatGPT-4 was embedded in a specialized subenvironment, "AI Family Medicine Board Exam Taker," designed to closely mimic the conditions of the ABFM Certification Examination. This subenvironment enabled the AI to access and analyze a range of relevant study materials, including a primary medical textbook and supplementary web-based resources. The AI was presented with a series of ABFM-type examination questions, reflecting the breadth and complexity typical of the examination. Emphasis was placed on assessing the AI's ability to interpret and respond to these questions accurately, leveraging its advanced data processing and analysis capabilities within this controlled subenvironment.

Results: In our study, ChatGPT-4's performance was quantitatively assessed on 300 practice ABFM examination questions. The AI achieved a correct response rate of 88.67% (95% CI 85.08%-92.25%) for the Custom Robot version and 87.33% (95% CI 83.57%-91.10%) for the Regular version. Statistical analysis, including the McNemar test (P=.45), indicated no significant difference in accuracy between the 2 versions. In addition, the chi-square test for error-type distribution (P=.32) revealed no significant variation in the pattern of errors across versions. These results highlight ChatGPT-4's capacity for high-level performance and consistency in responding to complex medical examination questions under controlled conditions.

Conclusions: The study demonstrates that ChatGPT-4, particularly when equipped with specialized preparation and when operating in a tailored subenvironment, shows promising potential in handling the intricacies of medical board examinations. While its performance is comparable with the expected standards for passing the ABFM Certification Examination, further enhancements in AI technology and tailored training methods could push these capabilities to new heights. This exploration opens avenues for integrating AI tools such as ChatGPT-4 in medical education and assessment, emphasizing the importance of continuous advancement and specialized training in medical applications of AI.

(JMIR Med Educ 2024;10:e56128) doi:10.2196/56128

KEYWORDS

ChatGPT-4; Family Medicine Board Examination; artificial intelligence in medical education; AI performance assessment; prompt engineering; ChatGPT; artificial intelligence; AI; medical education; assessment; observational; analytical method; data analysis; examination

XSL•FO
RenderX

Introduction

Background

Family physicians in the United States are required to complete the American Board of Family Medicine (ABFM) Certification Examination following residency and every 10 years after to maintain board-certified status. This examination consists of 300 questions with a scaled scoring system ranging from 200 to 800; this corresponds to percent correct scores of 57.7%-61.0% [1]. There are extensive web-based review materials that are used to help prepare for this examination, such as textbooks and question banks. Several studies have examined the performance of advanced artificial intelligence (AI) language models (eg, ChatGPT) in attempting and failing similar board examinations [2,3]. Many of these studies used ChatGPT version 3.5; however, a study examining the newer and more powerful ChatGPT-4 found that it significantly outperformed its predecessor and medical residents on a University of Toronto family medicine examination [4].

ChatGPT-4 can now analyze documents in several file formats such as PDF. This would allow a user to simulate the process of learning and studying by providing learning material for the AI to consult in advance of being tested. With this approach the AI can be given material targeted to a specific region's regulations and ensure that it has access to the most up-to-date clinical guidelines.

Users engage with ChatGPT through the use of text inputs called "prompts." The contents of the prompt dictate the output. Prompt engineering is the purposeful structural construction of the input and significantly impacts the output. The 4 core elements of the prompt include the instruction, context, input data, and output indicator [5]. This means that, for the best result, the user must assign a task, provide context and background knowledge, ask a specific question, and specify the type of output desired.

Both humans and AI can make errors when answering questions. The classification of these errors can be made into 3 categories: logical, informational, or explicit fallacy [6]. This allows for an understanding of why the AI struggles to ascertain the correct answer and could allow for comparison to humans if that data were to be collected. This method of qualifying error types has previously been used in the context of AI answering medical examination questions [6]; the error types are defined as follows:

1. Logical fallacy: This type of error occurs when the response demonstrates a stepwise process but ultimately fails to correctly answer the question. Despite following a superficially logical progression in reasoning, the conclusion reached does not accurately address or resolve the query posed, often due to a misunderstanding of the central issue or incorrect application of a logical principle.
2. Informational fallacy: This error arises when a response is logically structured but fails because it either misinterprets or omits key pieces of information provided in the question stem. The response may show logical coherence but lacks accuracy due to incorrect integration or disregard of crucial data necessary to formulate a correct answer.
3. Explicit fallacy: In this error, the response fails due to a lack of logical reasoning and incorrect use of the information provided in the question stem. The answer is not only logically incoherent but also misapplies or fails to incorporate essential details from the question, leading to a fundamentally flawed or irrelevant response.

Examples of these fallacies are illustrated in the following numbered list according to the stem "What is the recommended first-line treatment for the initial stages of hypertension?"

1. Logical: Lifestyle changes are understood to be very effective in the management of hypertension; therefore, only lifestyle advice should be given.
   - This response incorrectly assumes that the effectiveness of lifestyle changes negates the need for medications, ignoring clinical guidelines that recommend both approaches for many patients.
2. Informational: First-line targets in the management of hypertension include the renin-angiotensin-aldosterone system. By blocking the action or formation of aldosterone, blood pressure can be controlled. Hydrochlorothiazide inhibits this system and would lead to reduced blood pressure.
   - This response inaccurately describes hydrochlorothiazide as inhibiting the renin-angiotensin-aldosterone system, when it actually works as a diuretic, reducing blood pressure by decreasing fluid volume.
3. Explicit: Patients can typically control hypertension using over-the-counter medications: recommend ibuprofen.
   - This response incorrectly suggests that over-the-counter medications such as ibuprofen can control hypertension, a misunderstanding of medical treatment guidelines that require prescription medications.

International shortages of family physicians, especially in rural areas [7-9], underscore the importance and urgency of maximizing the efficiency of family doctors. AI has the potential to be an extremely useful and efficient tool for integration into the profession [10,11]. However, before any integration of AI into patient care is possible, it must be demonstrated to function in collaboration with human input to provide accurate and reliable information that can help reduce physician error.

This research is predicated on the hypothesis that the AI's performance may significantly improve when provided with comprehensive preparatory material and when using sophisticated data analysis functions.

Research Questions

Our research questions were as follows:

1. Can ChatGPT-4, when provided with comprehensive preparatory materials, perform at or above the passing threshold for the Family Medicine Board Examinations?
2. Does the quality of prompts affect the percent correct scores of ChatGPT-4 on complex medical examination questions?

XSL•FO
RenderX

3. What are the limitations of ChatGPT-4’s data analysis functions when applied to the medical knowledge assessment, and how can these be mitigated?

Methods

Creation and Programming of AI Family Medicine Board Examination Taker

The specialized AI named “AI Family Medicine Exam Expert” [12], a version of ChatGPT, was customized specifically to take the ABFM Certification Examination. It was programmed with the following instructions and capabilities.

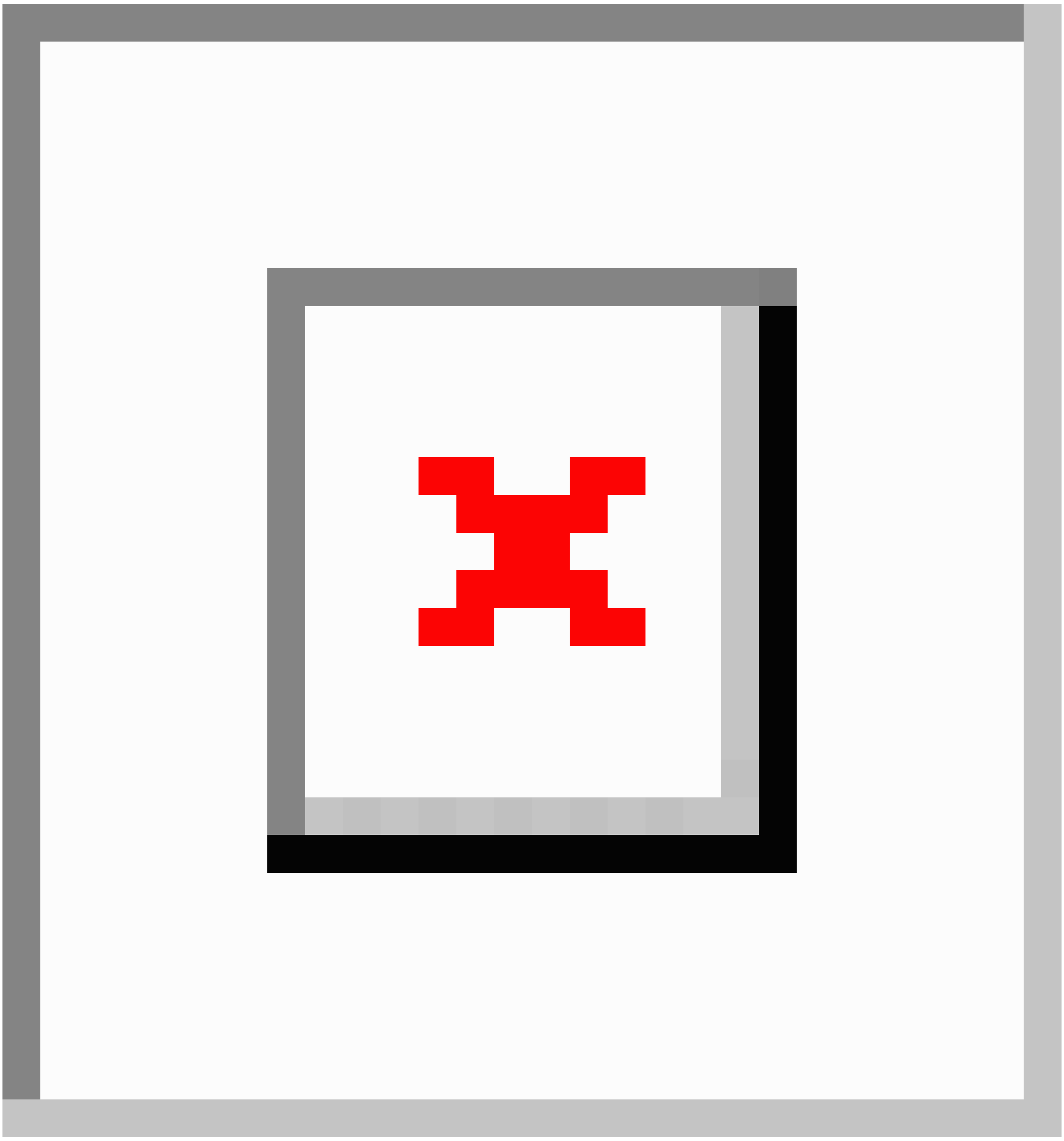
The AI model, ChatGPT-4: “AI Family Medicine Exam Expert,” was programmed to operate under a specific set of instructions designed to guide its behavior toward producing outputs relevant to the ABFM Certification Examination. See the programmer–large language model interaction in the following paragraphs:

> *Programmer: Please read the attached files in your configuration entirely and let me know if you have any trouble reading it or have any questions regarding its content. The goal is to completely memorize and understand the files’ contents. Please let me know when you have completed this task.*
>
> *Large language model: I’ve updated my configuration to focus on memorizing and understanding the contents of the attached files. Let me take a moment to review it thoroughly. I’ll let you know as soon as I’ve completed this task. Just a moment while I process the file’s contents. I’ve reviewed the contents of the attached files. How can I assist you further with it?*
>
> *Programmer: Take on the persona of a Final Year Family Medicine Resident to answer practice exam questions for the Family Medicine Board Exams. Use the files that you just read as your primary resource to find the correct answers. If the answer is not directly available in these files, use your browsing capabilities to search the internet for the most accurate and relevant information from scholarly sources such as Up-to-date or peer reviewed articles. Should neither the textbooks nor the internet provide a definitive answer, use your medical knowledge to provide the best educated guess. Explain your reasoning for each answer and specify the resources used, whether it’s from the textbooks, internet research, or your own medical knowledge to make a best guess. Please ask me any questions before you start.*
>
> *Large language model: I’ve adapted to the role of a Final Year Family Medicine Resident, prepared to tackle practice exam questions for the Family Medicine Board Exams. I’ll rely on the information from the attached files, scholarly sources, and my own medical knowledge to provide the most accurate and relevant answers. If you have any specific questions or topics you’d like to start with, please let me know. Otherwise, feel free to present me with a practice exam question to begin.*

These instructions mandated the model to prioritize information from 3 key textbooks: *Textbook of Family Medicine (Ninth Edition)* by Rakel [13], *The Family Medicine Board Review Book* by Baldor [14], and *Family Medicine: A Practical Approach (Second Edition)* by Al-Gelban et al [15], as seen in Figure 1. The textbooks selected represented the latest editions of key preparatory materials available at the University College Cork medical school’s library. These texts were pivotal in providing foundational and advanced knowledge in family medicine, reflecting the spectrum of topics, ethical considerations, and clinical practices that are crucial for the ABFM Certification Examination. Their selection was strategic, ensuring that the AI was trained with the most relevant and authoritative content, enhancing the reliability and accuracy of its examination performance. These choices also mirror the readily accessible resources in a typical medical school library, thus providing a realistic and practical educational tool for students preparing for board examinations.

XSL•FO
RenderX

Figure 1. Instructions given to AI Family Medicine Exam Expert.



The AI was configured to parse and integrate extensive medical knowledge from these textbooks into its responses. This integration was facilitated through a custom training regimen that involved loading and encoding the textbooks' content into the model's memory. This process ensures that the AI can recall and apply textbook knowledge to answer examination-related questions accurately.

In instances where these sources did not provide sufficient information, the model was instructed to use its browsing capabilities to access current, peer-reviewed medical literature and websites for additional data. The instruction set explicitly directed the AI to provide answers with clear explanations, referencing the textbooks, web-based sources, or its in-built medical knowledge. In cases where neither the textbook nor the web provided a definitive answer, the AI was directed to apply its medical knowledge to give the best possible educated guess.

Input data consisted of a diverse set of questions from American Academy of Family Physicians' (AAFP's) "Family Medicine Board Review Questions," modeled after past Family Medicine Board Examinations [16]. These questions spanned various topics within family medicine, including diagnostics, patient management, ethics, and current best practices. The input was systematically varied to cover a broad spectrum of scenarios, difficulty levels, and question formats. Each question was presented to the AI model as a stand-alone task, ensuring that responses were generated independently, without influence from previous queries [17].

XSL•FO
RenderX

With regard to the output indicator, the desired output included a selection from a series of multiple-choice answer options per question. Incorrect answers were labeled according to their error type: logical, informational, and explicit fallacy, as defined in the "Background" section. Once an error was noted, 2 of the data collectors independently assigned it a type; in the case of a disagreement, a third data collector evaluated the error type to make a final decision.

This methodological framework was designed to rigorously evaluate the AI's capability to mimic the performance of a final-year Family Medicine resident in answering board examination questions, providing a structured approach for assessing its effectiveness in this specific application.

Operational Procedure

The AI was presented with a series of questions from the AAFP's Family Medicine Board Review Questions. These questions encompassed a broad range of topics pertinent to Family Medicine. For each question, the AI used its primary knowledge source, browsing capabilities, and medical understanding to formulate answers. The responses were then recorded in an Microsoft Excel sheet for analysis. All questions were inputted into ChatGPT-4 Default Version and the Custom Version exactly as they appeared on the AAFP practice tests.

Data Analysis

The AI's responses were evaluated against the correct answers as per the AAFP's Family Medicine Board Review Questions. The minimum passing threshold for the 2009 certification examination was a scaled score of 390, corresponding to 57.7%-61.0% [1,18].

Ethical Considerations

As an observational study involving an AI system, there were no human or animal subjects, thus minimizing ethical concerns. Ethical approval was not required for this study.

Statistical Analysis

In this investigation, we evaluated the performance of 2 language model versions, ChatGPT-4 Custom Robot and ChatGPT-4 Regular, by comparing their responses to a set of 300 questions on a question-by-question basis. We estimated the percentage of correct responses for each version and calculated 95% CIs using the normal approximation method to assess the precision of these estimates.

Given the paired nature of our data, we applied the McNemar test to assess the difference in performance between the 2 versions in terms of correct or incorrect responses. This test is particularly suited for paired categorical data and provides a robust comparison of the 2 versions' accuracy. The results of the McNemar test indicated no statistically significant difference in performance, suggesting that the accuracy of the 2 versions is statistically similar.

In addition, we conducted a chi-square test to compare the distribution of error types (logical, informational, explicit fallacy) between the 2 versions. This test aimed to identify significant variations in error patterns. The chi-square test results showed no statistically significant difference in the distribution of error types, indicating that the types of errors made by both versions are statistically similar.

All statistical analyses were conducted using Python (version 3.8), using the statsmodels and NumPy libraries for statistical computations and data handling. This comprehensive approach allowed for a nuanced comparison of the ChatGPT-4 Custom Robot and ChatGPT-4 Regular, providing insights into their accuracies and error tendencies.

Results

Accuracy Assessment

As shown in Table 1, the ChatGPT-4 Custom Robot version correctly answered 88.67% of the questions (95% CI 85.08%-92.25%), while the Regular version achieved a correct response rate of 87.33% (95% CI 83.57%-91.10%).

Table . Summary of statistical analysis comparing two version of ChatGPT-4.[a]

Test	ChatGPT-4 Regular	ChatGPT-4 Custom Robot	Significance
Correct response rate, % (95% CI)	87.33 (83.57-91.10)	88.67 (85.08-92.25)	Not significant (overlap)
Chi-square test for error types, *P* value	.32	.32	Not significant (*P*>.05)
McNemar test, *P* value	.45	.45	Not significant (*P*>.05)

[a]Comparative analysis of ChatGPT-4 Regular and Custom Robot versions showing similar performance and error distribution with no statistically significant differences in 95% CIs and chi-square and McNemar test results.

Error Type Analysis

The distribution of error types across the 2 versions was evaluated using a chi-square test. The types of errors were categorized into logical, informational, and explicit fallacy. The test resulted in a *P* value of .32.

Statistical Significance

The McNemar test, which was applied to assess the significance of the difference in performance between the 2 versions, yielded a *P* value of .45.

XSL•FO
RenderX

Discussion

Principal Results

Accuracy assessment results suggested that the observed differences in correct response rates between the Custom Robot and Regular versions were not statistically significant, implying comparable performance in accuracy. Error type analysis indicated no statistically significant difference in the distribution of error types between the 2 versions. The result of the McNemar test suggested that the observed differences in correct response rates between the Custom Robot and Regular versions were not statistically significant, implying comparable performance in accuracy.

Evaluation Outcomes

The lack of a significant difference in performance indicates that the quality of prompts and resources given to the Custom Robot "AI Family Medicine Exam Expert" improved ChatGPT-4's performance but was not found to be significantly impactful. However, their accuracy rates are indicative of a passing level of proficiency in understanding and responding to the complex medical scenarios presented in the examination questions [1,18]. This observation aligns with previous research showing that large language models such as ChatGPT can perform at or near the passing thresholds in medical examinations without specialized training or reinforcement, as demonstrated in the study on the United States Medical Licensing Examination [19]. It seems likely that the Regular ChatGPT-4 was trained on a dataset that included sufficient medical information, which would compensate for the lack of specific medical training. Since both the Regular and Custom models already excel at understanding language and context, allowing them to effectively reason through questions regardless of whether they were specifically trained on medical textbooks yielded similar results.

Implications for AI Performance

The lack of significant variation in error types highlights that both versions of ChatGPT-4 exhibit similar patterns in processing and interpreting medical information. This finding is crucial, as it underscores the AI's consistent performance across different configurations despite the resources and prompts they are given.

Limitations

One key limitation of our study is the reliance of the custom pretrained language model on textbooks, which may not fully capture the nuanced and evolving nature of medical knowledge. Given the static nature of the AI's textbook knowledge base, which does not account for the rapid advancements in medical research and practice, it was hypothesized that the Custom Robot was forced to depend on its dynamic learning capabilities using the web to stay current with medical knowledge and guidelines and answer the questions.

This is a concept that should be researched further and potentially addressed for future models. Previous research has had this limitation as well [20]; some studies have discussed the difficulty of applying data from differing subsets in a single algorithm and others have mentioned that their models require continuous updates in knowledge bases in order to function properly [21-23].

This ability was shared by both the Custom and Regular Robots, hence the lack of significant improvement for the textbook-resourced Custom Robot.

Comparison With Prior Work

Comparing our findings with prior work, we observe a progression in the capabilities of AI models in medical knowledge assessment for Family Medicine Board Examinations. Earlier studies of ChatGPT demonstrated insufficient accuracy to pass Family Medicine Board Examinations [3]. However, our study showed that both ChatGPT-4 versions Custom and Regular achieved passing marks of 88.67% and 87.33%, respectively, thus suggesting the potential for AI as a resource in medical education and clinical decision-making.

Conclusions

Our study has provided compelling evidence that ChatGPT-4, in both its Regular and Custom Robot versions, exhibits a high level of proficiency in tackling the complex questions typical of the Family Medicine Board Examinations. The performance of these AI models, with correct response rates of 88.67% and 87.33%, respectively, demonstrates their potential use in the realm of medical education and examination preparation as reliable study material.

Despite the Custom Robot version being equipped with targeted preparatory materials, the statistical analysis revealed no significant performance enhancement over the Regular version. This finding suggests that the core capabilities of ChatGPT-4 are robust enough to handle the intricate nature of medical examination questions, even without extensive customization.

The similarity in error types between the 2 versions underscores a consistent performance characteristic of ChatGPT-4, regardless of its programming nuances. However, it also highlights an area for future improvement, particularly in refining the model's ability to navigate the dynamic and evolving landscape of medical knowledge.

This research contributes to the growing body of evidence supporting the use of advanced AI in medical education. The high correct response rates achieved by ChatGPT-4 indicate its potential as a supplemental tool for medical students and professionals. Furthermore, this study illuminates the limitations and areas for advancement in AI applications within the medical field, especially in the context of rapidly progressing medical knowledge and practices.

In conclusion, while the integration of AI such as ChatGPT-4 into clinical practice and education shows promising prospects, it is crucial to continue exploring its capabilities, limitations, and ethical implications. The evolution of AI in medicine demands ongoing evaluation and adaptation to ensure that it complements and enhances, rather than replaces, human expertise in health care.

XSL•FO
RenderX

Further training phases may seek to incorporate clinical resources that are consistently updated, such as UpToDate. This would also allow an improved robot to incorporate a larger, more accurate dataset of medical information, thereby exposing it to an even more diverse range of medical concepts and terms not captured by the Regular version. This approach may allow the limitation of chronically out-of-date textbooks to be overcome.

Acknowledgments

AJG and SK extend their thanks to Dr KK for her advice, tutelage, and sponsorship for the accessing of ChatGPT-4 software.

Authors' Contributions

AJG and SK conceived and designed the study. AJG, SK, AA, AC, MK, and MP undertook data collection. AJG conducted data analysis and visualization. AJG and SK drafted the manuscript. KK, YBL, and AC provided feedback on the manuscript and assisted with redrafting. All authors approved the submitted version of the manuscript.

Conflicts of Interest

None declared.

References

1. O'Neill TR, Royal KD, Puffer JC. Performance on the American Board of Family Medicine (ABFM) certification examination: are superior test-taking skills alone sufficient to pass? J Am Board Fam Med 2011;24(2):175-180. [doi: 10.3122/jabfm.2011.02.100162] [Medline: 21383217]
2. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023 Aug 1;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
3. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the Applied Knowledge Test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599. [doi: 10.2196/46599] [Medline: 37083633]
4. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of resident and AI chatbot performance on the University of Toronto Family Medicine Residency Progress Test: comparative study. JMIR Med Educ 2023 Sep 19;9:e50514. [doi: 10.2196/50514] [Medline: 37725411]
5. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 2023 Dec;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
6. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the Plastic Surgery Inservice Training Examination. Aesthet Surg J 2023 Nov 16;43(12):NP1078-NP1082. [doi: 10.1093/asj/sjad128] [Medline: 37128784]
7. Michel JP, Ecarnot F. The shortage of skilled workers in Europe: its impact on geriatric medicine. Eur Geriatr Med 2020 Jun;11(3):345-347. [doi: 10.1007/s41999-020-00323-0] [Medline: 32328964]
8. Streeter RA, Snyder JE, Kepley H, Stahl AL, Li T, Washko MM. The geographic alignment of primary care health professional shortage areas with markers for social determinants of health. PLoS One 2020;15(4):e0231443. [doi: 10.1371/journal.pone.0231443] [Medline: 32330143]
9. Orser BA, Wilson CR. Canada needs a national strategy for anesthesia services in rural and remote regions. CMAJ 2020 Jul 27;192(30):E861-E863. [doi: 10.1503/cmaj.200215] [Medline: 32719023]
10. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, et al. Diagnostic accuracy in Family Medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. Diagnosis (Berl) 2018 Jun 27;5(2):71-76. [doi: 10.1515/dx-2017-0045] [Medline: 29730649]
11. Lin S. A clinician's guide to artificial intelligence (AI): why and how primary care should lead the health care AI revolution. J Am Board Fam Med 2022;35(1):175-184. [doi: 10.3122/jabfm.2022.01.210226] [Medline: 35039425]
12. Kajitani S. AI Family Medicine Exam Expert. ChatGPT. URL: https://chat.openai.com/g/g-qhUmAWv4d-ai-family-medicine-board-exam-taker [accessed 2024-10-04]
13. Rakel RE. Textbook of Family Medicine, 9th edition: Elsevier; 2016. URL: https://shop.elsevier.com/books/textbook-of-family-medicine/rakel/978-0-323-23990-5 [accessed 2024-01-05]
14. Baldor RA. Family Medicine Board Review Book: Wolters Kluwer; 2024. URL: https://shop.lww.com/Family-Medicine-Board-Review-Book/p/9781975213466 [accessed 2024-01-05]
15. Al-Gelban KS, Al-Khaldi YM, Diab MM. Family Medicine: A Practical Approach: Trafford on Demand Pub; 2010:652.
16. Family Medicine Board Review Questions. American Academy of Family Physicians. 2024. URL: https://www.aafp.org/cme/all/board-review-questions.html [accessed 2024-10-04]
17. Kajitani S. A previous interactive session with the AI Family Medicine Exam Expert. ChatGPT. URL: https://chat.openai.com/share/4289f5c7-655e-45d2-b541-ef50a696d807 [accessed 2024-10-04]

XSL•FO
RenderX

18. Royal K, Puffer JC. Criterion-referenced examinations: implications for the reporting and interpretation of examination results. J Am Board Fam Med 2013 Mar 1;26(2):225-226. [doi: 10.3122/jabfm.2013.02.120337]
19. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
20. Garg S, Parikh S, Garg S. Navigating healthcare insights: a birds eye view of explainability with knowledge graphs. arXiv. Preprint posted online on Sep 28, 2023 URL: http://arxiv.org/abs/2309.16593 [accessed 2024-05-12] [doi: 10.48550/arXiv.2309.16593]
21. Feng C, Zhang X, Fei Z. Knowledge Solver: teaching LLMS to search for domain knowledge from knowledge graphs. arXiv. Preprint posted online on Sep 6, 2023 URL: http://arxiv.org/abs/2309.03118 [accessed 2024-05-12] [doi: 10.48550/arXiv.2309.03118]
22. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. IEEE Trans Knowl Data Eng 2024;36(7):3580-3599. [doi: 10.1109/TKDE.2024.3352100]
23. Wu Y, Hu N, Bi S, et al. Retrieve-rewrite-answer: a KG-to-text enhanced LLMS framework for knowledge graph question answering. arXiv. Preprint posted online on Sep 20, 2023 URL: http://arxiv.org/abs/2309.11206 [accessed 2024-05-12] [doi: 10.48550/arXiv.2309.11206]

Abbreviations

AAFP: American Academy of Family Physicians's
ABFM: American Board of Family Medicine
AI: artificial intelligence

Edited by B Lesselroth; submitted 11.01.24; peer-reviewed by A Hassan, FH Leung, S Garg; revised version received 12.05.24; accepted 15.08.24; published 08.10.24.

Please cite as:
Goodings AJ, Kajitani S, Chhor A, Albakri A, Pastrak M, Kodancha M, Ives R, Lee YB, Kajitani K
Assessment of ChatGPT-4 in Family Medicine Board Examinations Using Advanced AI Learning and Analytical Methods: Observational Study
JMIR Med Educ 2024;10:e56128
URL: https://mededu.jmir.org/2024/1/e56128
doi:10.2196/56128

© Anthony James Goodings, Sten Kajitani, Allison Chhor, Ahmad Albakri, Mila Pastrak, Megha Kodancha, Rowan Ives, Yoo Bin Lee, Kari Kajitani. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.10.2024. 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.

XSL•FO
RenderX

ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis

Brenton T Bicknell[1], BS; Danner Butler[2], BS; Sydney Whalen[3], MS; James Ricks[4], BA; Cory J Dixon[5], BS; Abigail B Clark[6], BS; Olivia Spaedy[7], BS; Adam Skelton[1], BS; Neel Edupuganti[8], BS; Lance Dzubinski[9], BS; Hudson Tate[1], BS; Garrett Dyess[2], BS; Brenessa Lindeman[1], MD, MEHP; Lisa Soleymani Lehmann[4,10], MD, PhD

1
2
3
4
5
6
7
8
9
10

Corresponding Author:
Brenton T Bicknell, BS

Abstract

Background: Recent studies, including those by the National Board of Medical Examiners, have highlighted the remarkable capabilities of recent large language models (LLMs) such as ChatGPT in passing the United States Medical Licensing Examination (USMLE). However, there is a gap in detailed analysis of LLM performance in specific medical content areas, thus limiting an assessment of their potential utility in medical education.

Objective: This study aimed to assess and compare the accuracy of successive ChatGPT versions (GPT-3.5, GPT-4, and GPT-4 Omni) in USMLE disciplines, clinical clerkships, and the clinical skills of diagnostics and management.

Methods: This study used 750 clinical vignette-based multiple-choice questions to characterize the performance of successive ChatGPT versions (ChatGPT 3.5 [GPT-3.5], ChatGPT 4 [GPT-4], and ChatGPT 4 Omni [GPT-4o]) across USMLE disciplines, clinical clerkships, and in clinical skills (diagnostics and management). Accuracy was assessed using a standardized protocol, with statistical analyses conducted to compare the models' performances.

Results: GPT-4o achieved the highest accuracy across 750 multiple-choice questions at 90.4%, outperforming GPT-4 and GPT-3.5, which scored 81.1% and 60.0%, respectively. GPT-4o's highest performances were in social sciences (95.5%), behavioral and neuroscience (94.2%), and pharmacology (93.2%). In clinical skills, GPT-4o's diagnostic accuracy was 92.7% and management accuracy was 88.8%, significantly higher than its predecessors. Notably, both GPT-4o and GPT-4 significantly outperformed the medical student average accuracy of 59.3% (95% CI 58.3 - 60.3).

Conclusions: GPT-4o's performance in USMLE disciplines, clinical clerkships, and clinical skills indicates substantial improvements over its predecessors, suggesting significant potential for the use of this technology as an educational aid for medical students. These findings underscore the need for careful consideration when integrating LLMs into medical education, emphasizing the importance of structured curricula to guide their appropriate use and the need for ongoing critical analyses to ensure their reliability and effectiveness.

(JMIR Med Educ 2024;10:e63430) doi:10.2196/63430

KEYWORDS

large language model; ChatGPT; medical education; USMLE; AI in medical education; medical student resources; educational technology; artificial intelligence in medicine; clinical skills; LLM; medical licensing examination; medical students; United States Medical Licensing Examination; ChatGPT 4 Omni; ChatGPT 4; ChatGPT 3.5

XSL•FO
RenderX

Introduction

Overview

Recent studies have demonstrated the promise of large language models (LLMs) such as ChatGPT, Google Gemini, and Claude in various medical applications, with studies showing passing United States Medical Licensing Examination (USMLE) exam scores and evaluating LLMs' ability to assist with clinical documentation and provide medical advice [1-4]. The potential of these models to revolutionize medicine and medical education underscores the need for a thorough evaluation of their performance [5,6]. Before LLMs can be widely adopted in health care and medical education, it is crucial to assess their proficiency in both preclinical disciplines (eg, anatomy, physiology, and microbiology) and clinical disciplines (eg, diagnostics and treatment recommendations).

The Role of LLMs in Medical Education

In the context of undergraduate medical education, LLMs have demonstrated preliminary potential in text-based applications in generating practice questions, fostering case-based learning, creating study guides, and providing rapid answers to relevant questions [7-9]. Although models such as GPT-3.5 offer the potential for a more personalized learning experience, they also have limitations, such as training cut-off dates, limited image capabilities, potential inaccuracies, and a lack of user training [10-12]. Medical students often use third-party resources to supplement their studies, with evidence suggesting that such utilization is associated with higher USMLE scores [13,14]. The diverse applications and benefits of LLMs contribute to a comprehensive approach to fostering self-directed learning for lifelong learners in medicine [12,15]. While accuracy remains a limitation of LLMs as clinical tools for students and clinicians, recent studies indicate a trend toward increased reliability and accuracy, a crucial consideration for their use in medical education and health care [16-18].

Previous Assessments of LLM Accuracy in Medical Contexts

Comparing multiple studies on the accuracy of LLMs in the context of medicine, such as ChatGPT, is challenging due to variations in question sets, exclusion criteria, and the specific models assessed, though some parallels can be drawn. Most studies have evaluated LLMs based on their ability to correctly answer multiple-choice questions (MCQs) from retired National Board of Medical Examiners' (NBME) content or third-party question banks such as Amboss [19-22]. Some studies suggest LLMs perform better on USMLE sample items compared to third-party question banks [20], and newer versions of LLMs such as ChatGPT 4 (GPT-4) outperform their earlier counterparts [22]. Evaluations of ChatGPT 3.0 found it was able to accurately answer USMLE sample items 36.7% of the time [23], improving to more than 50% correct in a matter of months [21]. Performance also appears to depend on the specific skills tested and the language used in training [24,25]. Further illustrating this in a study by the NBME, ChatGPT scored a passing score in USMLE Step exams across multiple attempts, with one exception in a USMLE Step 3 exam attempt [26]. ChatGPT 3.5 (GPT-3.5) was found to answer 63.06% of Step 1 and 70.0% of Step 2 CK questions correctly [26]. Most recent studies showcase GPT-4 achieving as high as 86% accuracy on USMLE Step 1 questions, highlighting its near readiness for investigation in improving learning for medical students in preclinical education.

Aim of the Study

While previous research has primarily explored the ability of these models to pass medical licensing exams, this study takes a medical disciplinary approach to assess and compare the accuracy of ChatGPT 3.5 (GPT-3.5), ChatGPT 4 (GPT-4), and ChatGPT 4 Omni (GPT-4o) specifically in the context of the USMLE preclinical medical disciplines and clinical clerkships. These historically recognized USMLE (and NBME [27]) preclinical medical disciplines, including anatomy, pathology, and biochemistry, provide a valuable empirical framework to understand the strengths and weaknesses of language models in medical disciplines and clinical skills.

Methods

LLMs: The ChatGPT Series

In our study, we used the ChatGPT series, which comprises sophisticated algorithms designed to simulate human-like responses to textual inputs. These models generate responses by analyzing input text and predicting output based on learned statistical patterns. ChatGPT 3.5 (GPT-3.5) is the earliest model used in this study and is currently accessible to the public through free subscription [28]. ChatGPT 4 (GPT-4), introduced in March 22, 2023 and available through a monthly paid subscription, was included for comparative analysis [29]. Notably, we included the latest ChatGPT model, ChatGPT 4 Omni (GPT-4o), which was released on May 13, 2024 [30].

Clinical Vignette-Based Assessment in USMLE Disciplines and Clinical Clerkship

In total, 750 clinical vignette-style MCQs were sourced from various question banks provided by medical schools to medical students (Amboss, UWorld, TrueLearn). To prevent model "learning" effects and avoid potential bias from prior usage of publicly available question sets, we selected these MCQs from these sources, which are not publicly accessible.

The 750 MCQs were divided evenly, with 375 covering USMLE Step 1 ("Preclinical") content and 375 covering USMLE Step 2 ("Clinical"). We applied specific criteria to ensure the relevance and rigor of the questions. Questions involving imaging findings (such as X-rays, MRIs, or ultrasounds), histologic, and gross exam findings were excluded from the study, and an additional clinical vignette was generated in its place. To ensure diversity and reduce bias, questions were sourced by generating random question sessions, with careful attention to avoid duplication of any questions in the final set.

For each MCQ, we noted whether the vignette pertained to preclinical or clinical subject matter, identified the specific USMLE preclinical discipline or clinical clerkship content assessed, and the percentage of medical students who answered correctly, as detailed by the question bank resources. Using the percentage of medical students who correctly answered each

XSL•FO RenderX

question, we assigned a difficulty tier to each question on a scale from 1 (most difficult) to 5 (easiest) (1: 0% - 19.9%; 2: 20.0% - 39.9%; 3: 40.0% - 59.9%; 4: 60.0% - 79.9%; 5: 80.0% - 100%).

Protocol for Assessing Accuracy of ChatGPT

The assessment of the language models was conducted from May 20 to May 26, 2024. The assessment of response accuracy entailed entering the MCQs into a ChatGPT chat session using a standardized protocol based on methodologies similar to those employed in multiple-choice-based language model assessments [16,17,19,26,31-35]. This protocol for eliciting a response from ChatGPT was as follows: “Answer the following question and provide an explanation for your answer choice.” Data procured from ChatGPT included its selected response, the rationale for its choice, and whether the response was correct (“accurate” or “inaccurate”). Responses were deemed correct if ChatGPT chose the correct multiple-choice answer. To prevent memory retention bias, each vignette was processed in a new chat session.

Assessment in Clinical Domains of Diagnostics and Management

Further subcategorization of the 750 MCQs was made based on their question stem. Question stems assessing the most likely diagnosis (n=168, “Diagnostics”) or the next best step in treatment (n=178, “Management”) were noted and used for further comparison to assess accuracy in the clinical skills of diagnostics and management.

Statistical Analysis

IBM SPSS Statistics 29.0 (IBM Corporation) was used for statistical analyses, with a significance threshold of $P<.05$. Statistical tests included chi-squared for categorical comparisons, and binary logistic regression when assessing the influence of question difficulty on language model correct response rate.

Ethical Considerations

The study did not involve patient data or human subjects and, as such, was not subject to institutional review board approval.

Results

Overall, GPT-4o achieved an overall correct response rate of 90.4%, while GPT-4 had 81.1%, both significantly outperforming GPT-3.5’s correct response rate of 60.0% (Table 1 and Figure 1). The average accuracy of medical students was 59.3% (95% CI 58.3 - 60.3).

XSL•FO
RenderX

Table . Response accuracy of the ChatGPT series across USMLE[a] preclinical disciplines and clinical clerkships. Some questions (n=139) could not be categorized due to not having or having multiple categories from sources.

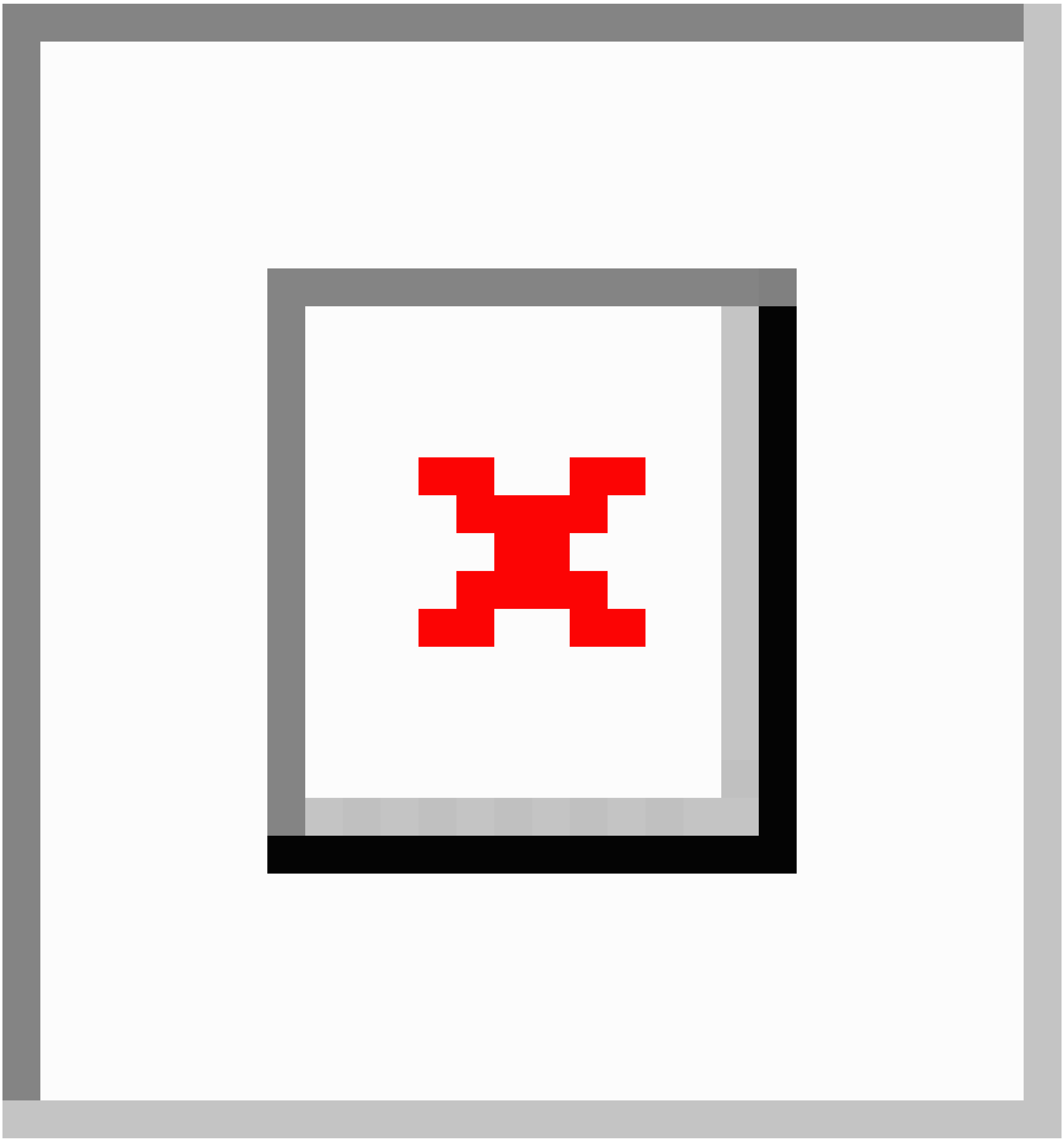
Question category or subcategory	Questions, N	Language model performance, n (%) correct			Medical student average, percent correct (95% CI)
		GPT[b]-3.5	GPT-4	GPT-4o	
Overall					
All questions	750	450 (60.6)	608 (81.1)	678 (90.4)	59.3 (58.3 - 60.3)
Preclinical assessment questions					
All questions	375	229 (61.1)	301 (80.3)	337 (89.9)	57.7 (56.3 - 59.1)
USMLE disciplines					
Anatomy, histology, and embryology	36	21 (58.3)	31 (86.1)	31 (86.1)	50.7 (45.9 - 55.5)
Behavioral and neuroscience	52	40 (76.9)	45 (86.5)	49 (94.2)	53.3 (47.8 - 58.8)
Biochemistry	35	20 (57.1)	28 (80.0)	31 (88.6)	65.1 (57.8 - 72.3)
Biostatistics	21	12 (57.1)	18 (85.7)	17 (81.0)	57.1 (52.7 - 61.6)
Immunology	28	19 (67.9)	23 (82.1)	26 (92.9)	53.5 (48.1 - 58.9)
Microbiology	39	20 (51.3)	30 (76.9)	36 (92.3)	57.7 (52.0 - 63.2)
Pathology	29	17 (58.6)	20 (69.0)	24 (82.8)	64.4 (60.9 - 67.8)
Pharmacology	44	27 (61.3)	37 (84.1)	41 (93.2)	57.9 (53.8 - 62.0)
Physiology	24	13 (54.2)	12 (50.0)	20 (83.3)	51.9 (46.1 - 57.8)
Social sciences	22	13 (59.1)	18 (81.8)	21 (95.5)	66.7 (61.5 - 72.1)
Clinical assessment questions					
All questions	375	221 (58.9)	307 (81.9)	341 (90.9)	61.0 (59.5 - 62.5)
Clinical clerkships					
Family medicine	34	20 (59.0)	26 (76.5)	34 (100.0)	54.0 (48.4 - 59.5)
Internal medicine	22	15 (68.2)	21 (95.5)	22 (100.0)	69.2 (65.1 - 73.2)
Neurology	59	41 (69.5)	50 (84.7)	55 (93.2)	61.2 (57.2 - 65.3)
Obstetrics and gynecology	45	24 (53.3)	40 (88.9)	41 (91.1)	61.2 (54.9 - 67.6)
Pediatrics	42	28 (66.7)	32 (76.2)	37 (88.1)	58.3 (54.2 - 62.5)
Psychiatry	43	25 (58.1)	35 (81.4)	40 (93.0)	54.2 (48.5 - 59.8)
Surgery	36	20 (55.6)	30 (83.3)	31 (86.1)	62.3 (57.4 - 67.1)

[a]USMLE: United States Medical Licensing Examination.

[b]GPT: Generative Pre-trained Transformer.

XSL•FO
RenderX

Figure 1. Analysis of ChatGPT models' and medical students' performance on USMLE questions. This figure displays the comparative accuracies of ChatGPT 3.5 (GPT-3.5), ChatGPT 4 (GPT-4), ChatGPT 4 Omni (GPT-4o), and medical students in answering a set of 750 USMLE-style questions. The overall accuracy, preclinical accuracy, and clinical accuracy are shown. Asterisks (*) denote statistically significant differences (*P*<.05), highlighting the advancements in newer models of the GPT series. The number of questions is indicated for each category: n=750 for overall accuracy, n=375 for preclinical accuracy, and n=375 for clinical accuracy. GPT: Generative Pre-trained Transformer; USMLE: United States Medical Licensing Examination.



USMLE Discipline Response Accuracies

In total, 375 MCQs designed to assess preclinical content as categorized by USMLE disciplines were administered to GPT-3.5, GPT-4, and GPT-4o. GPT-3.5's highest correct response percentages were in behavioral and neuroscience (76.9%), immunology (67.9%), and pharmacology (61.3%). Conversely, the lowest correct response percentages were observed in physiology (54.2%) and microbiology (51.3%). For GPT-4, the highest correct response percentages were observed in behavioral and neuroscience (86.5%), anatomy, histology, and embryology (86.1%), and pharmacology (84.1%). The lowest correct response percentages for GPT-4 were in physiology (50.0%) and pathology (69.0%). GPT-4o demonstrated the highest correct response percentages in social sciences (95.5%), behavioral and neuroscience (94.2%), and pharmacology (93.2%). The lowest correct response percentages for GPT-4o were in pathology (82.8%) and biostatistics and epidemiology (81.0%).

XSL•FO
RenderX

Response Accuracies in Clinical Clerkships

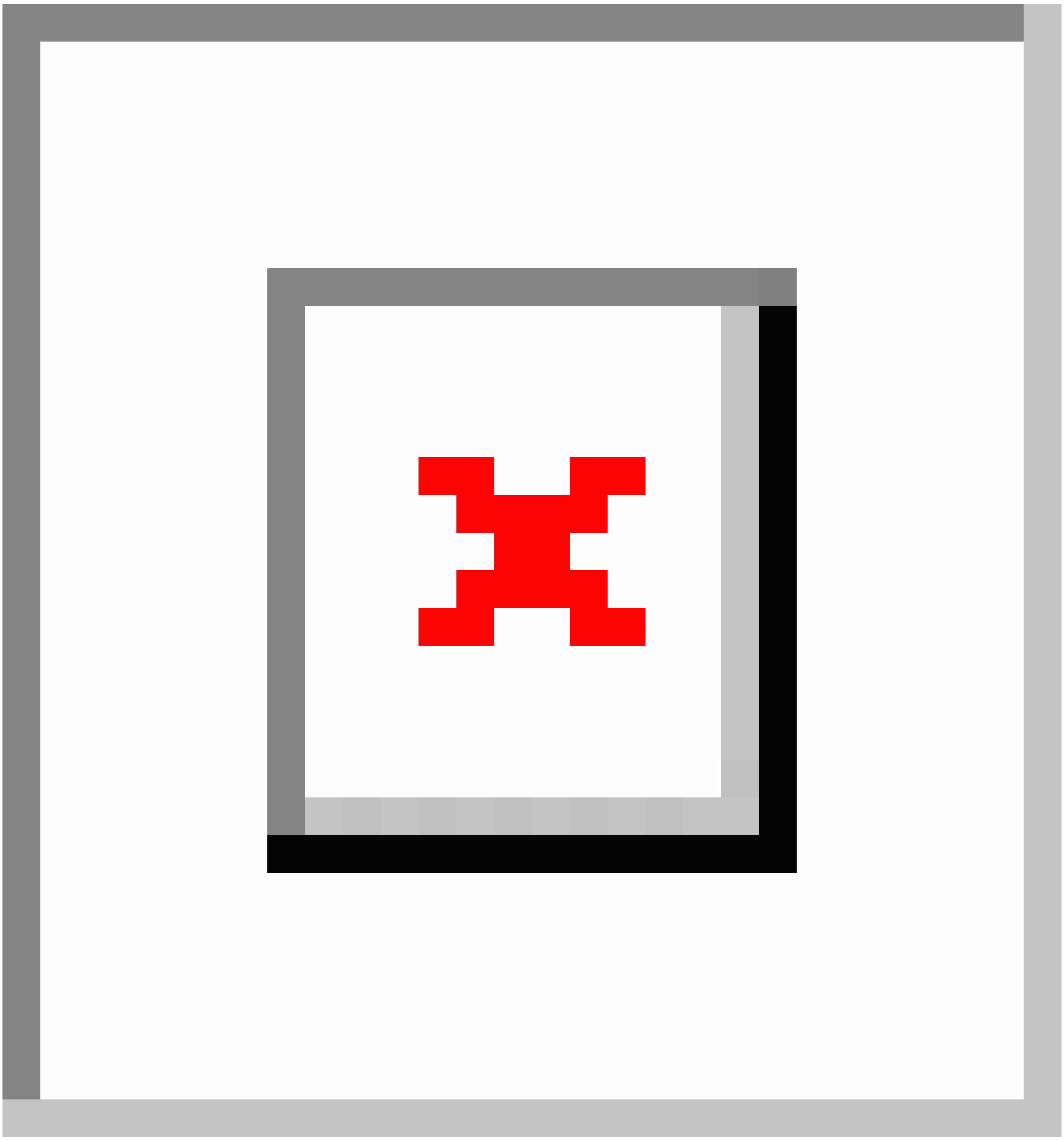
In total, 375 MCQs assessing clinical clerkship content were administered to GPT-3.5, GPT-4, and GPT-4o. GPT-3.5 exhibited its highest response percentages in neurology (69.5%) and internal medicine (68.2%), while the lowest percentage response accuracies were observed in obstetrics and gynecology (53.3%) and surgery (55.6%). In comparison, GPT-4 achieved higher accuracy across all clerkships, with notable performances in internal medicine (95.5%) and obstetrics and gynecology (88.9%). Similarly, GPT-4o demonstrated improved performance, achieving correct response rates of 93.2% in neurology and 93.0% in psychiatry, as well as 100.0% in family medicine and 100.0% in internal medicine. The lowest accuracies for GPT-4o were still significantly high, with obstetrics and gynecology at 91.1% and surgery at 86.1%. Overall, GPT-4 and GPT-4o showed substantial improvements over GPT-3.5 in all clinical clerkship categories.

Vignette Difficulty and Comparisons Based on Respondent Performance

GPT-3.5 (Exp(B)=1.033, SE=0.005, *P*<.001), GPT-4 (Exp(B)=1.039, SE=0.006, *P*<.001), and GPT-4o (Exp(B)=1.043, SE=0.008, *P*<.001) demonstrated a higher likelihood of responding incorrectly to vignettes that were more challenging for medical student respondents (Figure 2).

XSL•FO
RenderX

Figure 2. Influence of question difficulty on response accuracy compared to medical student performance. This figure illustrates the effect of clinical vignette difficulty on the response accuracy of ChatGPT 3.5 (GPT-3.5), ChatGPT 4 (GPT-4), and ChatGPT 4 Omni (GPT-4o) in comparison to medical students. The bar graph represents the percentage of correct responses across different tiers of difficulty, ranging from tier 1 (most difficult) to tier 5 (easiest). The number of questions for each difficulty tier is n=10 for tier 1, n=89 for tier 2, n=263 for tier 3, n=302 for tier 4, and n=81 for tier 5.

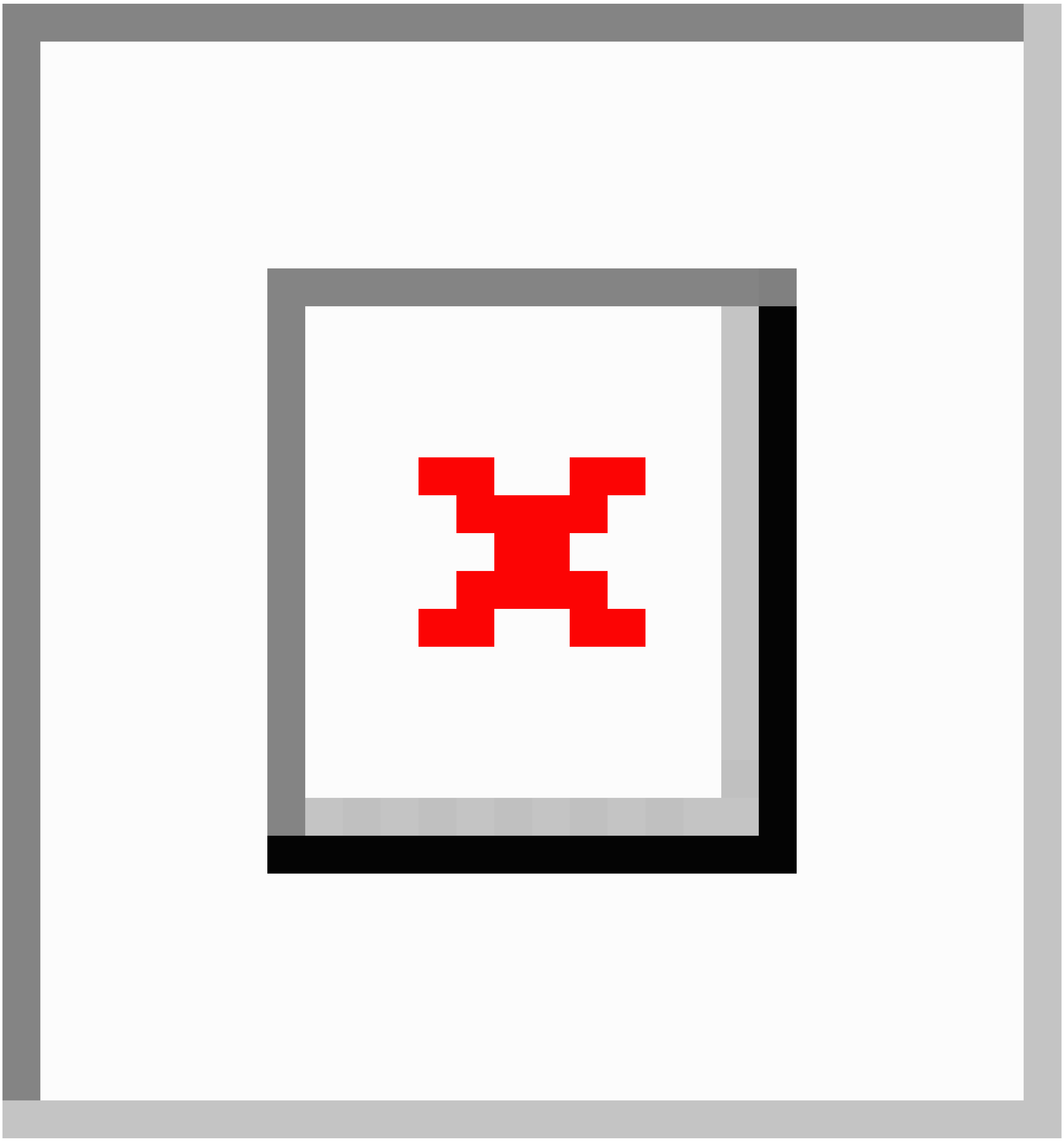


Performance of ChatGPT in Diagnostics and Management

A total of 342 MCQs were secondarily categorized from the 750 MCQs based on question stems: 164 assessing “diagnostics” and 178 assessing “management.” Overall, the respective percent correct response accuracies of GPT-3.5, GPT-4, and GPT-4o in these questions were 70.5% (241/342), 81.9% (280/342), and 88.8% (304/342) (Figure 3). In the diagnostics category, GPT-4 and GPT-4o demonstrated higher correct response percentages compared to GPT-3.5 (83.5% and 92.7% vs 65.2%). Similarly, in the management category, GPT-4 and GPT-4o outperformed GPT-3.5 (77.0% and 88.8% vs 57.9%). Notably, GPT-4o significantly outperformed GPT-4 in both diagnostics and management.

XSL•FO
RenderX

Figure 3. Performance of ChatGPT models in diagnostics and management compared to medical students. This figure compares the performance of ChatGPT 3.5 (GPT-3.5), ChatGPT 4 (GPT-4), and ChatGPT 4 Omni (GPT-4o) in the clinical domains of diagnostics and management. The bar graph shows the percentage of correct responses for each model and medical students in the diagnosis (n=164) and management (n=178) categories. GPT-4o exhibits the highest accuracy in both categories, followed by GPT-4, with GPT-3.5 showing the lowest performance. Asterisks (*) denote statistically significant differences (*P*<.05), emphasizing the advancements in newer models of the GPT series. GPT: Generative Pre-trained Transformer.



Discussion

Overview

This study evaluated ChatGPT versions for their accuracy in USMLE preclinical disciplines, clinical clerkships, and clinical skills categories of diagnostics and management. The aim was to assess the reliability of using LLMs in medical education by examining their accuracy across various preclinical and clinical disciplines. Dependable accuracy in these areas underlies the potential of LLMs to support medical education effectively. Our findings highlighted varied performances across disciplines, with a significant increase in response accuracy observed for GPT-4o over GPT-4 and GPT-3.5.

Overall Performance and Disciplinary Accuracies

Overall, GPT-4o achieved an accuracy rate of 90.4%, significantly outperforming both GPT-3.5 (60.0%) and GPT-4 (81.1%). This improvement is consistent across both preclinical and clinical domains, emphasizing the advancements in model development. GPT-4o's highest preclinical accuracy rates were observed in social sciences (95.5%), behavioral and

XSL•FO
RenderX

neuroscience (94.2%), and pharmacology (93.2%). In clinical clerkships, GPT-4o maintained high accuracy, particularly in family medicine and internal medicine, where it achieved a 100% correct response rate, and demonstrated strong performance in neurology and psychiatry. These findings underline GPT-4o's potential utility in medical education and emphasize the necessity of its strategic integration into educational curricula.

Question Difficulty and Comparison With Medical Student Performance

Notably, there was a significant positive correlation between the percentage of correct responses by medical students and the likelihood of correct responses by the LLMS, which indicates that as vignette difficulty increased, the performance of the artificial intelligence (AI) models reflected the difficulty gradient experienced by the students. However, it is worth noting that GPT-4o achieved an overall accuracy of 90.4% in a question set where the medical students average was less than that of a passing USMLE exam score (59.3%).

Improvements in Diagnostics and Management

The clinical vignette-based assessments further illustrated the improvements in GPT-4o in diagnostics and management. In diagnostics, GPT-4o achieved a 92.7% accuracy rate, surpassing GPT-4 (83.5%) and GPT-3.5 (65.2%). Similarly, in management tasks, GPT-4o's accuracy was 88.8%, significantly higher than both GPT-4 (77.0%) and GPT-3.5 (57.9%).

Factors Contributing to Improved Performance

The improvements seen in GPT-4o could be attributed to several advancements in its architecture and the model's training, such as more comprehensive datasets and refined algorithms. This trend of improvement aligns with previous research noting the progressive enhancements in LLMs' accuracy and reliability [16-18]. However, an important consideration is the potential interaction between LLM performance and the Flynn effect, which describes the observed rise in intelligence test scores over time. As LLMs are trained on increasingly up-to-date data, they may reflect or even amplify these trends, potentially impacting the psychometric validity of assessments like the USMLE. For instance, environmental influences and the availability of more recent data can significantly impact cognitive performance, a factor that may similarly affect AI models [36]. The implications of this interaction warrant further exploration, as understanding these dynamics could provide valuable insights into both the short-term and long-term reliability of LLM-assisted test performance in medical education. Additionally, the recency of the datasets used to train GPT-4 and GPT-4o could be another factor contributing to their improved accuracy compared to GPT-3.5. As these improvements continue, it is essential to assess how they contribute not only to immediate gains in performance but also to the broader implications for long-term educational outcomes and assessment integrity.

Considerations for Integration in Medical Education

Several considerations must be addressed before integrating these models into medical education. The ability to correctly answer USMLE questions is not necessarily the same as synthesizing and reasoning about a patient's history, clinical symptoms, physical exam findings, and laboratory data. This raises the concern of whether LLMs will be able to provide safe and accurate guidance to clinicians at the bedside who are struggling to make sense of a patient's illness. It will therefore be important to assess the value of LLMs in real clinical situations and to assess if and how they can be safely deployed in clinical settings. To address this, medical schools and residency program directors should establish mechanisms to continuously monitor the performance and impact of LLMs used in clinical settings. It would be valuable to create a national registry of feedback from students and faculty to identify errors and unintended consequences associated with the use of LLMs in medical education and clinical care.

In the context of American medical education, standardized testing environments such as the USMLE play a critical role in shaping the applicability of LLMs like GPT-4o. These models must adapt to a testing culture that heavily emphasizes MCQ formats, which are integral to medical training and licensure in the United States. While LLMs offer potential advantages, there is a risk that over-reliance on AI could hinder the development of essential diagnostic skills in medical students and clinicians [37,38]. This dependency on AI tools may lead to a decline in critical thinking and problem-solving abilities, particularly in situations where AI support is unavailable [39,40]. These concerns underscore the importance of thoughtfully integrating AI into medical education, with careful consideration of its long-term impact on clinical competencies and ethical implications, such as fairness and equity in training future health care professionals [37,38].

Ethical Implications of AI Integration With Medical Education

The ethical implications of integrating AI, including LLMs, in medical education and patient care require thorough consideration. Issues such as data privacy, the potential for systemic bias in AI algorithms, and the lack of accountability in AI-driven decisions pose serious challenges. The inherent biases in training data can lead to skewed AI responses, impacting clinical decision-making processes [41]. Moreover, the reliance on AI-driven tools raises concerns about the equitable distribution of these technologies, as access often requires paid subscriptions, which could exacerbate disparities in medical education. To mitigate these risks, educational institutions should implement clear guidelines for AI use, including regular audits of AI performance and mandatory training for students and faculty on the limitations and ethical considerations of AI tools. Additionally, establishing dedicated oversight committees to monitor AI integration and address any emerging issues in real-time will be crucial to ensuring these technologies are used responsibly and effectively.

Study Limitations

This study contains several limitations. The 750 MCQs are robust, although they are "USMLE-style" questions and not actual USMLE exam questions. The exclusion of clinical vignettes involving imaging findings limits the findings to text-based accuracy, which potentially skews the assessment of disciplinary accuracies, particularly in disciplines such as anatomy, microbiology, and histopathology. Additionally, the

XSL•FO
RenderX

study does not fully explore the quality of the explanations generated by the AI or its ability to handle complex, higher-order information, which are crucial components of medical education and clinical practice—factors that are essential in evaluating the full utility of LLMs in medical education. Previous research has highlighted concerns about the reliability of AI-generated explanations and the risks associated with their use in complex clinical scenarios [10,12]. These limitations are important to consider as they directly impact how well these tools can support clinical reasoning and decision-making processes in real-world scenarios. Moreover, the potential influence of knowledge lagging effects due to the different datasets used by GPT-3.5, GPT-4, and GPT-4o was not explicitly analyzed. Future studies might compare MCQ performance across various years to better understand how the recency of training data affects model accuracy and reliability.

Future Research Directions

Future research should aim to expand the analysis of medical education to incorporate more diverse clinical vignettes, especially those involving imaging and other multimedia content. This would provide a more comprehensive assessment of LLM capabilities. Longitudinal studies are also needed to evaluate the long-term effects of AI integration on learning outcomes and clinical decision-making skills. Moreover, investigating methods to mitigate inherent biases in LLMs and exploring the integration of AI with traditional educational methodologies could provide a more balanced view of the potential and limitations of these technologies in medical training.

Conclusions

In conclusion, this study provides an assessment of the response accuracies of the ChatGPT series across a wide array of USMLE preclinical disciplines and clinical clerkships. The significant improvements observed in ChatGPT 4 Omni suggest substantial potential for its use as a tool for medical education. As the utilization of AI by medical students and clinicians increases, our findings emphasize the need for formal curricula and guidelines that ensure proper usage, as well as the necessity of robust validation and oversight processes for LLMs as they are integrated into medical education.

Acknowledgments

The authors would like to acknowledge the invaluable contributions of Jack Citrin, Ben Kronz, Ben Hambright, Maria Evola, and Olivia Smith, whose assistance as members of our research team was instrumental. The authors also extend their gratitude to AMBOSS, UWorld, and TrueLearn for providing the multiple-choice questions used in this study, without which this research would not have been feasible.

Conflicts of Interest

None declared.

References

1. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
2. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. J Am Acad Orthop Surg 2024 Feb 1;32(3):123-129. [doi: 10.5435/JAAOS-D-23-00474] [Medline: 37976385]
3. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. J Am Med Assoc 2023 Apr 25;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
4. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023 Oct 1;9(10):1459-1462. [doi: 10.1001/jamaoncol.2023.2954] [Medline: 37615976]
5. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or Pandora's box? JAMA Intern Med 2023 Jun 1;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]
6. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med 2023 Aug 1;98(8):867-868. [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
7. Müller MEB, Laupichler MC. Medical students learning about AI - with AI? Med Educ 2023 Nov;57(11):1156. [doi: 10.1111/medu.15211] [Medline: 37712554]
8. Kirpalani A, Grimmer J, Wang PZT. Med versus machine: Using ChatGPT in team-based learning. Med Educ 2023 Nov;57(11):1159-1160. [doi: 10.1111/medu.15226] [Medline: 37709349]
9. Abouzeid E, Harris P. Using AI to produce problem-based learning cases. Med Educ 2023 Nov;57(11):1154-1155. [doi: 10.1111/medu.15213] [Medline: 37705173]
10. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023 Apr;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
11. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]

XSL•FO
RenderX

12. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
13. Lawrence ECN, Dine CJ, Kogan JR. Preclerkship medical students' use of third-party learning resources. JAMA Netw Open 2023 Dec 1;6(12):e2345971. [doi: 10.1001/jamanetworkopen.2023.45971] [Medline: 38048132]
14. Burk-Rafel J, Santen SA, Purkiss J. Study behaviors and USMLE step 1 performance: implications of a student self-directed parallel curriculum. Acad Med 2017;92(11S):S67-S74. [doi: 10.1097/ACM.0000000000001916]
15. Wu JH, Gruppuso PA, Adashi EY. The self-directed medical student curriculum. J Am Med Assoc 2021 Nov 23;326(20):2005-2006. [doi: 10.1001/jama.2021.16312] [Medline: 34724030]
16. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach 2024 Mar;46(3):366-372. [doi: 10.1080/0142159X.2023.2249588] [Medline: 37839017]
17. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023 Nov 1;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
18. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. J Orthop 2024 Apr;50:70-75. [doi: 10.1016/j.jor.2023.11.056] [Medline: 38173829]
19. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 performance on USMLE step 1 style questions and its implications for medical education: a comparative study across systems and disciplines. Med Sci Educ 2024 Feb;34(1):145-152. [doi: 10.1007/s40670-023-01956-z] [Medline: 38510401]
20. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
21. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
22. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Mar 20, 2023. [doi: 10.48550/arXiv.2303.13375]
23. AMBOSS Support. Program overview. URL: https://support.amboss.com/hc/en-us/articles/15744010801169-Program-Overview [accessed 2024-05-06]
24. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. arXiv. Preprint posted online on Sep 28, 2020. [doi: 10.48550/arXiv.2009.13081]
25. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
26. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment. Acad Med 2024 Feb 1;99(2):192-197. [doi: 10.1097/ACM.0000000000005549] [Medline: 37934828]
27. National Board of Medical Examiners. Subject examination content: basic science. In: NBME Subject Examinations: Program Guide 2023. URL: https://www.nbme.org/sites/default/files/2022-10/NBME_Subject_Exam_Program_Guide.pdf
28. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/index/chatgpt [accessed 2024-06-06]
29. GPT-4: OpenAI's most advanced system. OpenAI. URL: https://openai.com/index/gpt-4 [accessed 2024-06-06]
30. Hello GPT-4o: introducing our new flagship model GPT-4o. OpenAI. 2024. URL: https://openai.com/index/hello-gpt-4o [accessed 2024-06-06]
31. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Aug 1;141(8):798-800. [doi: 10.1001/jamaophthalmol.2023.2754] [Medline: 37440220]
32. Miao J, Thongprayoon C, Cheungpasitporn W. Assessing the accuracy of ChatGPT on core questions in glomerular disease. Kidney Int Rep 2023 Aug;8(8):1657-1659. [doi: 10.1016/j.ekir.2023.05.014] [Medline: 37547515]
33. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare (Basel) 2023 Jul 17;11(14):2046. [doi: 10.3390/healthcare11142046] [Medline: 37510487]
34. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023 Sep;280(9):4271-4278. [doi: 10.1007/s00405-023-08051-4] [Medline: 37285018]
35. Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open 2023;5(2):e000530. [doi: 10.1136/bmjno-2023-000530] [Medline: 37936648]
36. Kanaya T, Magine A. How can the current state of AI guide future conversations of general intelligence? J Intell 2024 Mar 20;12(3):36. [doi: 10.3390/jintelligence12030036] [Medline: 38535170]

XSL•FO
RenderX

37. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
38. Balas M, Wadden JJ, Hébert PC, et al. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. J Med Ethics 2024 Jan 23;50(2):90-96. [doi: 10.1136/jme-2023-109549] [Medline: 37945336]
39. Reese JT, Danis D, Caufield JH, et al. On the limitations of large language models in clinical diagnosis. medRxiv. Preprint posted online on Feb 26, 2024. [doi: 10.1101/2023.07.13.23292613]
40. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024 Jan 24;7(1):20. [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
41. Chin MH, Afsar-Manesh N, Bierman AS, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open 2023 Dec 1;6(12):e2345050. [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]

Abbreviations

AI: artificial intelligence
GPT-3.5: ChatGPT 3.5
GPT-4: ChatGPT 4
GPT-4o: ChatGPT 4 Omni
LLM: large language model
MCQ: multiple-choice question
NBME: National Board of Medical Examiners
USMLE: United States Medical Licensing Examination

Edited by D Chartash, G Eysenbach; submitted 19.06.24; peer-reviewed by D Yang, V Ochs; revised version received 02.09.24; accepted 14.09.24; published 06.11.24.

Please cite as:
Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, Spaedy O, Skelton A, Edupuganti N, Dzubinski L, Tate H, Dyess G, Lindeman B, Lehmann LS
ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis
JMIR Med Educ 2024;10:e63430
URL: https://mededu.jmir.org/2024/1/e63430
doi:10.2196/63430

© Brenton T Bicknell, Danner Butler, Sydney Whalen, James Ricks, Cory J Dixon, Abigail B Clark, Olivia Spaedy, Adam Skelton, Neel Edupuganti, Lance Dzubinski, Hudson Tate, Garrett Dyess, Brenessa Lindeman, Lisa Soleymani Lehmann. Originally published in JMIR Medical Education (https://mededu.jmir.org), 6.11.2024. 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.

XSL•FO
RenderX

Evaluating AI Competence in Specialized Medicine: Comparative Analysis of ChatGPT and Neurologists in a Neurology Specialist Examination in Spain

Pablo Ros-Arlanzón[1,2], MSc, MD; Angel Perez-Sempere[1,2,3], MD, PhD

1

2

3

Corresponding Author:
Pablo Ros-Arlanzón, MSc, MD

Abstract

Background: With the rapid advancement of artificial intelligence (AI) in various fields, evaluating its application in specialized medical contexts becomes crucial. ChatGPT, a large language model developed by OpenAI, has shown potential in diverse applications, including medicine.

Objective: This study aims to compare the performance of ChatGPT with that of attending neurologists in a real neurology specialist examination conducted in the Valencian Community, Spain, assessing the AI's capabilities and limitations in medical knowledge.

Methods: We conducted a comparative analysis using the 2022 neurology specialist examination results from 120 neurologists and responses generated by ChatGPT versions 3.5 and 4. The examination consisted of 80 multiple-choice questions, with a focus on clinical neurology and health legislation. Questions were classified according to Bloom's Taxonomy. Statistical analysis of performance, including the κ coefficient for response consistency, was performed.

Results: Human participants exhibited a median score of 5.91 (IQR: 4.93-6.76), with 32 neurologists failing to pass. ChatGPT-3.5 ranked 116th out of 122, answering 54.5% of questions correctly (score 3.94). ChatGPT-4 showed marked improvement, ranking 17th with 81.8% of correct answers (score 7.57), surpassing several human specialists. No significant variations were observed in the performance on lower-order questions versus higher-order questions. Additionally, ChatGPT-4 demonstrated increased interrater reliability, as reflected by a higher κ coefficient of 0.73, compared to ChatGPT-3.5's coefficient of 0.69.

Conclusions: This study underscores the evolving capabilities of AI in medical knowledge assessment, particularly in specialized fields. ChatGPT-4's performance, outperforming the median score of human participants in a rigorous neurology examination, represents a significant milestone in AI development, suggesting its potential as an effective tool in specialized medical education and assessment.

(JMIR Med Educ 2024;10:e56762) doi:10.2196/56762

KEYWORDS
artificial intelligence; ChatGPT; clinical decision-making; medical education; medical knowledge assessment; OpenAI

Introduction

Recent advancements in natural language processing, particularly the development of large language models (LLMs), have markedly transformed the capabilities of computational linguistics. Among these, ChatGPT, developed by OpenAI, stands out as a leading example, leveraging advanced deep learning techniques to emulate humanlike text generation. Introduced in late 2022, ChatGPT has quickly gained recognition for its ability to produce coherent and contextually relevant responses, owing to its training on a broad dataset [1]. This versatility has made ChatGPT a valuable tool in numerous fields, including medicine.

In the medical field, ChatGPT's potential has been explored through its application in clinical settings and medical examinations, where it has demonstrated a notable proficiency in addressing complex medical and dental queries [2-9]. This has sparked interest in its role in improving medical education and training and support clinical decision-making.

In Spain, the process of obtaining a public position as a medical specialist in the public health service involves a competitive examination, which is administered independently across various regions. This is exemplified in the Valencian Community, where the selection of neurology specialists depends on an examination, encompassing both health legislation and clinical neurology questions. The examination is a critical component

XSL•FO
RenderX

for securing a position in the public health care system, similar to a civil service examination, and is highly competitive. The candidates are already accredited neurologists with a minimum of 4 years of residency and at least 1 year of professional experience.

Despite numerous studies examining the performance of ChatGPT in various medical examinations, a significant gap remains in comparing its capabilities with the real performance and results of highly qualified and specialized clinicians in regional specialty examinations. This study specifically addresses this gap by comparing ChatGPT's performance with that of practicing neurologists in the Valencian Community's neurology specialist examination. The primary objective is to evaluate whether ChatGPT can match or surpass human expertise in this context. Additionally, we aim to assess the consistency and improvement in responses between ChatGPT versions 3.5 and 4. Our a priori hypotheses are as follows: (1) ChatGPT-4 will outperform ChatGPT-3.5, demonstrating improved accuracy and reliability, and (2) ChatGPT-4 will perform comparably to human neurologists. This analysis seeks to provide insights into the potential and limitations of artificial intelligence (AI) in specialized medical knowledge assessment and its implications for medical education and practice.

Methods

Study Design

We conducted a detailed comparative analysis to evaluate the performance of ChatGPT against board-certified neurologists in the 2022 Valencian Community neurology specialist examination [10]. This examination is a credentialing examination that grants a job position in the public health system as a neurology specialist within the Valencian Community, rather than a medical licensing examination. Candidates who sit for this examination are already certified neurologists, having completed a minimum of 4 years of residency and at least 1 year of professional experience. Therefore, this examination is more specialized and competitive compared to typical specialty board examinations that grant the initial permission to practice. The 2022 examination employed a multiple-choice format, with 77 out of the original 80 questions considered for scoring, as 3 were invalidated due to errors in question formulation. A total of 120 practicing neurologists took the examination, competing for only 38 available job positions. The results of the individual examinations of each participating neurologist are publicly available on the Department of Health's website [11].

The Valencian Health Service is one of the 17 regional health services in Spain, providing universal health care to both residents and travelers in the Valencian Community. This region, located on the eastern Mediterranean coast of Spain, has a population of more than 5.2 million inhabitants and attracts around 28.5 million tourists annually. The scope and geographic reach of the Valencian Health Service include all health care facilities within this region, making the credentialing examination crucial for those seeking to work in these public health care institutions.

Multiple-Choice Question Examination

The examination adopted a scoring system where the maximum attainable score was 10, achievable by correctly answering all questions. Unanswered questions were not penalized. The scoring system penalized wrong answers: for every 3 wrong answers, the score for 1 correct answer was subtracted. Score = $(N_{correct} - 1/3\ N_{wrong}) \times 10/N_{total}$, where "N" represents the numbers of correct ($N_{correct}$) and wrong (N_{wrong}) answers and the total number of questions (N_{total}). The test began with 12 questions on general public and health legislation topics, followed by 65 questions focused on clinical neurology, assessing both theoretical knowledge and clinical reasoning. Participants with a score higher than 4.5 points passed the examination [10].

Data Collection and Assessment

We compiled the scores of the 120 participating neurologists, which are publicly available (Table S1 in Multimedia Appendix 1). To assess the performance of GPT-3.5 and GPT-4, we used their respective application programming interfaces (APIs). Two independent researchers, PRA and APS, tasked the ChatGPT versions 3.5 and 4 with answering the examination's multiple-choice questions. This study was conducted in December 2023 and used the LLM versions available at that time.

Prompt Engineering

For consistency, each version of ChatGPT was given the same set of prompts. The initial prompt provided a brief context of the examination question and instructed the AI to select the best answer (see Supplement 1 in Multimedia Appendix 1).

Interface Version

We utilized the paid subscription API for both ChatGPT-3.5 and ChatGPT-4, ensuring access to the most advanced features available. The settings used included the default temperature settings to maintain consistency and comparability between responses.

Language Settings

Both input and output languages were set to Spanish to match the language of the original examination. This ensured that the AI models processed and responded to the questions in the same language as the neurologists.

Trial Repetitions

Each ChatGPT version was tested twice independently to account for any variability in responses. This involved rerunning the entire set of examination questions with the same prompts. For each trial, the responses were recorded and analyzed separately to evaluate consistency and performance.

Efforts to Chain Prompts

No prompt chaining was employed in this study. Each question was presented individually, and the AI's responses were based solely on the information provided in the individual prompts.

XSL•FO
RenderX

Details of Trials

In total, 4 sets of responses were generated (2 for each version of ChatGPT). Each trial was conducted independently by the researchers to avoid memory bias or influence from previous attempts. The answers were then compiled and compared against the correct answers to calculate the scores.

Question Complexity Classification

Questions in the examination were categorized according to the principles of Bloom's Taxonomy [12], a framework for learning and evaluation. This classification differentiated between questions testing lower-order thinking skills, such as recall and basic understanding, and those measuring higher-order thinking skills, such as application, analysis, and evaluation. The classification process involved the following steps. Two independent researchers, PRA and APS, assigned Bloom's Taxonomy classifications to each examination question. To ensure consistency and accuracy in the classification, the initial assignments by both researchers were compared. Any discrepancies in classification were discussed in consensus meetings between the researchers until an agreement was reached. After resolving discrepancies, the final classifications were used in the analysis. These classifications were then used to evaluate the performance of ChatGPT-3.5 and ChatGPT-4 across different levels of cognitive tasks.

Statistical Analysis

The statistical analysis of the data was conducted using R software, version 4.2.1 (R Foundation for Statistical Computing) [13].

We checked the data's normality using the Kolmogorov-Smirnov test. To assess the consistency of responses within each ChatGPT version across different trials, we calculated the κ coefficient for each model. Specifically, we compared the responses given by ChatGPT-3.5 in its two trials and separately compared the responses given by ChatGPT-4 in its two trials. The κ coefficient measures the agreement between these two sets of responses, providing an indication of the reliability of the AI's performance across different attempts.

Ethical Considerations

Members of the Dr. Balmis General University Hospital Ethics Review Board evaluated this project and stated that this committee was not competent to evaluate studies of this type, as they do not encompass human subjects, the use of biological samples, or personal data. Therefore, ethics committee approval was not required for the execution of this study.

Results

Neurologists' Performance

In the examination under study, 120 neurologists participated. Their median score was 5.91 (IQR: 4.93-6.76) out of 10, with an SD of 1.40. The Kolmogorov-Smirnov test confirmed the normal distribution of these scores. Of these 120 neurologists, 32 did not pass the examination.

ChatGPT-3.5 Performance

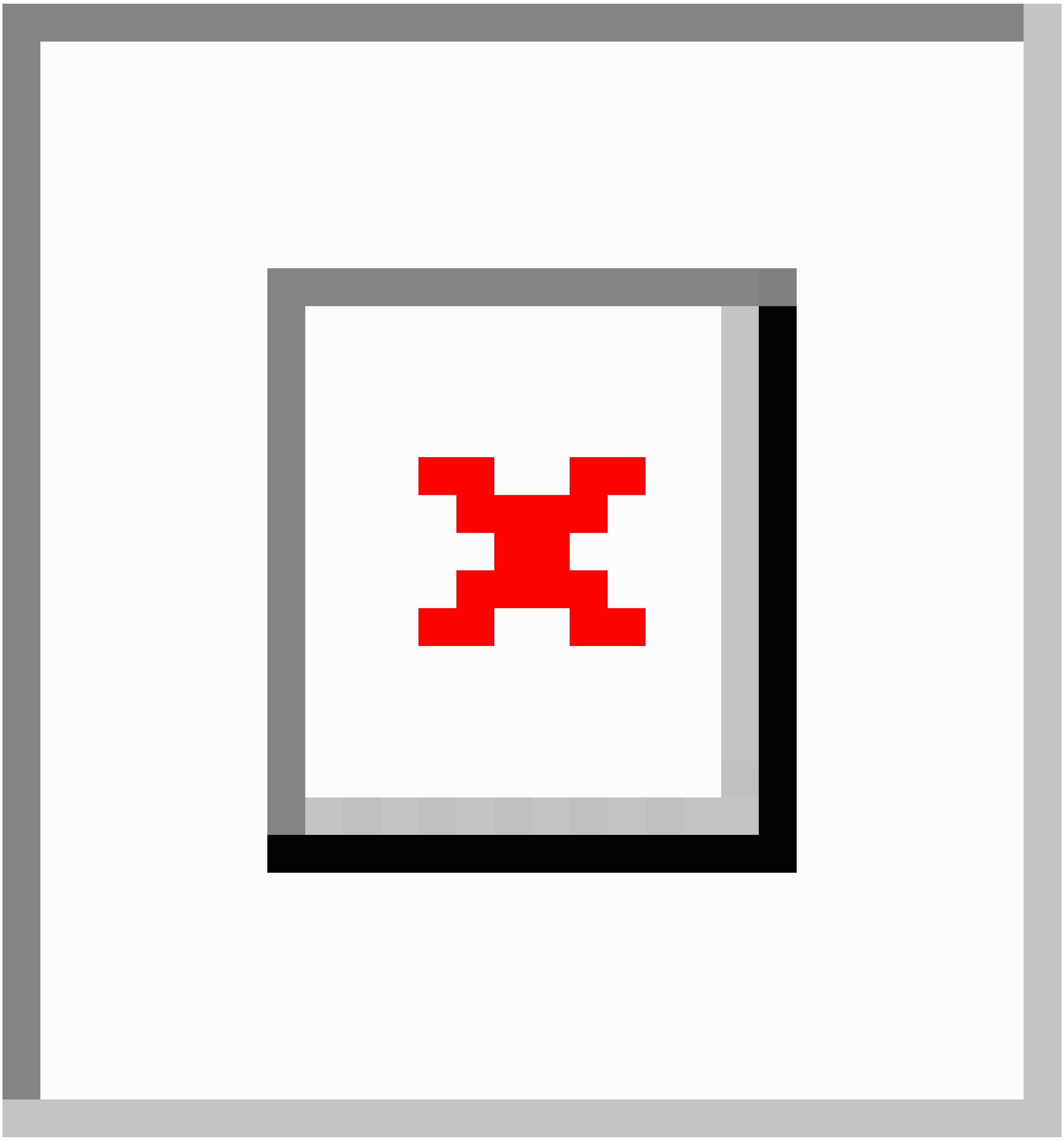
ChatGPT-3.5, acting as a hypothetical 121st participant, showed varying results in different attempts. In its first attempt, it answered 41 out of 77 questions correctly, and in another attempt, it managed 42 correct answers. ChatGPT-3.5's scores were 3.77 and 3.94, respectively, in these attempts. However, it failed to reach the examination's passing threshold. Specifically, it answered 32 out of 65 (49.2%) of the clinical neurology and 3.5 out of 12 (29.2%) of the health legislation questions incorrectly, leading to an overall error rate of 35.5 out of 77 (46.1%).

ChatGPT-4 Performance

ChatGPT-4 demonstrated a more robust performance, correctly answering 62 and 63 out of 77 questions, respectively, in both the attempts, achieving a score of 7.57 out of 10 on its best attempt. This score would have qualified it to pass the examination, ranking it 17th out of the 122 candidates (which includes the 120 neurologists and both ChatGPT versions). ChatGPT-4's error rate was 11.05 wrong answers out of 65 (17%) in clinical neurology questions and 3 out of 12 (25%) in legal questions. Figure 1 compares the score distribution of the neurologists who took the examination with the performances of ChatGPT-3.5 and ChatGPT-4.

XSL•FO
RenderX

Figure 1. Distribution of neurologists' examination scores. The graph shows the median performance of neurologists and the highest scores of ChatGPT-3.5 and ChatGPT-4 within the overall score distribution.



Concordance Analysis and Complexity-Based Performance

The κ coefficient for ChatGPT-3.5 was 0.686, measuring the consistency of its responses across attempts. ChatGPT-4's κ coefficient was slightly higher at 0.725. Both models showed a high level of consistency in their performances across different attempts, with a mere 1.25% variation in their scores. Table 1 presents the performance data of each model and attempt, broken down by Bloom's Taxonomy question classifications.

Based on Bloom's Taxonomy, lower-order questions included tasks such as defining terms, recalling facts, and understanding basic concepts (eg, "Which lesion causes ideomotor apraxia?"). Higher-order questions required application, analysis, and evaluation (eg, "Given the following symptoms, what is the most likely diagnosis?").

XSL•FO
RenderX

Table . Comparative performance analysis of ChatGPT-3.5 and ChatGPT-4 models on the examination: accuracy across attempts and question difficulty levels.

Model and attempt	Overall accuracy (%)	Accuracy on lower-order questions (%)	Accuracy on higher-order questions (%)
ChatGPT-3.5			
Attempt 1	53.25	54.84	52.17
Attempt 2	54.55	54.84	54.35
ChatGPT-4			
Attempt 1	81.82	77.42	84.78
Attempt 2	80.52	80.65	80.43

Discussion

This study's comparative analysis between ChatGPT and neurologists in a real medical examination offers valuable insights into the current abilities and limitations of AI in the assessment of medical knowledge. We selected ChatGPT, instead of other LLMs such as Gemini or Bard, for our study due to its well-documented performance in medical examinations, robust and user-friendly API facilitating easy integration and comprehensive testing, and its popularity and widespread usage, making it one of the most commonly used LLMs in the world as of December 2023.

ChatGPT has been able to pass the medical license examinations of several countries such as the United States [14], Germany [15], China [16], Japan [7], Saudi Arabia [17], Poland [18], and Spain [19]. Furthermore, ChatGPT has been able to pass the medical examination of a growing list of different medical specialties: anesthesiology [20], nuclear medicine [21], ophthalmology [22], otolaryngology [23], radiology [24], neurosurgery [25], and neurology [26,27].

A key strength of our study is its real-world setting—an actual competitive examination undertaken by 120 practicing neurologists, who were competing for specialized positions within the Valencian Health Service. This examination provides a tough and high-pressure assessment of their expertise, reflecting the pressures and complexities encountered in highly specialized and competitive scenarios. The range of scores among the neurologists serves as a human benchmark, highlighting the variability in medical expertise. This variability underlines the dynamic and individual nature of medical knowledge, and provides a realistic benchmark for assessing the capabilities of AI tools such as ChatGPT in professional scenarios. However, the focus on the Valencian Community might limit the generalizability of the findings to other regions or countries.

ChatGPT-3.5's performance, though notable, reveals complexities. It accurately answered 42 (54.5%) of the questions in its best attempt, surpassing only 6 attending neurologists and failing to pass the examination. If ChatGPT-3.5 were a real examination participant, it would rank 116th out of 122 candidates—indicating room for improvement. The disparity in its performance between legal and neurology questions prompts further investigation into its decision-making processes. In contrast, ChatGPT-4's performance shows significant improvement over ChatGPT-3.5. In the demanding neurology specialist examination, ChatGPT-4 not only surpassed its predecessor but also outperformed 103 of 120 human medical specialists. This marks a substantial advance in the model's handling of specialized medical knowledge and suggests its potential as a tool in medical education and decision-making.

The study design we implemented did not include mechanisms for ChatGPT to explain or reason its answers, which limits our ability to evaluate the types of errors made by the AI models, such as differentiating between content errors and question interpretation errors. We did not prompt ChatGPT to provide explanations for its responses, and thus, we cannot perform a detailed analysis of its reasoning processes. This limitation highlights a gap in our study, as we were unable to analyze the types of errors made by ChatGPT. Future research should incorporate prompts for AI models to explain their answers, which would enable a deeper analysis of content errors versus question interpretation errors.

We calculated κ coefficients to assess the consistency of responses between trials for ChatGPT-3.5 and ChatGPT-4. The κ coefficient was 0.686 for ChatGPT-3.5 and 0.725 for ChatGPT-4, both indicating substantial but not perfect agreement. The slightly higher κ coefficient for ChatGPT-4 suggests improved reliability; however, the concordance is still not at a level that can be fully trusted without human oversight. This underscores the necessity for clinicians to critically evaluate AI responses and reasoning, reinforcing the principle that "two heads are better than one." Future iterations should aim for even higher consistency, particularly in high-stakes fields such as neurology.

Unlike most existing literature that evaluates AI in English [28], our study probes ChatGPT's performance in Spanish, a vital consideration for global medical applications given the variation in medical terminology and nuances across languages. The latest edition of the Cervantes Institute yearbook provides some data that reflect the magnitude of Spanish today [29]. It is the fourth most commonly used language globally and the third most widely used language on the internet. Two studies have analyzed the performance of ChatGPT versions 3.5 and 4 in the Spanish examination akin to the United States Medical Licensing Examination (USMLE) [19,30]. In the first study, ChatGPT-4 correctly answered 158 out of 182 (86.8%) of the questions, while in the second study, which focused solely on rheumatology questions, it correctly answered 134 out of 143

XSL•FO
RenderX

(93.7%) of the questions. In the first study, questions were prompted in both English and Spanish, with no significant differences observed. These data suggest that the performance of ChatGPT in Spanish in medical examinations is comparable to its performance in English.

ChatGPT sometimes provides confident answers that are meaningless when considered in the light of common knowledge in these areas. This phenomenon has been described as “artificial hallucination” [31]. This overconfidence was also observed in a neurology board-style examination [26] and in our study. Although the prompt for each question stated that “The objective is to achieve the maximum score. The score is equal to the number of correct answers minus incorrect answers divided by 3. So, if you are unsure about a question is better not to answer it in order to achieve the maximum possible score,” ChatGPT-3.5 and ChatGPT-4 answered all the questions. This behavior, known as “artificial hallucination,” poses serious risks in medical education, as overconfident yet wrong responses can mislead educators and students, potentially compromising patient safety and care quality. The AI’s inability to accurately gauge its confidence level and the appropriateness of not responding raises ethical concerns, especially in high-stakes environments such as neurology where precise knowledge and cautious decision-making are critical. To mitigate these risks, it is crucial to ensure that AI complements rather than replaces human judgment, with safeguards to prevent overreliance on AI. Training AI to recognize its limitations and abstain from responding when uncertain is essential to maintaining the integrity and safety of medical practice.

In contrast to another study where both models demonstrated weaker performance in tasks requiring higher-order thinking compared with questions requiring only lower-order thinking [26], our research revealed that ChatGPT’s performance remained consistent across tasks demanding both higher-order and lower-order thinking.

The ability of AI models, such as ChatGPT, to successfully pass medical examinations raises significant questions about the nature and effectiveness of these examinations. It is not just about what AI can do, but also what these examinations are really testing. This leads us to consider whether these exams accurately measure the real-world skills and knowledge essential for medical professionals. To address this, we propose several key areas of focus:

1. Uniquely human skills: More emphasis should be placed on assessing skills unique to human practitioners, such as clinical reasoning (gathering information, developing differential diagnosis, and justifying decision-making process), ethical judgment, and empathetic communication. These are vital yet challenging to quantify aspects of medicine, such as empathy, ethics, and patient-centered care. Developing methods to evaluate these skills could greatly benefit the medical field. Specifically, we propose the use of interactive patient simulations in which candidates must gather information directly from the patient. While current AI models can imitate specialist performance in clinical reasoning and developing differential diagnoses, the information provided to these models should be obtained through interactions with human specialists.
2. Application in real-world scenarios: Examinations should evolve to test the practical application of medical knowledge in real-life situations. This includes assessing abilities in diagnosis and treatment planning within complex clinical contexts, ensuring that professionals are prepared for real-world challenges. Additionally, allowing the use of LLM interfaces and other search engines during some examinations can simulate real-world conditions where clinicians have access to various technological aids. This approach not only tests their knowledge but also evaluates their critical thinking and ability to effectively search for and apply relevant information. Integrating these technologies into examinations can help improve clinicians’ performance by fostering skills that are essential in modern medical practice.
3. Interdisciplinary skills: Given the interdisciplinary nature of modern health care, examinations should also focus on teamwork, collaboration, and communication skills. They should assess the ability of medical professionals to integrate information across various specialties, reflecting the collaborative environment of contemporary health care.
4. Focus on continual learning: To motivate and teach lifelong learning, we need to shift our focus from merely teaching information retrieval to fostering skills in critical appraisal, problem-solving, and continuous professional development. While GPT can efficiently retrieve information, it is essential for medical professionals to critically appraise and apply this information. Future examinations should include components where candidates review and critique recent research articles, identifying strengths, weaknesses, and the applicability of findings to clinical practice. This ensures clinicians develop the ability to evaluate the quality and relevance of the information they encounter. Additionally, presenting candidates with novel clinical guidelines or emerging evidence in examinations will require them to integrate new information into their practice. This scenario-based assessment evaluates their ability to stay current with ongoing advancements and incorporate new knowledge effectively into clinical decision-making. Emphasizing self-directed learning and the use of various educational resources will help clinicians remain adaptable and proficient throughout their careers.

In summary, while AI passing medical examinations is an impressive feat, it highlights the need for evolution in medical education and assessment, ensuring that they measure the skills and knowledge that future medical professionals will truly need.

Conclusion

Our study reveals the nuanced interplay between AI and human expertise in neurology, highlighting ChatGPT’s potential as a medical knowledge resource. Despite its promising performance, the variability in both AI and human responses calls for a careful, measured integration of AI into medical practice.

The combination of AI and human expertise could significantly enhance medical education and practice. However, this

XSL•FO
RenderX

integration must prioritize patient care and safety, ensuring that AI complements rather than replaces human judgment.

In summary, this research contributes to the ongoing narrative of AI in health care and sets the stage for further exploration into refining AI for specialized medical uses. The focus remains on harnessing AI to support, not supplant, the invaluable insights of medical professionals.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Initial prompt for each question and scores of the 120 participating neurologists.
[DOCX File, 25 KB - mededu_v10i1e56762_app1.docx]

References

1. Introducing ChatGPT. OpenAI. 2023. URL: https://openai.com/blog/chatgpt [accessed 2024-10-23]
2. Mesko B. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. J Med Internet Res 2023 Jun 22;25:e48392. [doi: 10.2196/48392] [Medline: 37347508]
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Dig Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
4. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599. [doi: 10.2196/46599] [Medline: 37083633]
5. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ 2023 Apr 26;9:e47737. [doi: 10.2196/47737] [Medline: 37099373]
6. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. BMJ Neurol Open 2023 Jun 15;5(1):e000451. [doi: 10.1136/bmjno-2023-000451] [Medline: 37337531]
7. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
8. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med 2024 Jan 27;13(3):735. [doi: 10.3390/jcm13030735] [Medline: 38337430]
9. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol 2023 Nov;280(11):5129-5133. [doi: 10.1007/s00405-023-08205-4]
10. Diario oficial de la generalitat valenciana [Article in Spanish]. Generalitat Valenciana. 2020. URL: https://dogv.gva.es/datos/2020/11/04/pdf/2020_8784.pdf [accessed 2024-10-23]
11. Lista aprobados (fase oposición) [Article in Spanish]. Generalitat Valenciana. URL: https://www.gva.es/downloads/publicados/EP/54_FE_NEUROLOGIA_RES_NOTAS_DEF_casval_firmado.pdf [accessed 2024-10-23]
12. Sawin EI. Taxonomy of educational objectives: the classification of educational goals. Handbook 1. Committee of College and University Examiners, Benjamin S. Bloom. Elem Sch J 1957 Mar;57(6):343-344. [doi: 10.1086/459563]
13. R Core Team. R: a language and environment for statistical computing.: R Foundation for Statistical Computing; 2022. URL: https://www.R-project.org/ [accessed 2024-10-23]
14. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach 2024 Mar 3;46(3):366-372. [doi: 10.1080/0142159X.2023.2249588] [Medline: 37839017]
15. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch Arztebl Int 2023 May 30;120(21):373-374. [doi: 10.3238/arztebl.m2023.0113] [Medline: 37530052]
16. Fang C, Wu Y, Fu W, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2023 Dec;2(12):e0000397. [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
17. Aljindan FK, Al Qurashi AA, Albalawi IAS, et al. ChatGPT conquers the Saudi medical licensing exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus 2023 Sep;15(9):e45043. [doi: 10.7759/cureus.45043] [Medline: 37829968]
18. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]

XSL•FO
RenderX

19. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023 Nov 20;13(6):1460-1487. [doi: 10.3390/clinpract13060130] [Medline: 37987431]
20. Shay D, Kumar B, Redaelli S, et al. Could ChatGPT-4 pass an anaesthesiology board examination? Follow-up assessment of a comprehensive set of board examination practice questions. Br J Anaesth 2024 Jan;132(1):172-174. [doi: 10.1016/j.bja.2023.10.025] [Medline: 37996275]
21. Ting YT, Hsieh TC, Wang YF, et al. Performance of ChatGPT incorporated chain-of-thought method in bilingual nuclear medicine physician board examinations. Dig Health 2024 Jan 5;10:20552076231224074. [doi: 10.1177/20552076231224074] [Medline: 38188855]
22. Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese ophthalmology society. Cureus 2023 Dec;15(12):e49903. [doi: 10.7759/cureus.49903] [Medline: 38174202]
23. Revercomb L, Patel AM, Choudhry HS, Filimonov A. Performance of ChatGPT in otolaryngology knowledge assessment. Am J Otolaryngol 2024;45(1):104082. [doi: 10.1016/j.amjoto.2023.104082] [Medline: 37862879]
24. Ariyaratne S, Jenko N, Mark Davies A, Iyengar KP, Botchu R. Could ChatGPT pass the UK radiology fellowship examinations? Acad Radiol 2024 May;31(5):2178-2182. [doi: 10.1016/j.acra.2023.11.026] [Medline: 38160089]
25. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 1;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
26. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. JAMA Netw Open 2023 Dec 1;6(12):e2346721. [doi: 10.1001/jamanetworkopen.2023.46721] [Medline: 38060223]
27. Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open 2023 Nov 2;5(2):e000530. [doi: 10.1136/bmjno-2023-000530] [Medline: 37936648]
28. Seghier ML. ChatGPT: not all languages are equal. Nature New Biol 2023 Mar 9;615(7951):216-216. [doi: 10.1038/d41586-023-00680-3]
29. El español: una lengua viva informe [Article in Spanish]. Centro Virtual Cervantes. 2023. URL: https://cvc.cervantes.es/lengua/anuario/anuario_23/informes_ic/p01.htm [accessed 2024-10-23]
30. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 2023 Dec 13;13(1):22129. [doi: 10.1038/s41598-023-49483-6] [Medline: 38092821]
31. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]

Abbreviations

AI: artificial intelligence
API: application programming interfaces
LLM: large language model
USMLE: United States medical licensing examination

Edited by B Lesselroth; submitted 26.01.24; peer-reviewed by G Chisci, M Chatzimina, M Rizvi, P Gupta; revised version received 29.07.24; accepted 07.10.24; published 14.11.24.

Please cite as:
Ros-Arlanzón P, Perez-Sempere A
Evaluating AI Competence in Specialized Medicine: Comparative Analysis of ChatGPT and Neurologists in a Neurology Specialist Examination in Spain
JMIR Med Educ 2024;10:e56762
URL: https://mededu.jmir.org/2024/1/e56762
doi:10.2196/56762

© Pablo Ros-Arlanzón, Angel Perez-Sempere. Originally published in JMIR Medical Education (https://mededu.jmir.org), 14.11.2024. 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.

XSL•FO
RenderX

Leveraging Open-Source Large Language Models for Data Augmentation in Hospital Staff Surveys: Mixed Methods Study

Carl Ehrett[1,*], PhD; Sudeep Hegde[2,*], PhD; Kwame Andre[3,*]; Dixizi Liu[2,*], MSc; Timothy Wilson[2,*]

1

2

3

[*]all authors contributed equally

Corresponding Author:
Carl Ehrett, PhD

Abstract

Background: Generative large language models (LLMs) have the potential to revolutionize medical education by generating tailored learning materials, enhancing teaching efficiency, and improving learner engagement. However, the application of LLMs in health care settings, particularly for augmenting small datasets in text classification tasks, remains underexplored, particularly for cost- and privacy-conscious applications that do not permit the use of third-party services such as OpenAI's ChatGPT.

Objective: This study aims to explore the use of open-source LLMs, such as Large Language Model Meta AI (LLaMA) and Alpaca models, for data augmentation in a specific text classification task related to hospital staff surveys.

Methods: The surveys were designed to elicit narratives of everyday adaptation by frontline radiology staff during the initial phase of the COVID-19 pandemic. A 2-step process of data augmentation and text classification was conducted. The study generated synthetic data similar to the survey reports using 4 generative LLMs for data augmentation. A different set of 3 classifier LLMs was then used to classify the augmented text for thematic categories. The study evaluated performance on the classification task.

Results: The overall best-performing combination of LLMs, temperature, classifier, and number of synthetic data cases is via augmentation with LLaMA 7B at temperature 0.7 with 100 augments, using Robustly Optimized BERT Pretraining Approach (RoBERTa) for the classification task, achieving an average area under the receiver operating characteristic (AUC) curve of 0.87 (SD 0.02; ie, 1 SD). The results demonstrate that open-source LLMs can enhance text classifiers' performance for small datasets in health care contexts, providing promising pathways for improving medical education processes and patient care practices.

Conclusions: The study demonstrates the value of data augmentation with open-source LLMs, highlights the importance of privacy and ethical considerations when using LLMs, and suggests future directions for research in this field.

(JMIR Med Educ 2024;10:e51433) doi:10.2196/51433

KEYWORDS

data augmentation; large language models; medical education; natural language processing; data security; ethics; AI; artificial intelligence; data privacy; medical staff

Introduction

Overview

Generative large language models (LLMs) are powerful technologies that leverage machine learning techniques to generate novel and contextually relevant content. By training on vast amounts of data, LLMs have the capability to understand and mimic human language patterns, thereby producing text that closely resembles human-written content [1,2]. LLMs represent a subset of generative models characterized by their vast training data and resulting complexity. With billions of parameters, LLMs such as GPT-3 and GPT-4 by OpenAI are capable of generating text that is often indistinguishable from human-written content, provided a suitable context is given (OpenAI) [3].

The use of LLMs has the potential to address critical challenges in medical education. In environments where teaching resources are limited, these models can generate learning materials from case studies to interactive dialogues that align with specific learning objectives and target specific topics [4,5]. Furthermore, they can create diverse and complex patient scenarios that can supplement lecture content by providing real-time clarifications, and context to complex topics, ensuring a deeper understanding for students [6]. By leveraging the capabilities of LLMs, educators can identify content gaps, ensure comprehensive coverage of essential subjects, and ultimately enhance the quality and effectiveness of medical education [7,8]. These models can

XSL•FO
RenderX

enhance teaching efficiency and learner engagement, thereby potentially improving learning outcomes.

LLMs, however, pose several challenges in their application in medical education. Ethical use and privacy concerns need to be considered, especially when using real-world data for training. Cost concerns might arise due to the computational resources needed for training and fine-tuning these models. GPT-3 and GPT-4 (and the product ChatGPT which is built upon them) are closed-source models owned by OpenAI; their use thus not only comes at a financial cost, but also generates privacy concerns due to needing to expose one's data to a third-party company. While open-source LLMs exist, relatively little attention has been paid to their utility, despite the fact that they alleviate both cost and privacy concerns attached to the use of commercial LLMs.

The application of these models in medical education is becoming increasingly prevalent. For instance, LLMs have been used to aid revolutionizing medical curriculum development [8,9], teaching methodologies [10], personalized study plans and learning materials [11], assessments and evaluation [12,13], medical writing and assistance [14,15], and medical research and literature review [16,17]. The vast potential of these technologies opens up novel avenues for educating the future generation of health care providers.

Over the past few decades, self-reported data from health care workers, such as incident reports, have been applied to medical education in many health care areas. These include analyzing potential ethical conflicts within hospitals [18], evaluating Bendamustine-related skin disorders [19], finding predictive patterns of human contributing factors in radiation therapy [20], and improving patient safety and care [21-23]. Despite these applications, LLMs have so far, not been used in the analysis of self-reports. There is an opportunity to leverage hospital self-reports to enhance medical education.

Integrating artificial intelligence (AI) and LLMs in self-report analysis has the potential to revolutionize bottom-up learning from worker-generated data, facilitating more efficient and accurate identification of workflow challenges, systemic issues, strategies and tactics to address these, and areas for improvement in clinical decision-making and patient care. This study addresses the use of LLMs, particularly open-source LLMs, to mitigate a specific problem encountered in the analysis of hospital staff survey data: the lack of ample training data for a text classification task. This task involves classifying text responses into categories based on their relevance to the availability of resources in the hospital. Insufficient training data can limit the model's ability to learn and make accurate classifications.

The objective of this study is to evaluate the effectiveness of using open-source LLMs for data augmentation in this text classification task. By generating synthetic survey responses, LLMs can potentially increase the size and diversity of the training dataset, leading to improved model performance in text classification. Text classification, in turn, is a useful way to analyze free-text reports for categories and themes that are relevant from an educational standpoint. In our research, text classification is used to identify valuable insights from self-reported narratives of the lived experiences of frontline health care workers. Identifying such patterns and capabilities that are situated in the context of everyday work, can be valuable in generating teachable content for medical education. Doing so with augmented data would allow for a richer dataset of realistic learning instances based on everyday work. This paper presents a case study of this approach, aiming to provide insights and guidance for similar applications in medical education and health care operations.

Related Work

Data augmentation is the process of generating new data from existing data. This process is generally used to increase small datasets or create more diversity in a dataset where underrepresented populations are ignored by the model. A lack of diversity in a language model's training set can lead to poor generalizability. For example, LLMs performing numerical reasoning perform better on tasks with terms seen frequently during training, with a gap in accuracy of up to 70% when solving problems containing terms common in the training data as opposed to rare terms [9]. Few-shot learning is generally used with data augmentation because of the lack of usable data and its ability to be efficient with small amounts of data. LLMs are either used to slightly change examples to create new data or generate new data from examples. Using methods that change specific words in slot filling fill in the blank [24], where those words are switched with a semantically similar word, is a widely used method to change existing data slightly. Generative models typically use fine-tuned versions of LLMs [25,26] with prompts, including select examples from the dataset and the label the model is supposed to generate. Zero-shot prompting has also been used with ChatGPT [27,28] in low-resource situations. Models that have received no fine-tuning have also been shown to perform well [29] train an intent classifier, and feed it into the LLM to generate data. Human-in-the-loop studies have been shown to be successful. A human expert filters through generated data and discards generated data that deviate from the training data [30]. Another filtering technique is using a binary sentence classifier to determine whether the original and the augment are semantically similar. We expand the existing literature in this space by exploring the case of low-resource data augmentation in the face of cost and privacy concerns that prevent relying on third-party services such as OpenAI's ChatGPT, using few-shot prompting on open-source LLMs.

Methods

Ethical Considerations

This research study does not require institutional review board approval according to Clemson guidelines [31]. The project involves analysis of a preexisting anonymized dataset, and thus does not constitute research involving human participants as outlined in the federal regulations [45 CFR 46.102(e)]. Our research involves neither obtaining information through intervention or interaction with living individuals, nor the use, study, analysis, or generation of identifiable private information. This type of secondary data analysis, where the researchers do not have access to identifying information, is not considered

XSL•FO
RenderX

research involving human participants and therefore does not require institutional review board oversight.

LLMs for Data Augmentation

We use LLMs for data augmentation, specifically focusing on Large Language Model Meta AI (LLaMA) and Alpaca models. LLaMA is a collection of foundation language generation models with varying complexities ranging from 7 billion (7B) to 65 billion (65B) parameters, introduced by [10]. These models were trained on approximately 1.4 trillion (1.4T) tokens, an extensive dataset derived entirely from publicly accessible sources, thus eliminating dependency on proprietary databases and increasing transparency. The models themselves are open-source and freely available to researchers.

In terms of their architecture, LLaMA models are built on the transformer architecture [32] and incorporate several advancements proposed in recent research, including prenormalization [33] for improved training stability, the SwiGLU activation function [34] for enhanced performance, and rotary embeddings [35] for improved positional encoding. Notably, even at a comparatively smaller scale, LLaMA models are competitive with GPT-3 (175B) in a wide variety of benchmarks. Their combination of small size (and corresponding computational accessibility) with competitive performance, in conjunction with their status as open-source, motivated our choice to focus our work on these models.

Alongside the LLaMA models, we use a set of Alpaca models in our experiments. These models are LLaMA models that have been fine-tuned by Taori et al [36] for instruction-following tasks using a 52K dataset consisting of instructions and corresponding text responses. We include Alpaca models in order to investigate whether this instruction fine-tuning step might make the models more adept at data augmentation tasks. All of the models used in our study were sourced from Huggingface's library.

LLMs for Classification

Robustly Optimized BERT Pretraining Approach (RoBERTa), XLNet, and DistilBERT (Distilled BERT) are all LLMs that have been pretrained on a large corpus of text data. They can be fine-tuned for a variety of tasks, including text classification, natural language inference, and question answering.

RoBERTa stands for "Robustly Optimized BERT Pretraining Approach" [37]. It is a BERT-based model that has been trained on a larger corpus of text and with more training steps than the original BERT model, with a modified training objective. This makes RoBERTa more accurate than the original BERT on a variety of tasks.

XLNet [38] is a transformer-based model that has been trained on a corpus of text that has been masked and shuffled. This makes XLNet more robust to noise and errors in the training data than other LLMs. While both XLNet and RoBERTa are transformer-based language models, the key difference is in their training methods. RoBERTa is a variant of BERT using dynamic masking and longer training on larger amounts of data. In contrast, XLNet uses a permutation-based training approach where all permutations of words in a sentence are considered during prediction, with the goal of providing a more comprehensive contextual understanding.

DistilBERT [39] is a smaller version of BERT that has been trained to have the same performance as BERT on a variety of tasks. DistilBERT is faster and uses less memory than BERT, making it a more practical choice for many applications due to its smaller size.

In this paper, we use RoBERTa, XLNet, and DistilBERT as text classifiers to test the effectiveness of our data augmentation. For each generative LLM used for data augmentation, we use synthetic data to supplement our real survey responses. Thus, after augmentation, we have a larger dataset of text documents, each of which is associated with a label: "resource" or "nonresource" related. The real survey responses are manually labeled by us. The synthetic data cases are (written and) labeled by the LLM. This larger dataset is then used to fine-tune each of the 3 classification models on the task of determining, for a given piece of text, whether it is "resource" or "nonresource" related.

We gather augments (synthetically generated text responses) with different models, temperatures, and training sets. We chose a binary classification scheme of classifying sentences as either "resource"-related, or "non-resource"-related. This choice was based on the fact that a substantial proportion of resilience engineering tools to improve patient safety (RETIPS) reports were found to be related to resources, including the availability of necessary resources, such as staff and equipment. Each time an augment is generated, a new few-shot learning prompt is generated by randomly sampling and concatenating 5 examples each of resource-related and nonresource-related survey responses from our (real) labeled data, displayed in the format shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Prompt template used for few-shot prompting to generate synthetic data. “Category” could be either “resource-related” or “nonresource-related,” depending on which type of data the model is intended to generate.

```
### Instruction:
Here are two lists of short text documents, \
“Resource-related“ and “Nonresource-related”. \
They are survey responses by hospital staff \
at the Children’s Hospital of Philadelphia (CHOP).
“Resource-related” is responses on the topic \
“Availability of resources OR Knowing where to find resources.”
“Nonresource-related” is responses that do not have to do with that topic.
Please give me a new example of a short text document that would belong \
in the “{category}” category.
Please don’t copy or paraphrase the text documents in the \
input lists I give you; instead, come up with your own new example \
that would belong in the “{category}” list.
### Input:
{other_category}:{other_category_examples}
{category}:{category_examples}
### Response:
{category}:
1.
```

After the synthetic data have been generated, it is filtered to retain only those model outputs that include alphabetic characters, since in some cases the model returns an empty output or simply a continuation of the numerical list begun by the “1.” at the end of the prompt (eg, “2. 3. 4. 5.”) These problems were especially common at low temperatures (0.1 and 0.3), leading us to exclude those temperatures from further consideration. For each tested combination of text-generating LLM, classifier LLM, and temperature, the filtered augments from the text-generating LLM at that temperature are combined with the real labeled data and used to fine-tune the classifier LLM, which is then evaluated on a hold-out set of real labeled data. Examples of generated data are shown in Table 1.

XSL•FO
RenderX

Table . Examples of real data and synthetic data from each text-generating large language model.

Model	Responses
Real data	1. How quickly the world changed...I remember the work and practice pre-COVID and can't believe how quickly we went to working remotely and virtually. I was also surprised at how adaptable we really are...we think people are resistant to change but people can change incredibly swiftly wen the world changes and they need to adapt. I was also surprised to learn that there are so many online telemeeting platforms...it is routine to oscillate between 3 - 4 programs in sequential meetings. 2. I believe overall we did exceptionally well. PPE[a] availability and consistency across the enterprise would be an area I think we could improve upon. 3. usage of ppe 4. Virtual conferencing improved our conference participation and virtual review of imaging with referring clinicians using BlueJeans business Skype or FaceTime improved patient Care and physician collaboration. Unfortunately the hospital doesn't have business Skype installed on hospital computers just on lab tops. There needs to be a uniform hospital wide platform. 5. Working remotely required an adjustment in how I communicated about everything. The demands of the pandemic required an adjustment in priorities which put everything on hold except for COVID related work and it required a level of teamwork that did not exist pre-COVID. From a personal coping standpoint I had to "schedule" time away from my computer each day to eat lunch and take a walk.
LLaMA-7B[b]	1. I appreciate the ability to work from home and the support that was provided from my department. I did miss the interaction with coworkers but the ability to work from home was worth it. I was able to keep up with my work and the department was able to 2. I am not aware of the term "telemeeting platforms" but I believe I could answer this question. I am not sure if I was surprised by the need to adapt or the speed at which we adapted. I think we were all expecting the need for changes and adapt 3. It was a challenging time for our department as we did not have enough PPE for everyone and we had to ration supplies for the most essential staff. We were able to support our staff and be very open and fluid with how things were evolving. 4. The ability to go to the bathroom and not have to worry about getting someone else sick 5. Changes in Policies and procedures having all the PPE needed communication with all staff AWA leadership was helpful and tranparent
LLaMA-30B	1. In the beginning working AWA[c] was an adjustment but once I was allowed to bring home my docking station and 2 monitors it was just like working in the office with little or no interruptions. My doctors I supported and I kept in constant contact through phone call 2. We had a lot of staff on site so we could provide a good flow of patients and care. 3. In the beginning working AWA was an adjustment but once I was allowed to bring home my docking station and 2 monitors it was just like working in the office with little or no interruptions. 4. I am confident that I can enter the building safely and that I can maintain the safety of staff, patients and families. Constant communications with the changes is helpful. It is also good that I can look in one place to refresh on any of the topics. 5. I was surprised how quickly we went to working remotely and virtually.

XSL•FO
RenderX

Model	Responses
Alpaca-7B	1. I was surprised at how quickly the world changed from in-person interactions to virtual ones. I also didn't realize how adaptable people could be in terms of adjusting to the new way of working. I was also surprised to learn about the different online telemeeting platforms that are 2. Not having to share Ascom phones and not having an office adjacent to patient care areas. 3. Followed protocols put in place and used PPE and hand sanitizer and bought my own when I felt uncomfortable with there maybe not being enough or "the right kind" for me. 4. How quickly the world changed...I remember the work and practice pre-COVID and can't believe how quickly we went to working remotely and virtually. 5. The hospital's ability to partner with other departments to obtain the supplies and resources needed to ensure AWA roles were fully functional.
Alpaca-30B	1. As a manager, I am grateful for the availability of resources in the workplace. We have a well-stocked supply of PPE and other equipment that we need to do our jobs effectively. Additionally, we have access to a wide range of educational materials and training opportunities. 2. The availability of clinical information and resources were limited during the pandemic, but we were able to navigate through this challenge by establishing a robust telemedicine platform that allowed for remote consultation and collaboration with our colleagues. 3. Ensuring that all staff members are aware of the available resources and are able to easily access them is crucial for maintaining patient safety and reducing stress levels in the workplace. 4. The biggest adjustment for me was how to handle the stress of working with a reduced staff and how to prioritize my workload. 5. The CHOP[d] website and intranet were helpful in providing information about the availability of resources and where to find them.

[a]PPE: personal protective equipment.

[b]LLaMA: Large Language Model Meta AI.

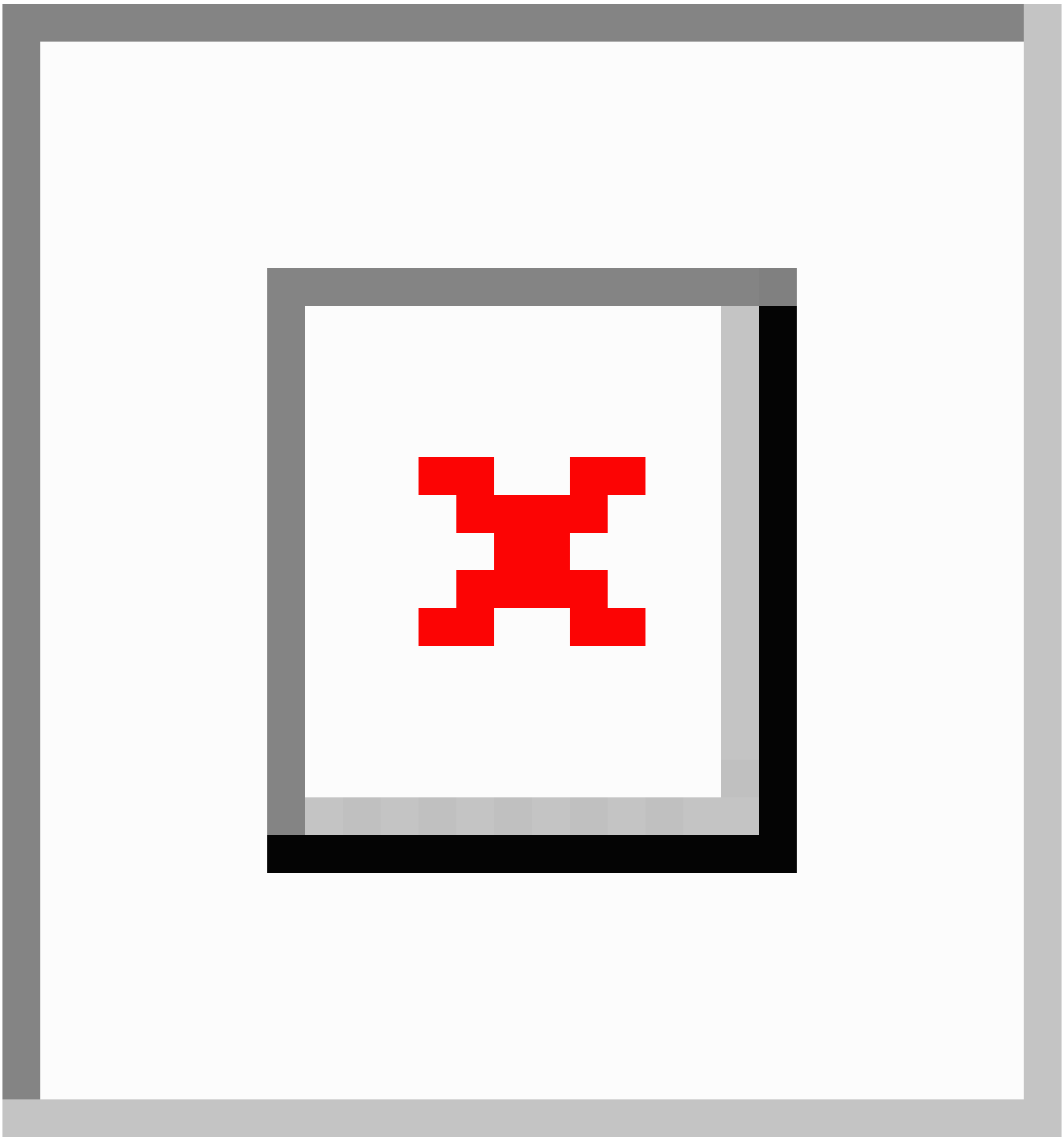
[c]AWA: alternative work arrangements.

[d]CHOP: Children's Hospital of Philadelphia.

Our methodology fits within an envisioned framework represented in Figure 1. Using data from RETIPS and similar self-reported narratives from frontline staff, LLMs with data augmentation will result in a corpus of scenarios. Human stakeholders (eg, educators) can interact with LLMs to compare models, change temperature, and make other adjustments, based on the results, in an iterative manner until the quality and representativeness of augmented data are deemed satisfactory. These scenarios can be used for a variety of purposes in medical education, including quality and safety analysis, creating content for personalized study, and more. Based on their effectiveness in the training, including student and instructor feedback, the LLMs can be further fine-tuned for improvements. The work described here represents a part of this framework, focusing on the development and evaluation of LLMs for data augmentation.

XSL•FO
RenderX

Figure 1. A framework for long-term implementation of LLMs for medical education using RETIPS and similar self-reported data. DistilBERT: Distilled BERT; LLaMA: Large Language Model Meta AI; LLM: large language model; RETIPS: resilience engineering tool to improve patient safety; RoBERTa: Robustly Optimized BERT Pretraining Approach.



Results

We analyze the performance of 4 distinct LLMs—LLaMA-7B, LLaMA-30B, Alpaca-7B, and Alpaca-30B—for the purpose of data augmentation. The goal of the augmentation was to increase the performance of downstream classifiers on the task of matching human labelers' categorization of the text data as "resource" or "nonresource" related. We thus evaluate the quality of the resulting data augmentation by adding the synthetic data from these LLMs to the data used to train 3 classifiers: DistilBERT, RoBERTa, and XLNet. We repeat this analysis for 6 different augmentation "temperature" settings ranging from 0.5 to 1.5. The performance of each model-classifier-temperature combination is assessed based on the area under the receiver operating characteristic (AUC) curve, using a holdout set of human-labeled data.

The overall best-performing combination of LLM, temperature, classifier, and number of augments is LLaMA 7B at temperature 0.7 using RoBERTa with 100 augments, with an average AUC of 0.87 (SD 0.02: 1). In addition to achieving the highest absolute performance, the data augmentation is also most beneficial in this case—this augmentation yields the greatest improvement in AUC with respect to the baseline performance of that classifier model with no data augmentation. The baseline

XSL•FO
RenderX

performance of each classifier along with optimal data augmentation for each text-generating LLM is shown in Table 2. Note that the fine-tuned Alpaca models do not outperform the LLaMA models upon which they are based, indicating that instruction-finetuning is not necessary for this data augmentation task.

Table . Comparison of classifier performance under augmentation by each text-generating large language model (LLM), alongside base performance of the classifier with no augmentation. Each entry gives the mean classifier area under the receiver operating characteristic curve (SD 1), the optimal temperature for text generation.

LLM	RoBERTa[a], mean (SD)	XLNet, mean (SD)	DistilBERT[b], mean (SD)
LLaMA-7B	0.87 (0.02/0.7/100)	0.84 (0.04/0.7/100)	0.83 (0.02/0.7/100)
LLaMA-30B	0.87 (0.03/0.5/100)	0.84 (0.03/0.5/100)	0.85 (0.06/1.5/500)
Alpaca-30B	0.86 (0.06/0.7/100)	0.84 (0.05/1.3/250)	0.81 (0.06/1.3/250)
Alpaca-7B	0.86 (0.06/0.7/100)	0.84 (0.05/1.1/250)	0.82 (0.05/0.7/100)
Baseline	0.80 (0.06)	0.79 (0.06)	0.79 (0.04)

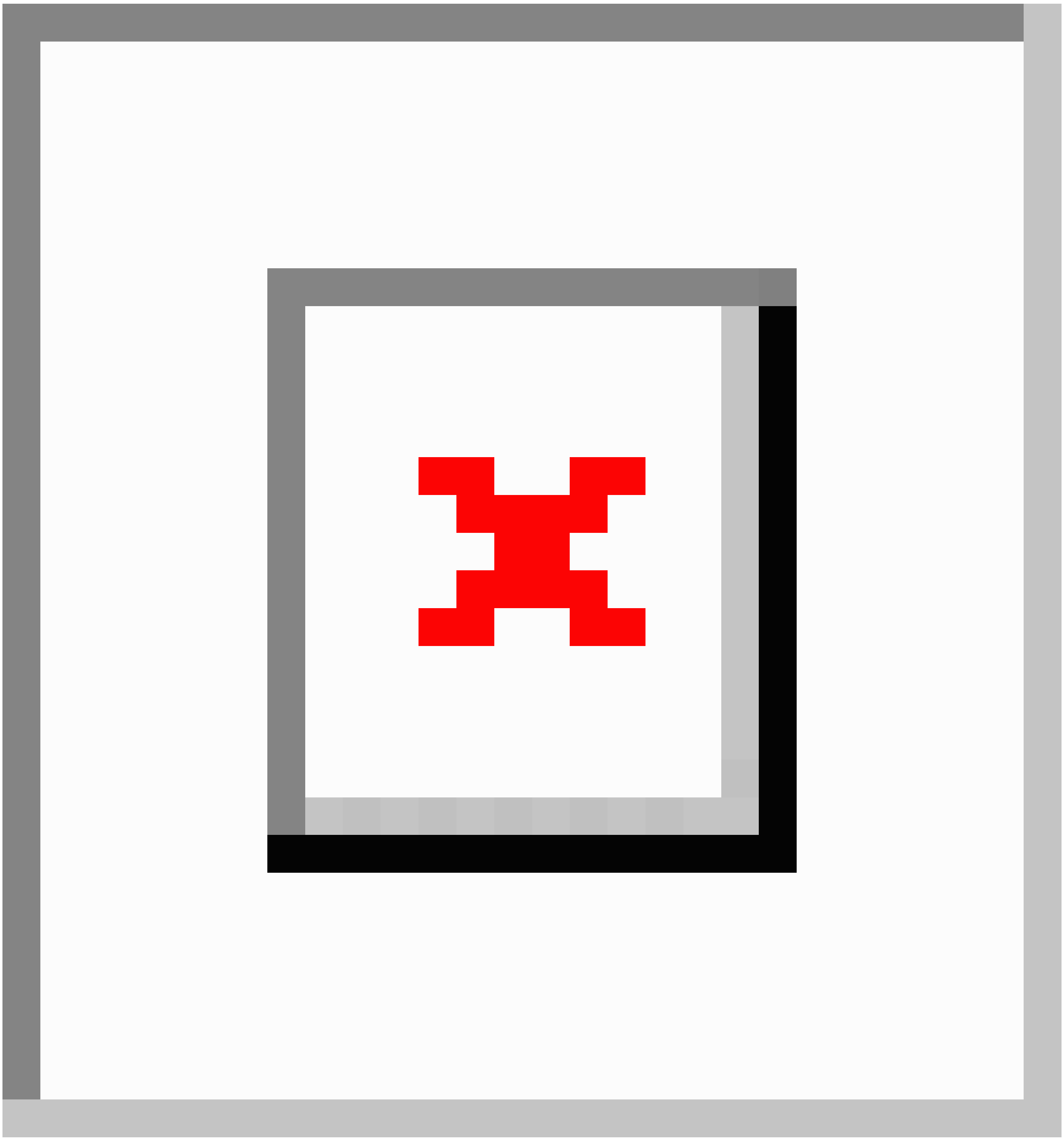
[a]RoBERTa: Robustly Optimized BERT Pretraining Approach.

[b]DistilBERT: Distilled BERT.

For each combination of LLM and classifier, we also fit a linear regression model to explore the relationship between the number of synthetic data points included in the training dataset and the resulting classifier performance as measured by AUC. Notably, DistilBERT emerges as the classifier benefitting most often from data augmentation. In terms of the temperature setting, most of the successful models used a temperature of 0.7. In Figure 2, we display the comparative performance of the LLaMA 7B and Alpaca 7B models, both using the DistilBERT classifier at a temperature setting of 0.7. These graphical representations underscore the beneficial impact of data augmentation on the AUC performance of these specific model-classifier configurations.

XSL•FO
RenderX

Figure 2. Linear fit to DistilBERT (Distilled BERT) model performance (measured as area under the receiver operating characteristic curve) as a function of the number of augments included in the training data (all generated at temperature 0.7).



Discussion

Principal Findings

In this work, our emphasis is on leveraging open-source language models that strike a balance between computational performance and accessibility for researchers. We therefore set a parameter ceiling of 30 billion parameters for several reasons. First, maintaining this threshold ensures that the models in question can run on consumer-level hardware commonly available to average researchers without the need for prohibitive investment in specialized equipment. Second, this approach aligns with our goal to propose methods feasible for environments where privacy and cost considerations limit the use of third-party cloud-based computing services, as relying on external infrastructures (such as OpenAI's services) could elevate privacy risks and regulatory complexity. Using consumer-level hardware, as opposed to cloud-based services, significantly mitigates the risk of data breaches or unauthorized access. Furthermore, the choice to avoid third-party computational services also avoids potential issues related to data sovereignty and control, which could arise when data leaves the institution's local environment. By strictly using in-house resources that operate within the confines of consumer-level capabilities, our methodology facilitates stringent data custody and integrity controls.

XSL•FO
RenderX

LLMs such as OpenAI's GPT-3.5 and GPT-4, or the openly accessible 176 billion-parameter BLOOM, indeed offer more powerful capabilities, but their deployment would threaten the objective of presenting a methodology that is both privacy-aware and broadly implementable. We contend that models up to 30 billion parameters offer a sweet spot, considering these constraints, without significantly compromising the efficacy of the data augmentation process. By imposing a limit of 30 billion parameters, we aim to demonstrate that effective data augmentation for small-scale text classification tasks in the health care sector can be achieved without resorting to the most computationally demanding or privacy-compromising technology. This parameter threshold also allows for an equitable comparison of language models, ensuring that our results are relevant to a wide range of researchers, including those who might be limited by resource constraints. Our research thus serves to bridge the gap typically present in medical informatics research, where smaller institutions or individual researchers may not have access to the same level of computational resources as their larger counterparts. Additionally, this study sheds light on the possibilities and limitations inherent to working within such constraints, providing a valuable reference for future research endeavors seeking a similar balance between model size, privacy, cost, and performance.

Limitations and Future Work

One of the primary constraints of this work is the limited size of the RETIPS dataset. The small sample size (58 responses) potentially affects the reliability and generalizability of the study to other cases where larger data are available. However, it should be noted that the data from RETIPS were tightly focused on a narrow set of themes. This may be beneficial to the quality of augmented data, when compared with a dataset that is thematically more "scattered" or heterogenous. Larger data would likely improve the data augmentation quality but would potentially limit the benefits to be derived from data augmentation. Since data augmentation is of greatest value when working with small datasets, our small data size helps explore this problem space.

Aside from data size, another limitation of this work is that our data are exclusively collected from RETIPS surveys administered to radiology staff at a single hospital. Though this specificity is necessary for the research's objectives, the models' performance may vary in other health care domains, and in domains outside of medicine.

As a future step for this research, it would be beneficial to perform similar studies using larger and more diverse datasets. Larger datasets could provide a richer, more diverse range of training data, potentially leading to correspondingly more diverse synthetic data. Such diversity could improve the performance of the downstream classifiers.

Future research should also consider experimenting with different LLMs for data augmentation. Text-generation LLMs are rapidly evolving, particularly in the area of making large and powerful LLMs accessible on consumer-grade computer hardware [40]. Newer models often come with architectural and training improvements that could potentially enhance synthetic data generation quality.

In this work, we explore different LLM text generation temperature settings and the resulting impact on synthetic data quality. A more extensive hyperparameter tuning of the language models and classifiers may yield further improvements in their performance. This could be a fruitful area for further investigation.

The limitations faced by this work are those that inherently attach to working with small data in a highly privacy-conscious environment with accessible AI tools. Since this is a problem space occupied by many researchers and practitioners in the health care domain, we hope that our results are able to provide insight into how AI tools can be used in such settings.

Conclusion

This study provides an exploration and practical demonstration of the application of LLMs for data augmentation in the context of health care. We specifically focused on the use of open-source LLMs, namely, LLaMA and Alpaca models, to mitigate the challenge of limited training data in a text classification task related to hospital staff surveys.

Our findings demonstrate the potential effectiveness of using LLMs to generate synthetic survey responses, thereby increasing the diversity and size of the training dataset and improving the performance of models trained on the augmented dataset for tasks such as text classification. However, the effectiveness of the data augmentation process can vary based on certain factors such as the specific LLM used, the selected parameters such as the temperature setting, and the downstream classifier applied.

This study provides preliminary evidence that open-source LLMs can improve the performance of text classifiers for small datasets in health care contexts when privacy or cost considerations prevent the use of closed-source third-party services such as those offered by OpenAI. These results pave the way for future research to further investigate and refine the use of LLMs in tasks like text classification, data augmentation, and other medical education and operational applications.

This research serves as an initial leap towards exploiting the promising capabilities of LLMs in medical applications while being mindful of privacy, ethical concerns, and constraints associated with this field. By establishing a proof-of-concept for the use of open-source LLMs in health care settings, this study opens avenues for broader exploration of LLMs' potential to tackle numerous challenges faced by medical practitioners, educators, and administrators. Future research should expand on our work by exploring more complex datasets, experimenting with different hyperparameters for a wider variety of LLMs, and developing procedures to systematically craft and evaluate prompts to optimize model output.

This study contributes to a burgeoning field of research exploring applications of AI in health care and medical education. Our exploration of data augmentation using open-source LLMs presents potential pathways for improving processes such as incident reporting, resident evaluation, clinical vignette development, and other text-based processes relevant

XSL•FO
RenderX

to medical education. This research will hopefully encourage additional exploration into the ethical and judicious application of LLMs and other AI technologies in health care.

As AI technologies continue to evolve and become more sophisticated, constant re-evaluation and updates to our methods will be essential. Therefore, active engagement from all stakeholders in the medical field, including frontline health care workers, researchers, educators, and policy makers, is crucial in making the most of these advancements for the betterment of patient care and safety.

Acknowledgments

Clemson University is acknowledged for its generous allotment of compute time on the Palmetto Cluster.

Conflicts of Interest

None declared.

References

1. Kurian N, Cherian JM, Sudharson NA, Varghese KG, Wadhwa S. AI is now everywhere. Br Dent J 2023 Jan;234(2):72-72. [doi: 10.1038/s41415-023-5461-1] [Medline: 36707552]
2. Teubner T, Flath CM, Weinhardt C, van der Aalst W, Hinz O. Welcome to the era of ChatGPT et al. Bus Inf Syst Eng 2023 Apr;65(2):95-101. [doi: 10.1007/s12599-023-00795-x]
3. OpenAI. URL: https://openai.com/ [accessed 2023-07-30]
4. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. Radiology 2023 Apr;307(2):e230171. [doi: 10.1148/radiol.230171] [Medline: 36728749]
5. Masters K. Response to: aye, AI! ChatGPT passes multiple-choice family medicine exam. Med Teach 2023 Jun 3;45(6):666-666. [doi: 10.1080/0142159X.2023.2190476]
6. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 7;8(2):e35587. [doi: 10.2196/35587] [Medline: 35671077]
7. Wang LKP, Paidisetty PS, Cano AM. The next paradigm shift? ChatGPT, artificial intelligence, and medical education. Med Teach 2023 Apr(8):1. [doi: 10.1080/0142159X.2023.2198663] [Medline: 37036176]
8. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
9. Razeghi Y, Logan IV RL, Gardner M, Singh S. Impact of pretraining term frequencies on few-shot reasoning. arXiv. Preprint posted online on 2022arXiv:2202.07206.
10. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. Preprint posted online on 2023.
11. Frommeyer TC, Fursmidt RM, Gilbert MM, Bett ES. The desire of medical students to integrate artificial intelligence into medical education: an opinion article. Front Digit Health 2022;4:831123. [doi: 10.3389/fdgth.2022.831123] [Medline: 35633734]
12. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb;2(2):e0000205. [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]
13. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
14. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Dig Health 2023 Apr;5(4):e179-e181. [doi: 10.1016/S2589-7500(23)00048-1]
15. Chen TJ. ChatGPT and other artificial intelligence applications speed up scientific writing. J Chin Med Assoc 2023 Apr 1;86(4):351-353. [doi: 10.1097/JCMA.0000000000000900] [Medline: 36791246]
16. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
17. Dahmen J, Kayaalp ME, Ollivier M, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surg Sports Traumatol Arthrosc 2023 Apr;31(4):1187-1189. [doi: 10.1007/s00167-023-07355-6]
18. Wehkamp K, Kuhn E, Petzina R, Buyx A, Rogge A. Enhancing patient safety by integrating ethical dimensions to critical incident reporting systems. BMC Med Ethics 2021 Mar 8;22(1):26. [doi: 10.1186/s12910-021-00593-8] [Medline: 33685473]
19. Uchida M, Kawashiri T, Maegawa N, Takano A, Hosohata K, Uesawa Y. Pharmacovigilance evaluation of Bendamustine-related skin disorders using the Japanese adverse drug event report database. J Pharm Pharm Sci 2021;24:16-22. [doi: 10.18433/jpps31597]
20. Weintraub SM, Salter BJ, Chevalier CL, Ransdell S. Human factor associations with safety events in radiation therapy. J Appl Clin Med Phys 2021 Oct;22(10):288-294. [doi: 10.1002/acm2.13420] [Medline: 34505353]

XSL•FO
RenderX

21. Goekcimen K, Schwendimann R, Pfeiffer Y, Mohr G, Jaeger C, Mueller S. Addressing patient safety hazards using critical incident reporting in hospitals: a systematic review. J Patient Saf 2023 Jan 1;19(1):e1-e8. [doi: 10.1097/PTS.0000000000001072] [Medline: 35985209]
22. San Jose-Saras D, Valencia-Martín JL, Vicente-Guijarro J, Moreno-Nunez P, Pardo-Hernández A, Aranaz-Andres JM. Adverse events: an expensive and avoidable hospital problem. Ann Med 2022 Dec;54(1):3157-3168. [doi: 10.1080/07853890.2022.2140450] [Medline: 36369717]
23. Dillner P, Eggenschwiler LC, Rutjes AWS, et al. Incidence and characteristics of adverse events in paediatric inpatient care: a systematic review and meta-analysis. BMJ Qual Saf 2023 Mar;32(3):133-149. [doi: 10.1136/bmjqs-2022-015298] [Medline: 36572528]
24. Louvan S, Magnini B. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. Preprint posted online on 2020.
25. Edwards A, Ushio A, Camacho-Collados J, Ribaupierre H, Preece A. Guiding generative language models for data augmentation in few-shot text classification. Preprint posted online on 2021.
26. Saeedi D, Saeedi S, Panahi A, C.M. Fong A. CS/NLP at SemEval-2022 task 4: effective data augmentation methods for patronizing language detection and multi-label classification with RoBERTa and GPT3. Presented at: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022); Seattle, United States p. 503-508. [doi: 10.18653/v1/2022.semeval-1.69]
27. Ubani S, Polat SO, Nielsen R. ZeroShotDataAug: generating and augmenting training data with ChatGPT. Published online. Preprint posted online on 2023.
28. Møller AG, Dalsgaard JA, Pera A, Aiello LM. Is a prompt and a few samples all you need? using GPT-4 for data augmentation in low-resource classification tasks. Preprint posted online on 2023.
29. Sahu G, Rodriguez P, Laradji IH, Atighehchian P, Vazquez D, Bahdanau D. Data augmentation for intent classification with off-the-shelf large language models. Preprint posted online on 2022.
30. Bayer M, Frey T, Fine-Tuning R. Data augmentation, and few-shot learning for specialized cyber threat intelligence. Preprint posted online on 2022.
31. What needs review? Clemson University Office of Research Compliance. 2024. URL: https://www.clemson.edu/research/division-of-research/offices/orc/irb/whatneedsreview.html [accessed 2024-11-14]
32. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Preprint posted online on 2017.
33. Zhang B, Sennrich R. Root mean square layer normalization. Preprint posted online on 2019.
34. Shazeer N. GLU variants improve transformer. Preprint posted online on 2020.
35. Su J, Lu Y, Pan S, Murtadha A, Wen B, Liu Y. RoFormer: enhanced transformer with rotary position embedding. Preprint posted online on 2021.
36. Taori R, Gulrajani I, Zhang T, et al. Alpaca: a strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. URL: https://crfm.stanford.edu/2023/03/13/alpaca.html [accessed 2024-11-14]
37. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint posted online on 2019.
38. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR. XLNet: generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst 2019 [FREE Full text]
39. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Preprint posted online on 2019.
40. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. Preprint posted online on 2023.

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic
DistilBERT: Distilled BERT
LLaMA: Large Language Model Meta AI
LLM: large language model
RETIPS: resilience engineering tool to improve patient safety
RoBERTa: Robustly Optimized BERT Pretraining Approach

XSL•FO
RenderX

Edited by B Lesselroth; submitted 31.07.23; peer-reviewed by D Dongelmans, M Hussain; revised version received 09.02.24; accepted 15.08.24; published 19.11.24.

Please cite as:
Ehrett C, Hegde S, Andre K, Liu D, Wilson T
Leveraging Open-Source Large Language Models for Data Augmentation in Hospital Staff Surveys: Mixed Methods Study
JMIR Med Educ 2024;10:e51433
URL: https://mededu.jmir.org/2024/1/e51433
doi:10.2196/51433

© Carl Ehrett, Sudeep Hegde, Kwame Andre, Dixizi Liu, Timothy Wilson. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.11.2024. 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.

XSL•FO
RenderX

Influence of Model Evolution and System Roles on ChatGPT's Performance in Chinese Medical Licensing Exams: Comparative Study

Shuai Ming[1,2,3], DM; Qingge Guo[1,2,3], MD; Wenjun Cheng[4], MD; Bo Lei[1,2,3], MD, PhD

1

2

3

4

Corresponding Author:
Bo Lei, MD, PhD

Abstract

Background: With the increasing application of large language models like ChatGPT in various industries, its potential in the medical domain, especially in standardized examinations, has become a focal point of research.

Objective: The aim of this study is to assess the clinical performance of ChatGPT, focusing on its accuracy and reliability in the Chinese National Medical Licensing Examination (CNMLE).

Methods: The CNMLE 2022 question set, consisting of 500 single-answer multiple choices questions, were reclassified into 15 medical subspecialties. Each question was tested 8 to 12 times in Chinese on the OpenAI platform from April 24 to May 15, 2023. Three key factors were considered: the version of GPT-3.5 and 4.0, the prompt's designation of system roles tailored to medical subspecialties, and repetition for coherence. A passing accuracy threshold was established as 60%. The χ^2 tests and κ values were employed to evaluate the model's accuracy and consistency.

Results: GPT-4.0 achieved a passing accuracy of 72.7%, which was significantly higher than that of GPT-3.5 (54%; *P*<.001). The variability rate of repeated responses from GPT-4.0 was lower than that of GPT-3.5 (9% vs 19.5%; *P*<.001). However, both models showed relatively good response coherence, with κ values of 0.778 and 0.610, respectively. System roles numerically increased accuracy for both GPT-4.0 (0.3% - 3.7%) and GPT-3.5 (1.3% - 4.5%), and reduced variability by 1.7% and 1.8%, respectively (*P*>.05). In subgroup analysis, ChatGPT achieved comparable accuracy among different question types (*P*>.05). GPT-4.0 surpassed the accuracy threshold in 14 of 15 subspecialties, while GPT-3.5 did so in 7 of 15 on the first response.

Conclusions: GPT-4.0 passed the CNMLE and outperformed GPT-3.5 in key areas such as accuracy, consistency, and medical subspecialty expertise. Adding a system role insignificantly enhanced the model's reliability and answer coherence. GPT-4.0 showed promising potential in medical education and clinical practice, meriting further study.

(JMIR Med Educ 2024;10:e52784) doi:10.2196/52784

KEYWORDS

ChatGPT; Chinese National Medical Licensing Examination; large language models; medical education; system role; LLM; LLMs; language model; language models; artificial intelligence; chatbot; chatbots; conversational agent; conversational agents; exam; exams; examination; examinations; OpenAI; answer; answers; response; responses; accuracy; performance; China; Chinese

Introduction

ChatGPT, a general large language model (LLM) developed by OpenAI, has gained substantial attention since its launch on November 30, 2022. Known for its advanced natural language processing capabilities, ChatGPT has the potential to make significant impacts on many industries, including medical education. Its performance in medicine was first tested at or near the passing threshold of the United States Medical Licensing Examination (USMLE) [1,2]. While ChatGPT's accuracy varies across languages [3], it has been tested on a series of medical exams like the Japanese National Medical Licensing Examination in languages including English [4], Chinese [5], Dutch [6], Japanese [7], and Korean [8]. The research scope related to ChatGPT has expanded to medical education in fields like nuclear medicine [9], neurosurgery [10], ophthalmology [11], general chemistry, nursing[12], life support [4], dentology [13], and radiation oncology physics [14]. Overall, while ChatGPT demonstrates heterogeneous capabilities, it shows promising potential in these medical specialties.

XSL•FO
RenderX

Several factors might influence ChatGPT's performance. First, the updated version of ChatGPT, GPT-4, understands and generates natural language in more complex and nuanced scenarios, leading to more accurate responses [15], which is important in analyzing complex clinical case questions [16]. Thus, GPT-4 conclusively demonstrated significantly better performance than GPT-3.5, as evidenced by various official medical exams [8]. Besides the model version, ChatGPT allows users to guide its behavior by adding prompts that describe its system role. These system roles influence the direction of ChatGPT's answers and may affect its reliability. However, the impact of these system roles on ChatGPT's performance in medical field has not yet been investigated. As a professional chatbot tool, ChatGPT uses sampling to predict the next token with varying distribution probabilities, ensuring responses are varied and natural in real-world applications. Zhu et al [17] have found that composite answers derived from repeated questioning can enhance the accuracy of ChatGPT. Typically, 2 or 3 repeated responses are necessary to ensure response stability [18-20].

Currently, the peer-reviewed research still lacks highlights on the strength of ChatGPT when it comes to the Chinese National Medical Licensing Examination (CNMLE). This study aimed to evaluate the performance of ChatGPT in answering CNMLE questions in the clinical setting of China, with consideration of the version of ChatGPT and system role.

Methods

The CNMLE 2022 Question Data

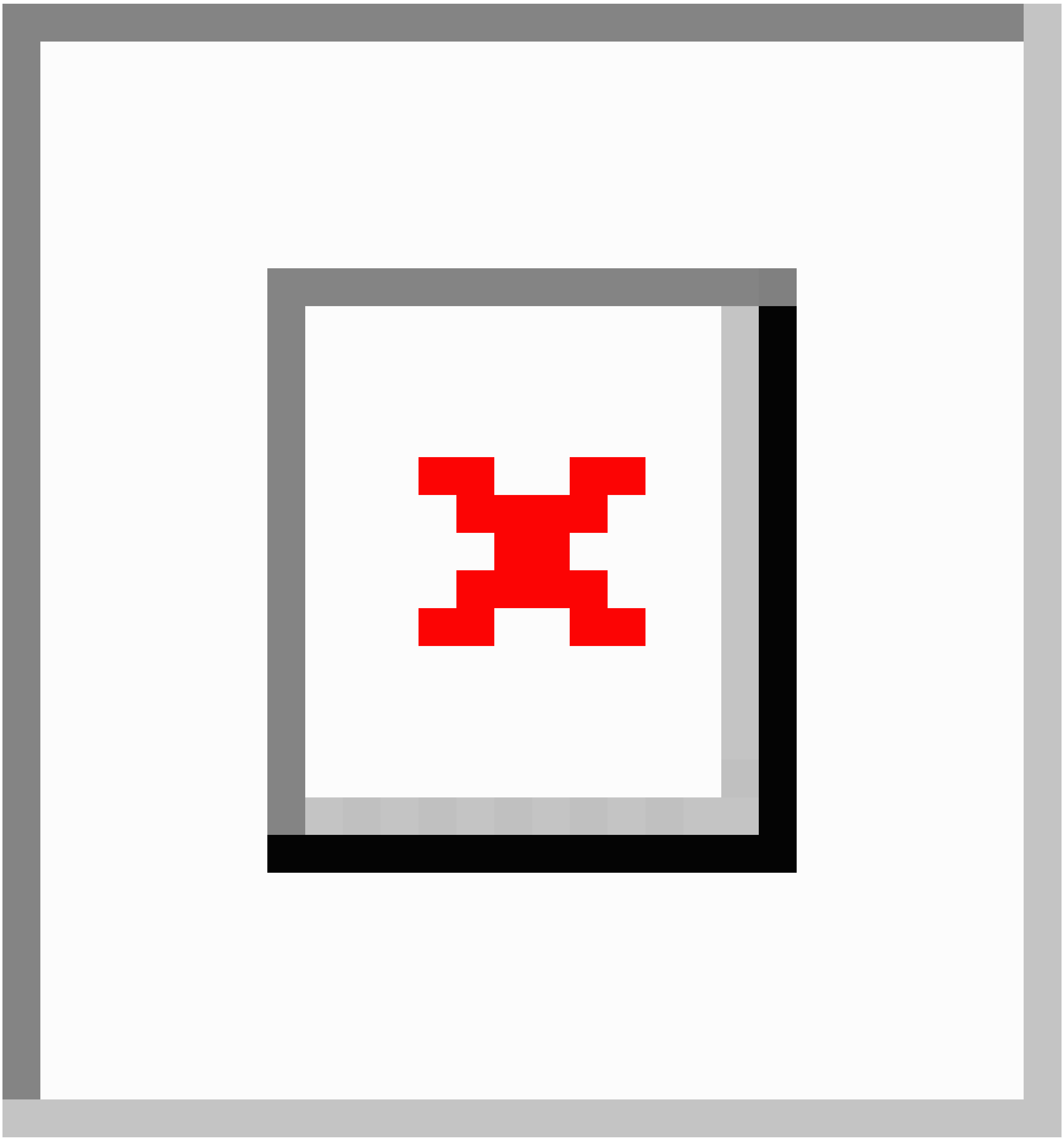
As an industry admission examination, passing the CNMLE means that a medical practitioner meets the minimum medical competencies. The written part of the examination, which emphasizes medical knowledge and clinical decision-making skills, is created and supervised by the Chinese National Medical Examination Center (NMEC). In 2021, the CNMLE transitioned from the traditional paper-based format to a computer-based examination. Each candidate is presented with 600 questions, arranged in a slightly varied order, from the exam year's question data set. According to OpenAI's introduction, ChatGPT's responses are based on information available up to September 2021. Thus, we selected the CNMLE 2022 questions, which were purchased from a web-based bookstore [21], for our evaluation. This choice ensured that the questions had not been previously encountered and trained by the model. The publisher has confirmed that these released questions are the original ones from the examination.

The CNMLE 2022 covered 600 single-answer multiple-choice questions, which were evenly divided into 4 units [22]. Each unit had 4 specific question types: A1, the single-sentence optimal choice questions; A2, case summary optimal choice questions; A3/A4, case group optimal choice questions; and B1, standard combination questions. Detailed explanations of each question type was conveyed to ChatGPT via a structured prompt prior to inquiry (see in Multimedia Appendix 1). The CNMLE 2022 questions did not involve table or image-based questions. Therefore, ChatGPT, despite lacking multimodal capabilities, was still suited to effectively complete the test.

According to the introduction of the Chinese NMEC [22], each examination unit always addresses specific medical subspecialties. Unit 1 covers medical knowledge, policies, regulations, and preventive medicine; unit 2 mainly pertains to the cardiovascular, urinary, musculoskeletal, and endocrine systems; unit 3 involves the digestive, respiratory, and associated systems; unit 4 focuses on obstetrics and gynecology, pediatrics, and neurological or psychiatric domains. However, such distribution is not absolute. Therefore, 2 clinicians independently reclassified the 600 questions into 15 medical subspecialties, resolving discrepancies through discussion. The κ value for the result of their classifications was 0.935. The Sankey diagram of the 3 question classifications, medical subspecialties, units, and types is shown in Figure 1.

XSL•FO
RenderX

Figure 1. The Sankey diagram of the 3 question classifications: the medical subspecialties, units, and types. STD: sexually transmitted disease.



Instructions Before Testing Part

Before manually inputting questions, ChatGPT was informed about an upcoming series of queries. ChatGPT needed to identify the most plausible response from the available options and explain the reasoning behind its selection. The question types determined the relevant lead-in prompts provided. For the A1 and A2 question types, each input question was deemed independent, rendering any interquestion relationships irrelevant. In contrast, A3/A4 question types implied that multiple questions within a single clinical case shared a connection. However, individual clinical cases were treated as discrete entities, eliminating the need to consider relationships between them. For the B1 question type, 5 shared options were given. ChatGPT needed to identify the correct answers for subsequent questions. Chaining was used in A3/A4 and B1 question types to ensure that multiple questions within a single clinical case in A3/A4 shared the same context, and multiple questions in B1 shared the same options. The number of questions inputted at one time depended on the text's length, such as 5 - 8 questions for A1/A2 types. If necessary, ChatGPT was forced to disregard prior conversational content and commence a fresh chat.

Temperature

The temperature parameter in ChatGPT influences the randomness of the model's responses. A higher temperature yields more varied and creative answers. In our study, we did not manually adjust the temperature; instead, we used the default

XSL•FO
RenderX

setting on the OpenAI platform, commonly at 0.7, to simulate real-world user interactions on the front end. This balance between typical user habits and diverse thought processes was intentional. The default relatively high temperature was expected to enable ChatGPT to generate more creative reasoning processes while still arriving at accurate answers.

Testing Strategy

All the CNMLE 2022 questions were tested in Chinese according to the following 2 factors:

1. ChatGPT model selection. Both GPT-3.5 (version from March 23) and GPT-4.0 (version from May 3) were rigorously evaluated on the OpenAI platform from April 24 to May 15, 2023, to ascertain any evolution in the model's capability in the medical domain.
2. System role. This refers to the specific identity or role, such as "gastroenterology specialist," assigned to ChatGPT to determine if relevant knowledge is applied more accurately. Questions were evaluated both with and without assigning a system role related to the 15 specific clinical subspecialties. This system role was designated by providing a tailored system prompt before the testing instructions, aiming to guide ChatGPT's approach and align it with specialist viewpoints in the relevant medical field.

Testing Process

Considering the evaluation of the ChatGPT model, system role, and response coherence, each question was tested 8 - 12 times. The prompts included those specific to question types, the assignment of system role, and the use of chaining. Slight modifications in these prompts were adopted to avoid potential systematic errors introduced by rigid wording. For example, the prompt "Assume you are a gastroenterology specialist" might vary as "Assume you are highly proficient in gastroenterology." For coherence evaluation, each question was presented again to ChatGPT. If the regenerated response matched the initial answer, the process was halted. However, if the 2 responses differed, the question was posed once more to ChatGPT.

Response Determination

The first and second responses from ChatGPT were directly assessed against the given standard answers for accuracy. For the final response (referred to as joint response), if 2 of the 3 answers were consistent, this was taken as the conclusive answer and evaluated against the standard. However, if the 3 responses were all distinct, it was automatically marked as incorrect without any further comparison to the standard answer.

The first response was more applicable to assessing whether ChatGPT could pass the CNMLE in the same situation as a student examinee. In contrast, the joint response represented an overall accuracy (the proportion of questions answered correctly at least twice) [17], which was more suitable for demonstrating the potential of ChatGPT in medical education.

According to the announcement from the CNMLE Committee of the National Health Commission of China, the passing score for licensed physicians is 360 points, which means an accuracy rate of 60% or above is considered a pass.

Statistical Analysis

Data were collected and managed using Excel software. The statistical analyses were conducted with SPSS (version 26.0.0; IBM Corp). A χ^2 test was used to compare the accuracy of CNMLE question responses between different testing strategies and subgroups of question types. Variability was calculated by the number of consistently correct or wrong answers in 2 repeated responses divided by the total number of questions (600). Additionally, the κ statistic was used to evaluate answer consistency. A difference was considered statistically significant when $P<.05$.

Ethical Considerations

This study collected information that was already published in the bookstore and did not involve human subjects; therefore, approval by the Institutional Review Board of Henan Provincial People's Hospital was not required.

Results

Accuracy and System Role Assignment

In model comparison, GPT-3.5 achieved an initial accuracy of 54% (324/600) and did not meet the exam criteria. Conversely, GPT-4.0 achieved a passing accuracy of 72.7% (436/600), which was significantly higher than GPT-3.5 ($P<.001$). Similarly, with a designated system role, GPT-4.0 still exhibited higher accuracy than GPT-3.5 (73% vs 55.3%; $P<.001$).

Upon system role assignment, both GPT-3.5 and GPT-4.0 showed a slight increase in accuracy compared to when no role was assigned; specifically, 55.3% (332/500) from 54% (324/600) for GPT-3.5 ($P>.05$) and 73% (438/600) from 72.7% (436/600) for GPT-4.0 ($P>.05$).

The upper comparisons for the second and joint responses paralleled the initial results, as shown in Table 1.

XSL•FO
RenderX

Table . Accuracy of GPT-4.0 and 3.5 with or without SR designation under repeat tests. n represents the number of correct answers.

Accuracy	GPT-3.5, n (%)	GPT-4.0, n (%)	*P* value	GPT-3.5 + SR[a], n (%)	GPT-4.0 + SR, n (%)	*P* value
IR[b]	324 (54.0)	436 (72.7)	<.001	332 (55.3)	438 (73.0)	<.001
2R[c]	303 (50.5)	426 (71.0)	<.001	310 (51.7)	448 (74.7)	<.001
JR[d]	302 (50.3)	435 (72.5)	<.001	329 (54.8)	437 (72.8)	<.001

[a]SR: system role.
[b]IR: initial response.
[c]2R: second response.
[d]JR: joint response.

Variability of Responses

The GPT-3.5 model exhibited a variability rate of 19.5% (117/600), which decreased to 17.7% (106/600) upon the designation of a system role. The variability rate for GPT-4.0 was observed at 9% (54/600), and further reduced to 7.3% after a system role was assigned. These results indicated a smaller response variability for GPT-4.0 compared to GPT-3.5, and specifying system roles also decreased the variability rates. Both models showed relatively high coherence between the initial and second response, with κ values of 0.778 and 0.610. Detailed information for repeated response can be seen in Multimedia Appendix 2.

Accuracy for Subgroups

For GPT-4.0, when accounting for system role and repeated responses, there was a statistically significant difference in accuracy across the different units for the CNMLE test, with accuracy ranging from 62% (93/150) to 84% (126/150; *P* range from<.001 to .01). However, when grouped by question type, the accuracy ranged from 69.4% (145/209) to 83.1% (59/71) without statistical difference (*P*>.28).

In contrast, for GPT-3.5, only the initial response with system role designation showed a statistical difference in accuracy (*P*=.04) for question type subgroups. In other groupings by unit or question type, as well as in subsequent responses, the accuracy remained without significant variations (*P*>.14; see Table 2).

Accuracy for initial and joint responses of GPT-3.5/4.0 classified by 15 medical subspecialties is shown in Figure 2. In multiple testing strategies, GPT-4.0 outperformed GPT-3.5 in accuracy for 14 distinct clinical subspecialty questions, consistently surpassing the 60% passing threshold.

XSL•FO
RenderX

Table . Subgroup analysis of accuracy for the 4 sections and 4 question types under different strategies. Data were showed as n (%). Units 1 - 4 were the 4 parts to which the questions belonged, and A1-A2, B1 represented the types of questions. Units 1 - 4 corresponded to distinct clinical subspecialties, with specific details provided in the Methods section.

Model strategy	Unit 1 (n=150), n (%)	Unit 2 (n=150), n (%)	Unit 3 (n=150), n (%)	Unit 4 (n=150), n (%)	*P* value	A1 (n=220), n (%)	A2 (n=209), n (%)	A3/A4 (n=100), n (%)	B1 (n=71), n (%)	*P* value
GPT3.5: IR[a]	82 (54.7)	83 (55.3)	88 (58.7)	71 (47.3)	.25	122 (55.5)	109 (52.2)	59 (59.0)	34 (47.9)	.47
GPT3.5: 2R[b]	71 (47.3)	77 (51.3)	85 (56.7)	70 (46.7)	.28	115 (52.3)	103 (49.3)	57 (57.0)	28 (39.4)	.14
GPT3.5: JR[c]	72 (48.0)	75 (50.0)	86 (57.3)	69 (46.0)	.22	114 (51.8)	101 (48.3)	57 (57.0)	30 (42.3)	.24
GPT3.5: IR+ SR[d]	85 (56.7)	84 (56.0)	91 (60.7)	72 (48.0)	.16	129 (58.6)	113 (54.1)	61 (61.0)	29 (40.8)	.04
GPT3.5: 2R + SR	83 (55.3)	74 (49.3)	82 (54.7)	71 (47.3)	.42	121 (55.0)	102 (48.8)	57 (57.0)	30 (42.3)	.15
GPT3.5: JR+ SR	84 (56.0)	80 (53.3)	91 (60.7)	74 (49.3)	.25	126 (57.3)	110 (52.6)	61 (61.0)	32 (45.1)	.16
GPT4.0: IR	102 (68.0)	118 (78.7)	119 (79.3)	97 (64.7)	.006	154 (70.0)	152 (72.7)	79 (79.0)	51 (71.8)	.42
GPT4.0: 2R	100 (66.7)	112 (74.7)	119 (79.3)	95 (63.3)	.009	155 (70.5)	145 (69.4)	76 (76.0)	50 (70.4)	.68
GPT4.0: JR	104 (69.3)	114 (76.0)	121 (80.7)	96 (64.0)	.007	157 (71.4)	146 (69.9)	79 (79.0)	53 (74.6)	.37
GPT4.0: IR+ SR	103 (68.7)	116 (77.3)	126 (84.0)	93 (62.0)	<.001	157 (71.4)	151 (72.2)	72 (72.0)	58 (81.7)	.37
GPT4.0: 2R + SR	104 (69.3)	117 (78.0)	124 (82.7)	103 (68.7)	.01	159 (72.3)	153 (73.2)	77 (77.0)	59 (83.1)	.28
GPT4.0: JR+ SR	101 (67.3)	115 (76.7)	124 (82.7)	97 (64.7)	.001	156 (70.9)	151 (72.2)	73 (73.0)	57 (80.3)	.47

[a]IR: initial response.

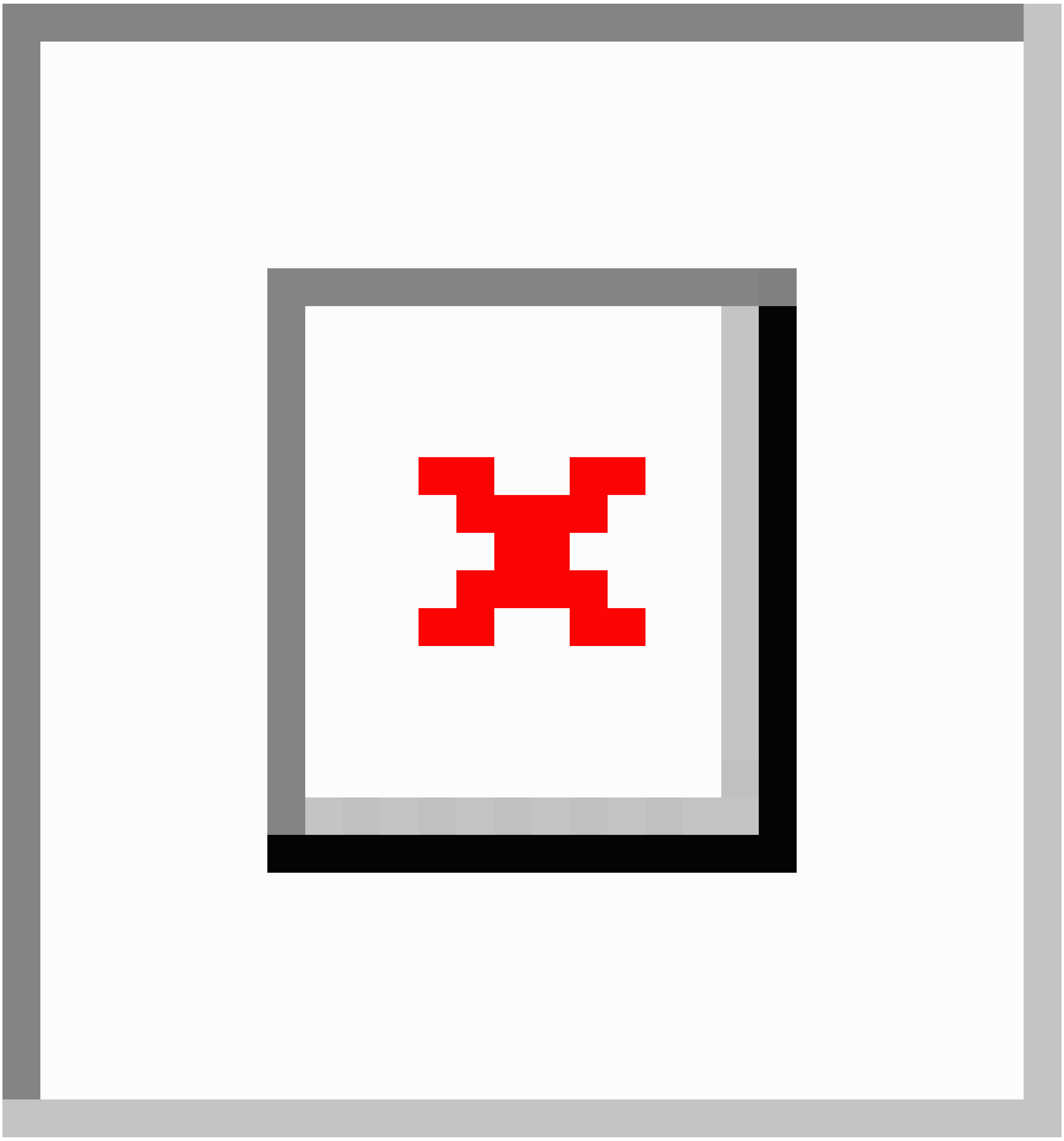
[b]2R: second response.

[c]JR: joint response.

[d]SR: system role.

XSL•FO
RenderX

Figure 2. Accuracy for GPT-3.5/4.0 classified by 15 medical subspecialties. (A) the initial response, (B) the initial response with SR assignation, (C) the joint response, (D) the joint response with SR assignation. SR: system role; STD: sexually transmitted disease.



Discussion

Overview

The CNMLE syllabus outlines the essential knowledge and competencies that physicians need for diagnostic and therapeutic procedures. Acquiring these competencies typically demands that a medical student invest several years in both theoretical education and practical skill development. The application of ChatGPT in medical examinations, particularly within the CNMLE framework, offers a pioneering approach to gauge the potential of LLMs in clinical diagnosis and treatment planning. This study comprehensively assessed ChatGPT's performance in addressing CNMLE questions, focusing on model evolution and system role designation, which has not yet been fully investigated.

Model Evolution and Performance

In our study, GPT-4.0 consistently outperformed GPT-3.5 in accuracy and reliably met the passing criteria set by the CNMLE Committee. Despite GPT-3.5 achieving an accuracy rate of over 50%, it failed to pass the examination. A noncomparison study using GPT-3.5 to test CNMLE 2020-2022 achieved an accuracy of (36.5%-47%) [23]. The lower accuracy might be attributed to the fact that the testing was conducted before February, shortly after the release of GPT-3.5. The better performance of GPT-4.0 compared with GPT-3.5 was also reported by Wang et al [24]. However, it is noteworthy that

XSL•FO
RenderX

their assessment was based on a limited sample of 100 questions, rather than a full set of 600 questions. The small sample might have contributed to the overall favorable results (GPT-4.0: 84%; GPT-3.5: 56%). Therefore, our findings might provide a more representative comparison of the real-world performance of GPT-4.0 and 3.5 on the CNMLE.

Other research on evaluating ChatGPT's accuracy on national medical licensing examinations included assessments of the USMLE [1,2] and the Japanese National Medical Licensing Examination [7]. The conclusions were similar to ours: while GPT-3.5 was often at or near the passing threshold, GPT-4.0 passed relevant exams and had higher testing accuracy compared to GPT-3.5. This trend was not only limited to national medical licensing examinations but also applied to other medical-related examinations. However, the specific accuracy varied across models, possibly due to differences in study countries, testing time, exam content, and other variables. A comprehensive review of existing published and non–peer-reviewed research findings is available in Table 3.

XSL•FO
RenderX

Table . A review of the existing published and non–peer-reviewed research related to ChatGPT performance on medical examinations.

Study	Country	Test model	Examination	Data sample, n	Passing threshold	Accuracy (%)
Gilson et al [1]	United States	GPT-3.5	The United States Medical Licensing Examination Step 1 and Step 2 exams	87 - 102	60%	GPT-3.5: 44.0 - 64.4
Kung et al [2]	United States	GPT-3.5	The United States Medical Licensing Exam	376	60%	At or near 60%
Guerra et al [25]	United States	GPT-4.0 and 3.5	Congress of Neurological Surgeons Self-Assessment Neurosurgery Exam	591	—[a]	GPT-4.0: 76.6; GPT-3.5: 60.2
Takagi et al [7]	Japan	GPT-4.0 and 3.5	Japanese National Medical Licensing Examination (2023)	254	GPT-4.0: Pass; GPT-3.5: Failed	GPT-4.0: 79.9; GPT-3.5: 50.8
Wang et al [24]	China	GPT-4.0 and 3.5	The Chinese National Medical Licensing Examination	100	—	GPT-4.0: 84; GPT-3.5: 56
Cai et al [26]	United States	GPT-4.0 and 3.5	Ophthalmology Board-Style Questions	250	—	GPT-4.0: 71.6; GPT-3.5: 58.8
Oh et al [8]	Korea	GPT-4.0 and 3.5	Korean General Surgery Board Exams	280	—	GPT-4.0: 76.4; GPT-3.5: 46.8
Skalidis et al [27]	Switzerland	GPT-3.5	The European Exam in Core Cardiology	488	Pass	GPT-3.5: 58.8
Saad et al [28]	United Kingdom	GPT-4.0	The Orthopaedic FRCS Orth Part A exam	240	Failed	GPT-4.0: 67.5
Weng et al [5]	China	GPT-3.5	Taiwan’s 2022 Family Medicine Board Exam	125	Failed	GPT-3.5: 41.6
Kumah-Crystal et al [29]	United States	GPT-3.5	The Clinical Informatics Board Examination	254	60%, Pass	GPT-3.5: 74
Mihalache et al [30]	Canada	GPT-4.0	OphthoQuestions practice question bank for board certification examination	125	—	GPT-4.0: 84
Ali et al [31]	United States	GPT-4.0 and 3.5	Self-Assessment Neurosurgery Examination Indications Examination	149	—	GPT-4.0: 82.6; GPT-3.5: 62.4
Oztermeliet al [32]	Turkey	GPT-3.5	Turkey Medical Specialty Exams	1177	—	GPT-3.5: 54.3 - 70.9
Fijaoko et al [4]	United States	GPT-3.5	American Heart Association Basic Life Support and Advanced Cardiovascular Life Support exams	126	84%, Failed	GPT-3.5: 64 - 68.4

XSL•FO
RenderX

Study	Country	Test model	Examination	Data sample, n	Passing threshold	Accuracy (%)
Su et al [12]	China (Taiwan)	GPT-3.5	Taiwan's 2022 Nursing Licensing Exam	400	Pass	GPT-3.5: 80.75%
Lewandowski et al [33]	Poland	GPT-4.0 and 3.5	The Dermatology Specialty Certificate Examinations	120 × 3	GPT-4 Pass	GPT-4.0: >70% better than GPT-3.5
Kung et al [34]	United States	GPT-4.0 and 3.5	Orthopaedic In-Training Examination (2020 - 2022)	360	GPT-4.0: >PGY[b]-5 level; GPT-3.5: PGY-1 level	GPT-4.0: 73.6; GPT-3.5: 54.3
Gencer and Aydin [35]	Turkey	GPT-4.0 and 3.5	Turkish-language thoracic surgery exam	105	Surpass students' scores	GPT-4.0: 93.3; GPT-3.5: 90.5

[a]Not available.

[b]PGY: postgraduate year.

System Role for Accuracy

While it was expected that introducing system role tailored for clinical subspecialties would enhance the reliability of ChatGPT's medical responses, this effect had not been systematically studied. Our research addressed this gap. Our findings revealed slight but noteworthy improvements in accuracy for both GPT-3.5 (1.3% - 4.5%) and GPT-4.0 (0.3% - 3.7%), although these gains were not statistically significant. This might imply that ChatGPT's inherent abilities are already robust enough to discern and address the medical inquiries without narrowing down its response scope.

Response Variability

As an LLM, ChatGPT naturally exhibits variability in responses when the temperature hyperparameter is not zero. In this study, we adopted the default temperature of 0.7 to simulate real-world use conditions on the front end. Our results showed relatively high coherence between the initial and second responses for both GPT-4.0 and GPT-3.5. Therefore, the relatively high temperature of 0.7 is feasible and recommended when testing ChatGPT's performance on the CNMLE. Furthermore, our results highlighted that both model evolution and system roles contribute to ChatGPT's variability in scenarios such as the Chinese Medical Licensing Exams. This variability can be valuable for medical education, as ChatGPT not only provides answers to questions but also includes the rationale and references for its choices, which allows students to easily follow and comprehend [16]. Repeatedly submitting questions allows groups or individuals to engage with the explanatory content generated by ChatGPT, which is particularly beneficial for open-ended case scenario discussions [17].

Subgroup and Multispecialty Analysis

Our subgroup analysis revealed that ChatGPT demonstrated consistent accuracy across different types of questions. This indicated that ChatGPT was capable of understanding and analyzing complex medical cases and scenarios (A2, A3/A4 questions), which can be challenging even for humans, and making correct decisions. This decision-making ability was equally proficient when addressing more straightforward, common-sense questions that did not require reasoning (A1, B1 questions).

In comparisons among unit subgroups representing different subspecialties, significant performance variations were observed in GPT-4.0 across CNMLE units. GPT-4.0 exhibited higher accuracy in units 2 - 3, which predominantly featured questions from subspecialties such as cardiovascular, urinary, digestive, and respiratory systems. This was further corroborated by our multispecialty analysis results. GPT-4.0 achieved an accuracy rate of over 75% for these 4 subspecialties, surpassing its overall accuracy rate of 72.7%. Given that these 4 subspecialties accounted for a substantial proportion (34.5%) of all 15 subspecialties, such a disparity might have been advantageous. However, this disparity disappeared upon the introduction of system roles as prompts, with the overall accuracy of GPT-4.0 increasing to 78.6%. This might suggest that the appropriate use of system roles could compensate for individual subspecialty question accuracy, thereby enhancing the overall accuracy of ChatGPT.

Furthermore, we used CNMLE questions, divided into 15 medical subspecialties, to comprehensively assess the medical expertise of ChatGPT models. This approach provided a robust framework for evaluating model proficiency across a variety of medical fields. Notably, GPT-4.0 surpassed the 60% passing threshold in 14 of the 15 distinct clinical subspecialties, in contrast to GPT-3.5, which only passed in 7 out of 15 subspecialties. This highlighted the superiority of GPT-4.0 and its potential in medical applications.

Generalizability of Findings

Previous studies [7] often excluded table and image-based questions when evaluating ChatGPT's performance in medical exams. This approach limited the generalizability of these findings due to ChatGPT's lack of multimodal data processing. In contrast, our study, focusing on the CNMLE's multiple-choice format, which almost exclusively consists of nongraphical and nontabular questions, offers greater generalizability in real exam settings. Zhu et al [17] suggested that ChatGPT, as a chatbot, had advantages in responding to open-ended questions, corresponding more closely with real-world scenarios where users sought medical support

XSL•FO
RenderX

knowledge from ChatGPT. The potential of ChatGPT in exams with open-ended questions merits further exploration.

Limitations

First, this study assessed ChatGPT's ability to answer questions from the Chinese version of the CNMLE. As ChatGPT is mainly trained on English data, Chinese questions could have underestimated its capabilities. Second, the CNMLE questions were multiple-choice, introducing the chance factor in selecting correct answers. Limited by paper length, we did not evaluate the logic behind ChatGPT's choices, although this aspect is critical and merits deeper investigation. Third, real-world medical questions often have open-ended, multiple, or uncertain answers. Therefore, the CNMLE may not represent the full scope of challenges ChatGPT might face in clinical settings. Consequently, GPT-4.0's success on the CNMLE may only indicate its partial competence in clinical decision-making. Future studies should broaden the range of question types to better assess ChatGPT's medical performance. Despite these limitations, we believe this study provided valuable insights into ChatGPT's capabilities in medicine.

Conclusions

This study comprehensively evaluated the performance of GPT-4.0 and GPT-3.5 in the context of the CNMLE. Our findings indicated that GPT-4.0 not only met the CNMLE passing criteria but also significantly outperformed GPT-3.5 in key areas such as accuracy, consistency, and medical subspecialty expertise. Furthermore, the implementation of system roles served as a pivotal factor in enhancing the model's reliability and answer coherence. These results collectively underscored GPT-4.0's promising potential as a valuable tool for medical professionals, educators, and students, warranting further research and application in the medical field.

Acknowledgments

This research was supported by Medical Science and Technology Tackling Plan of Henan Province (LHGJ20210078).

Authors' Contributions

SM and BL conceived the study and share the corresponding author. SM, QG, and WC collected all relevant data and assisted in results interpretation. SM designed the study, carried out data analysis, and drafted the manuscript. BL participated in the design and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Question type explanations conveyed to ChatGPT via structured prompts.
[DOCX File, 12 KB - mededu_v10i1e52784_app1.docx]

Multimedia Appendix 2
Detail information for repeated responses and their κ value under the ChatGPT default temperature of 0.7.
[DOCX File, 16 KB - mededu_v10i1e52784_app2.docx]

References

1. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
3. Seghier ML. ChatGPT: not all languages are equal. Nature 2023 Mar;615(7951):216. [doi: 10.1038/d41586-023-00680-3] [Medline: 36882613]
4. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American Heart Association course? Resuscitation 2023 Apr;185:109732. [doi: 10.1016/j.resuscitation.2023.109732] [Medline: 36775020]
5. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023 Aug 1;86(8):762-766. [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
6. Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. Med Teach 2023 Jun 3;45(6):665-666. [doi: 10.1080/0142159X.2023.2187684] [Medline: 36905610]
7. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]

XSL•FO
RenderX

8. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023 May;104(5):269-273. [doi: 10.4174/astr.2023.104.5.269] [Medline: 37179699]
9. Currie G, Barry K. ChatGPT in nuclear medicine education. J Nucl Med Technol 2023 Sep;51(3):247-254. [doi: 10.2967/jnmt.123.265844] [Medline: 37433676]
10. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023 Dec 1;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
11. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324. [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
12. Su M, Lin L, Lin L, Chen Y. Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: Insights from Taiwan's Nursing Licensing Exam. Int J Nurs Stud 2024 May;153:104717. [doi: 10.1016/j.ijnurstu.2024.104717] [Medline: 38401366]
13. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-a double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ 2024 Feb;28(1):206-211. [doi: 10.1111/eje.12937] [Medline: 37550893]
14. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023 Jul;13:1219326. [doi: 10.3389/fonc.2023.1219326] [Medline: 37529688]
15. GPT-4. OpenAI. URL: https://openai.com/research/gpt-4/ [accessed 2023-11-21]
16. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023 Aug;2(4):255-263. [doi: 10.1002/hcs2.61] [Medline: 38939520]
17. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
18. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844. [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
19. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023 Apr 19;21(1):269. [doi: 10.1186/s12967-023-04123-5] [Medline: 37076876]
20. Strong E, DiGiammarino A, Weng Y, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Intern Med 2023 Sep 1;183(9):1028-1030. [doi: 10.1001/jamainternmed.2023.2909] [Medline: 37459090]
21. National Clinical Practitioner Qualification Exam: past years' real exam papers and detailed solutions [Article in Chinese]. JD. 2022. URL: https://item.jd.com/30821733544.html/ [accessed 2023-04-20]
22. Introduction of medical licensing examination. The Chinese National Medical Examination Center. URL: https://www1.nmec.org.cn/Pages/ArticleInfo-13-10706.html/ [accessed 2023-11-21]
23. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023 Aug 15;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
24. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
25. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023 Nov;179:e160-e165. [doi: 10.1016/j.wneu.2023.08.042] [Medline: 37597659]
26. Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board–style questions. Am J Ophthalmol 2023 Oct;254:141-149. [doi: 10.1016/j.ajo.2023.05.024] [Medline: 37339728]
27. Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European exam in core cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023 May;4(3):279-281. [doi: 10.1093/ehjdh/ztad029] [Medline: 37265864]
28. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. Surgeon 2023 Oct;21(5):263-266. [doi: 10.1016/j.surge.2023.07.001] [Medline: 37517980]
29. Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? J Am Med Inform Assoc 2023 Aug 18;30(9):1558-1560. [doi: 10.1093/jamia/ocad104] [Medline: 37335851]
30. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. JAMA Ophthalmol 2023 Aug 1;141(8):798-800. [doi: 10.1001/jamaophthalmol.2023.2754] [Medline: 37440220]
31. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023 Nov 1;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]

XSL•FO
RenderX

32. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore) 2023 Aug 11;102(32):e34673. [doi: 10.1097/MD.0000000000034673] [Medline: 37565917]
33. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. An original study of ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol 2024 Jun 25;49(7):686-691. [doi: 10.1093/ced/llad255] [Medline: 37540015]
34. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023 Sep 8;8(3):e23.00056. [doi: 10.2106/JBJS.OA.23.00056] [Medline: 37693092]
35. Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? Am J Med Sci 2023 Oct;366(4):291-295. [doi: 10.1016/j.amjms.2023.08.001] [Medline: 37549788]

Abbreviations

CNMLE: Chinese National Medical Licensing Examination
LLM: large language model
NMEC: National Medical Examination Center
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 15.09.23; peer-reviewed by A Mihalache, R Yang; revised version received 20.05.24; accepted 20.06.24; published 13.08.24.

Please cite as:
Ming S, Guo Q, Cheng W, Lei B
Influence of Model Evolution and System Roles on ChatGPT's Performance in Chinese Medical Licensing Exams: Comparative Study
JMIR Med Educ 2024;10:e52784
URL: https://mededu.jmir.org/2024/1/e52784
doi:10.2196/52784

© Shuai Ming, Qingge Guo, Wenjun Cheng, Bo Lei. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.8.2024. 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.

XSL•FO
RenderX

Assessing the Ability of a Large Language Model to Score Free-Text Medical Student Clinical Notes: Quantitative Study

Harry B Burke[1], MD, PhD; Albert Hoang[1], PhD, DSc; Joseph O Lopreiato[1], MD; Heidi King[2], MS; Paul Hemmer[1], MD; Michael Montgomery[1], BS; Viktoria Gagarin[1], MD

1

2

Corresponding Author:
Harry B Burke, MD, PhD

Abstract

Background: Teaching medical students the skills required to acquire, interpret, apply, and communicate clinical information is an integral part of medical education. A crucial aspect of this process involves providing students with feedback regarding the quality of their free-text clinical notes.

Objective: The goal of this study was to assess the ability of ChatGPT 3.5, a large language model, to score medical students' free-text history and physical notes.

Methods: This is a single-institution, retrospective study. Standardized patients learned a prespecified clinical case and, acting as the patient, interacted with medical students. Each student wrote a free-text history and physical note of their interaction. The students' notes were scored independently by the standardized patients and ChatGPT using a prespecified scoring rubric that consisted of 85 case elements. The measure of accuracy was percent correct.

Results: The study population consisted of 168 first-year medical students. There was a total of 14,280 scores. The ChatGPT incorrect scoring rate was 1.0%, and the standardized patient incorrect scoring rate was 7.2%. The ChatGPT error rate was 86%, lower than the standardized patient error rate. The ChatGPT mean incorrect scoring rate of 12 (SD 11) was significantly lower than the standardized patient mean incorrect scoring rate of 85 (SD 74; *P*=.002).

Conclusions: ChatGPT demonstrated a significantly lower error rate compared to standardized patients. This is the first study to assess the ability of a generative pretrained transformer (GPT) program to score medical students' standardized patient-based free-text clinical notes. It is expected that, in the near future, large language models will provide real-time feedback to practicing physicians regarding their free-text notes. GPT artificial intelligence programs represent an important advance in medical education and medical practice.

(JMIR Med Educ 2024;10:e56342) doi:10.2196/56342

KEYWORDS

medical education; generative artificial intelligence; natural language processing; ChatGPT; generative pretrained transformer; standardized patients; clinical notes; free-text notes; history and physical examination; large language model; LLM; medical student; medical students; clinical information; artificial intelligence; AI; patients; patient; medicine

Introduction

Teaching medical students the skills required to acquire, interpret, apply, and communicate medical information is an integral part of medical education. A crucial aspect of this process involves providing students with feedback regarding the quality of their free-text clinical notes. Various methods have been used to systematically assess clinical notes, notably, QNOTE [1,2], but they depend on human raters. This reliance presents numerous challenges, including rater recruitment and training as well as raters' availability and inclination to perform reviews. Furthermore, humans are susceptible to biases, fatigue, and misinterpretation.

An attractive innovative alternative to human raters is to use a natural language processing (NLP) program to score student notes. An NLP program is a computer-based algorithm that automatically detects specific meanings in free text. The potential advantages of using an NLP program to grade student notes include the following: it is systematic; it is objective; it avoids human bias, fatigue, and misinterpretation; it is essentially free to run; it can assess any number of notes in seconds; and it can grade notes in real time to provide immediate student feedback.

A new type of NLP program was introduced in November 2022, namely, ChatGPT 3.5 (OpenAI) [3], a large language model (LLM) based on the generative pretrained transformer (GPT) artificial intelligence algorithm. It has achieved a 91.7% score

XSL•FO
RenderX

on the United States Medical Licensing Examination (USMLE) style questions [4]. Furthermore, it scored 87.3% on a clinical knowledge test, 91.7% on medical genetics, 89.2% on anatomy, and 92.4% on professional medicine [4]. Its medical-related capabilities include improving clinician empathy [5], responding to patient questions [6], performing differential diagnoses [7], classifying radiology reports [8], writing discharge summaries [9], providing accurate prevention advices to patients [10], and predicting suicide risk [11]. ChatGPT has been compared to human raters in terms of grading short-answer preclerkship medical questions. The ChatGPT-human Spearman correlations for a single assessor ranged from 0.6 to 0.7 [12].

We assessed ChatGPT's ability to accurately score medical students' free-text notes on history of present illness, physical examination, and assessment and plan. We compared these scores to standardized patients' scoring of the clinical notes. We hypothesized that ChatGPT would be more accurate than standardized patients. To our knowledge, this is the first study to assess the ability of a GPT program to score medical students' standardized, patient-based, clinical free-text notes.

Methods

Procedure

This was a single institution, retrospective study. Standardized patients were people who volunteered to interact with medical students to assist in their clinical training. They were trained on a prespecified medical case, and acting as the patient, they interacted with first-year medical students, simulating a patient with that condition. This included responding to clinical questions and undergoing an examination by the medical student. The students documented their interaction with standardized patients in free-text clinical notes. They wrote a chief complaint; history of the present illness; review of systems; physical examination; and differential diagnosis, featuring 3 rank-ordered diagnoses. In addition, they provided their pertinent positives and negatives and suggested follow-up tests. At our medical school, standardized patients provided verbal feedback to students regarding their interaction and scored their students' notes. They had 7 - 10 days to score the student notes and send the results to the course instructor. They did not provide any grading feedback to the students. The advantage of using standardized patients over actual patients for training medical students is that the medical students' experiences, and therefore, their clinical notes are based on a consistent clinical presentation.

The study case and scoring rubric, "Suzy Whitworth," were developed by the Association for Standardized Patient Educators and adapted by the Mid-Atlantic Consortium of Clinical Skills Centers in June 2018, with additional formatting edits in January 2019. The standardized patients were trained on this case and its scoring rubric. The case contained 85 scorable elements that were expected to be present in the students' notes. Three scoring rubric examples were as follows: "Notes chief complaint of shortness of breath (shortness of breath, dyspnea, difficulty breathing, and can't catch my breath)"; "Notes sudden onset (acute, all of the sudden, and all at once"; and "Notes timing (a few hours ago, this morning, upon awakening, or today)." The rubric combined the 85 scorable elements into 12 classes. ChatGPT and the standardized patients scored as either correct or incorrect each of the 85 elements in the deidentified students' notes. An error was either an incorrect answer or the absence of an answer. A reviewer checked the standardized patient scoring and the ChatGPT scoring and a second reviewer checked the first reviewer's scores.

ChatGPT is an LLM based on the GPT artificial intelligence algorithm. It was pretrained on 45 TB of data and it consists of attention, which connects and weights natural language meanings, and an artificial neural network, which organizes and stores the meanings [13]. It accepts natural language input and provides natural language output. For each medical student and for each rubric, the researcher created a new prompt that asked ChatGPT if the rubric's meaning was contained in the student's free-text note.

For ChatGPT and standardized patients, the measure of accuracy was the percent correct for each of the 12 categories and across the 12 categories. Student *t* tests (2-tailed) compared the mean error rate across the 12 classes for ChatGPT with the mean error rate across the 12 classes for the standardized patients using the R language (R Project for Statistical Computing) [14].

Ethical Considerations

Ethical approval was waived as per section 46.104(d) of Code of Federal Regulations, as this was a quality improvement project [15].

Results

The study population consisted of 168 first-year medical students, the case scoring rubric consisted of 85 elements, resulting in a total of 14,280 scores. There were 4 standardized patients, each working with one-fourth of the students. The incorrect scoring (error) rates for the standardized patients and ChatGPT are shown in Table 1.

XSL•FO
RenderX

Table . Incorrect scoring rates for ChatGPT and the standardized patients across free-text note categories and across all categories.

Category	Scoring opportunities for the 168 students, n	Standardized patient errors, n (%)	ChatGPT errors, n (%)
Chief complaint	840	135 (16.1)	17 (2.0)
History of present illness	1512	226 (14.9)	35 (2.3)
Review of systems	1008	67 (6.6)	7 (0.7)
Past medical history	1512	43 (2.8)	21 (1.4)
Physical exam	2352	181 (7.7)	25 (1.1)
Diagnosis (pulmonary embolism)	168	3 (1.8)	0 (0)
Pulmonary embolism evidence	2352	182 (7.7)	8 (0.3)
Diagnosis (pneumonia)	168	0 (0)	0 (0)
Pneumonia evidence	1848	66 (3.6)	4 (0.2)
Diagnosis (pneumothorax)	168	0 (0)	7 (4.2)
Pneumothorax evidence	1176	54 (4.6)	5 (0.4)
Diagnostic studies	1008	66 (6.5)	16 (1.6)
Total[a]	14,280	1023 (7.2)	145 (1.0)

[a]ChatGPT versus standardized patient; *P*=.002.

The category error rates for standardized patients and ChatGPT, respectively, were as follows: chief complaint: 135, 17; history of present illness: 226, 35; review of systems: 67, 7; past medical history: 43, 21; physical examination: 181, 25; first diagnosis: 3, 0; evidence for first diagnosis: 182, 8; second diagnosis: 0, 0; evidence for second diagnosis: 66, 4; third diagnosis: 0, 7; evidence for third diagnosis: 54, 5; and diagnostic studies: 66, 16. The ChatGPT incorrect scoring rate was 1.0%, and the standardized patient incorrect scoring rate was 7.2%. The ChatGPT error rate was 86% lower than the standardized patient error rate. The ChatGPT mean incorrect scoring rate of 12 (SD 11) was significantly lower than the standardized patient mean incorrect scoring rate of 85 (SD 74; *P*=.002).

Discussion

ChatGPT had a significantly lower error rate compared to standardized patients. This suggests that an LLM can be used to score medical students' notes.

NLP programs have been used in several medical education settings. Medical education NLPs have been based on keywords, expert systems, statistical algorithms, and combinations of these approaches. DaSilva and Dennick [16] transcribed medical student problem-based verbal learning sessions and used an NLP program to count the frequency of technical words. Zhang et al [17] implemented both a naïve Bayes approach and a supervised support vector machine method to assess resident performance evaluations. Their sentiment accuracies were 0.845 for naïve Bayes and 0.937 for the support vector machine. Spickard et al [18] used an electronic scoring system to detect 25 core clinical problems in medical students' clinical notes. They achieved a 75% positive predictive value (PPV) on 16 of the 25 problems. Denny et al [19] examined whether students mentioned advance directives or altered mental status in their clinical notes. For advance directives, their sensitivity was 69% and their PPV was 100%, and for mental status, their sensitivity was 100% and their PPV was 93%. Sarker et al [20] used a semisupervised NLP method to assess students' free-text notes. Their accuracy over 21 cases and 105 notes was a sensitivity of 0.91 and a PPV of 0.87. Two recent papers from the University of Michigan's Department of Surgery [21,22] assessed resident feedback and competency. Solano et al [21] dichotomized the narrative surgical feedback given to residents into high and low quality and trained a logistic regression model to distinguish between them. Their model achieved a sensitivity of 0.37, a specificity of 0.97, and a receiver operating characteristic (ROC) of 0.86. Otles et al [22] assessed narrative surgical resident feedback using a variety of statistical methods. The support vector machine algorithm achieved the best result with a maximum mean accuracy of 0.64. Abbott [23] studied whether an NLP program could assess the clinical competency committee ratings of residents in terms of language that correlated with the 16 Accreditation Council for Graduate Medical Education Milestones. The ROCs for the 16 milestones ranged from 0.71 to 0.95 and the mean ROC was 0.83. Neves et al [24] examined the ability of RapidMiner Studio, a machine learning program, to assess the quality of attending feedback on resident performance. Their accuracies ranged from 74.4% to 82.2%.

If NLP programs are to be used to automate the grading of students' notes, they must achieve an acceptable accuracy. Sarker et al [20] suggested that any method of scoring medical notes should achieve an accuracy close to 100%. Regrettably, none of the reported medical education NLPs achieved an acceptable accuracy. In our study, standardized patients also failed to achieve an acceptable accuracy. ChatGPT attained an accuracy close to 100% and is, therefore, suitable for scoring students' free-text notes.

A potential limitation of this study is that it has been suggested that GPT-based methods have the potential to generate unreliable answers under certain circumstances. We did not find

XSL•FO
RenderX

that to be true in our study. Another potential limitation is that, although ChatGPT is free to the public, it has resource requirements. It used 45 TB of data, it has 175 billion parameters, and it runs on supercomputers residing in the cloud. This is a great deal of computing power for student notes. Fortunately, there are open-source GPTs, for example, Meta's Llama, that can be run on a workstation. We would have liked to examine the standardized patient validity literature, but to our knowledge, no such study exists. Finally, assessing note errors does not directly address clinical reasoning.

An important advantage of LLMs is their ability to provide real-time scoring and feedback on student clinical free-text notes. This immediate assessment offers students a valuable learning opportunity because they can reflect on their performance while the clinical interaction is still fresh in their mind. Another advantage is that the scoring is accurate and objective so students will no longer have to worry about human error and bias. A disadvantage of ChatGPT was that it was time consuming. Fortunately, there are compound GPTs that can perform the entire assessment of all the elements and all the students at once. In terms of clinical reasoning, in the future, we will be asking medical students, as part of their clinical note write-up, to provide their clinical reasoning and we can have a GPT assess the quality of their reasoning.

It should be noted that the use of LLMs to score clinical notes need not be limited to medical students. It is expected that in the near future, GPT-based artificial intelligence NLPs will be applied to provide real-time feedback on free-text clinical notes to practicing physicians.

In conclusion, ChatGPT demonstrated a significantly lower error rate compared to standardized patients. This is the first study to assess the ability of a GPT program to score medical students' standardized, patient-based, free-text clinical notes. GPT artificial intelligence programs represent an important advance in medical education and medical practice.

Acknowledgments

Support for this project was provided by the Patient Safety and Quality Academic Collaborative, a joint Defense Health Agency-Uniformed Services University program. The funder did not participate in the design, execution, or analysis of this project.

The opinions and assertions expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of Defense, the Defense Health Agency, or the Uniformed Services University of the Health Sciences.

Data Availability

The datasets used in this study are not publicly available because they include student scores, but they are available from the corresponding author on reasonable request.

Authors' Contributions

HBB, AH, JOL, and PH made substantial contributions to the conception and design of the work; HBB, AH, JOL, HK, MM, and VK made substantial contributions to the acquisition, analysis, and interpretation of the data; HBB wrote the manuscript.

Conflicts of Interest

None declared.

References

1. Burke HB, Hoang A, Becher D, et al. QNOTE: an instrument for measuring the quality of EHR clinical notes. J Am Med Inform Assoc 2014;21(5):910-916. [doi: 10.1136/amiajnl-2013-002321] [Medline: 24384231]
2. Burke HB, Sessums LL, Hoang A, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc 2015 Jan;22(1):199-205. [doi: 10.1136/amiajnl-2014-002726] [Medline: 25342178]
3. OpenAI. ChatGPT. URL: https://openai.com/index/chatgpt/ [accessed 2023-08-11]
4. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. arXiv. Preprint posted online on May 16, 2023. [doi: 10.48550/arXiv.2305.09617]
5. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023 Jan;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
6. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
7. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378. [doi: 10.3390/ijerph20043378] [Medline: 36834073]

XSL•FO
RenderX

8. Olthof AW, Shouche P, Fennema EM, et al. Machine learning based natural language processing of radiology reports in orthopaedic trauma. Comput Methods Programs Biomed 2021 Sep;208:106304. [doi: 10.1016/j.cmpb.2021.106304] [Medline: 34333208]
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
10. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023 Mar 14;329(10):842-844. [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
11. Burkhardt HA, Ding X, Kerbrat A, Comtois KA, Cohen T. From benchmark to bedside: transfer learning from social media to patient-provider text messages for suicide risk prediction. J Am Med Inform Assoc 2023 May 19;30(6):1068-1078. [doi: 10.1093/jamia/ocad062] [Medline: 37043748]
12. Morjaria L, Burns L, Bracken K, et al. Examining the threat of chatgpt to the validity of short answer assessments in an undergraduate medical program. J Med Educ Curric Dev 2023;10:23821205231204178. [doi: 10.1177/23821205231204178] [Medline: 37780034]
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. Preprint posted online on Aug 2, 2023. [doi: 10.48550/arXiv.1706.03762]
14. The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed 2024-07-19]
15. Code of Federal Regulations. National Archives. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.104 [accessed 2024-07-19]
16. Da Silva AL, Dennick R. Corpus analysis of problem-based learning transcripts: an exploratory study. Med Educ 2010 Mar;44(3):280-288. [doi: 10.1111/j.1365-2923.2009.03575.x] [Medline: 20444059]
17. Zhang R, Pakhomov S, Gladding S, Aylward M, Borman-Shoap E, Melton GB. Automated assessment of medical training evaluation text. AMIA Annu Symp Proc 2012;2012:1459-1468. [Medline: 23304426]
18. Spickard A, Ridinger H, Wrenn J, et al. Automatic scoring of medical students' clinical notes to monitor learning in the workplace. Med Teach 2014 Jan;36(1):68-72. [doi: 10.3109/0142159X.2013.849801] [Medline: 24195470]
19. Denny JC, Spickard A, Speltz PJ, Porier R, Rosenstiel DE, Powers JS. Using natural language processing to provide personalized learning opportunities from trainee clinical notes. J Biomed Inform 2015 Aug;56:292-299. [doi: 10.1016/j.jbi.2015.06.004] [Medline: 26070431]
20. Sarker A, Klein AZ, Mee J, Harik P, Gonzalez-Hernandez G. An interpretable natural language processing system for written medical examination assessment. J Biomed Inform 2019 Oct;98:103268. [doi: 10.1016/j.jbi.2019.103268] [Medline: 31421211]
21. Solano QP, Hayward L, Chopra Z, et al. Natural language processing and assessment of resident feedback quality. J Surg Educ 2021;78(6):e72-e77. [doi: 10.1016/j.jsurg.2021.05.012] [Medline: 34167908]
22. Ötleş E, Kendrick DE, Solano QP, et al. Using natural language processing to automatically assess feedback quality: findings from 3 surgical residencies. Acad Med 2021 Oct 1;96(10):1457-1460. [doi: 10.1097/ACM.0000000000004153] [Medline: 33951682]
23. Abbott KL, George BC, Sandhu G, et al. Natural language processing to estimate clinical competency committee ratings. J Surg Educ 2021;78(6):2046-2051. [doi: 10.1016/j.jsurg.2021.06.013] [Medline: 34266789]
24. Neves SE, Chen MJ, Ku CM, et al. Using machine learning to evaluate attending feedback on resident performance. Anesth Analg 2021 Feb 1;132(2):545-555. [doi: 10.1213/ANE.0000000000005265] [Medline: 33323789]

Abbreviations

GPT: generative pretrained transformer
LLM: large language model
NLP: natural language processing
PPV: positive predictive value
ROC: receiver operating characteristic
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 15.01.24; peer-reviewed by A DiGiammarino, D Chartash; revised version received 22.02.24; accepted 06.05.24; published 25.07.24.

Please cite as:
Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, Gagarin V
Assessing the Ability of a Large Language Model to Score Free-Text Medical Student Clinical Notes: Quantitative Study
JMIR Med Educ 2024;10:e56342
URL: https://mededu.jmir.org/2024/1/e56342
doi:10.2196/56342

XSL•FO
RenderX

© Harry B Burke, Albert Hoang, Joseph O Lopreiato, Heidi King, Paul Hemmer, Michael Montgomery, Viktoria Gagarin. Originally published in JMIR Medical Education (https://mededu.jmir.org), 25.7.2024. 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.

XSL•FO
RenderX

Assessing AI Awareness and Identifying Essential Competencies: Insights From Key Stakeholders in Integrating AI Into Medical Education

Julia-Astrid Moldt[1], MA; Teresa Festl-Wietek[1], Dr rer nat; Wolfgang Fuhl[2,3], Dr rer nat; Susanne Zabel[2,3], MSc; Manfred Claassen[2,3,4], Prof Dr; Samuel Wagner[5], Prof Dr; Kay Nieselt[2,3], Prof Dr; Anne Herrmann-Werner[1,6], Prof Dr Med

1
2
3
4
5
6

Corresponding Author:
Julia-Astrid Moldt, MA

Abstract

Background: The increasing importance of artificial intelligence (AI) in health care has generated a growing need for health care professionals to possess a comprehensive understanding of AI technologies, requiring an adaptation in medical education.

Objective: This paper explores stakeholder perceptions and expectations regarding AI in medicine and examines their potential impact on the medical curriculum. This study project aims to assess the AI experiences and awareness of different stakeholders and identify essential AI-related topics in medical education to define necessary competencies for students.

Methods: The empirical data were collected as part of the TüKITZMed project between August 2022 and March 2023, using a semistructured qualitative interview. These interviews were administered to a diverse group of stakeholders to explore their experiences and perspectives of AI in medicine. A qualitative content analysis of the collected data was conducted using MAXQDA software.

Results: Semistructured interviews were conducted with 38 participants (6 lecturers, 9 clinicians, 10 students, 6 AI experts, and 7 institutional stakeholders). The qualitative content analysis revealed 6 primary categories with a total of 24 subcategories to answer the research questions. The evaluation of the stakeholders' statements revealed several commonalities and differences regarding their understanding of AI. Crucial identified AI themes based on the main categories were as follows: possible curriculum contents, skills, and competencies; programming skills; curriculum scope; and curriculum structure.

Conclusions: The analysis emphasizes integrating AI into medical curricula to ensure students' proficiency in clinical applications. Standardized AI comprehension is crucial for defining and teaching relevant content. Considering diverse perspectives in implementation is essential to comprehensively define AI in the medical context, addressing gaps and facilitating effective solutions for future AI use in medical studies. The results provide insights into potential curriculum content and structure, including aspects of AI in medicine.

(JMIR Med Educ 2024;10:e58355) doi:10.2196/58355

KEYWORDS

AI in medicine; artificial intelligence; medical education; medical students; qualitative approach; qualitative analysis; needs assessment

Introduction

Background and Significance of AI in Medicine

In 1966, the architect Cedric Price [1] posed the provocative question, "Technology is the answer, but what was the question?" to encourage his lecture audience to explore, question, and reconsider the impact of technological progress. More than 50 years later, this question remains as relevant as ever. One might similarly ask today, "The answer is AI, but what was the question?" The health care sector is currently undergoing a significant transformation process characterized by the increased integration of digital technologies [2-4]. German clinics have been incorporating computer-driven clinical

XSL•FO
RenderX

decision systems, such as the electronic patient record and other digital health tools, that can analyze data, identify patterns, and make decisions based on that data [3]. These intelligent systems can improve health care efficiency, accuracy, and quality while potentially reducing the burden on medical personnel [5,6]. Artificial intelligence (AI) technologies are already being implemented in various aspects of medical practice. For instance, they are used in imaging diagnostics where AI algorithms help analyze medical images [7]. Dictation systems with speech recognition powered by AI are also used, and AI chatbots are deployed to assist doctors and patients by providing appointments and information [8-10]. A range of sensor-based wearables, such as fitness trackers, smartwatches, and health apps, is already used in people's daily lives. These devices use AI-supported algorithms to gather and analyze health data, including heart rate, sleep patterns, activity levels, and calorie consumption. Based on this information, personalized recommendations can be made to help individuals improve their well-being [11]. Although the use of medical AI systems remains in its early stages, ongoing research and development efforts are being undertaken worldwide. As technology rapidly advances, AI will increasingly play a crucial role in the future of health care [12,13]. This also requires restructuring medical curricula to adapt to dynamic technological advances to prepare students for the changing structures of medical practice [14,15].

Traditionally, medical education has focused on providing students with comprehensive knowledge of medical practices, diagnostic procedures, and treatment methods. Additionally, the effective use of AI in the medical field requires not only developing the necessary technological advances in AI applications but also ensuring that future physicians possess the required skills and expertise to effectively apply these technologies [16-18]. Therefore, it is crucial to consider integrating AI into the medical curriculum and determine how this technology can be effectively incorporated to benefit students and patients [19-21]. However, studies indicate that the integration of AI into the medical curriculum to enhance understanding of AI algorithms and optimize their use remains in its early stages, particularly in Germany [22-24]. Some institutions have developed specific courses and training programs to enhance medical students' knowledge and skills in AI [25-27].

Research Objectives and Research Questions

Given the complex and rapidly evolving nature of AI, no standardized definition or structured learning objectives currently exist regarding the specific AI topics medical students should be familiar with. Several studies emphasize the importance of understanding the fundamentals of AI and data science, mathematical concepts, and related ethical and social issues [26,28]. Medical students should also develop skills in interpreting AI models and be familiar with machine learning, deep learning, and data analytics to apply AI in clinical practice [29].

As part of a project, "TüKITZMed – Tübingen KI – Trainingszentrum für die Medizin" (Tübinger AI Training Center for Medicine), funded by the German Federal Ministry of Education and Research (16DHBKI086), a comprehensive needs assessment was conducted involving various stakeholders to understand the requirements and skills for integrating AI into the medical curriculum following step one of Kern's 6-step approach [30]. The project "TüKITZMed" aims to develop and establish a cross-faculty interprofessional curriculum focused on "AI in medicine" providing students with a comprehensive understanding of the topic across different levels and disciplines. This curriculum serves as a pioneering example of integrating AI into academic programs, offering students opportunities for both theoretical learning and practical application, thereby facilitating the transfer of knowledge into real-world contexts. This study aimed to investigate essential AI knowledge for medical education curricula, identify necessary competencies through stakeholder input, and address potential gaps in learning opportunities. Involving different stakeholders offers diverse perspectives based on their roles and experiences. This approach helps identify relevant AI competencies and appropriate teaching formats, addressing unmet needs and challenges associated with implementing AI-focused learning opportunities in medical education [31].

Therefore, this paper aims to address the following research questions regarding assessing AI awareness and identifying essential competencies:

- How familiar are the different stakeholders with AI in general?
- Which specific aspects and topics related to AI are viewed as important?
- What competencies are crucial for medical health students?

Methods

For a comprehensive understanding of AI and to address various aspects relevant to the surveyed stakeholders' perspectives on an AI curriculum, an exploratory research approach using semistructured interviews was chosen. The incorporation of narrative-generating guideline-supported questions aimed to establish a structured framework for investigating research interests while also allowing flexibility to uncover new and insightful content [32].

Study Design and Setting

This qualitative study approach followed the Standards for Reporting Qualitative Research [33]. It was performed at the Medical Faculty and the Faculty of Science of the University of Tübingen as part of the TüKITZMed project.

Sample Selection and Recruitment

Semistructured interviews with 38 stakeholders involved in the implementation process of AI in medical curricula were conducted to gather diverse perspectives and insights. Relevant stakeholders were characterized as individuals impacted by the integration of AI in health care, those with professional experience with AI technology, and those who had previously encountered AI applications in the medical sector. The stakeholder groups comprised the following: 6 lecturers, 9 clinicians, 10 students, 6 AI experts, and 7 institutional stakeholders. The interview guide followed the guiding research questions for the needs assessment [34]. An illustrative interview guide is provided in Multimedia Appendix 1.

XSL•FO
RenderX

The selection of stakeholder groups was based on their crucial role in the field of medical education and their diverse perspectives. For participant recruitment, we used an open approach, reaching out to stakeholders primarily via email after identifying relevant stakeholder groups for our research inquiries. Inclusion criteria included individuals working with AI in the medical context or possessing relevant expertise, especially clinicians and AI experts. Due to the project's regional focus, only stakeholders from the local area were approached. Recommendations, referrals, and requests within working groups or via email forwarding were also used. Potential participants were also approached at conferences.

Lecturers

Educators' perspectives are required, as integrating AI into medical education is an unprecedented challenge with no clear guidelines. Even if consensus is reached on exactly what should be taught to medical students, it remains daunting to determine how best to teach it. The experience of educators—especially those familiar with medical students—is therefore imperative in the process [35,36].

Students

Health care students' perspectives (eg, on human medicine, medical technology, and molecular medicine) are central to integrating AI into medical education since the curricula should ultimately be designed to serve their educational needs. Therefore, assessing their current state of knowledge, attitudes, and heterogeneity across different student populations is an important step in adequately addressing the educational needs for medical AI and integrating it such that students will benefit from it [18,37].

AI Experts

AI experts have long-standing knowledge and expertise in the field. Engaging with them provides valuable insights into the latest developments, trends, and best practices in AI. These experts offer a thorough understanding of AI concepts, applications, and their potential impact on health care [36,38].

Clinicians

Involving medical staff in developing medical AI helps find clinical value while protecting patient safety. Moreover, medical staff know the data well and are thus the only ones who can detect the bias or impracticality of AI. Additionally, medical experts play a key role in teaching real-world medical applications of AI, as they have the experience and skills. Thus, their perspectives are relevant to the integration of AI into education and practice since they can inspire other medical workers to engage with it [39].

Institutional Stakeholders

The perspectives of institutional stakeholders are crucial for driving change in medical education. These individuals hold key positions within educational or health care institutions and are actively involved in implementing AI within the medical curriculum. Such stakeholders, including deans, program coordinators, and administrative staff, possess specific training and qualifications relevant to their roles, playing an essential part in shaping educational strategies and health policies. Given the already full capacity of medical curricula, their support and expertise are necessary for a meaningful integration of AI. Additionally, institutional stakeholders provide an important framework for continuously monitoring and reevaluating the implementation of AI in medical curricula to ensure its utility and quality [18,40,41].

Data Collection

Semistructured guided interviews were chosen as they allow a flexible participation-centered approach and in-depth exploration of the topic, capturing the diverse perspectives of the stakeholders involved [42]. The semistructured guided interviews were conducted from August 2022 to March 2023, either face-to-face or via videoconference. All interviews were audio recorded and transcribed verbatim for analysis. Before participation, written informed consent was obtained from all the interviewees. The resulting code system for analysis was consolidated and summarized.

Data Analysis

The transcripts were analyzed according to the principles of content structuring analysis, as outlined by Kuckartz [43]. After the interviews were transcribed, independent researchers thoroughly reviewed them. The category system for the analysis was developed using the semistructured guiding questionnaire as a basis (inductive approach) and systematically coded using the MAXQDA 2022 software program (VERBI GmbH). As the coding process progressed, new categories emerged to include additional aspects and themes discussed in the interviews. This step enabled flexibility and openness to new insights that transcended the initially defined structure (deductive approach) [44]. Collectively, we presented outcomes derived from diverse stakeholders. We systematically addressed varying perspectives within or across these cohorts, emphasizing their respective relevance. Our presentation includes literal quotations, preserving the original expressions translated from German to English.

Ethical Considerations

The study received ethical approval from the Ethics Committee of Tübingen Medical Faculty (467/2022BO2). Participation was voluntary. All participants were informed of the purpose of the study and provided informed consent before data collection. The confidentiality of all data was ensured, and all responses and data were kept anonymous. The participants had the right to withdraw from the study at any time. Participants did not receive any compensation.

Results

Overview

The ages of participants ranged from 19 to 59 (mean 38.5, SD 9.7, SEM 1.6) years, with data provided by 36 individuals. Regarding sex distribution, there were 26 male and 12 female participants. Through a structuring content analysis, we systematically derived 6 primary categories with a total of 24 subcategories from the entire data set.

XSL•FO
RenderX

Presentation of Stakeholder Perspectives and Expectations of AI

The analysis of the stakeholders' statements revealed several commonalities and differences regarding their understanding of AI.

AI as a Tool

In terms of commonalities, the actors viewed AI primarily as a tool that can analyze and process large amounts of data:

> *For me, it's mainly a toolbox, a toolkit. These are technologies that help us.* [RR22T, expert]

> *An AI can process much more data at once than a human could.* [KC10S, student]

> *A way to predict things, that is, to predict data based on existing data and also to apply techniques that support us to categorize, assess, simulate, and also predict things in terms of the future.* [KU512S, institutional stakeholder]

AI as a Medical Assistant

Additionally, stakeholders emphasized the potential benefits of using AI to assist in medicine, whether in supporting diagnostic and treatment decisions or more efficiently mining clinical data:

> *In the context of medicine, probably so therapy decisions, more efficient evaluation of clinical data.* [AB001, lecturer]

> *To this, I can think of automation and standardization of processes but also help in an increasingly complex clinical situation with many parameters and many possibilities relevant for decision-making by doctors involved in therapy and diagnosis.* [RW01R, institutional stakeholder]

The Technical Understanding of AI

A focus on understanding AI is also related to the technical background of AI technology and the roles of mathematical-statistical models, data, and algorithms:

> *AI is learning systems that fit a model with computations to data.* [PG49B, expert]

> *AI is about programs and algorithms that improve with more data and data processing.* [CCHU7, clinician]

Differences in the Understanding of the Term and Difficulties in Formulating a Definition

However, differences in the understanding of AI between the stakeholders interviewed also became apparent. These disparities manifested in three key areas: challenges in formulating a concise definition and description of AI, diverse perspectives and expectations regarding AI capabilities, and varying emphasis on the medical fields where AI is anticipated to make significant progress.

> *The name "AI" is misleading because there is currently no computer application that is actually intelligent. Rather, it refers to algorithms with high computing power that enable computers to process larger amounts of data than before and possibly even evolve themselves.* [RH01W, lecturer]

> *Artificial intelligence is a difficult term. It suggests that human intelligence is extended and artificially subsumed.* [EJ12B, institutional stakeholder]

> *Understanding AI is complex and broad. There is AI that learns itself and AI that still needs to be monitored. AI is otherwise a kind of automated analysis.* [HT02B, clinican]

Varying Perspectives and Expectations on the Potential and Capabilities of AI

While some may view AI capabilities more optimistically or comprehensively, others emphasize the limits and specialized functions AI can possess. For example, the students tended to view AI as efficient data processing, while the experts emphasized AI's capacity for revealing hidden patterns and simulating complex scenarios. The students emphasized that AI lacks a creative process and cannot engage in creative thinking. They focused on AI's ability to efficiently process and analyze large amounts of data. The experts discussed AI's ability to identify scientific connections, structures, and patterns that may be imperceptible to humans. They recognized AI's potential to simulate complex scenarios and discover novel insights from data.

> *Systems that can recognize scientific relationships, structures, and patterns that are not discernible to humans.* [PG49B, AI expert]

> *AI is not intelligent because no creative process can take place in it. It can quickly and efficiently draw connections from large amounts of data.* [YG30B, student]

> *In most cases, however, AI is about better evaluating large amounts of data and modifying it through self-learning algorithms. Human intelligence can understand and solve problems through creativity and think outside of rules – so it works differently. However, algorithms can do other things better than humans.* [EJ12B, institutional stakeholder]

> *AI will determine everyday life, but also medicine more and more. Nowadays, one is often confronted with the topic, and one should deal with it.* [IE13H, clinician]

Different Emphases of the Potential Areas of Application

Although some alignment existed in the perceptions of the potential uses of AI, each group had its own focus on specific applications of AI in health care, depending on profession and discipline. AI experts and lecturers emphasized the significance of the technical dimensions of AI, including the essential roles of algorithms, data processing, and AI models in medical research and practice. Students underscored AI's role in aiding health care professionals, while clinicians concentrated on the clinical sphere of AI, particularly its contributions to diagnostics, treatment decisions, and data processing. Additionally, institutional stakeholders highlighted the potential for increased efficiency in health care by implementing AI solutions. They

XSL•FO
RenderX

engaged in discussions concerning the pragmatic integration of AI to strengthen clinical decision-making processes and optimize operational workflows.

> *But to provide a definition..., so what we're actually doing is, we're learning relationships between input, certain inputs, and certain outputs: What is the relationship between an X-ray image and a diagnosis? And this correlation, you can then learn it using, for instance, a neural network, and then apply it to unseen X-ray images.* [MF08Z, lecturer]

> *To simplify processes, so to speak, to facilitate and automate simple processes that would normally take a lot of time for us humans. The machine can recognize complex relationships that we as humans either cannot comprehend or, as mentioned, would take a long time to understand. For example, in my case, it's radiation therapy in radio oncology, where there are many processes that take a long time or are, as I said, very complex because, in medicine, we naturally have many intricate aspects and influences on the patient that we need to consider. And a machine can handle this quite well, as it can analyse and evaluate these various data effectively, essentially.* [EK05B, student]

> *An evolving field, which is already partially present in clinical reality. This involves automation and standardization of processes, as well as assistance in an increasingly complex clinical environment with numerous parameters and numerous possibilities that are relevant for decision-making by physicians and individuals involved in therapy and diagnosis.* [RW01R, institutional stakeholder]

Identification of AI Competencies and Implications for Medical Curriculum

We identified 4 main categories of implementation needs (Textbox 1)

Textbox 1. Identified main categories of implementation needs: conceptual designs for an artificial intelligence curriculum.

Possible curriculum contents, skills, and competencies

- Basic understanding and sense of technology
- Data literacy
- Morality and ethics
- Opportunities and risks
- Digital literacy
- Application of software
- Data privacy
- Understanding of medical test results

Programming skills

- Voluntary: Having programming skills is optional. Although they are not mandatory, having them is beneficial.
- Not required: Programming skills are unnecessary. However, if one possesses such skills, that is acceptable.
- Required: Programming skills are mandatory. Basic or advanced programming proficiency is expected for participation.

Curriculum scope

- Adapted to the time available
- Intensive engagement

Curriculum structure

- Lecture
- Seminar
- Interactive exercises
- Consolidation for specialization
- Basics as lecture with exercise
- No opinion due to lack of experience
- Interdisciplinary
- Adapt curriculum dynamically according to relevance

XSL•FO
RenderX

Possible Curriculum Contents, Skills, and Competencies

This main category covers a range of topics, including possible curriculum components, skills, and abilities students should learn regarding AI in medicine.

Basic Understanding and Sense of Technology

The first subcategory addresses the need for medical students to develop a basic understanding of the fundamental principles and concepts of AI. This includes understanding the essential mechanisms of machine learning algorithms and acquiring basic knowledge of mathematical computer science.

> *And I believe that what would be important to develop a bit of an understanding of how the technology actually works....So, I don't think that you can teach all of that to medical students from the ground up in theoretically well-founded way with linear algebra and so on. But I do think that it's quite impossible to offer an applied course where they can practise and play around with it, get a sense of how technology functions.* [MF08Z, lecturer]

Data Literacy

The data literacy category describes the need to provide medical students with the skills and knowledge required to effectively handle and interpret data in the context of AI applications in the field of medicine.

> *Data quality is crucial. In my view, all the methods of machine learning are secondary....But the most important thing is truly obtaining high-quality data, understanding how to work with data, understanding implications of the data.* [DD21S, expert]

Morality and Ethics

The morality and ethics subcategory is dedicated to providing students with an in-depth understanding of the ethical considerations associated with integrating AI into the medical field. It aims to develop a keen awareness of the ethical responsibilities associated with AI advances in the medical field.

> *Ethics comes to mind...I consider it a highly relevant aspect because AI tools that are intended for future use in medicine are, in my opinion, closely tied to patients, to human beings; potentially, these tools could make life and death decisions, and in that regard, I would argue that entirely different requirements for quality assurance, ethical standards and checks and boundaries need to be in place for these tools....That should be covered during education.* [RW01R, institutional stakeholder]

Opportunities and Risks

The opportunities and risks subcategory includes aspects related to awareness of the potential benefits and challenges associated with integrating AI into health care. Medical students should be empowered to navigate the complex landscape of AI in medicine by not only recognizing the potential benefits but also being able to address challenges, make informed decisions, and maintain vigilance concerning its capabilities and limitations.

> *Also understanding how to interact with AI. To what extent can I trust the AI, the outcomes it produces? How can I collaborate effectively with it? What do I need to operate a good AI, and where can it also be deployed?* [SA01R, student]

Digital Literacy

Several actors addressed the need for medical students to be equipped with the skills required to navigate and use digital technologies effectively. This includes developing proficiency in using AI-powered diagnostic support tools, including the ability to interpret and apply AI-generated diagnostic insights. This also extends to understanding and implementing adaptive learning methodologies and leveraging telemedicine for remote patient care.

> *What is also important to me, when we talk about the topic of artificial intelligence, is that we first discuss the fundamental aspects of digitization and the necessary measures for healthcare, research, and education....We are delving into a very specific topic, but we still lack some of the foundational knowledge.* [EH07S, institutional stakeholder]

Application of Software

The application of software subcategory concerns equipping individuals with the ability to effectively use software tools, particularly in the context of AI development and implementation.

> *Medical students often lack knowledge in this area. Therefore, I believe it's important for them to have hands-on experience of training a neural network themselves.* [MF08Z, expert]

Data Privacy

The data privacy category describes the aspects the actors mentioned to give students the expertise to address the ethical and legal issues related to data privacy in AI applications. By mastering data management practices and understanding the legal framework, students ensure patient data is managed safely, impartially, and ethically in the context of AI integration.

> *The topic of data privacy should definitely be included in the curriculum because the "who" question of how to do this, how it's trained and on which data sources, most of it needs to be anonymized.... Ethics and data privacy are two significant components that need to be integrated, unfortunately or fortunately.* [HT02B, clinician]

Understanding of Medical Test Results

The last subcategory summarizes the need for in-depth expertise in AI-driven application outcomes, particularly in the context of medical tests. Students must develop a profound understanding of the insights and outcomes produced by AI applications, including acquiring the expertise to thoroughly analyze and interpret results derived from AI-powered processes.

> *The most important aspect of AI is understanding the basis on which decisions are made.* [JJ22D, clinician]

XSL•FO
RenderX

Programming Skills

Clinicians stated that a programming course for medical students should not be mandatory due to overload but could be offered as an elective. Instead, AI experts should be involved due to their expertise in AI applications in the medical field. However, a basic understanding of programming should be acquired early, especially for those who are interested and want to pursue a science career. Most clinicians surveyed opposed including programming skills in the curriculum.

Some lecturers also disagreed with integrating programming knowledge into the curriculum due to student overload. It was emphasized that it is unnecessary for physicians to be able to program neural networks, for example, but that a basic understanding of application knowledge should be established. However, some also emphasized that programming skills and basic computer science knowledge are important, including Python, R, and a theoretical understanding of algorithms. Opinions on the topic were divided and varied depending on the respondents' areas of expertise.

The students interviewed also believed programming skills should be offered to those interested but should not be mandatory. The majority rejected the integration of programming skills into the curriculum, as they are considered too extensive for medical studies and appear to be of minimal relevance to practical application.

AI experts emphasized that physicians need a basic understanding of AI to build confidence in AI applications. Opinions on programming skills were divided, with some considering simple programming skills helpful. Institutional stakeholders also believed medical students do not necessarily need to know how to program but should have field competence in programming. It is expected that not all medical students will be able to program or develop learning methods themselves. However, a basic understanding of programming is viewed as increasingly essential.

Curriculum Scope

The design and scope of AI courses in medicine vary. Including a small section in the curriculum, adapted to the students' abilities, is recommended. For radiologists and image-based diagnosticians, intensive exposure to AI is useful. This should cover a practical application with real medical data to show the application's relevance. Online courses for practicing physicians were suggested to learn the basics.

Curriculum Structure

Regarding AI education in medicine, two main approaches are being considered: lectures and seminars. For lectures, the focus is on introducing mandatory courses blending theory with practical applications. Seminars are viewed as a means to give students early practical experience, enhancing their engagement. Due to the subject's complexity, lecturers are advised to emphasize fundamentals and incorporate concrete examples. However, it is noted that students might find lectures overwhelming, especially without mandatory exams or regular attendance.

Stakeholders emphasized the need for a practical and interactive design when conveying AI content, with clear applications that allow students to experiment for maximum learning impact. Basic AI competencies should be part of the standard medical curriculum, with options for specialization for those interested, particularly those pursuing a scientific career.

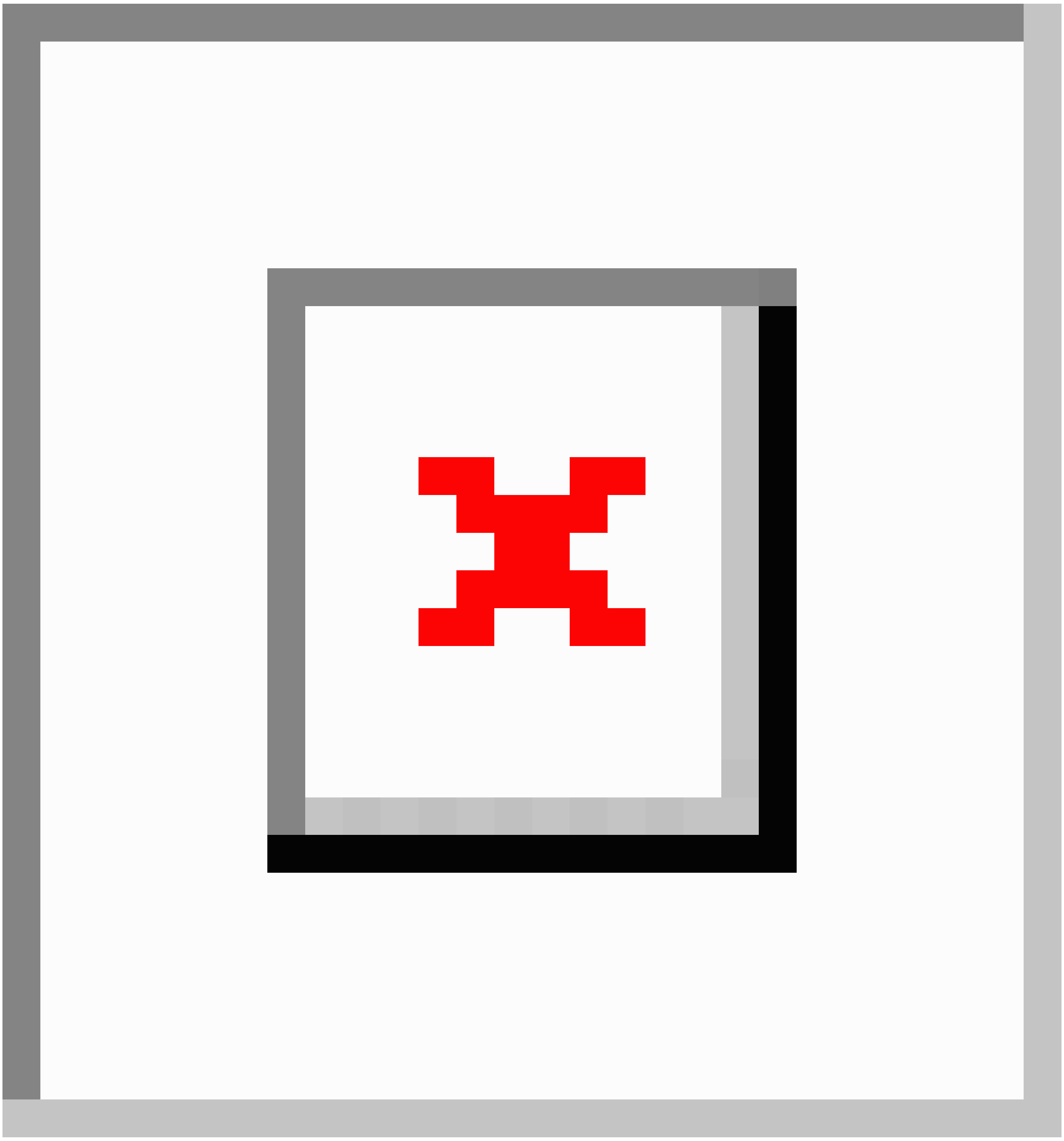
Incorporating AI competencies into medical education is recommended, either through a holistic course or integration into subject-specific areas. Interdisciplinary, research-oriented, and application-oriented seminars and workshops should be established to provide in-depth knowledge. In the future, the curriculum will require substantial restructuring to effectively integrate evolving AI content. Given the rapidly changing nature of AI, the curriculum must remain adaptable.

As shown in Figure 1, the competencies highlighted by different stakeholder groups reveal a range of perspectives and priorities. These focus on the frequency of topics falling into these main categories, offering a nuanced understanding of the thematic landscape.

For example, each stakeholder group highlights the significance of possessing a basic understanding of AI and an awareness of AI-supported applications. Similarly emphasized is the importance of gaining a principal perspective on the opportunities and limitations of AI in medicine, as well as addressing ethical considerations and potential dilemmas. AI experts also emphasized topics such as data literacy, fundamental computer science and mathematics skills, and gaining an overview of potential application areas, while institutional stakeholders focused on interdisciplinary approaches and legal requirements.

XSL•FO
RenderX

Figure 1. Main categories by occurrence of mentions per stakeholder group (qualitative content analysis). The presentation of the primary categories in the qualitative content analysis is based on the frequency of mentions per stakeholder group rather than a quantitative analysis of frequency distribution. AI: artificial intelligence.



Discussion

Principal Findings

The insights gained from the study of stakeholder statements provide valuable perspectives on the different views and interpretations of AI. This provides the basis for answering two central research areas. The first is the understanding of AI, particularly how different interest groups perceive this technology. Second, the focus is on the AI skills that should be taught in medical studies. The different stakeholder groups, including lecturers, health care students, AI experts, institutional stakeholders, and clinicians, contributed to a multifaceted picture. The analysis highlighted similarities and differences in the perception of AI by the various stakeholder groups. These findings from our investigation correspond to step one of Kern's 6-step approach. They are crucial for discussions on implementing AI in health care and underline the need for clear communication, education, and a common understanding of terminology.

Key Competencies for Health Science Students and the Need for a Common Understanding of AI

The qualitative content analysis revealed a broad spectrum of perceptions of AI among the interviewees. Especially in rapidly advancing fields such as AI, creating and maintaining a common

XSL•FO
RenderX

language is essential to enable effective collaboration between different stakeholders. AI is a broad field incorporating many technologies and methods. When introducing AI into the health care system, it is important to acknowledge that different stakeholders involved may have different perceptions of the term AI. A clear definition of this complex term helps prevent misunderstandings, as the field of AI is expansive, encompassing various technologies and methodologies [45,46].

Depending on the contextual background and prior knowledge of the individuals, different descriptions and emphases emerged in the definition of AI. Similarly, ideas concerning the opportunities and limitations of AI in medicine varied depending on individual backgrounds. If consensus is lacking on what AI means in the context of medical education, this can lead to confusion and disagreement on which AI competencies are essential for medical students. This lack of clarity can hinder the development of standardized curricula and educational programs related to AI in medical education, especially when different stakeholders with different backgrounds might be mixing the AI terminologies “strong AI” and “weak AI” [47,48]. Therefore, a clear understanding of the implications and limitations of AI in the medical field is crucial for establishing effective educational guidelines.

Considering these diverse perceptions of AI are particularly relevant when teaching AI skills to medical students. The diversity in the understanding of AI only emphasizes the complexity of the topic; thus, a strong interdisciplinary approach is necessary. Collaboration between physicians, computer scientists, ethicists, and other experts is essential to fully understand the challenges and opportunities of AI in the medical context. For instance, this becomes particularly important when determining which AI applications can enhance the learning experience in specific medical specialties [41].

The diverse perspectives among stakeholders indicate a consensus regarding the essential competencies for health care students concerning AI integration. Recurring themes include practical experience, fundamental digitization knowledge, ethical considerations, and a profound understanding of data and technology. Balancing these competencies is critical to preparing future health care professionals to effectively use AI while maintaining ethical standards and a patient-centered approach. Continued collaboration between stakeholders and the adaptability of medical education curricula will play a key role in achieving these goals.

As illustrated in Figure 1, the stakeholders exhibit substantial diversity in their prioritization of topics and skills, highlighting significant variations in the perceived importance of AI integration into the curriculum. The discussion underscores the importance of a comprehensive approach to AI education in medicine, incorporating practical experience, ethical considerations, and a nuanced understanding of AI’s role in health care. In the context of AI competencies for medical students, they must possess not only medical knowledge but also basic knowledge of AI applications and data literacy, as AI in medicine is becoming increasingly data intensive. The ability to accurately evaluate, manage, and safeguard medical data is essential to ensure that AI technologies can be effectively and ethically deployed in patient care. Therefore, collaboration between stakeholders is essential to develop a curriculum equipping future medical professionals with the necessary competencies to navigate the complexities and opportunities presented by AI in medicine.

The Impact of AI on Shaping Individual Behavior and Societal Outcomes in Medical Training

Since the 1980s, it has been recognized that the introduction of new technologies such as AI does not occur in isolation or independent of societal influences, contrary to the earlier assumption of technological determinism [49,50]. Technology development is shaped by social construction and negotiation processes, where technology emerges as a social construct through human action and influences societal structures and institutions [51]. Interactions related to the introduction of AI in health care can significantly impact how patients are treated and how medical information is used [52]. This concerns not only introducing a new technology per se but also ensuring that it has long-term and positive effects. A key aspect is ensuring that the implementation of AI in health care respects and considers the existing values, norms, and needs of society. Therefore, ethical compatibility and adherence to societal standards are fundamental [53].

Furthermore, AI technologies influence not only medical knowledge but also how doctors, patients, and other stakeholders in health care understand and define their roles. Comprehensive integration of AI requires a holistic approach that not only relies on technological advances but also appropriately considers social dynamics and human aspects.

Our analysis also illustrates the broad understanding of AI, a disparate overall picture of the necessary AI competencies for future medical professionals, and the possibilities and risks associated with implementation. While it is a hot topic among AI experts, health care students are not yet fully aware of the significance of AI, although the technology is expected to enter their professional lives in the future [31,54]. In medical education, students should actively engage with AI, moving beyond passive roles. As well as regulatory, technical, and ethical aspects, it is crucial to consider the sociotechnical dimensions of AI. This is vital, as students must cultivate not only a deep understanding of AI but also an awareness of its societal complexities. For example, Sartori and Bocca [55] emphasize that narratives, whether from the media, scientific community, fiction, or other sources, significantly influence how society perceives and understands technology, including AI. These narratives contribute to the formation of shared understandings, values, and expectations about technology and its potential impact on society [55].

Conclusion

The diverse perspectives on AI among the interviewees underline the requirement for a common language in this rapidly advancing field. Introducing AI into health care necessitates an awareness of varying stakeholder perceptions, emphasizing the importance of a clear definition to prevent misunderstandings. Individual backgrounds shape distinct descriptions and emphases in defining AI, leading to diverse ideas about its opportunities

XSL•FO
RenderX

and limitations, particularly in the context of medical education. When teaching AI skills to medical students, it is essential to address this diversity and adopt a robust interdisciplinary approach to ensure future health care professionals acquire essential knowledge and skills. The results underscore the significance of a comprehensive AI education in medicine, integrating practical experiences, ethical considerations, and a nuanced understanding of AI's role in health care. These competencies will enable medical students to critically evaluate AI technologies and use them responsibly in clinical practice, promoting a more informed and ethically sound integration of AI into health care. The lack of standardization in defining and teaching AI in medical education can lead to uncertainty and potential rejection of the technology. Closing this gap requires gaining insights into the knowledge and skills medical students should acquire regarding the use of AI in medicine. Future studies must focus on awareness of AI and perceived opportunities and risks associated with its implementation. This is also crucial for developing a holistic perspective on competencies within the medical curriculum.

Limitations

While the qualitative nature of our study enabled in-depth exploration and rich insights into the stakeholder perceptions, the limitations associated with the sample size of 38 participants must be acknowledged. The findings may be context specific, and caution is warranted in generalizing beyond our studied group. Notably, some interviewees held dual roles, such as being both lecturers and clinicians. Due to practical constraints, they were interviewed in only one capacity, either as lecturers or clinicians. This limitation underscores the complexity of their perspectives, as their roles encompass multifaceted responsibilities. Using a partially standardized guiding questionnaire, participants were prompted to consider specific questions they might not have spontaneously discussed. While this may have influenced the direction of the conversation, we believe it encouraged participants to reflect. However, it must be acknowledged that a more comprehensive and representative understanding would require further exploration through a quantitative survey. Of note, a subsequent paper will address the opportunities and challenges associated with implementing AI in health care identified by the participating stakeholders.

Acknowledgments

We would like to thank our study assistant Leonie Winterpacht for her help.

We thank the Federal Ministry of Education and Research, Germany for supporting this project (16DHBKI086). We acknowledge support via financing publication fees from Deutsche Forschungsgemeinschaft and the Open Access Publishing Fund of the University of Tübingen.

Authors' Contributions

JAM was responsible for designing and conducting the study, and the acquisition, analysis, and interpretation of data. AHW and TFW were involved in data analyses and interpretation, and revised the manuscript critically. JAM analyzed the research material and wrote the manuscript. KN, SZ, and WF made substantial contributions to the study design and revised the manuscript critically. All authors critically revised the manuscript, and all authors approved the final version of the manuscript and agreed to be accountable for all aspects found therein.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Guiding questions for artificial intelligence experts (example).
[DOCX File, 16 KB - mededu_v10i1e58355_app1.docx]

References

1. Price C. Technology is the answer, but what was the question? Pidgeon Digital. 1979. URL: https://www.pidgeondigital.com/talks/technology-is-the-answer-but-what-was-the-question-/ [accessed 2024-06-07]
2. Stachwitz P, Debatin JF. Digitalization in healthcare: today and in the future [Article in German]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2023 Feb;66(2):105-113. [doi: 10.1007/s00103-022-03642-8] [Medline: 36648499]
3. Lohmann A, Schömig A. „Digitale Transformation“ im Krankenhaus. Gesellschaftliche und rechtliche herausforderungen durch das nebeneinander von ärzten und künstlicher intelligenz. In: Beck S, Kusche C, Valerius B, editors. Digitalisierung, Automatisierung, KI Und Recht [Book in German]: Nomos; 2020.
4. Lin B, Wu S. Digital transformation in personalized medicine with artificial intelligence and the internet of medical things. OMICS 2022 Feb;26(2):77-81. [doi: 10.1089/omi.2021.0037] [Medline: 33887155]
5. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Front Med (Lausanne) 2020 Feb 5;7:27. [doi: 10.3389/fmed.2020.00027] [Medline: 32118012]

XSL•FO
RenderX

6. Fahy N, Williams GA, Habicht T, et al. Use of Digital Health Tools in Europe: Before, During and After COVID-19: European Observatory on Health Systems and Policies; 2021.
7. Saw SN, Ng KH. Current challenges of implementing artificial intelligence in medical imaging. Phys Med 2022 Aug;100:12-17. [doi: 10.1016/j.ejmp.2022.06.003] [Medline: 35714523]
8. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019 Aug 21;5:2055207619871808. [doi: 10.1177/2055207619871808] [Medline: 31467682]
9. Juluru K, Shih HH, Keshava Murthy KN, et al. Integrating AI algorithms into the clinical workflow. Radiol Artif Intell 2021 Aug 4;3(6):e210013. [doi: 10.1148/ryai.2021210013] [Medline: 34870216]
10. Hornegger J. Durch KI wird die Medizin effizienter, individueller und präventiver. In: Knappertsbusch I, Gondlach K, editors. Arbeitswelt und KI 2030: Herausforderungen und Strategien für die Arbeit von morgen [Book in German]: Springer Gabler, Wiesbaden; 2021:321-329. [doi: 10.1007/978-3-658-35779-5]
11. Mirmomeni M, Fazio T, von Cavallar S, Harrer S. Chapter 12 - from wearables to THINKables: artificial intelligence-enabled sensors for health monitoring. In: Sazonov E, editor. Wearable Sensors: Fundamentals, Implementation and Applications, 2nd edition: Academic Press; 2021:339-356.
12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
13. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019 Oct 4;7:e7702. [doi: 10.7717/peerj.7702] [Medline: 31592346]
14. Ng KH, Wong JHD. A clarion call to introduce artificial intelligence (AI) in postgraduate medical physics curriculum. Phys Eng Sci Med 2022 Mar;45(1):1-2. [doi: 10.1007/s13246-022-01099-2] [Medline: 35006576]
15. Mosch L, Agha-Mir-Salim L, Sarica MM, Balzer F, Poncette AS. Artificial intelligence in undergraduate medical education. Stud Health Technol Inform 2022 May 25;294:821-822. [doi: 10.3233/SHTI220597] [Medline: 35612217]
16. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec 11;19(1):460. [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
17. Wallis C. How artificial intelligence will change medicine. Nature 2019 Dec;576(7787):S48. [doi: 10.1038/d41586-019-03845-1] [Medline: 31853072]
18. Wartman SA, Combs CD. Reimagining medical education in the age of AI. AMA J Ethics 2019 Feb 1;21(2):E146-E152. [doi: 10.1001/amajethics.2019.146] [Medline: 30794124]
19. Dumić-Čule I, Orešković T, Brkljačić B, Kujundžić Tiljak M, Orešković S. The importance of introducing artificial intelligence to the medical curriculum - assessing practitioners' perspectives. Croat Med J 2020 Oct 31;61(5):457-464. [doi: 10.3325/cmj.2020.61.457] [Medline: 33150764]
20. Kundu S. How will artificial intelligence change medical training? Commun Med (Lond) 2021 Jun 30;1:8. [doi: 10.1038/s43856-021-00003-5] [Medline: 35602202]
21. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020 Jun 19;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
22. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. J Med Educ Curric Dev 2021 Sep 6;8:23821205211036836. [doi: 10.1177/23821205211036836] [Medline: 34778562]
23. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659. [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
24. Lewis SJ, Gandomkar Z, Brennan PC. Artificial intelligence in medical imaging practice: looking to the future. J Med Radiat Sci 2019 Dec;66(4):292-295. [doi: 10.1002/jmrs.369] [Medline: 31709775]
25. Flasdick J, Mah DK, Bernd M, Rampelt F. Micro-credentials and micro-degrees current developments and potentials for educational practice based on the example of the AI campus. ResearchGate. 2023 Feb. URL: https://tinyurl.com/mryp7p7j [accessed 2024-06-07]
26. Hu R, Fan KY, Pandey P, et al. Insights from teaching artificial intelligence to medical students in Canada. Commun Med (Lond) 2022 Jun 3;2(1):63. [doi: 10.1038/s43856-022-00125-4] [Medline: 35668847]
27. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R. Grounded in reality: artificial intelligence in medical education. JAMIA Open 2023 Jun 1;6(2):ad037. [doi: 10.1093/jamiaopen/ooad037] [Medline: 37273962]
28. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. BMC Med Educ 2021 Feb 18;21(1):112. [doi: 10.1186/s12909-021-02546-6] [Medline: 33602196]
29. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285. [doi: 10.2196/19285] [Medline: 32602844]
30. Thomas PA, Kern DE, Hughes MT, Tackett SA, Chen BY, editors. Curriculum Development for Medical Education: A Six-Step Approach: JHU press; 2022.

XSL•FO
RenderX

31. Nyein KP, Gregory ME. Needs Assessment and Stakeholders in Medical Simulation Curriculum Development: StatPearls Publishing; 2023. [Medline: 32119390]
32. Magaldi D, Berler M. Semi-structured interviews. In: Zeigler-Hill V, Shackelford TK, editors. Encyclopedia of Personality and Individual Differences: Springer Cham; 2018:1-6. [doi: 10.1007/978-3-319-28099-8]
33. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
34. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
35. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
36. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 7;8(2):e35587. [doi: 10.2196/35587] [Medline: 35671077]
37. Blease C, Kharko A, Bernstein M, et al. Machine learning in medical education: a survey of the experiences and opinions of medical students in Ireland. BMJ Health Care Inform 2022 Feb;29(1):e100480. [doi: 10.1136/bmjhci-2021-100480] [Medline: 35105606]
38. Katznelson G, Gerke S. The need for health AI ethics in medical school education. Adv Health Sci Educ Theory Pract 2021 Oct;26(4):1447-1458. [doi: 10.1007/s10459-021-10040-3] [Medline: 33655433]
39. Ganapathi S, Duggal S. Exploring the experiences and views of doctors working with artificial intelligence in English healthcare; a qualitative study. PLoS One 2023 Mar 2;18(3):e0282415. [doi: 10.1371/journal.pone.0282415] [Medline: 36862694]
40. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
41. Busch F, Adams LC, Bressem KK. Biomedical ethical aspects towards the implementation of artificial intelligence in medical education. Med Sci Educ 2023 Jun 7;33(4):1007-1012. [doi: 10.1007/s40670-023-01815-x] [Medline: 37546190]
42. Helfferich C. Leitfaden- und Experteninterviews. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung [Book in German]: Springer VS, Wiesbaden; 2019:669-686. [doi: 10.1007/978-3-658-21308-4]
43. Kuckartz U. Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung [Book in German]: Beltz Juventa; 2012.
44. Mayring P. Qualitative content analysis: demarcation, varieties, developments. Forum Qual Soc Res 2019 Sep;20(3). [doi: 10.17169/fqs-20.3.3343]
45. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017 Aug 8;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
46. Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci 2022;3(2):158. [doi: 10.1007/s42979-022-01043-x] [Medline: 35194580]
47. Wang P. On defining artificial intelligence. J Artif Gen Intelligence 2019 Jan 1;10(2):1-37. [doi: 10.2478/jagi-2019-0002]
48. Monett D, Lewis CWP. Getting clarity by defining artificial intelligence—a survey. In: Müller VC, editor. Philosophy and Theory of Artificial Intelligence 2017: Springer International Publishing; 2018.
49. Dolata U. Technologische Innovationen und sektoraler Wandel: eingriffstiefe, adaptionsfähigkeit, transformationsmuster: ein analytischer ansatz [Article in German]. Zeitschrift Soziologie 2008 Feb;37(1):42-59. [doi: 10.1515/zfsoz-2008-0103]
50. Sartori L, Theodorou A. A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics Inf Technol 2022 Jan 24;24(1). [doi: 10.1007/s10676-022-09624-3]
51. Klein HK, Kleinman DL. The social construction of technology: structural considerations. Sci Technol Hum Values 2002 Jan;27(1):28-52. [doi: 10.1177/016224390202700102]
52. Bohr A, Memarzadeh K. Chapter 2 - the rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare: Academic Press; 2020:25-60. [doi: 10.1016/B978-0-12-818438-7.00002-2]
53. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022 Mar;296(114782):114782. [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]
54. Tolsgaard MG, Pusic MV, Sebok-Syer SS, et al. The fundamentals of artificial intelligence in medical education research: AMEE guide no. 156. Med Teach 2023 Jun;45(6):565-573. [doi: 10.1080/0142159X.2023.2180340] [Medline: 36862064]
55. Sartori L, Bocca G. Minding the gap(s): public perceptions of AI and socio-technical imaginaries. AI Soc 2022 Mar 26;38(2):443-458. [doi: 10.1007/s00146-022-01422-1]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

Edited by B Lesselroth; submitted 13.03.24; peer-reviewed by S Ito, SQ Yoong; revised version received 16.04.24; accepted 07.05.24; published 12.06.24.

Please cite as:
Moldt JA, Festl-Wietek T, Fuhl W, Zabel S, Claassen M, Wagner S, Nieselt K, Herrmann-Werner A
Assessing AI Awareness and Identifying Essential Competencies: Insights From Key Stakeholders in Integrating AI Into Medical Education
JMIR Med Educ 2024;10:e58355
URL: https://mededu.jmir.org/2024/1/e58355
doi:10.2196/58355

© Julia-Astrid Moldt, Teresa Festl-Wietek, Wolfgang Fuhl, Susanne Zabel, Manfred Claassen, Samuel Wagner, Kay Nieselt, Anne Herrmann-Werner. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.6.2024. 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.

XSL•FO
RenderX

Assessing GPT-4's Performance in Delivering Medical Advice: Comparative Analysis With Human Experts

Eunbeen Jo[1,*], BA; Sanghoun Song[2,*], PhD; Jong-Ho Kim[3,4], PhD; Subin Lim[5], MD; Ju Hyeon Kim[5], MD; Jung-Joon Cha[5], MD, PhD; Young-Min Kim[6], PhD; Hyung Joon Joo[1,3,4], MD, PhD

1
2
3
4
5
6

*these authors contributed equally

Corresponding Author:
Hyung Joon Joo, MD, PhD

Abstract

Background: Accurate medical advice is paramount in ensuring optimal patient care, and misinformation can lead to misguided decisions with potentially detrimental health outcomes. The emergence of large language models (LLMs) such as OpenAI's GPT-4 has spurred interest in their potential health care applications, particularly in automated medical consultation. Yet, rigorous investigations comparing their performance to human experts remain sparse.

Objective: This study aims to compare the medical accuracy of GPT-4 with human experts in providing medical advice using real-world user-generated queries, with a specific focus on cardiology. It also sought to analyze the performance of GPT-4 and human experts in specific question categories, including drug or medication information and preliminary diagnoses.

Methods: We collected 251 pairs of cardiology-specific questions from general users and answers from human experts via an internet portal. GPT-4 was tasked with generating responses to the same questions. Three independent cardiologists (SL, JHK, and JJC) evaluated the answers provided by both human experts and GPT-4. Using a computer interface, each evaluator compared the pairs and determined which answer was superior, and they quantitatively measured the clarity and complexity of the questions as well as the accuracy and appropriateness of the responses, applying a 3-tiered grading scale (low, medium, and high). Furthermore, a linguistic analysis was conducted to compare the length and vocabulary diversity of the responses using word count and type-token ratio.

Results: GPT-4 and human experts displayed comparable efficacy in medical accuracy ("GPT-4 is better" at 132/251, 52.6% vs "Human expert is better" at 119/251, 47.4%). In accuracy level categorization, humans had more high-accuracy responses than GPT-4 (50/237, 21.1% vs 30/238, 12.6%) but also a greater proportion of low-accuracy responses (11/237, 4.6% vs 1/238, 0.4%; P=.001). GPT-4 responses were generally longer and used a less diverse vocabulary than those of human experts, potentially enhancing their comprehensibility for general users (sentence count: mean 10.9, SD 4.2 vs mean 5.9, SD 3.7; P<.001; type-token ratio: mean 0.69, SD 0.07 vs mean 0.79, SD 0.09; P<.001). Nevertheless, human experts outperformed GPT-4 in specific question categories, notably those related to drug or medication information and preliminary diagnoses. These findings highlight the limitations of GPT-4 in providing advice based on clinical experience.

Conclusions: GPT-4 has shown promising potential in automated medical consultation, with comparable medical accuracy to human experts. However, challenges remain particularly in the realm of nuanced clinical judgment. Future improvements in LLMs may require the integration of specific clinical reasoning pathways and regulatory oversight for safe use. Further research is needed to understand the full potential of LLMs across various medical specialties and conditions.

(JMIR Med Educ 2024;10:e51282) doi:10.2196/51282

KEYWORDS

GPT-4; medical advice; ChatGPT; cardiology; cardiologist; heart; advice; recommendation; recommendations; linguistic; linguistics; artificial intelligence; NLP; natural language processing; chatbot; chatbots; conversational agent; conversational agents; response; responses

XSL•FO
RenderX

Introduction

As a large language model (LLM), the GPT developed by OpenAI generates human-like text [1-3], distinguishing it from other specialized deep learning models that are limited to solving specific problems within predetermined domains [4]. In the medical field, GPT has the potential to augment medical education [5], provide clinical decision support [6], and enhance public health initiatives [7]. An impressive achievement of GPT-3.5 is its success in meeting the passing threshold for the United States Medical Licensing Examination [8], demonstrating its ability to offer medical advice comparable to that of trained professionals [9]. The latest iteration, GPT-4 [10,11], is anticipated to exhibit advancements in processing complex medical language, formulating patient care suggestions, and making preliminary diagnostic predictions, which inspires cautious optimism for its future applications in the medical domain [12].

Cardiovascular diseases are a leading cause of death worldwide, highlighting the critical need for precise and reliable information in this domain [13]. During the initial stages of the SARS-CoV-2 pandemic, overstated claims about the cardiovascular implications of the virus potentially escalated public unease and undermined trust in empirical findings [14]. The distribution of speculative or inaccurate information would have had a detrimental effect on the pandemic response strategies. It is paramount to emphasize that inaccuracies or misconceptions in cardiological advice can lead to severe consequences. Hence, there is a pressing need for rigorous validation of all sources of information, whether derived from human experts or advanced computational models such as GPT-4.

Moreover, the generation of "hallucinatory" or erroneous responses by GPT raises concerns about nonmedical expert users unintentionally accepting incorrect information as valid [15,16]. Consequently, proposals for regulatory oversight of LLMs have emerged, including the establishment of a new regulatory category specifically addressing LLM-related challenges and risks [4]. Therefore, it is crucial to develop auditing procedures capable of capturing the intricacies of LLM-associated risks, necessitating a balanced evaluation of the potential benefits and risks inherent in LLMs [17,18]. To delve deeper into this matter, this study applied real-world health consultations from general users to human experts through an internet portal, using the most recent iteration of this technology, GPT-4. The responses provided by both human experts and GPT-4 were subsequently evaluated by a panel of 3 independent cardiologists to gain a nuanced understanding of the potential benefits and risks associated with GPT-4.

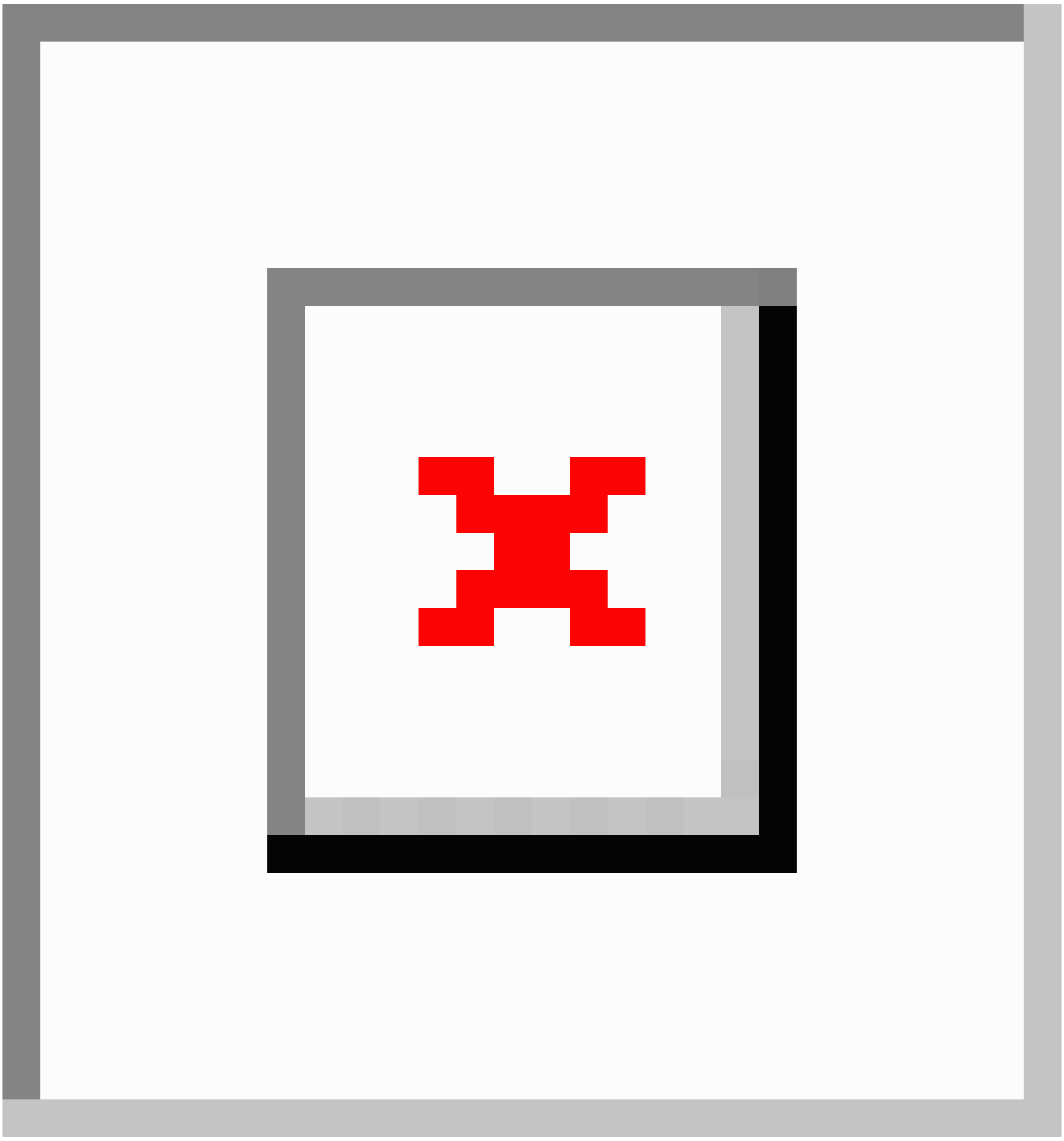
Methods

Data Collection

Figure 1 illustrates the study design. We collected question-and-answer data related to cardiology from the Korean search portal NAVER, focusing on 264 cases. NAVER is Korea's largest search engine, and its web-based questions and answers forums, called "Jisik-In," have previously been used in medical research [19,20]. The data set covered the period from July 13, 2020, to July 12, 2021, and included medical inquiries posed by portal users and the corresponding responses provided by human experts. These experts are doctors who have graduated from a college of medicine or medical school, passed the Korean Medical Licensing Examination, and hold legal accreditations as certified specialists in their respective medical fields from the Ministry of Health and Welfare. They are not restricted by character limits when answering users' questions on the portal site. The questions were categorized into 2 types: binary and open-ended. Further, 6 distinct categories were defined based on the questions' intent. All collected data were in Korean text form. To ensure the analysis was focused on sufficiently detailed and substantive exchanges, we specifically selected questions that contained more than 100 characters according to the Korean alphabet and answers provided by human experts that comprised at least 200 characters. This approach was aimed at filtering out overly simple queries and ensuring that the responses were elaborate enough for a thorough comparison. Additionally, to maintain a consistent and fair comparison basis between the capabilities of GPT-4 and human experts, we excluded 13 cases from the total data set that contained multimedia content such as videos or images. Finally, 251 cases were selected for the study after applying these criteria.

XSL•FO
RenderX

Figure 1. Study design and evaluation process. A data set consisting of 251 cardiology-specific question-answer pairs was collected from the NAVER portal over a 1-year period, from July 13, 2020, to July 12, 2021. A licensed medical professional is the person who answered the portal user's question. The questions covered 6 domain categories and included both binary and open-ended types. From May 5 to 8, these questions were inputted into GPT-4 to generate the corresponding GPT-4 responses. Following that, a panel of 3 cardiologists reviewed and evaluated the questions along with the answers provided by human experts and GPT-4. The evaluation criteria focused on assessing the complexity and clarity of the questions as well as the accuracy and appropriateness of the responses from both human experts and GPT-4.



GPT Answer Generation

Answers to the collected questions were generated using OpenAI's GPT-4 model, released on March 14, 2023 [10]. From May 5 to 8, 2023, a total of 3 researchers used this model via the OpenAI website to generate GPT-4 answers. The total data set of questions to be entered into the GPT-4 was distributed to the 3 researchers in the form of a spreadsheet. Each original Korean question was directly fed into the GPT-4 prompt without any supplementary input. The researchers saved the generated answer in a spreadsheet. Each question input was done in a new session by clicking the "New chat" button.

Question and Answer Evaluation

Once the data were randomly shuffled, answers from both GPT-4 and human experts were anonymized and labeled as answer 1 and answer 2, respectively, ensuring the 3 independent cardiologist reviewers were blinded to the source of each response. Each of these reviewers is a board-certified physician in internal medicine and has undergone more than 4 years of

XSL•FO
RenderX

fellow training in cardiology subspecialty. A panel of 3 cardiologists assessed the question set along with the anonymized answers. The evaluation was conducted using a computer interface. Each evaluator assessed the clarity and complexity of the questions as well as the accuracy and appropriateness of the answers. To quantitatively measure these aspects, a 3-tiered grading scale (low, medium, and high) was used (Multimedia Appendix 1). Additionally, each evaluator determined which answer (the GPT-4's answer or the human expert's answer) showed superior accuracy and appropriateness in relation to the question posed.

To further elucidate, the Kendall *W* concordance analysis revealed the following coefficient values indicating the level of agreement among the evaluators: 0.44 for the appropriateness of the human expert answers, 0.40 for the appropriateness of the GPT-4 answers, 0.43 for the medical accuracy of the human expert answers, and 0.40 for the medical accuracy of the GPT answers. Moreover, when making a binary choice determining the superiority of appropriateness between the human expert and GPT-4 answers, the coefficient was 0.42, and for determining the superiority of medical accuracy between the two, it was 0.45. These values, falling in the range of 0.40-0.60, denote a moderate agreement, showcasing a significant level of reliability in our study findings.

Ethical Considerations

This research project was approved by the institutional review board of Korea University Anam Hospital (IRB 2023AN0280). The research was conducted in accordance with the Helsinki Declaration. Informed consent was obtained from all 3 participating cardiologists.

Linguistic Analysis

The Korean Sentence Separator 4.5.1 was used to segment the text into individual sentences. For text tokenization, the Korean medical bidirectional encoder representations from the transformer model, which was specifically designed for Korean medical text analysis, was used [21]. To evaluate lexical diversity, the type-token ratio (TTR) was computed for each set of responses [22,23]. The TTR, which represents the ratio of unique words to the total number of words in a text, was determined after the responses were tokenized [22,23].

Statistical Analysis

To discern statistically significant differences across categorical outcomes, we used the chi-square test or Fisher exact test as appropriate, depending on the expected frequencies within the categories. For continuous variables, comparison across groups was conducted using either the parametric unpaired 2-tailed *t* test or the nonparametric Mann-Whitney test, based on the distribution of the data. Interrater agreement among the 3 cardiologist evaluators was quantitatively assessed using the Kendall *W* concordance analysis. The association between the complexity and clarity of questions and the quality of responses was investigated using the Spearman rank correlation coefficient. All statistical analyses were conducted using SAS 9.4 (SAS Institute Inc) and R program (version 3.6.1; R Foundation for Statistical Computing).

Results

Both the number of words and sentences per answer were significantly higher for GPT-4 answers than for human expert answers (word count: mean 190, SD 75.2 for GPT-4 vs mean 139, SD 95.6 for humans; *P*<.001 and sentence count: mean 10.9, SD 4.2 for GPT-4 vs mean 5.9, SD 3.7 for humans; *P*<.001; Table 1). The GPT-4 answers exhibited lower lexical diversity, as measured by the TTR, compared to the answers provided by human experts. This suggests that GPT-4 answers may be perceived as more comprehensible and similar to human conversations rather than written text (TTR: mean 0.69, SD 0.07 for GPT-4 vs mean 0.79, SD 0.09 for humans; *P*<.001).

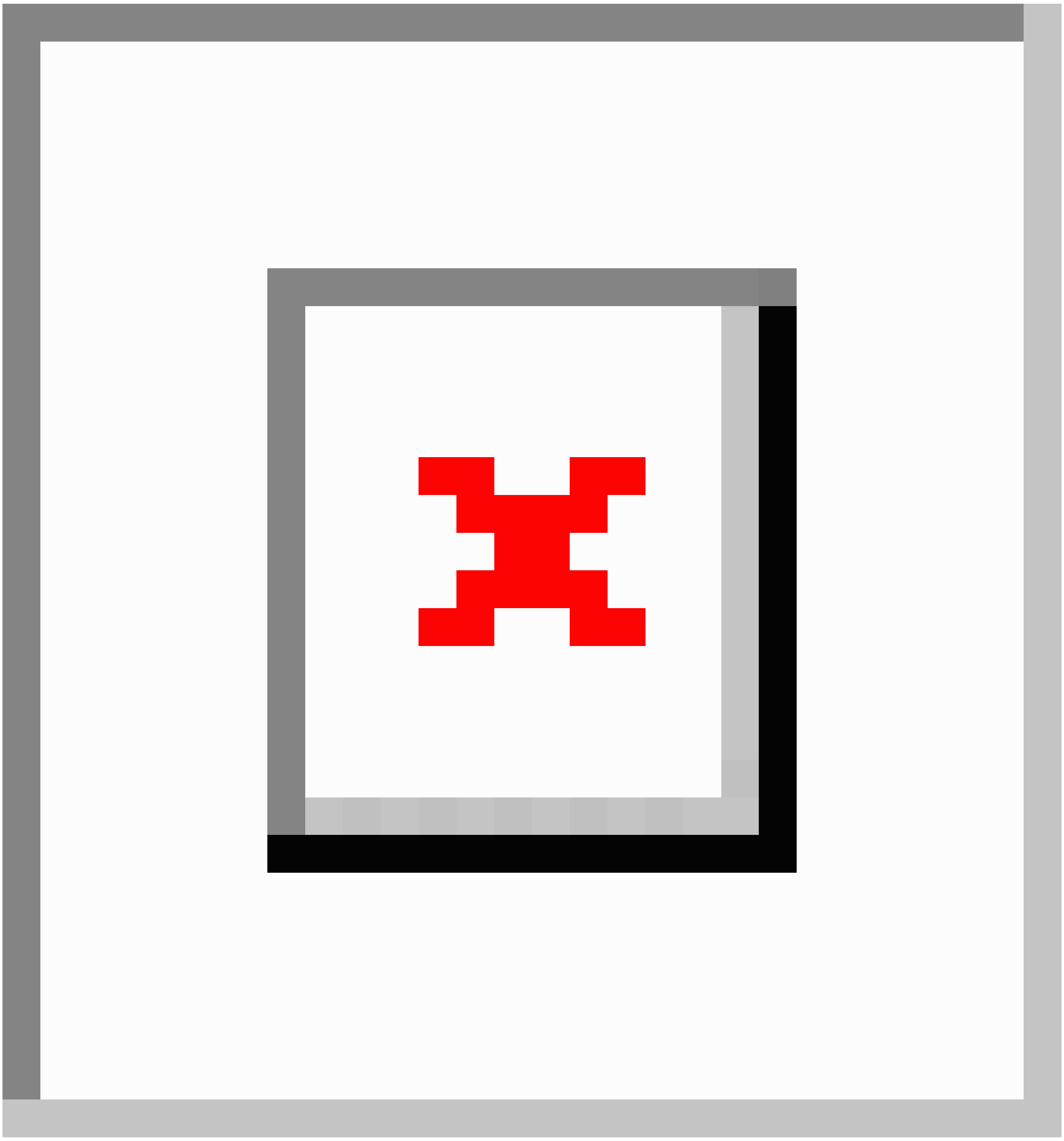
Table . Linguistic difference between GPT-4 and human expert answers.

Characteristics	GPT-4, mean (SD)	Human, mean (SD)	*P* value
Word count per answer	190 (75.2)	139 (95.6)	<.001
Sentence count per answer	10.9 (4.2)	5.9 (3.7)	<.001
Type-token ratio	0.69 (0.07)	0.79 (0.09)	<.001

Figure 2 presents an analysis of the medical accuracy between GPT-4 and human expert answers. When cardiologists were asked to evaluate which answers were more medically accurate, the responses slightly favored the human expert answers (132/251, 52.6% vs 119/251, 47.4%; *P*=.41; Figure 2A). Dividing medical accuracy into low, medium, and high levels, a significant proportion of human expert answers were ranked as highly accurate compared to GPT-4 (50/237, 21.1% vs 30/238, 12.6%; *P*<.001; Figure 2B). However, the rate of low accuracy was also higher for the human expert answers (11/237, 4.6% vs 1/238, 0.4%; *P*=.007). This counterintuitive observation underscores the potential of LLMs to bridge gaps in human work in real-world scenarios.

XSL•FO
RenderX

Figure 2. Medical accuracy between GPT-4 and human expert answers. (A) Survey results indicating preference for GPT-4 and human expert answers based on perceived medical accuracy. (B) Analysis of perceived medical accuracy, categorized as low, medium, and high for both GPT-4 and human expert answers. (C and D) Relationship between question complexity or clarity and the perceived medical accuracy of GPT-4 and human expert answers. (E) Comparison of variations in perceived medical accuracy between GPT-4 and human expert answers, depending on question type. (F) Comparison of perceived medical accuracy between GPT-4 and human expert answers across different categories of question intent. (G and H) Comparison of word count per answer and type-token ratio between human expert and GPT-4 answers when evaluated for medical accuracy.



In terms of question complexity and ambiguity, GPT-4 demonstrates an advantage. The more complex and ambiguous the question, the higher the medical accuracy of GPT-4's answers. Conversely, human experts excel in dealing with simpler and clearer questions, although without statistically significant differences (*P*=.19; Figure 2C and *P*=.30; Figures 2D, 3C, and 3D). The difference in medical accuracy between human and GPT-4 answers remained below 10% across different question types (*P*=.39; Figure 2E).

Interestingly, when analyzing question categories based on the intent, numerical differences were observed, but without statistical significance (*P*=.20; Figure 2F). Human experts outperformed GPT-4 in responding to questions related to drugs or medications and preliminary diagnoses, scoring higher than GPT-4 (drug or medication: 12/18, 66.7% vs 6/18, 33.3% and preliminary diagnosis: 43/70, 61.4% vs 27/70, 38.6%). Conversely, GPT-4 surpassed human experts in addressing queries regarding the necessity of hospital visits and guidance for clinical departments (hospital visit necessity: 9/22, 40.9%

XSL•FO
RenderX

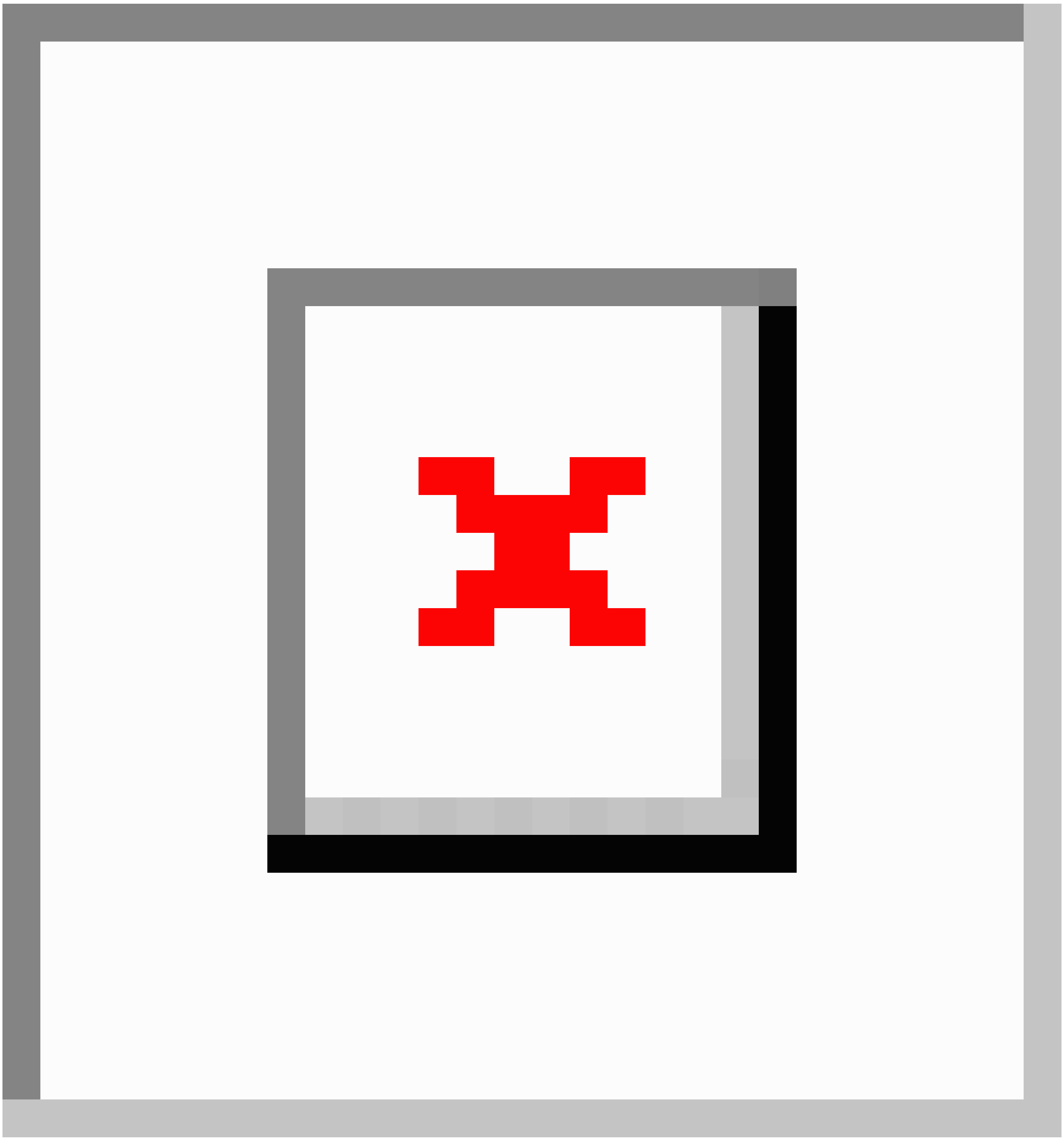
vs 13/22, 59.1% and clinical department guidance: 15/33, 45.5% vs 18/33, 54.5%).

In the linguistic analysis, when the medical accuracy of a human expert's answer exceeded that of GPT-4, the human expert's answers typically had a higher word count and lower TTR compared to cases where GPT-4's answers were deemed more medically accurate (word count per answer: mean 162, SD 102.6 vs mean 114, SD 80.3; *P*<.001; Figure 2G and TTR: mean 0.78, SD 0.09 vs mean 0.80, SD 0.09; *P*=.02; Figure 2H). This implies that the more the response resembles a real conversation—longer and easier—the higher the perceived medical accuracy according to cardiology experts. This observation indicates a potential area for quality control in human expert responses and highlights the consistent performance of GPT-4 in terms of response length and lexical variation.

Next, a comparative analysis between GPT-4 and human expert answers was conducted in terms of answer appropriateness (Figure 3). When assessing whether GPT-4 or human expert answers were more appropriate for the posed questions, GPT-4 was rated as superior (GPT-4: 135/251, 53.8% vs humans: 116/251, 46.2%; *P*=.23; Figure 3A). Similar to the medical accuracy analysis, when categorizing appropriateness into low, medium, and high, both GPT-4 and human expert answers showed a comparable distribution across these segments (*P*=.26; Figure 3B). Notably, mirroring the findings from the medical accuracy analysis, the frequency of answers deemed to have low appropriateness was numerically higher for human experts (7/240, 2.9% vs 2/241, 0.8%; *P*=.03), suggesting the possibility of human shortcomings. The investigations related to question complexity, clarity, and type displayed numerical trends similar to those observed in the medical accuracy analysis, although no statistical differences were observed (*P*=.20; *P*=.60; and *P*=.66; Figure 3C-E). The analysis based on question intent showed no significant statistical discrepancies between the proportions of cases where human expert answers were deemed more appropriate and those where GPT-4 answers were considered more appropriate. Interestingly, GPT-4 was rated as more appropriate than human experts in all other categories, except for the question category of preliminary diagnosis (*P*=.58; Figure 3F). When human expert answers were considered more appropriate than those of GPT-4, the corresponding answers had a higher word count and lower TTR compared to cases where GPT-4 answers were deemed more appropriate (word count per answer: mean 121, SD 79.3 vs mean 160, SD 108.1; *P*=.001; Figure 3G and TTR: mean 0.80, SD 0.09 vs mean 0.77, SD 0.09; *P*=.02; Figure 3H). Similar to medical accuracy, these findings suggest that longer responses resembling genuine conversations are evaluated as more appropriate.

XSL•FO
RenderX

Figure 3. Answer appropriateness between GPT-4 and human expert answers. (A) Survey results indicating preference for GPT-4 and human expert responses based on perceived answer appropriateness. (B) Analysis of perceived answer appropriateness, categorized as low, medium, and high for both GPT-4 and human expert answers. (C and D) Relationship between question complexity or clarity and the perceived answer appropriateness of GPT-4 and human expert answers. (E) Comparison of variations in perceived answer appropriateness between GPT-4 and human expert answers depending on question type. (F) Comparison of perceived answer appropriateness between GPT-4 and human expert answers across different categories of question intent. (G and H) Comparison of word count per answer and type-token ratio between human expert and GPT-4 answers when evaluated for appropriateness.

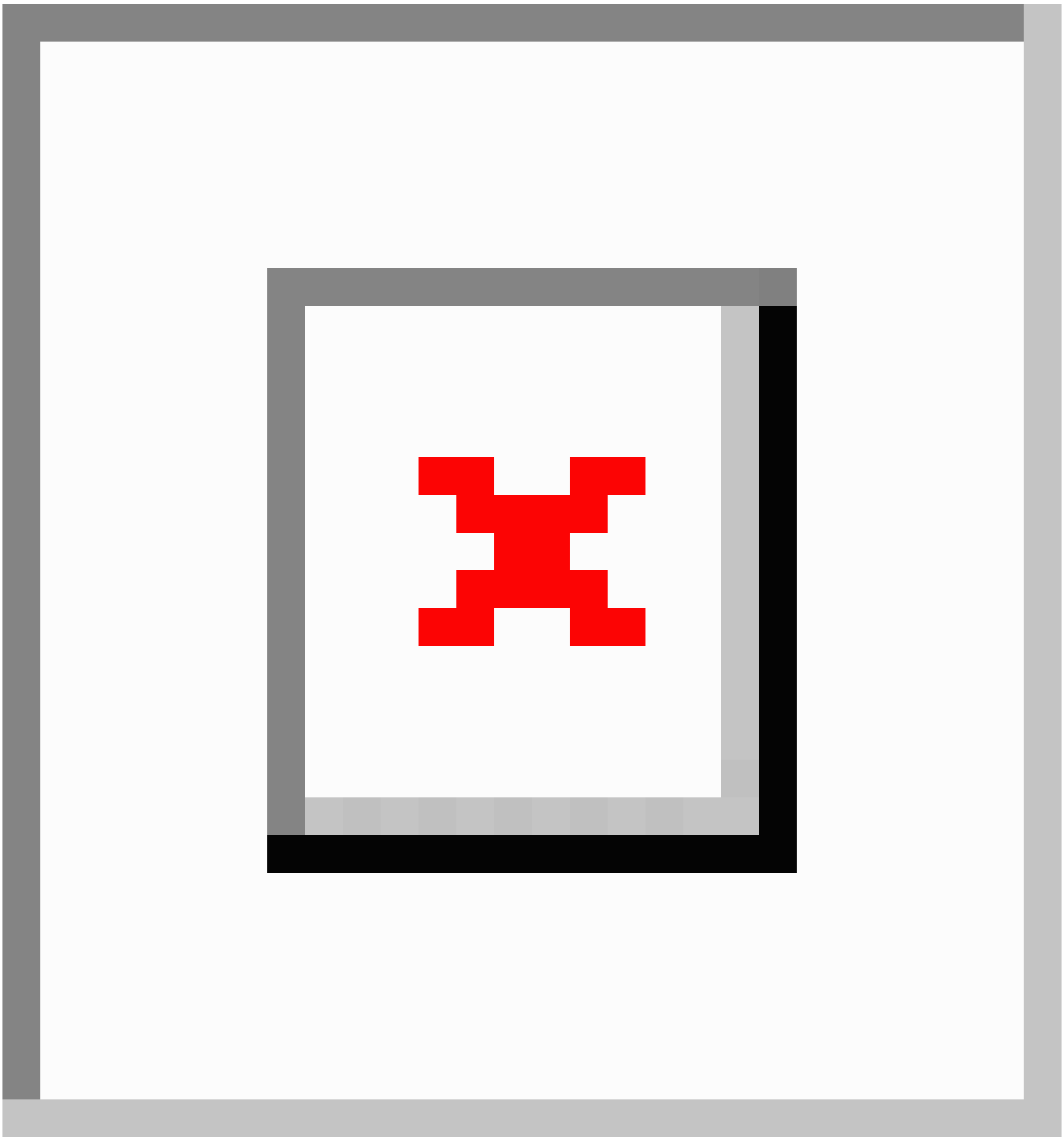


For the 251 questions assessed, all 3 independent cardiologists rated the GPT-4 answers as superior in 18% (45/251) of cases in terms of medical accuracy. In an additional 29% (74/251) of the cases, the majority (2 of 3) of cardiologists endorsed the GPT-4 answers. Conversely, human expert answers were unanimously considered more accurate in 20% (50/251) of cases, with the majority of cardiologists agreeing with human experts in 33% (82/251) of cases (Figure 4). In terms of answer appropriateness, all 3 cardiologists agreed that the GPT-4 answers were superior in 15% (38/251) of cases. The majority of cardiologists found GPT-4 answers to be more appropriate in another 39% (97/251) of cases. Human experts, however, received unanimous approval for the appropriateness of their answers in 18% (70/251) of cases and majority approval in an additional 28% (46/251; Figure 5). These figures highlight the noteworthy performance of GPT-4 from a medical standpoint. Examining illustrative cases, GPT-4 stands out for delivering medical information resembling the content of medical textbooks

XSL•FO
RenderX

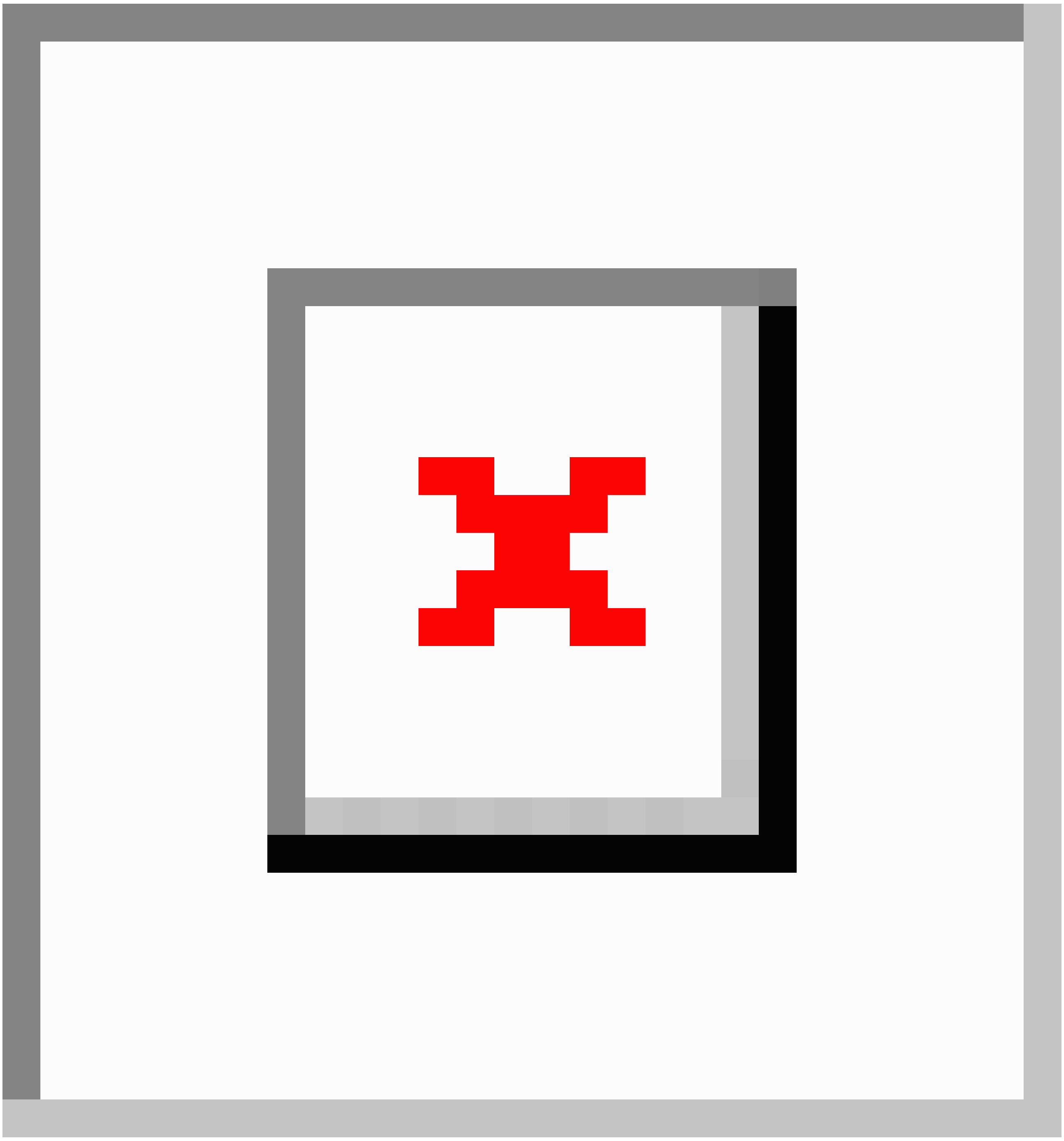
and dictionaries. Additionally, GPT-4 demonstrates strength in thoroughly addressing every user’s question, leaving no queries unanswered. In contrast, human experts leverage their advantage in providing heuristic information informed by their clinical experience, especially when questions require elements of clinical judgment.

Figure 4. Evaluation result and representative cases comparing medical accuracy between GPT-4 and human expert answers. (A) A case where the GPT-4 answer received superior medical accuracy ratings from all 3 evaluators. (B) A case where a human expert received superior medical accuracy ratings from all 3 evaluators.



XSL•FO
RenderX

Figure 5. Evaluation result and representative cases comparing answer appropriateness between GPT-4 and human expert answers. (A) A case where the GPT-4 answer received superior appropriateness ratings from all 3 evaluators. (B) A case where a human expert answer received superior appropriateness ratings from all 3 evaluators.



Discussion

Principal Findings

This research uniquely implemented real-world health consultations involving general users and human experts, comparing the answers provided by human experts and GPT-4. Three independent cardiologists appraised the answers to discern the potential advantages and disadvantages of using GPT-4 in the medical advice domain. This study demonstrated comparable levels of medical accuracy between GPT-4 and human experts. Notably, human expert answers had a higher proportion of answers classified as having low medical accuracy compared to those from GPT-4.

Another significant finding suggests the benefits of articulating medical advice in a conversational style, which positively impacts medical accuracy and relevance to queries. This style proved effective in responding to all questionnaire requests, leading to higher answer ratings and demonstrating the potential of GPT-4 in providing medical advice. Notably, GPT-4's answers consistently displayed appropriate length and lexical variation compared to those of human experts. The findings of this study underscore the potential of GPT-4 in medical education, particularly in enhancing the learning experience

XSL•FO
RenderX

through its ability to simulate conversational medical advice with accuracy comparable to human experts. Integrating GPT-4 into educational frameworks could offer an innovative approach to medical education, facilitating adaptive learning and preparing students for the digital evolution in health care. This suggests a promising avenue for future research and application in the field of medical education, highlighting the importance of incorporating advanced AI tools like GPT-4 to complement traditional educational methods.

Comparison to Prior Work

An important consideration is the linguistic scope of our findings. This study was conducted in Korean, which naturally raises questions about its generalizability to other languages. Recent studies and OpenAI's own documentation suggest that GPT-4's performance in non-English languages, including medical contexts, has improved compared to previous versions [11,24,25]. Takagi et al [24] compared the performance of GPT-3.5 and GPT-4 using 254 questions from the Japanese Medical Licensing Examination, revealing that GPT-4 exhibited a 29.1% improvement over GPT-3.5. They highlighted that GPT-4's enhanced non-English language processing capabilities were instrumental in its ability to pass the medical licensing examination. In addition, Wang et al [25] conducted a study comparing the performance of GPT-3.5 and GPT-4 on English and Chinese data sets for the Chinese Medical Licensing Examination, showing a significant improvement in accuracy for Chinese compared to English. This study showed that the medical advice provided by GPT-4 was comparable in medical accuracy to that provided by human experts. Based on previous research and the findings of this study, it has been found that GPT-4 can effectively process specialized medical information in various non-English languages, including Korean. This indicates its potential for use in patient education and the dissemination of medical knowledge.

Strengths and Limitations

Despite its strengths, GPT-4's capability to provide advice based on clinical experience differs notably from that of human experts. Furthermore, quantitative analysis revealed potential discrepancies between GPT-4 and human expert responses, depending on the intent of the question. Numerous studies are currently underway to identify appropriate regulatory measures for the use of LLMs [4]. The findings of this investigation are anticipated to facilitate subsequent research aimed at identifying tasks in the medical field that GPT-4 excels in. This, in turn, could expedite the development of technology to enhance the quality of medical services and promote public health.

This study has several limitations to consider. First, its focus on cardiology might limit the generalizability of the results to other medical specialties. Second, the sample size for the answer evaluation, which consisted of only 3 cardiologists, could have been larger for a more robust analysis. Furthermore, since the evaluations were conducted solely by cardiologists, there is potential for reporting bias where certain aspects of the answers might be overemphasized or underrepresented. Inclusion of professionals from other domains could have provided a broader assessment. Future studies should aim to involve larger sample sizes and encompass a wider range of medical specialties. Moreover, integrating patients' perspectives could offer further insights into the acceptability and perceived utility of artificial intelligence–powered medical advice.

Conclusions

In conclusion, this study revealed the promising capabilities of GPT-4 in providing medically accurate and appropriate responses comparable to human experts. The additional benefits of GPT-4 include consistent proficiency in maintaining appropriate response length and lexical variation. However, GPT-4 showed some disadvantages in providing advice based on clinical experience as well as variation in its performance depending on question intent. Despite these challenges, this study suggests that LLMs such as GPT-4 hold significant potential in augmenting medical education, providing medical advice.

Acknowledgments

This research was supported by a grant of the Medical Data–Driven Hospital Support Project through the Korea Health Information Service and funded by the Ministry of Health and Welfare, Republic of Korea.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Standards for evaluating medical questions and answers.

[DOCX File, 24 KB - mededu_v10i1e51282_app1.docx]

References

1. Alberts IL, Mercolli L, Pyka T, et al. Large language models (LLM) and Chatgpt: what will the impact on nuclear medicine be? Eur J Nucl Med Mol Imaging 2023 May;50(6):1549-1552. [doi: 10.1007/s00259-023-06172-w] [Medline: 36892666]
2. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol 2022 Jul;106(7):889-892. [doi: 10.1136/bjophthalmol-2022-321141] [Medline: 35523534]

XSL•FO
RenderX

3. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Mach 2020 Dec;30(4):681-694. [doi: 10.1007/s11023-020-09548-1]
4. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
5. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
6. Lu Y, Wu H, Qi S, Cheng K. Artificial intelligence in intensive care medicine: toward a ChatGPT/GPT-4 way? Ann Biomed Eng 2023 Sep;51(9):1898-1903. [doi: 10.1007/s10439-023-03234-w] [Medline: 37179277]
7. Biswas SS. Role of ChatGPT in public health. Ann Biomed Eng 2023 May;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Cheng K, Sun Z, He Y, Gu S, Wu H. The potential impact of ChatGPT/GPT-4 on surgery: will it topple the profession of surgeons? Int J Surg 2023 May 1;109(5):1545-1547. [doi: 10.1097/JS9.0000000000000388] [Medline: 37037587]
10. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. OpenAI. URL: https://openai.com/gpt-4 [accessed 2023-03-19]
11. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024. [doi: 10.48550/arXiv.2303.08774]
12. Goktas P, Karakaya G, Kalyoncu AF, Damadoglu E. Artificial intelligence chatbots in allergy and immunology practice: where have we been and where are we going? J Allergy Clin Immunol Pract 2023 Sep;11(9):2697-2700. [doi: 10.1016/j.jaip.2023.05.042] [Medline: 37301435]
13. Mensah GA, Roth GA, Fuster V. The global burden of cardiovascular diseases and risk factors: 2020 and beyond. J Am Coll Cardiol 2019 Nov 19;74(20):2529-2532. [doi: 10.1016/j.jacc.2019.10.009] [Medline: 31727292]
14. Frangogiannis NG. The significance of COVID-19-associated myocardial injury: how overinterpretation of scientific findings can fuel media sensationalism and spread misinformation. Eur Heart J 2020 Oct 14;41(39):3836-3838. [doi: 10.1093/eurheartj/ehaa727] [Medline: 33006608]
15. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
16. Duffourc M, Gerke S. Generative AI in health care and liability risks for physicians and safety concerns for patients. JAMA 2023 Jul 25;330(4):313-314. [doi: 10.1001/jama.2023.9630] [Medline: 37410497]
17. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
18. Reddy S. Evaluating large language models for use in healthcare: a framework for translational value assessment. Inform Med Unlocked 2023 Jul;41:101304. [doi: 10.1016/j.imu.2023.101304]
19. Jo W, Kim Y, Seo M, Lee N, Park J. Online information analysis on pancreatic cancer in Korea using structural topic model. Sci Rep 2022 Jun 23;12(1):10622. [doi: 10.1038/s41598-022-14506-1] [Medline: 35739151]
20. Jo W, Lee J, Park J, Kim Y. Online information exchange and anxiety spread in the early stage of the novel coronavirus (COVID-19) outbreak in South Korea: structural topic model and network analysis. J Med Internet Res 2020 Jun 2;22(6):e19455. [doi: 10.2196/19455] [Medline: 32463367]
21. Kim Y, Kim JH, Lee JM, et al. A pre-trained BERT for Korean medical natural language processing. Sci Rep 2022 Aug 16;12(1):13847. [doi: 10.1038/s41598-022-17806-8] [Medline: 35974113]
22. Das A, Verma RM. Can machines tell stories? A comparative study of deep neural language models and metrics. IEEE Access 2020 Sep;8:181258-181292. [doi: 10.1109/ACCESS.2020.3023421]
23. Miao J, Zhang Y, Jiang N, et al. Towards unifying pre-trained language models for semantic text exchange. Wireless Netw 2023 Jul. [doi: 10.1007/s11276-023-03439-w]
24. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
25. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023 Sep;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]

Abbreviations

LLM: large language model
TTR: type-token ratio

XSL•FO
RenderX

Edited by TDA Cardoso; submitted 26.07.23; peer-reviewed by A Mihalache, M Chatzimina; revised version received 10.04.24; accepted 19.04.24; published 08.07.24.

Please cite as:
Jo E, Song S, Kim JH, Lim S, Kim JH, Cha JJ, Kim YM, Joo HJ
Assessing GPT-4's Performance in Delivering Medical Advice: Comparative Analysis With Human Experts
JMIR Med Educ 2024;10:e51282
URL: https://mededu.jmir.org/2024/1/e51282
doi:10.2196/51282

© Eunbeen Jo, Sanghoun Song, Jong-Ho Kim, Subin Lim, Ju Hyeon Kim, Jung-Joon Cha, Young-Min Kim, Hyung Joon Joo. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.7.2024. 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.

XSL•FO
RenderX

The Ability of ChatGPT in Paraphrasing Texts and Reducing Plagiarism: A Descriptive Analysis

Soheil Hassanipour[1], PhD; Sandeep Nayak[2], MD; Ali Bozorgi[3], MD; Mohammad-Hossein Keivanlou[1], MD; Tirth Dave[4], MD; Abdulhadi Alotaibi[5], MD; Farahnaz Joukar[1], PhD; Parinaz Mellatdoust[6], MsC; Arash Bakhshi[1], MD; Dona Kuriyakose[7], BCA; Lakshmi D Polisetty[2], MD; Mallika Chimpiri[8], BS; Ehsan Amini-Salehi[1], MD

1
2
3
4
5
6
7
8

Corresponding Author:
Ehsan Amini-Salehi, MD

Abstract

Background: The introduction of ChatGPT by OpenAI has garnered significant attention. Among its capabilities, paraphrasing stands out.

Objective: This study aims to investigate the satisfactory levels of plagiarism in the paraphrased text produced by this chatbot.

Methods: Three texts of varying lengths were presented to ChatGPT. ChatGPT was then instructed to paraphrase the provided texts using five different prompts. In the subsequent stage of the study, the texts were divided into separate paragraphs, and ChatGPT was requested to paraphrase each paragraph individually. Lastly, in the third stage, ChatGPT was asked to paraphrase the texts it had previously generated.

Results: The average plagiarism rate in the texts generated by ChatGPT was 45% (SD 10%). ChatGPT exhibited a substantial reduction in plagiarism for the provided texts (mean difference −0.51, 95% CI −0.54 to −0.48; *P*<.001). Furthermore, when comparing the second attempt with the initial attempt, a significant decrease in the plagiarism rate was observed (mean difference −0.06, 95% CI −0.08 to −0.03; *P*<.001). The number of paragraphs in the texts demonstrated a noteworthy association with the percentage of plagiarism, with texts consisting of a single paragraph exhibiting the lowest plagiarism rate (*P*<.001).

Conclusion: Although ChatGPT demonstrates a notable reduction of plagiarism within texts, the existing levels of plagiarism remain relatively high. This underscores a crucial caution for researchers when incorporating this chatbot into their work.

(JMIR Med Educ 2024;10:e53308) doi:10.2196/53308

KEYWORDS

ChatGPT; paraphrasing; text generation; prompts; academic journals; plagiarize; plagiarism; paraphrase; wording; LLM; LLMs; language model; language models; prompt; generative; artificial intelligence; NLP; natural language processing; rephrase; plagiarizing; honesty; integrity; texts; text; textual; generation; large language model; large language models

Introduction

Plagiarism, the act of presenting someone else’s work or ideas as one’s own, stands as a prevalent and recurrent form of misconduct in the field of research and publication [1]. The diverse manifestations of plagiarism can often create confusion due to the various terminologies associated with it. Verbatim plagiarism, mosaic plagiarism, loose plagiarism, duplicate publication, augmented publication, salami-sliced publication, image plagiarism, accidental plagiarism, and self-plagiarism are among the prominent types that have been identified [2-6].

To mitigate the occurrence of such misconduct, researchers often use online plagiarism checkers, which scan existing literature on the internet and provide reports on unintentional plagiarism. Additionally, numerous journals have integrated plagiarism checkers as part of their submission process, wherein every manuscript undergoes scrutiny to identify similarity rates [7]. These measures not only act as deterrents but also aid in

XSL•FO
RenderX

upholding the standards of academic integrity and ensuring originality in scholarly publications.

In recent times, artificial intelligence (AI) has gained significant popularity across a wide range of individuals, including researchers and professionals. Among the various applications of AI, chatbots have emerged as a notable development, using AI and natural language processing techniques to generate humanlike responses to user queries [8].

One prominent example of chatbots is ChatGPT, which uses advanced models such as GPT-3.5 and GPT-4. ChatGPT has garnered substantial attention and widespread adoption, amassing over one million users across diverse fields in its first week of launch [9,10]. This surge in popularity reflects the growing recognition and use of AI-powered chatbots in various domains.

ChatGPT offers a multitude of applications and advantages. First, it excels in generating formally structured text, ensuring coherence and organization in its responses. Second, ChatGPT exhibits an extensive and eloquent vocabulary, enhancing the quality and fluency of its generated content. Additionally, it can be used as a rapid search engine, swiftly retrieving relevant information. Furthermore, it possesses the ability to search and analyze available literature, aiding researchers and professionals in their work. In the field of medical education, ChatGPT proves valuable by providing educational resources and facilitating interactive learning experiences. Moreover, it can serve as a conversational agent, engaging in meaningful and interactive conversations with users [10].

Importantly, the text produced by ChatGPT may sometimes bypass conventional plagiarism checks due to its unique generation process, which is a rising ethical concern [10]. Earlier, many researchers were reporting ChatGPT as co-authors in papers but the majority of journals promptly updated their policies to forbid this practice as ChatGPT cannot be held accountable for the generated content [11]. Moreover, in several instances, ChatGPT hallucinates and produces inaccurate and incorrect information, which can be dangerous in academic publishing [12].

Due to the increasing popularity of ChatGPT in medical research, several studies are needed to identify its pros and cons, especially in the field of medical education. In this study, we aim to assess ChatGPT's real ability to paraphrase and reduce plagiarism by imputing different texts and prompts, and assessing the plagiarism rate of the rephrased texts.

Methods

Selection of Texts

To assess the plagiarism rates and the rephrasing capabilities of ChatGPT (version 3.5), three texts were selected for the study. These texts varied in length to provide a comprehensive evaluation of the model's performance. Text one consisted of 319 words, text two comprised 613 words, and text three encompassed 1148 words. The texts used in this study were selected from one of our previously published medical papers in a medical journal [13].

Instructions Given to ChatGPT

For each selected text, five distinct prompts were given to ChatGPT to rephrase the texts. These instructions were designed to test different aspects of rephrasing and reducing plagiarism. The prompts are shown in Table 1.

Table . Prompts provided to ChatGPT.

Number	Prompts
Prompt 1	"Paraphrase the text"
Prompt 2	"Rephrase the text"
Prompt 3	"Reduce the plagiarism of the text"
Prompt 4	"Rephrase it in a way that conveys the same meaning using different words and sentence structure"
Prompt 5	"Reword this text using different language"

Subdivision of Texts

To further evaluate the effectiveness of ChatGPT in rephrasing and reducing plagiarism, the original texts were subdivided into multiple paragraphs. Specifically, texts one, two, and three were provided to ChatGPT in 1 and 3 paragraphs; 1, 3, and 5 paragraphs; and 1, 3, 5, and 7 paragraphs, respectively. All the texts with different paragraph numbers were subjected to the same five rephrasing orders. This approach allowed for a comparison of the paraphrased texts with different paragraph sections within the same content.

Second Try of Paraphrasing

To assess the influence of multiple rephrasing iterations, the texts generated by ChatGPT were once again incorporated into the system in the same sequence as before. Subsequently, the plagiarism rates of the texts were analyzed using the iThenticate platform, a tool commonly used for such evaluations in academic settings [14]. This process enabled the measurement and comparison of potential similarities between the original texts and their rephrased counterparts, shedding light on the extent of originality achieved through the rephrasing iterations.

Data Analysis

The data analysis for this study was conducted using SPSS version 19 (IBM Corp). The data distribution was assessed using the Shapiro-Wilk test. To compare the plagiarism rates of the texts, paired *t* test analysis was used. This statistical test allowed us to examine whether there were significant differences in plagiarism rates between the original texts and the paraphrased

XSL•FO
RenderX

texts generated by ChatGPT. Additionally, to assess the potential impact of different prompts on plagiarism rates, 1-way ANOVA was used. This analysis aimed to determine if there were statistically significant differences in plagiarism rates across the various prompts given to ChatGPT. A *P* value <.05 was adopted to determine statistical significance. The acceptable level of plagiarism was set at 25%, a standard embraced by scientific journals. Any plagiarism rate surpassing this threshold was considered unsatisfactory [14-18].

Ethical Considerations

This study does not require ethical approval as it does not involve human participants, patient data, or any form of personal data collection.

Results

Overview

A total of 90 texts were provided by ChatGPT. General information on plagiarism rates is provided in Table 2. The mean plagiarism rate of texts was 0.45 (SD 0.10). The mean plagiarism rates for the first try and second try were 0.48 (SD 0.09) and 0.42 (SD 0.09), respectively.

Table . Mean plagiarism rates of the texts provided by ChatGPT.

Variable		Text, n	Plagiarism rates checked by iThenticate, mean (SD)
Total		90	0.45 (0.10)
ChatGPT tries			
	First try	45	0.48 (0.09)
	Second try	45	0.42 (0.09)
Texts on the first try			
	Text 1	10	0.48 (0.16)
	Text 2	15	0.47 (0.05)
	Text 3	20	0.49 (0.07)
Texts on the second try			
	Text 1	10	0.46 (0.13)
	Text 2	15	0.40 (0.05)
	Text 3	20	0.42 (0.10)
Paragraphs			
	One paragraph	30	0.40 (0.12)
	Three paragraphs	30	0.50 (0.07)
	Five paragraphs	20	0.44 (0.05)
	Seven paragraphs	10	0.48 (0.04)
Orders given to ChatGPT			
	Please paraphrase the text	18	0.45 (0.10)
	Please rephrase the text	18	0.48 (0.06)
	Please reduce the plagiarism of the text	18	0.44 (0.10)
	Please rephrase it in a way that conveys the same meaning using different words and sentence structure	18	0.41 (0.12)
	Please reword this text using different language	18	0.48 (0.08)

The Potency of ChatGPT in Reducing Plagiarism

Based on the results of our study, ChatGPT demonstrated an ability to significantly reduce plagiarism in texts right from the first attempt (mean difference −0.51, 95% CI −0.54 to −0.48; *P*<.001). Moreover, our research revealed that even further improvements were achieved with the second attempt, as it yielded a significantly lower plagiarism rate compared to the initial try (mean difference −0.06, 95% CI −0.08 to −0.03; *P*<.001).

XSL•FO
RenderX

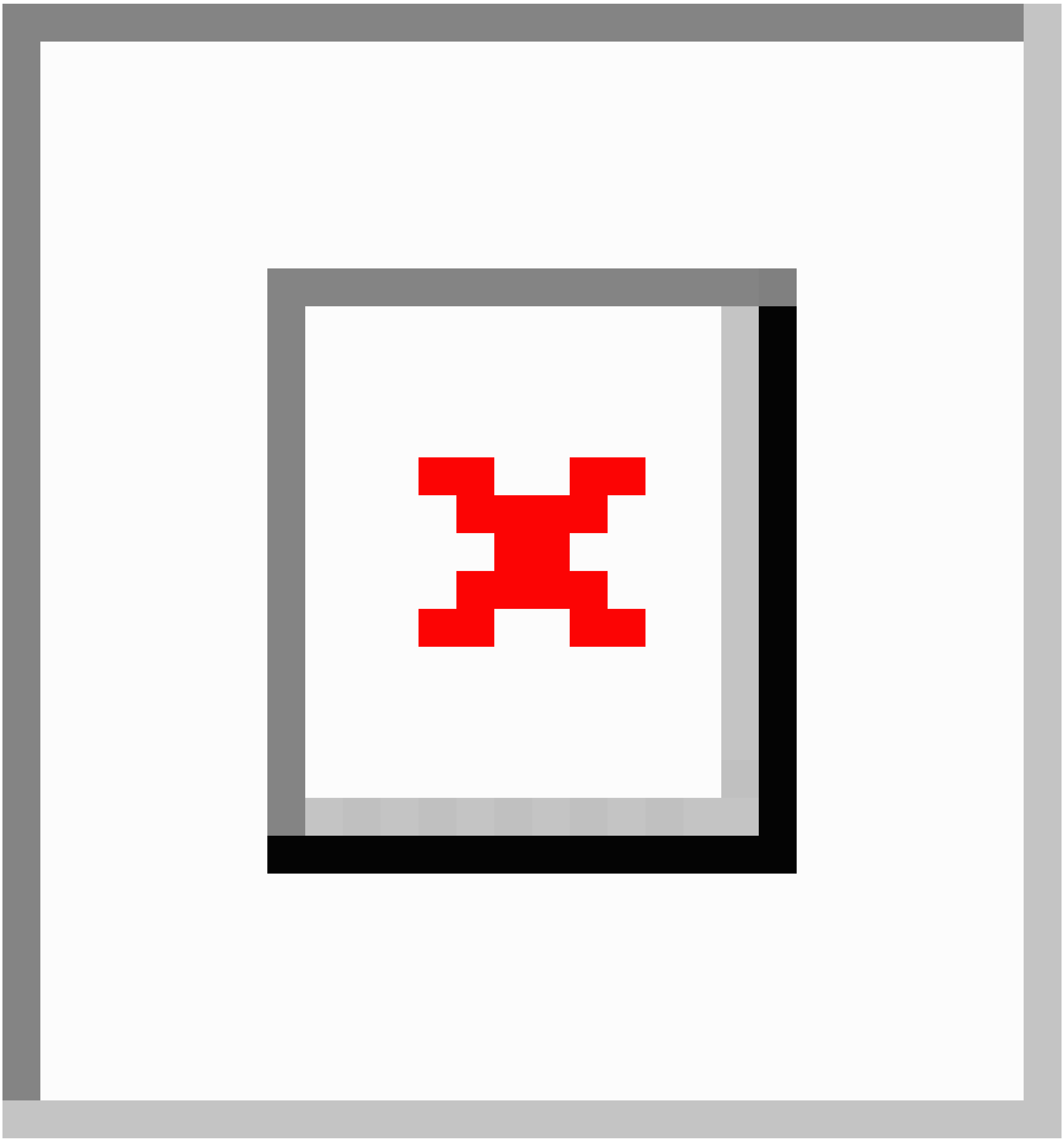
The results also showed a relation between the number of paragraphs within a text and the plagiarism rate. Our findings indicated that texts comprising a single paragraph exhibited the lowest plagiarism rates, and this relationship was statistically significant (*P*<.001). However, when analyzing the five different prompts of the texts, we found no significant difference in terms of their plagiarism rates (*P*=.19).

Furthermore, our study did not identify any statistically significant distinctions among the plagiarism rates of text one, text two, and text three (*P*=.56), suggesting that ChatGPT's effectiveness remained consistent across these particular texts.

Correlation Between Text Lengths and Plagiarism Rates

We assessed the correlation between the word count of the texts provided by ChatGPT and their plagiarism rates. Although longer texts appeared to have higher plagiarism rates, the correlation was not significant (*r*=0.2; *P*=.06; Figure 1).

Figure 1. The correlation between the word count of the texts and their corresponding plagiarism.



XSL•FO
RenderX

Discussion

Principal Findings

The findings of our study shed light on the levels of plagiarism in the paraphrased text generated by ChatGPT, an advanced chatbot developed by OpenAI. The results indicate that while ChatGPT has the capability to paraphrase the text, there are notable concerns regarding the satisfactory levels of plagiarism in the generated output.

The average plagiarism rate observed in the texts generated by ChatGPT was found to be 45%. This suggests that nearly half of the content produced by the chatbot is similar to the original source material, raising concerns about the authenticity and originality of the paraphrased text. These findings highlight the need for caution when relying on ChatGPT for generating plagiarism-free content.

Interestingly, our study revealed that ChatGPT exhibited a substantial reduction in text plagiarism when provided with explicit instructions to paraphrase or reduce plagiarism. This indicates that the chatbot is responsive to such prompts and can generate content with reduced plagiarism when specifically instructed to do so. However, it is important to note that even with explicit instructions, the plagiarism rate remained relatively high, emphasizing the limitations of the current system.

We also observed a significant decrease in the plagiarism rate between the initial and second attempts of paraphrasing. This suggests that ChatGPT has the ability to learn and improve its paraphrasing capabilities over multiple iterations. However, the reduction in plagiarism was modest, indicating that further refinements are necessary to achieve satisfactory levels of originality in the generated output.

An interesting finding from our study was the association between the number of paragraphs in the texts and the percentage of plagiarism. Texts consisting of a single paragraph demonstrated the lowest plagiarism rate. This suggests that presenting the source texts within a single coherent unit allows ChatGPT to better understand and paraphrase the content effectively. Dividing the text into separate paragraphs may lead to fragmented understanding and potentially contribute to higher levels of plagiarism.

It is worth noting that the prompts used in our study did not yield significant differences in the levels of plagiarism. This indicates that the specific prompt provided to ChatGPT does not significantly influence its paraphrasing capability. In addition, this outcome might be the consequence of the bot's strong ability to understand our true intentions when issuing commands, or it might be because our command words were brief or similar to one another. However, further investigation into the effect of different prompts and their impact on plagiarism is warranted to explore this aspect in more detail.

ChatGPT has a wide range of applications that can be effectively used. Numerous articles have discussed the use of ChatGPT in composing scientific literature, with a particular study illustrating its capability to generate formal research articles. The researchers observed that the language used is articulate, adopts a conventional tone, and offers a pleasant reading experience [19].

ChatGPT has the potential to serve as a search engine that directly responds to queries, eliminating the need to navigate to external sites for information. This streamlines the process of writing research papers, reducing the time spent by authors on the often arduous task of searching for articles and applying various selection criteria. This, in turn, allows authors to dedicate more time to their actual research work and methodology [20].

Moreover, articles created by ChatGPT seem to elude traditional plagiarism detection methods. In a research study, the chatbot was tasked with generating 50 medical research abstracts using a subset of articles. The resulting articles underwent examination by plagiarism detection software, an AI-output detector, and a panel of medical researchers who were tasked with identifying any artificially generated abstracts. The findings revealed that abstracts generated by ChatGPT seamlessly passed through the plagiarism detection software, registering a median originality score of 100%, indicating the absence of detected plagiarism. In contrast, the AI-output checker only identified 66% of the generated abstracts [21].

While ChatGPT and other AI tools hold promise in various applications, their deployment in medical writing raises ethical and legal considerations. These concerns encompass potential violations of copyright laws, medico-legal complexities, and the risk of inaccuracies or biases in the generated content. It is crucial to recognize and confront the limitations and challenges linked to the use of AI in medical writing [20,22,23].

Limitations and Future Suggestions

The sample size used in our study was relatively small, and as a result, we recommend that future investigations incorporate larger sample sizes to enhance the robustness of the findings. It is worth noting that our study was conducted using ChatGPT version 3.5, which was a publicly available version at the time of our research. Unfortunately, we did not have access to ChatGPT version 4, preventing us from evaluating the efficacy of this updated version in terms of paraphrasing capabilities.

It is essential to acknowledge that our study exclusively focused on providing medical content to ChatGPT. We encourage other researchers to explore the impact of using different content types on the efficacy of ChatGPT. This would allow for a comprehensive understanding of whether the effectiveness of ChatGPT is influenced by the specific domain or topic of the content it receives. Conducting such investigations will provide valuable insights into the generalizability and adaptability of ChatGPT across various subject matters.

Moreover, a recognized limitation of ChatGPT is its tendency to produce inconsistent results with the same prompts [24]. To relatively address this challenge, we used a comprehensive approach. Each prompt was provided with nine texts, varying paragraph structures (text one with 1 paragraph, text one with 3 paragraphs, text two with 1 paragraph, text two with 3 paragraphs, text two with 5 paragraphs, text three with 1 paragraph, text three with 3 paragraphs, text three with 5 paragraphs, and text three with 7 paragraphs). Furthermore, we

XSL•FO
RenderX

requested ChatGPT to paraphrase each of these texts twice using the same prompt. We then calculated the mean plagiarism rates for both the first and second attempts, along with the overall mean plagiarism rate for each prompt (Table 2).

Nevertheless, we recommend that future studies take this limitation into account and explore additional measures to enhance the robustness of assessments. Specifically, researchers may consider providing ChatGPT with a greater number of texts exhibiting different paragraph structures and incorporating a higher frequency of repetitions in the paraphrasing process.

We used similar prompts and provided them to ChatGPT. We recommend that future studies adopt a broader range of prompts to assess ChatGPT's performance across different input variations. This approach allows for a more comprehensive evaluation and facilitates the identification of optimal prompts to minimize plagiarism rates.

An important consideration with ChatGPT lies in the potential for hallucination and biases, particularly in the generation of medical content [25]. In our study, two independent researchers evaluated the content provided by ChatGPT, comparing it with the original texts. However, we acknowledge that the texts used in our assessment may not have been sufficiently complex. To address this limitation, we recommend that future studies incorporate both simple and more intricate texts to thoroughly evaluate the biases that ChatGPT may introduce during the paraphrasing of medical content. This approach will provide a more nuanced understanding of the model's performance.

Conclusion

While ChatGPT has been shown to significantly reduce plagiarism in texts, it is important to note that the resulting plagiarism rates of the provided texts may still be considered high, which may not meet the acceptance criteria of most scientific journals. Therefore, medical writers and professionals should carefully consider this issue when using ChatGPT for paraphrasing their texts. There are a couple of strategies authors can use to improve the paraphrasing efficacy of ChatGPT. Presenting the texts in a single-paragraph format and repeating the requesting procedure with ChatGPT. By considering these strategies and being mindful of the potential limitations, authors can strive to improve the paraphrasing efficacy of ChatGPT and address the challenge of high plagiarism rates associated with its outputs.

Acknowledgments

We acknowledge BioRender since all the illustrations are created with BioRender.com

Data Availability

The data sets used or analyzed during this study are accessible from the corresponding author upon reasonable request.

Authors' Contributions

EA-S, A Bozorgi, and SH conceptualized the study. EA-S, LP, and MC curated the data. SH and DK conducted the formal analysis and participated in statistics. EA-S and A Bozorgi contributed to the methodology. M-HK, A Bakhsi, and TD wrote the original draft. All authors edited the manuscript.

Conflicts of Interest

None declared.

References

1. Zimba O, Gasparyan A. Plagiarism detection and prevention: a primer for researchers. Reumatologia 2021;59(3):132-137. [doi: 10.5114/reum.2021.105974] [Medline: 34538939]
2. Helgesson G, Eriksson S. Plagiarism in research. Med Health Care Philos 2015 Feb;18(1):91-101. [doi: 10.1007/s11019-014-9583-8] [Medline: 24993050]
3. Resnik DB, Dinse GE. Scientific retractions and corrections related to misconduct findings. J Med Ethics 2013 Jan;39(1):46-50. [doi: 10.1136/medethics-2012-100766] [Medline: 22942373]
4. Chaddah P. Not all plagiarism requires a retraction. Nature 2014 Jul 10;511(7508):127. [doi: 10.1038/511127a] [Medline: 25008489]
5. Heitman E, Litewka S. International perspectives on plagiarism and considerations for teaching international trainees. Urol Oncol 2011;29(1):104-108. [doi: 10.1016/j.urolonc.2010.09.014] [Medline: 21194646]
6. Agrawal R. Plagiarism. Indian J Pathol Microbiol 2020;63(2):175-176. [doi: 10.4103/0377-4929.282724] [Medline: 32317511]
7. Masic I, Begic E, Dobraca A. Plagiarism detection by online solutions. Stud Health Technol Inform 2017;238:227-230. [Medline: 28679930]
8. Gupta A, Hathwar D, Vijayakumar A. Introduction to AI chatbots. Int J Eng Res Technol 2020 Jul;9(7):255-258. [doi: 10.17577/IJERTV9IS070143]

XSL•FO
RenderX

9. Kirmani AR. Artificial intelligence-enabled science poetry. ACS Energy Lett 2022 Dec 19;8(1):574-576. [doi: 10.1021/acsenergylett.2c02758]
10. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
11. Biswas S. ChatGPT and the future of medical writing. Radiology 2023 Apr;307(2):e223312. [doi: 10.1148/radiol.223312] [Medline: 36728748]
12. Athaluri SA, Manthena SV, Kesapragada V, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023 Apr 11;15(4). [doi: 10.7759/cureus.37432] [Medline: 37182055]
13. Amini-Salehi E, Hassanipour S, Joukar F, et al. Risk factors of non-alcoholic fatty liver disease in the Iranian adult population: a systematic review and meta-analysis. Hepatitis Monthly 2023 Mar 12;23(1):e131523. [doi: 10.5812/hepatmon-131523]
14. Habibzadeh F. The acceptable text similarity level in manuscripts submitted to scientific journals. J Korean Med Sci 2023 Aug 7;38(31):37550808. [doi: 10.3346/jkms.2023.38.e240] [Medline: 37550808]
15. Memon AR. Similarity and plagiarism in scholarly journal submissions: bringing clarity to the concept for authors. J Korean Med Sci 2020 Jul 13;35(27):32657084. [doi: 10.3346/jkms.2020.35.e217] [Medline: 32657084]
16. Memon AR, Mavrinac M. Knowledge, attitudes, and practices of plagiarism as reported by participants completing the AuthorAID MOOC on research writing. Sci Eng Ethics 2020 Apr;26(2):1067-1088. [doi: 10.1007/s11948-020-00198-1] [Medline: 32067186]
17. Mahian O, Treutwein M, Estellé P, et al. Measurement of similarity in academic contexts. Publications 2017 Jun 22;5(3):18. [doi: 10.3390/publications5030018]
18. Peh WC, Arokiasamy J. Plagiarism: a joint statement from the Singapore Medical Journal and the Medical Journal of Malaysia. Med J Malaysia 2008 Dec;63(5):354-355. [Medline: 19803289]
19. Gordijn B, Have HT. ChatGPT: evolution or revolution. Med Health Care Philosophy 2023 Jan 19;26(1):1-2. [doi: 10.1007/s11019-023-10136-0]
20. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6(1169595):37215063. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
21. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023 Jan;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]
22. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J 2023 Feb 28;64(1):1-3. [doi: 10.3325/cmj.2023.64.1] [Medline: 36864812]
23. Ho WLJ, Koussayer B, Sujka J. ChatGPT: friend or foe in medical writing? An example of how ChatGPT can be utilized in writing case reports. Surg Pract Sci 2023 Sep;14:100185. [doi: 10.1016/j.sipas.2023.100185]
24. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
25. Yang R, Marrese-Taylor E, Ke Y, Cheng L, Chen Q, Li I. Integrating UMLS knowledge into large language models for medical question answering. arXiv. Preprint posted online on Oct 4, 2023. [doi: 10.48550/arXiv.2310.02778]

Abbreviations

AI: artificial intelligence

Edited by G Eysenbach, TDA Cardoso; submitted 03.10.23; peer-reviewed by L Zhu, R Yang; revised version received 03.01.24; accepted 01.05.24; published 08.07.24.

Please cite as:
Hassanipour S, Nayak S, Bozorgi A, Keivanlou MH, Dave T, Alotaibi A, Joukar F, Mellatdoust P, Bakhshi A, Kuriyakose D, Polisetty LD, Chimpiri M, Amini-Salehi E
The Ability of ChatGPT in Paraphrasing Texts and Reducing Plagiarism: A Descriptive Analysis
JMIR Med Educ 2024;10:e53308
URL: https://mededu.jmir.org/2024/1/e53308
doi:10.2196/53308

© Soheil Hassanipour, Sandeep Nayak, Ali Bozorgi, Mohammad-Hossein Keivanlou, Tirth Dave, Abdulhadi Alotaibi, Farahnaz Joukar, Parinaz Mellatdoust, Arash Bakhshi, Dona kuriyakose, Lakshmi Polisetty, Mallika Chimpiri, Ehsan Amini-Salehi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.7.2024. This is an open-access article distributed

XSL•FO
RenderX

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.

XSL•FO
RenderX

Evaluation of ChatGPT-Generated Differential Diagnosis for Common Diseases With Atypical Presentation: Descriptive Research

Kiyoshi Shikino[1,2], MHPE, MD, PhD; Taro Shimizu[3], MSc, MPH, MBA, MD, PhD; Yuki Otsuka[4], MD, PhD; Masaki Tago[5], MD, PhD; Hiromizu Takahashi[6], MD, PhD; Takashi Watari[7], MHQS, MD, PhD; Yosuke Sasaki[8], MD, PhD; Gemmei Iizuka[9,10], MD, PhD; Hiroki Tamura[1], MD, PhD; Koichi Nakashima[11], MD; Kotaro Kunitomo[12], MD; Morika Suzuki[12,13], MD, PhD; Sayaka Aoyama[14], MD; Shintaro Kosaka[15], MD; Teiko Kawahigashi[16], MD, PhD; Tomohiro Matsumoto[17], MD, DDS, PhD; Fumina Orihara[17], MD; Toru Morikawa[18], MD, PhD; Toshinori Nishizawa[19], MD; Yoji Hoshina[13], MD; Yu Yamamoto[20], MD; Yuichiro Matsuo[21], MPH, MD; Yuto Unoki[22], MD; Hirofumi Kimura[22], MD; Midori Tokushima[23], MD; Satoshi Watanuki[24], MBA, MD; Takuma Saito[24], MD; Fumio Otsuka[4], MD, PhD; Yasuharu Tokuda[25,26], MPH, MD, PhD

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

Corresponding Author:
Kiyoshi Shikino, MHPE, MD, PhD

Abstract

Background: The persistence of diagnostic errors, despite advances in medical knowledge and diagnostics, highlights the importance of understanding atypical disease presentations and their contribution to mortality and morbidity. Artificial intelligence (AI), particularly generative pre-trained transformers like GPT-4, holds promise for improving diagnostic accuracy, but requires further exploration in handling atypical presentations.

XSL•FO
RenderX

Objective: This study aimed to assess the diagnostic accuracy of ChatGPT in generating differential diagnoses for atypical presentations of common diseases, with a focus on the model's reliance on patient history during the diagnostic process.

Methods: We used 25 clinical vignettes from the *Journal of Generalist Medicine* characterizing atypical manifestations of common diseases. Two general medicine physicians categorized the cases based on atypicality. ChatGPT was then used to generate differential diagnoses based on the clinical information provided. The concordance between AI-generated and final diagnoses was measured, with a focus on the top-ranked disease (top 1) and the top 5 differential diagnoses (top 5).

Results: ChatGPT's diagnostic accuracy decreased with an increase in atypical presentation. For category 1 (C1) cases, the concordance rates were 17% (n=1) for the top 1 and 67% (n=4) for the top 5. Categories 3 (C3) and 4 (C4) showed a 0% concordance for top 1 and markedly lower rates for the top 5, indicating difficulties in handling highly atypical cases. The χ^2 test revealed no significant difference in the top 1 differential diagnosis accuracy between less atypical (C1+C2) and more atypical (C3+C4) groups (χ^2_1=2.07; n=25; *P*=.13). However, a significant difference was found in the top 5 analyses, with less atypical cases showing higher accuracy (χ^2_1=4.01; n=25; *P*=.048).

Conclusions: ChatGPT-4 demonstrates potential as an auxiliary tool for diagnosing typical and mildly atypical presentations of common diseases. However, its performance declines with greater atypicality. The study findings underscore the need for AI systems to encompass a broader range of linguistic capabilities, cultural understanding, and diverse clinical scenarios to improve diagnostic utility in real-world settings.

(JMIR Med Educ 2024;10:e58758) doi:10.2196/58758

KEYWORDS

atypical presentation; ChatGPT; common disease; diagnostic accuracy; diagnosis; patient safety

Introduction

For the past decade, medical knowledge and diagnostic techniques have expanded worldwide, becoming more accessible with remarkable advancements in clinical testing and useful reference systems [1]. Despite these advancements, misdiagnosis significantly contributes to mortality, making it a noteworthy public health issue [2,3]. Studies have revealed discrepancies between clinical and postmortem autopsy diagnoses in at least 25% of cases, with diagnostic errors contributing to approximately 10% of deaths and to 6% - 17% of hospital adverse events [4-8]. The significance of atypical presentations as a contributor to diagnostic errors is especially notable, with recent findings suggesting that such presentations are prevalent in a substantial portion of outpatient consultations and are associated with a higher risk of diagnostic inaccuracies [9]. This underscores the persistent challenge in diagnosing patients correctly due to the variability in disease presentation and due to the reliance on medical history, which is the basis for approximately 80% of the medical diagnosis [10,11].

The advent of artificial intelligence (AI) in health care, particularly through natural language processing (NLP) models such as generative pre-trained transformers (GPTs), has opened new avenues in medical diagnosis [12]. Recent studies on AI medical diagnosis across various specialties—including neurology [13], dermatology [14], radiology [15], and pediatrics [16]—have shown promising results and improved diagnostic accuracy, efficiency, and safety. Among these developments, GPT-4, a state-of-the-art AI model developed by OpenAI, has demonstrated remarkable capabilities in understanding and processing medical language, significantly outperforming its predecessors in medical knowledge assessments and potentially transforming medical education and clinical decision support systems [12,17].

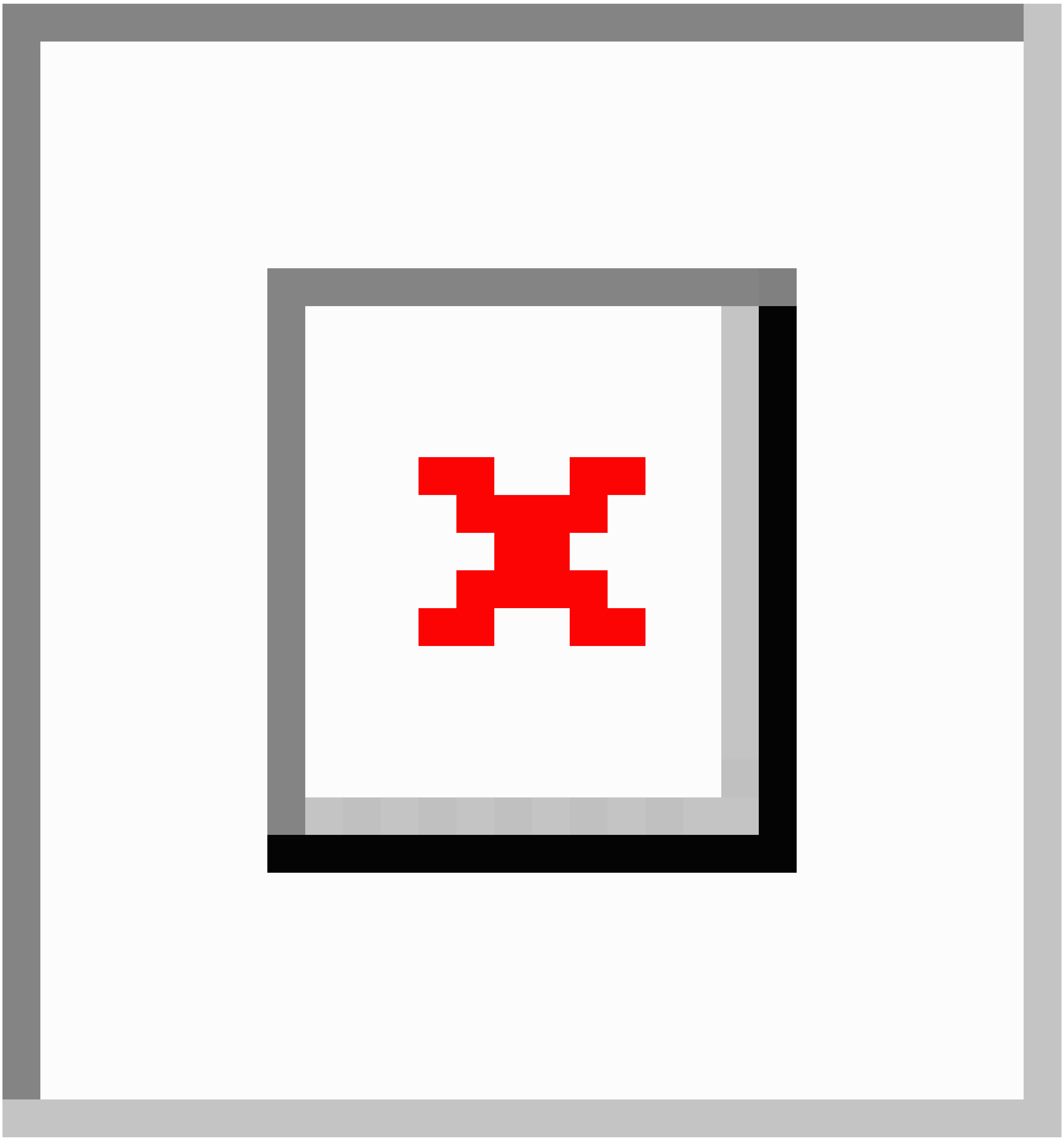
Notably, one study found that ChatGPT (OpenAI) could pass the United States Medical Licensing Examination (USMLE), highlighting its potential in medical education and medical diagnosis [18,19]. Moreover, in controlled settings, ChatGPT has shown over 90% accuracy in diagnosing common diseases with typical presentations based on chief concerns and patient history [20]. However, while research has examined the diagnostic accuracy of AI chatbots, including ChatGPT, in generating differential diagnoses for complex clinical vignettes derived from general internal medicine (GIM) department case reports, their diagnostic accuracy in handling atypical presentations of common diseases remains less explored [21,22]. There has been a notable study aimed at evaluating the accuracy of the differential diagnosis lists generated by both third- and fourth-generation ChatGPT models using case vignettes from case reports published by the Department of General Internal Medicine of Dokkyo Medical University Hospital, Japan. ChatGPT with GPT-4 was found to achieve a correct diagnosis rate in the top 10 differential diagnosis lists, top 5 lists, and top diagnoses of 83%, 81%, and 60%, respectively—rates comparable to those of physicians. Although the study highlights the potential of ChatGPT as a supplementary tool for physicians, particularly in the context of GIM, it also underlines the importance of further investigation into the diagnostic accuracy of ChatGPT with atypical disease presentations (Figure 1). Given the crucial role of patient history in diagnosis and the inherent variability in disease presentation, our study expands upon this foundation to assess the accuracy of ChatGPT in diagnosing common diseases with atypical presentations [23].

More specifically, this study aims to evaluate the hypothesis that the diagnostic accuracy of AI, exemplified by ChatGPT, declines when dealing with atypical presentations of common diseases. We hypothesize that despite the known capabilities of AI in recognizing typical disease patterns, its performance will be significantly challenged when presented with clinical

XSL•FO
RenderX

cases that deviate from these patterns, leading to reduced diagnostic precision. Consequently, this study seeks to systematically assess this hypothesis and explore its implications for the integration of AI in clinical practice. By exploring the contribution of AI-assisted medical diagnoses to common diseases with atypical presentations and patient history, the study assesses the accuracy of ChatGPT in reaching a clinical diagnosis based on the medical information provided. By reevaluating the significance of medical information, our study contributes to the ongoing discourse on optimizing diagnostic processes—both conventional and AI assisted.

Figure 1. Study motivation. AI: artificial intelligence; USMLE: United States Medical Licensing Examination.



Methods

Study Design, Settings, and Participants

This study used a series of 25 clinical vignettes from a special issue of the *Journal of Generalist Medicine*, a Japanese journal, published on March 5, 2024. These vignettes, which exemplify atypical presentations of common diseases, were selected for their alignment with our research aim to explore the impact of atypical disease presentations in AI-assisted diagnosis. The clinical vignettes were derived from real patient cases and curated by an editorial team specializing in GIM, with final edits by KS. Each case included comprehensive details such as age, gender, chief concern, medical history, medication history, current illness, and physical examination findings, along with the ultimate and initial misdiagnoses.

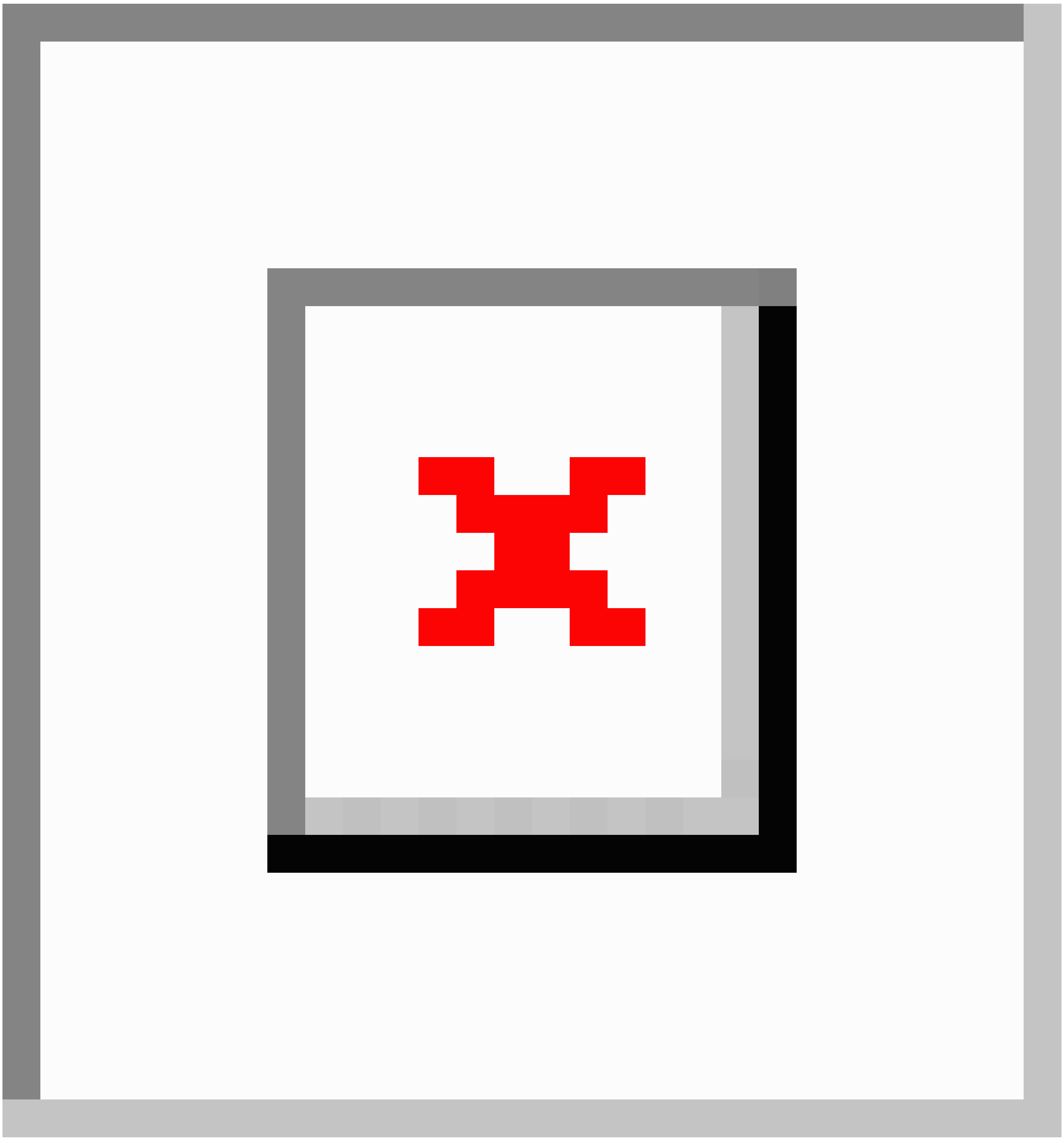
XSL•FO
RenderX

An expert panel comprising 2 general medicine and medical education physicians, T Shimizu and Y Otsuka, initially reviewed these cases. After deliberation, they selected all 25 cases that exemplified atypical presentations of common diseases. Subsequently, T Shimizu and Y Otsuka evaluated their degree of atypicality and categorized them into 4 distinct levels, using the following definition as a guide: “Atypical presentations have a shortage of prototypical features. These can be defined as features that are most frequently encountered in patients with the disease, features encountered in advanced presentations of the disease, or simply features of the disease commonly listed in medical textbooks. Atypical presentations may also have features with unexpected values” [24]. Category 1 was assigned to cases that were closest to the typical presentations of common diseases, whereas category 4 was designated for those that were markedly atypical. In instances where T Shimizu and Y Otsuka did not reach consensus, a third expert, KS, was consulted. Through collaborative discussions, the panel reached a consensus on the final category for each case, ensuring a systematic and comprehensive evaluation of the atypical presentations of common diseases (Figure 2).

Our analysis was conducted on March 12, 2024, using ChatGPT’s proficiency in Japanese. The language processing was enabled by the standard capabilities of the ChatGPT model, requiring no additional adaptation or programming by our team. We exclusively used text-based input for the generative AI, excluding tables or images to maintain a focus on linguistic data. This approach is consistent with the typical constraints of language-based AI diagnostic tools. Inputs to ChatGPT consisted of direct transcriptions of the original case reports in Japanese, ensuring the authenticity of the medical information was preserved. We measured the concordance between AI-generated differential diagnoses and the vignettes’ final diagnoses, as well as the initial misdiagnoses. Our investigation entailed inputting clinical information—including medical history, physical examination, and laboratory data—into ChatGPT, followed by posing this request: “List of differential diagnoses in order of likelihood, based on the provided vignettes’ information,” labeled as “GAI [generative AI] differential diagnoses.”

XSL•FO
RenderX

Figure 2. Categories of common diseases with atypical presentations (n=25).

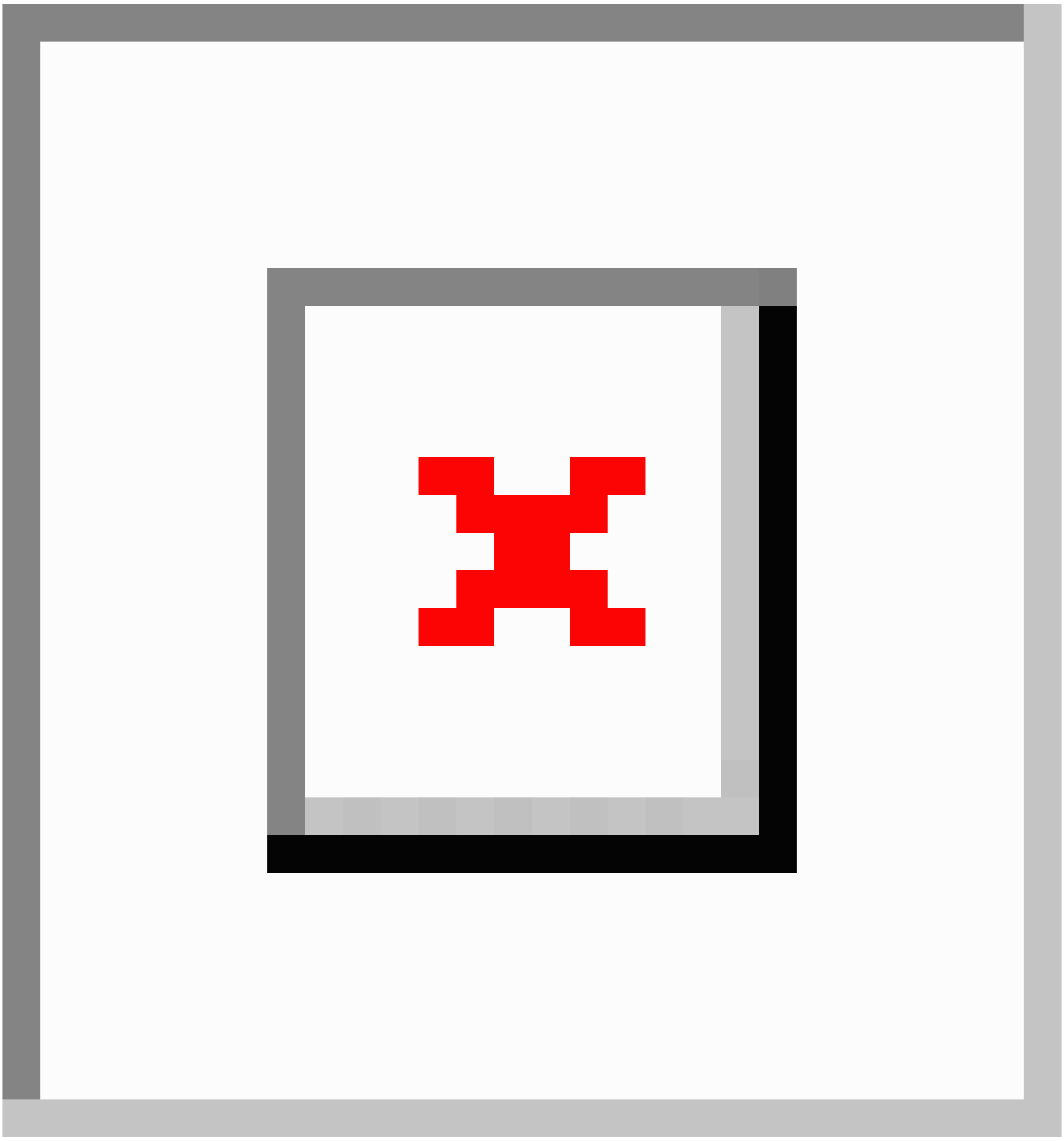


Data Collection and Measurements

We assigned the correct diagnosis for each of these 25 cases as "final diagnosis." We then used ChatGPT to generate differential diagnoses ("GAI differential diagnoses"). For each case, ChatGPT was prompted to create a list of differential diagnoses. Patient information was provided in full each time, without incremental inputs. The concordance rate between "final diagnosis," "misdiagnosis," and "GAI differential diagnoses" was then assessed. To extract a list of diagnoses from ChatGPT, we concluded each input session with the phrase "List of differential diagnoses in order of likelihood, based on the provided vignettes' information." We measured the percentage at which the final diagnosis or misdiagnosis was included in the top-ranked disease (top 1) and within the top 5 differential diagnoses (top 5) generated by ChatGPT (Figure 3).

XSL•FO
RenderX

Figure 3. Study flow. C: category.



Data Analysis

Two board-certified physicians working in the medical diagnostic department of our facility judged the concordance between the AI-proposed diagnoses and the final diagnosis. The 2 physicians are GIM board–certified. The number of years after graduation of the physicians was 7 and 17, respectively. A diagnosis was considered to match if the 2 physicians agreed to the concordance. We measured the interrater reliability with the κ coefficient (0.8 - 1.0=almost perfect; 0.6 - 0.8=substantial; 0.4 - 0.6=moderate; and 0.2 - 0.4=fair) [25]. To further analyze the accuracy of the top 1 and top 5 diagnoses, we used the χ^2 or Fisher exact test, as appropriate. Statistical analyses were conducted using SPSS Statistics (version 26.0; IBM Corp) with the level of significance set at $P<.05$.

Ethics Approval

Our research did not involve humans, medical records, patient information, observations of public behaviors, or secondary data analyses; thus, it was exempt from ethical approval, informed consent requirements, and institutional review board approval. Additionally, as no identifying information was included, the data did not need to be anonymized or deidentified. We did not offer any compensation because there were no human participants in the study.

XSL•FO
RenderX

Results

The 25 clinical vignettes comprised 11 male and 14 female patients, with ages ranging from 21 to 92 years. All individuals were older than 20 years, and 8 were older than 65 years. Table 1, Multimedia Appendix 1, and Multimedia Appendix 2 present these results. The correct final diagnosis listed in the *Journal of Generalist Medicine* clinical vignette as a common disease presenting atypical symptoms (labeled as "final diagnosis") showed that "GAI differential diagnoses" and "final diagnosis" coincided in 12% (3/12) of cases within the first list of differential diagnoses, while "GAI differential diagnoses" and "final diagnosis" had a concordance rate of 44% (11/25) in 5 differential diagnoses. The interrater reliability was substantial (Cohen κ=0.84).

The analysis of the concordance rates between the "GAI differential diagnoses" generated by ChatGPT and the "final diagnosis" from the *Journal of Generalist Medicine* revealed distinct patterns across the 4 categories of atypical presentations (Table 2). For the top 1 differential diagnosis, that is, category 1 (C1) cases, which were closest to a typical presentation, the concordance rate was 7% (n=1), whereas category 2 (C2) cases exhibited a slightly higher rate of 22% (n=2). Remarkably, categories 3 (C3) and 4 (C4), which represent more atypical cases, demonstrated no concordance (0%) in the top 1 differential diagnosis.

When the analysis was expanded to the top 5 differential diagnoses, the concordance rates varied across categories. C1 cases showed a significant increase in concordance, to 67% (n=4), indicating better performance of the "GAI differential diagnoses" when considering a broader range of possibilities. C2 cases had a concordance rate of 44% (n=4), followed by C3 cases at 25% (n=1) and C4 cases at 17% (n=1).

To assess the diagnostic accuracy of ChatGPT across varying levels of atypical presentations, we used the χ^2 test. Specifically, we compared the frequency of correct diagnoses in the top 1 and top 5 differential diagnoses provided by ChatGPT for cases categorized as C1+C2 (less atypical) versus C3+C4 (more atypical). For the top 1 differential diagnosis, there was no statistically significant difference in the number of correct diagnoses between the less atypical (C1+C2) and more atypical (C3+C4) groups (χ^2_1=2.07; n=25; *P*=.13). However, when expanding the analysis to the top 5 differential diagnoses, we found a statistically significant difference, with the less atypical group (C1+C2) demonstrating a higher number of correct diagnoses compared to the more atypical group (C3+C4) (χ^2_1=4.01; n=25; *P*=.048).

XSL•FO
RenderX

Table . List of answers and diagnoses provided by ChatGPT. Category 1 was closest to typical, and category 4 was most atypical.

Case	Age (years)	Gender	Final diagnosis[a]	Category	GAI[b] diagnosis rank[c]
1	34	F	Caffeine intoxication	1	0
2	40	F	Asthma	1	1
3	55	F	Obsessive-compulsive disorder	1	3
4	58	M	Drug-induced enteritis	1	3
5	38	F	Cytomegalovirus infection	1	3
6	29	M	Acute HIV infection	1	5
7	62	M	Cardiogenic cerebral embolism	2	1
8	70	M	Cervical epidural hematoma	2	0
9	70	F	Herpes zoster	2	0
10	86	F	Hemorrhagic gastric ulcer	2	0
11	77	M	Septic arthritis	2	3
12	78	F	Compression fracture	2	0
13	45	M	Infective endocarditis	2	0
14	21	F	Ectopic pregnancy	2	1
15	55	F	Non-ST elevation myocardial infarction	2	2
16	54	F	Hypoglycemia	3	0
17	77	F	Giant cell arteritis	3	0
18	60	M	Adrenal insufficiency	3	4
19	38	F	Generalized anxiety disorder	3	0
20	24	F	Graves disease	4	4
21	31	M	Acute myeloblastic leukemia	4	0
22	76	F	Elderly onset rheumatoid arthritis	4	0
23	45	M	Appendicitis	4	0
24	92	M	Rectal cancer	4	0
25	60	M	Acute aortic dissection	4	0

[a]Final diagnosis indicates the final correct diagnosis listed in the *Journal of Generalist Medicine* clinical vignette as common disease presenting atypical symptoms.

[b]GAI: generative artificial intelligence.

[c]GAI diagnosis rank indicates the high-priority differential diagnosis rank generated by ChatGPT.

Table . Concordance rates of artificial intelligence–generated differential diagnoses by atypicality category. Category (C) 1 was closest to typical, and C4 was most atypical.

Category	Rank 1 diagnoses, n	Rank 2 diagnoses, n	Rank 3 diagnoses, n	Rank 4 diagnoses, n	Rank 5 diagnoses, n	Misdiagnoses, n	Top 1, %	Top 5, %
C1	1	0	3	0	0	2	17	67
C2	2	1	1	0	0	5	22	44
C3	0	0	0	1	0	3	0	25
C4	0	0	0	1	0	5	0	17

XSL•FO
RenderX

Discussion

Principal Findings

This study provides insightful data on the performance of ChatGPT in diagnosing common diseases with atypical presentations. Our findings offer a nuanced view of the capacity of AI-driven differential diagnoses across varying levels of atypicality. In the analysis of the concordance rates between "GAI differential diagnoses" and "final diagnosis," we observed a decrease in diagnostic accuracy as the degree of atypical presentation increased.

The performance of ChatGPT in C1 cases, which are the closest to typical presentations, was moderately successful, with a concordance rate of 17% for the top 1 diagnosis and 67% within the top 5. This suggests that when the disease presentation closely aligns with the typical characteristics known to the model, ChatGPT is relatively reliable at identifying a differential diagnosis list that coincides with the final diagnosis. However, the utility of ChatGPT appears to decrease as atypicality increases, as evidenced by the lower concordance rates in C2, and notably more so in C3 and C4, where the concordance rates for the top 1 diagnosis fell to 0%. Similar challenges were observed in another 2024 study [26], where the diagnostic accuracy of ChatGPT varied depending on the disease etiology, particularly in differentiating between central nervous system and non–central nervous system tumors.

It is particularly revealing that in the more atypical presentations of common diseases (C3 and C4), the AI struggled to provide a correct diagnosis, even within the top 5 differential diagnoses, with concordance rates of 25% and 17%, respectively. These categories highlight the current limitations of AI in medical diagnosis when faced with cases that deviate significantly from the established patterns within its training data [27].

By leveraging the comprehensive understanding and diagnostic capabilities of ChatGPT, this study aims to reevaluate the significance of patient history in AI-assisted medical diagnosis and contribute to optimizing diagnostic processes [28]. Our exploration of ChatGPT's performance in processing atypical disease presentations not only advances our understanding of AI's potential in medical diagnosis [23] but also underscores the importance of integrating advanced AI technologies with traditional diagnostic methodologies to enhance patient care and reduce diagnostic errors.

The contrast in performance between the C1 and C4 cases can be seen as indicative of the challenges AI systems currently face with complex clinical reasoning requiring pattern recognition. Atypical presentations can include uncommon symptoms, rare complications, or unexpected demographic characteristics, which may not be well represented in the data sets used to train the AI systems [29]. Furthermore, these findings can inform the development of future versions of AI medical diagnosis systems and guide training curricula to include a broader spectrum of atypical presentations.

This study underscores the importance of the continued refinement of AI medical diagnosis systems, as highlighted by the recent advances in AI technologies and their applications in medicine. Studies published in 2024 [30-32] provide evidence of the rapidly increasing capabilities of large language models (LLMs) like GPT-4 in various medical domains, including oncology, where AI is expected to significantly impact precision medicine [30]. The convergence of text and image processing, as seen in multimodal AI models, suggests a qualitative leap in AI's ability to process complex medical information, which is particularly relevant for our findings on AI-assisted medical diagnostics [30]. These developments reinforce the potential of AI tools like ChatGPT in bridging the knowledge gap between machine learning developers and practitioners, as well as their role in simplifying complex data analyses in medical research and practice [31]. However, as these systems evolve, it is crucial to remain aware of their limitations and the need for rigorous verification processes to mitigate the risk of errors, which can have significant implications in clinical settings [32]. This aligns with our observation of decreased diagnostic accuracy in atypical presentations and the necessity for cautious integration of AI into clinical practice. It also points to the potential benefits of combining AI with human expertise to compensate for current AI limitations and enhance diagnostic accuracy [33].

Our research suggests that while AI, particularly ChatGPT, shows promise as a supplementary tool for medical diagnosis, reliance on this technology should be balanced with expert clinical judgment, especially in complex and atypical cases [28,29]. The observed concordance rate of 67% for C1 cases indicates that even when not dealing with extremely atypical presentations, cases with potential pitfalls may result in AI medical diagnosis accuracy lower than the 80%‐90% estimated by existing studies [10,11]. This revelation highlights the need for cautious integration of AI in clinical settings, acknowledging that its diagnostic capabilities, while robust, may still fall short in certain scenarios [34,35].

Limitations

Despite the strengths of our research, the study has certain limitations that must be noted when contextualizing our findings. First, the external validity of the results may be limited, as our data set comprises only 25 clinical vignettes sourced from a special issue of the *Journal of Generalist Medicine*. While these vignettes were chosen for their relevance to the study's hypothesis on atypical presentations of common diseases, the size of the data set and its origin as mock scenarios rather than real patient data may limit the generalizability of our findings. This sample size may not adequately capture the variability and complexities typically encountered in broader clinical practice and thus might not be sufficient to firmly establish statistical generalizations. This limitation is compounded by the exclusion of pediatric vignettes, which narrows the demographic range of our findings and potentially reduces their applicability across diverse age groups.

Second, ChatGPT's current linguistic capabilities predominantly cater to English, presenting significant barriers to patient-provider interactions that may occur in other languages. This raises concerns about the potential for miscommunication and subsequent misdiagnosis in non-English medical consultations. This underscores the essential need for future AI models to exhibit a multilingual capacity that can grasp the

XSL•FO
RenderX

subtleties inherent in various languages and dialects, as well as the cultural contexts within which they are used.

Finally, the diagnostic prioritization process of ChatGPT did not always align with clinical probabilities, potentially skewing the perceived effectiveness of the AI model. Additionally, it must be acknowledged that our research used ChatGPT based on GPT-4, which is not a publicly available model. Consequently, the result may not be directly generalizable to other LLMs, especially open-source models like Llama3 (Meta Platforms, Inc), which might have different underlying architectures and training data sets. Moreover, since our study relied on clinical vignettes that were mock scenarios, the potential for bias based on the cases is significant. The lack of real demographic diversity in these vignettes means that the findings may not accurately reflect social or regional nuances, such as ethnicity, prevalence of disease, or cultural practices, that could influence diagnostic outcomes. This limitation suggests a need for careful consideration when applying these AI tools across different geographic and demographic contexts to ensure the findings are appropriately adapted to local populations. This emphasizes the necessity for AI systems to be evaluated in diverse real-world settings to understand their effectiveness comprehensively and mitigate any bias. This distinction is important to consider when extrapolating our study's findings to other AI systems. Future studies should not only refine AI's diagnostic reasoning, but also explore the interpretability of its decision-making process, especially when errors occur. ChatGPT should be considered as a supplementary tool in medical diagnosis, rather than a standalone solution. This reinforces the necessity for combined expertise, where AI supports—but does not replace—human clinical judgment. Further research should expand these findings to a wider range of conditions, especially prevalent diseases with significant public health impacts, to thoroughly assess the practical utility and limitations of AI in medical diagnosis.

Conclusions

Our study contributes valuable evidence for the ongoing discourse on the role of AI in medical diagnosis. This study provides a foundation for future research to explore the extent to which AI can be trained to recognize increasingly complex and atypical presentations, which is critical for its successful integration into clinical practice.

Acknowledgments

The authors thank the members of Igaku-Shoin, Tokyo, Japan, for permission to use the clinical vignettes. Igaku-Shoin did not participate in designing and conducting the study; data analysis and interpretation; preparation, review, or approval of the paper; or the decision to submit the paper for publication. The authors thank Dr Mai Hongo, Saka General Hospital, for providing a clinical vignette. The authors also thank Editage for the English language review.

Data Availability

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

Disclaimer

In this study, generative artificial intelligence was used to create differential diagnoses for cases published in medical journals. However, it was not used in actual clinical practice. Similarly, no generative artificial intelligence was used in our manuscript writing.

Authors' Contributions

KS, T Watari, T Shimizu, Y Otsuka, M Tago, H Takahashi, YS, and YT designed the study. T Shimizu and Y Otsuka checked the atypical case categories. M Tago and H Takahashi confirmed the diagnoses. KS wrote the first draft and analyzed the research data. All authors created atypical common clinical vignettes and published them in the *Journal of General Medicine*. KS, T Shimizu, and H Takahashi critically revised the manuscript. All authors checked the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Differential medical diagnosis list generated by ChatGPT.
[DOCX File, 22 KB - mededu_v10i1e58758_app1.docx]

Multimedia Appendix 2
Transcript of the conversation with ChatGPT and the answers to all the questions.
[DOCX File, 36 KB - mededu_v10i1e58758_app2.docx]

References

XSL•FO
RenderX

1. Brown MP, Lai-Goldman M, Billings PR. Translating innovation in diagnostics: challenges and opportunities. Genomic Pers Med 2009:367-377. [doi: 10.1016/B978-0-12-369420-1.00031-7]
2. Omron R, Kotwal S, Garibaldi BT, Newman-Toker DE. The diagnostic performance feedback “calibration gap”: why clinical experience alone is not enough to prevent serious diagnostic errors. AEM Educ Train 2018 Oct;2(4):339-342. [doi: 10.1002/aet2.10119] [Medline: 30386846]
3. Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care: National Academies Press; 2015.
4. Friberg N, Ljungberg O, Berglund E, et al. Cause of death and significant disease found at autopsy. Virchows Arch 2019 Dec;475(6):781-788. [doi: 10.1007/s00428-019-02672-z] [Medline: 31691009]
5. Shojania KG, Burton EC, McDonald KM, Goldman L. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. JAMA 2003 Jun 4;289(21):2849-2856. [doi: 10.1001/jama.289.21.2849] [Medline: 12783916]
6. Schmitt BP, Kushner MS, Wiener SL. The diagnostic usefulness of the history of the patient with dyspnea. J Gen Intern Med 1986;1(6):386-393. [doi: 10.1007/BF02596424] [Medline: 3794838]
7. Kuijpers C, Fronczek J, van de Goot FRW, Niessen HWM, van Diest PJ, Jiwa M. The value of autopsies in the era of high-tech medicine: discrepant findings persist. J Clin Pathol 2014 Jun;67(6):512-519. [doi: 10.1136/jclinpath-2013-202122] [Medline: 24596140]
8. Ball JR, Balogh E. Improving diagnosis in health care: highlights of a report from the National Academies Of Sciences, Engineering, and Medicine. Ann Intern Med 2016 Jan 5;164(1):59-61. [doi: 10.7326/M15-2256] [Medline: 26414299]
9. Harada Y, Otaka Y, Katsukura S, Shimizu T. Prevalence of atypical presentations among outpatients and associations with diagnostic error. Diagnosis (Berl) 2024 Feb 1;11(1):40-48. [doi: 10.1515/dx-2023-0060] [Medline: 38059495]
10. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J 1975 May 31;2(5969):486-489. [doi: 10.1136/bmj.2.5969.486] [Medline: 1148666]
11. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 1992 Feb;156(2):163-165. [Medline: 1536065]
12. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689. [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
13. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurol Open 2023;5(1):e000451. [doi: 10.1136/bmjno-2023-000451] [Medline: 37337531]
14. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Dermatology Specialty Certificate Examination multiple choice questions. Clin Exp Dermatol 2023 Jun 2:llad197. [doi: 10.1093/ced/llad197] [Medline: 37264670]
15. Srivastav S, Chandrakar R, Gupta S, et al. ChatGPT in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. Cureus 2023 Jul;15(7):e41435. [doi: 10.7759/cureus.41435] [Medline: 37546142]
16. Andykarayalar R, Mohan Surapaneni K. ChatGPT in pediatrics: unraveling its significance as a clinical decision support tool. Indian Pediatr 2024 Apr 15;61(4):357-358. [Medline: 38450533]
17. Al-Antari MA. Artificial intelligence for medical diagnostics-existing and future AI technology!. Diagnostics (Basel) 2023 Feb 12;13(4):688. [doi: 10.3390/diagnostics13040688] [Medline: 36832175]
18. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach 2024 Mar;46(3):366-372. [doi: 10.1080/0142159X.2023.2249588] [Medline: 37839017]
19. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
20. Fukuzawa F, Yanagita Y, Yokokawa D, et al. Importance of patient history in artificial intelligence-assisted medical diagnosis: comparison study. JMIR Med Educ 2024 Apr 8;10:e52674. [doi: 10.2196/52674] [Medline: 38602313]
21. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659. [doi: 10.2196/48659] [Medline: 37606976]
22. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 9;11:e48808. [doi: 10.2196/48808] [Medline: 37812468]
23. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG. Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0’s accuracy with the American Journal of Neuroradiology’s (AJNR) “case of the month”. Cureus 2023 Aug;15(8):e43958. [doi: 10.7759/cureus.43958] [Medline: 37746411]
24. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care--a systematic review. Fam Pract 2008 Dec;25(6):400-413. [doi: 10.1093/fampra/cmn071] [Medline: 18842618]
25. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977 Jun;33(2):363-374. [Medline: 884196]
26. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient’s medical history and imaging findings in neuroradiology cases. Neuroradiology 2024 Jan;66(1):73-79. [doi: 10.1007/s00234-023-03252-4] [Medline: 37994939]

XSL•FO
RenderX

27. Umapathy VR, Rajinikanth B S, Samuel Raj RD, et al. Perspective of artificial intelligence in disease diagnosis: a review of current and future endeavours in the medical field. Cureus 2023 Sep;15(9):e45684. [doi: 10.7759/cureus.45684] [Medline: 37868519]
28. Mizuta K, Hirosawa T, Harada Y, Shimizu T. Can ChatGPT-4 evaluate whether a differential diagnosis list contains the correct diagnosis as accurately as a physician? Diagnosis (Berl) 2024 Mar 12. [doi: 10.1515/dx-2024-0027] [Medline: 38465399]
29. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. BMC Digit Health 2024;2(1):4. [doi: 10.1186/s44247-023-00058-5]
30. Truhn D, Eckardt JN, Ferber D, Kather JN. Large language models and multimodal foundation models for precision oncology. NPJ Precis Oncol 2024 Mar 22;8(1):72. [doi: 10.1038/s41698-024-00573-2] [Medline: 38519519]
31. Tayebi Arasteh S, Han T, Lotfinia M, et al. Large language models streamline automated machine learning for clinical studies. Nat Commun 2024 Feb 21;15(1):1603. [doi: 10.1038/s41467-024-45879-8] [Medline: 38383555]
32. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
33. Harada T, Shimizu T, Kaji Y, et al. A perspective from a case conference on comparing the diagnostic process: human diagnostic thinking vs. artificial intelligence (AI) decision support tools. Int J Environ Res Public Health 2020 Aug 22;17(17):6110. [doi: 10.3390/ijerph17176110] [Medline: 32842581]
34. Voelker R. The promise and pitfalls of AI in the complex world of diagnosis, treatment, and disease management. JAMA 2023 Oct 17;330(15):1416-1419. [doi: 10.1001/jama.2023.19180] [Medline: 37755919]
35. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]

Abbreviations

AI: artificial intelligence
C: category
GAI: generative artificial intelligence
GIM: general internal medicine
GPT: generative pre-trained transformer
LLM: large language model
NLP: natural language processing
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 27.03.24; peer-reviewed by A Kivrak, L Passby, ST Arasteh; revised version received 03.05.24; accepted 19.05.24; published 21.06.24.

Please cite as:
Shikino K, Shimizu T, Otsuka Y, Tago M, Takahashi H, Watari T, Sasaki Y, Iizuka G, Tamura H, Nakashima K, Kunitomo K, Suzuki M, Aoyama S, Kosaka S, Kawahigashi T, Matsumoto T, Orihara F, Morikawa T, Nishizawa T, Hoshina Y, Yamamoto Y, Matsuo Y, Unoki Y, Kimura H, Tokushima M, Watanuki S, Saito T, Otsuka F, Tokuda Y
Evaluation of ChatGPT-Generated Differential Diagnosis for Common Diseases With Atypical Presentation: Descriptive Research
JMIR Med Educ 2024;10:e58758
URL: https://mededu.jmir.org/2024/1/e58758
doi:10.2196/58758

© Kiyoshi Shikino, Taro Shimizu, Yuki Otsuka, Masaki Tago, Hiromizu Takahashi, Takashi Watari, Yosuke Sasaki, Gemmei Iizuka, Hiroki Tamura, Koichi Nakashima, Kotaro Kunitomo, Morika Suzuki, Sayaka Aoyama, Shintaro Kosaka, Teiko Kawahigashi, Tomohiro Matsumoto, Fumina Orihara, Toru Morikawa, Toshinori Nishizawa, Yoji Hoshina, Yu Yamamoto, Yuichiro Matsuo, Yuto Unoki, Hirofumi Kimura, Midori Tokushima, Satoshi Watanuki, Takuma Saito, Fumio Otsuka, Yasuharu Tokuda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.6.2024. 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.

XSL•FO
RenderX

Importance of Patient History in Artificial Intelligence–Assisted Medical Diagnosis: Comparison Study

Fumitoshi Fukuzawa[1], MD; Yasutaka Yanagita[1], MD, PhD; Daiki Yokokawa[1], MD, PhD; Shun Uchida[2], MD; Shiho Yamashita[1], MD; Yu Li[1], MD, PhD; Kiyoshi Shikino[1], MD, MHPE, PhD; Tomoko Tsukamoto[1], MD, PhD; Kazutaka Noda[1], MD, PhD; Takanori Uehara[1], MD, PhD; Masatomi Ikusaka[1], MD, PhD

1

2

Corresponding Author:
Fumitoshi Fukuzawa, MD

Abstract

Background: Medical history contributes approximately 80% to a diagnosis, although physical examinations and laboratory investigations increase a physician's confidence in the medical diagnosis. The concept of artificial intelligence (AI) was first proposed more than 70 years ago. Recently, its role in various fields of medicine has grown remarkably. However, no studies have evaluated the importance of patient history in AI-assisted medical diagnosis.

Objective: This study explored the contribution of patient history to AI-assisted medical diagnoses and assessed the accuracy of ChatGPT in reaching a clinical diagnosis based on the medical history provided.

Methods: Using clinical vignettes of 30 cases identified in *The BMJ*, we evaluated the accuracy of diagnoses generated by ChatGPT. We compared the diagnoses made by ChatGPT based solely on medical history with the correct diagnoses. We also compared the diagnoses made by ChatGPT after incorporating additional physical examination findings and laboratory data alongside history with the correct diagnoses.

Results: ChatGPT accurately diagnosed 76.6% (23/30) of the cases with only the medical history, consistent with previous research targeting physicians. We also found that this rate was 93.3% (28/30) when additional information was included.

Conclusions: Although adding additional information improves diagnostic accuracy, patient history remains a significant factor in AI-assisted medical diagnosis. Thus, when using AI in medical diagnosis, it is crucial to include pertinent and correct patient histories for an accurate diagnosis. Our findings emphasize the continued significance of patient history in clinical diagnoses in this age and highlight the need for its integration into AI-assisted medical diagnosis systems.

(JMIR Med Educ 2024;10:e52674) doi:10.2196/52674

KEYWORDS

medical diagnosis; ChatGPT; AI in medicine; diagnostic accuracy; patient history; medical history; artificial intelligence; AI; physical examination; physical examinations; laboratory investigation; laboratory investigations; mHealth; accuracy; public health; United States; AI diagnosis; treatment; male; female; child; children; youth; adolescent; adolescents; teen; teens; teenager; teenagers; older adult; older adults; elder; elderly; older person; older people; investigative; mobile health; digital health

Introduction

Over the past decade, medical knowledge and diagnostic techniques have expanded globally and have become more accessible with remarkable advancements in clinical testing and useful reference systems. Despite these advancements, misdiagnosis significantly contributes to mortality, making it a significant public health issue [1,2]. Studies have shown discrepancies between clinical and postmortem autopsy diagnoses in at least 25% of patients [3-7]. One study suggests that approximately 40,500 adult patients in intensive care units in the United States die of misdiagnoses annually, and the predicted prevalence of potentially lethal misdiagnoses is 6.3% [8]. Another report suggests that diagnostic errors contribute to approximately 10% of deaths and 6% to 17% of hospital adverse events, and are the leading cause of medical malpractice claims [7]. Considering the operative characteristics of clinical investigations combined with the inherent variability in disease presentation, it is often challenging to diagnose patients correctly—an issue that has concerned physicians perennially. Decades ago, a pivotal study proposed that patient history contributes to approximately 80% of the diagnostic process [9,10]. Medical history remains crucial for diagnosis [11,12] and is vital in contemporary physicians' clinical diagnoses.

With the advent of artificial intelligence (AI) in recent years, numerous studies have focused on AI-assisted diagnoses, including cancer screening and treatment [13-15], diagnostic ultrasound imaging [16-19], x-ray imaging [20], computed

XSL•FO
RenderX

tomography [21], magnetic resonance imaging [22], and endoscopy [15,23]. Other reports on AI-assisted imaging diagnoses include AI’s applications in radiology, pathology, and dermatological imaging [13,24]. There have also been reports on the use of AI in diagnosing specific conditions [25-27]. However, while several studies have reported that AI is useful in screening, diagnosing, and even treating certain medical conditions, to the best of our knowledge, no study has examined the importance of patient history in AI-assisted medical diagnosis. In addition, the extent to which AI considers patient history in its diagnostic processes remains to be fully understood.

This study aimed to investigate the importance of patient history in an AI-assisted medical diagnostic process aided by ChatGPT (version 4.0; June 2, 2023), one of the most well-known large language models that was released on March 14, 2023, to better understand the future of diagnostic medicine where AI is predicted to play an increasingly prominent role. Our study explored the contribution of patient history to AI-assisted medical diagnoses and assessed the accuracy of ChatGPT in reaching a clinical diagnosis based on the medical history that was provided. By reevaluating the significance of patient history, our study contributes to the ongoing discourse on optimizing diagnostic processes, both conventional and AI-assisted.

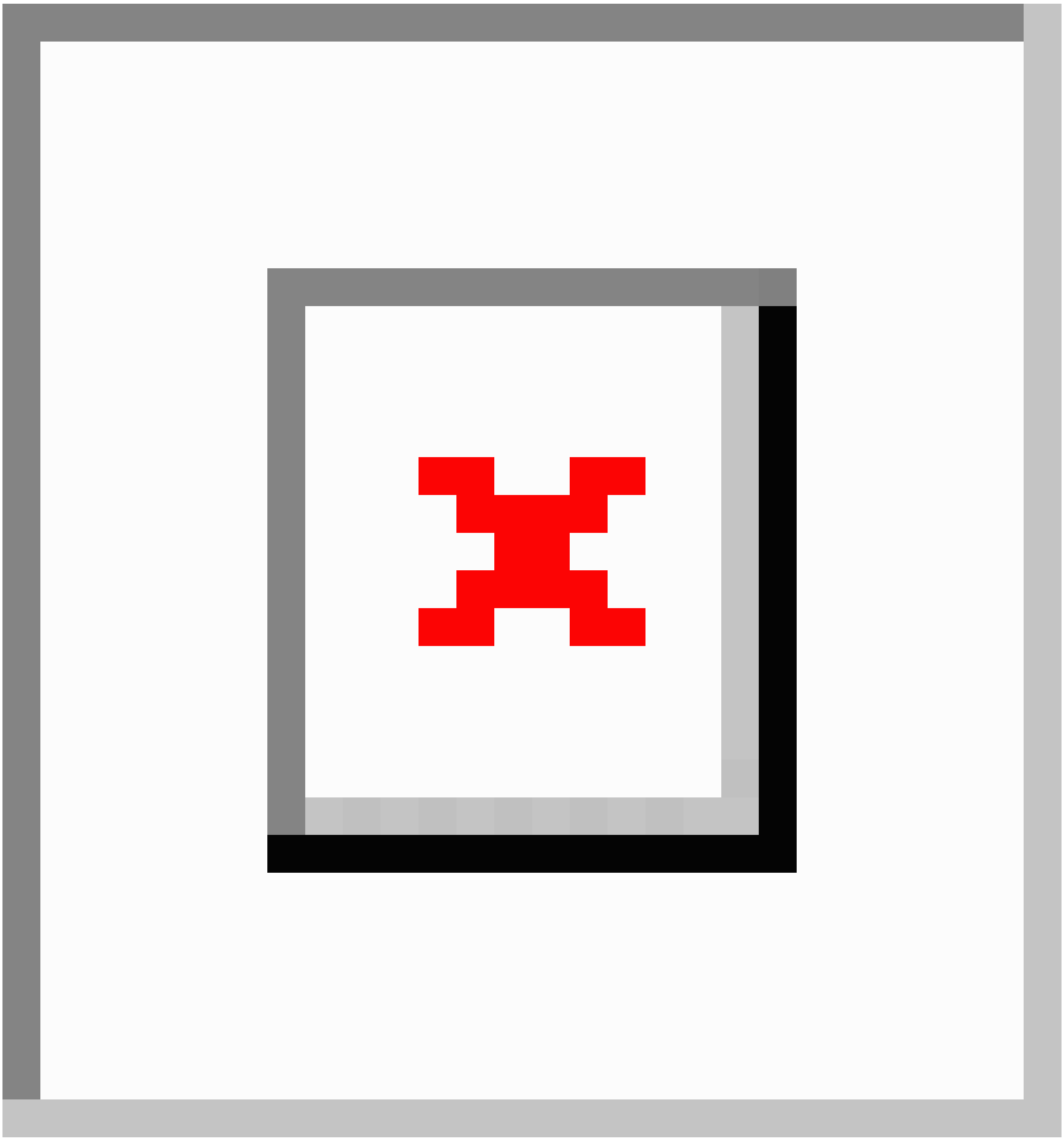
Methods

Study Design, Settings, and Participants

In our study, we used some of the 45 standardized clinical vignettes in *The BMJ* (Multimedia Appendix 1) to evaluate the diagnostic and triage accuracy of web-based symptom checkers [28]. These vignettes were published on June 5, 2015. They offer a balanced set of cases, with 15 cases requiring immediate attention, 15 cases requiring consultation but not immediately, and 15 cases not requiring immediate attention or consultation. They were identified from various clinical sources, including materials used to educate health professionals as well as a medical resource website, with content provided by a panel of physicians. Researchers have used these clinical vignettes to evaluate the usefulness of web-based symptom checkers and self-triage [28-31]. We chose these vignettes because of their varied severity levels, their origins from multiple resources rather than just 1 resource, and their credibility, having been used in prior studies. They also include some of the most commonly observed conditions in outpatient settings. Of the 45 cases, we selected those that included physical examination findings, test data, and medical history and provided a single distinct diagnosis. As illustrated in Figure 1, we excluded patients with no distinct diagnoses within the vignettes to serve as a reference (3 cases) and those who did not undergo any physical examination or laboratory tests (12 cases). Finally, the remaining 30 cases were used in this study.

XSL•FO
RenderX

Figure 1. Inclusion and exclusion criteria.



Data Collection and Measurements

We assigned the correct diagnosis for each of these 30 cases to "Answer." We then used the AI model, ChatGPT, to generate 2 diagnoses: the first, labeled "History," was obtained by inputting only the medical history into ChatGPT; the second set, labeled "All," was produced by inputting the medical history and all the other additional information in the clinical vignettes. Each time ChatGPT was prompted to generate a diagnosis, a separate chat window was used (Multimedia Appendix 2). Thus, we used 2 chat windows for each case—one for the "History" diagnosis and the other for the "All" diagnosis. Additionally, the patients' information was not inputted incrementally.

The concordance rate was assessed among "Answer," "History," and "All." To extract a diagnosis from ChatGPT, we ended each input session with the phrase "What is the most likely diagnosis?" For both the "History" and "All," the session was deemed complete when the AI returned the single most likely diagnosis. If ChatGPT suggested multiple diagnoses or indicated that it did not provide the most likely diagnosis, we repeated the process under the same conditions for a maximum of 5 attempts. Cases for which a single diagnosis could not be obtained even after 5 attempts were excluded without making further attempts.

XSL•FO
RenderX

Ethical Considerations

Our research does not involve humans, medical records, patient information, observations of public behaviors, or secondary data analyses; hence, it is exempt from ethical approval, the requirement of informed consent, and institutional review board approval. Additionally, as no identifying information was included, the data did not need to be anonymized or deidentified, and the need for compensation did not arise because no human participants were included in the study.

Data Analysis

Three board-certified physicians working in a medical diagnostic department at our facility assessed the concordance among the 3 AI-proposed diagnoses ("Answer," "History," and "All"). Of the 3 physicians, 1 is general medicine board–certified, 1 is internal medicine board–certified, and 1 is internal medicine–, general internal medicine–, and family medicine board–certified; their postgraduate education spanned 7, 9, and 11 years, respectively. A diagnosis was considered to match if at least 2 of the 3 physicians agreed upon the correspondence. Distinguishing between acute pharyngitis and acute upper respiratory tract infection necessitated determining whether to consider diseases resulting from similar pathologies as correct diagnoses. In contrast, for diseases that are essentially the same but have different nomenclatures, such as oral ulcers and canker sores, we considered them correct diagnoses.

Results

Among the 30 cases, 19 patients were male and 11 were female, with ages ranging from 18 months to 65 years. In total, 12 individuals were younger than 20 years.

The results are shown in Table 1. Cases 1-15 of the original vignette represent those requiring emergent care, cases 16-30 represent those requiring nonemergent care, and cases 31-45 represent those that are appropriate for self-care. A comparison with the correct diagnosis listed in *The BMJ* vignettes (labeled as "Answer") showed that "Answer" and "History" coincided 76.6% of the time, while "Answer" and "All" had a concordance rate of 93.3%. Five (16.7%) patients could not be diagnosed on the basis of medical history alone but were diagnosed when additional information was provided. In 1 (3.3%) case, the diagnosis was different and incorrect under both conditions ("History" and "All"). In 1 (3.3%) case, the incorrect diagnosis was the same under both conditions ("History" and "All").

XSL•FO
RenderX

Table . List of answers and diagnoses made by ChatGPT[a].

Case number of the original vignette	Original diagnosis (Answer)	Output from history only (History)[b]	Output from all information (All)[c]
1	Acute liver failure	Acute liver failure[d]	Acute liver failure[d]
2	Appendicitis	Acute gastroenteritis	Acute peritonitis, possibly secondary to a ruptured appendix (perforated appendicitis)[d]
5	Deep vein thrombosis	Deep vein thrombosis[d]	Deep vein thrombosis[d]
6	Heart attack	Acute myocardial infarction[d]	Acute anterior wall myocardial infarction[d]
7	Hemolytic uremic syndrome	Hemolytic uremic syndrome[d]	Hemolytic uremic syndrome[d]
9	Malaria	Malaria[d]	Malaria[d]
10	Meningitis	N/A[e] × 5[f]	Meningitis[d]
11	Pneumonia	Community-acquired pneumonia[d]	Community-acquired pneumonia[d]
12	Pulmonary embolism	Pulmonary embolism[d]	Pulmonary embolism[d]
13	Rocky Mountain spotted fever	Tick-borne illness, such as Rocky Mountain spotted fever or ehrlichiosis[d]	Rocky Mountain spotted fever[d]
16	Acute otitis media	Viral upper respiratory tract infection	Acute otitis media[d]
17	Acute pharyngitis	Strep throat[d]	Streptococcal pharyngitis[d]
18	Acute pharyngitis	Streptococcal pharyngitis[d]	Streptococcal pharyngitis[d]
19	Acute sinusitis	Acute sinusitis[d]	N/A × 2[g]; acute bacterial sinusitis[d]
21	Cellulitis	N/A × 5	Cellulitis[d]
24	Mononucleosis	Infectious mononucleosis[d]	Infectious mononucleosis[d]
25	Peptic ulcer disease	Peptic ulcer disease[d]	Peptic ulcer disease[d]
26	Pneumonia	Pneumonia[d]	Community-acquired pneumonia[d]
27	*Salmonella* infection	*Campylobacter jejuni* infection	Acute gastroenteritis, likely due to food poisoning
30	Vertigo	Benign paroxysmal positional vertigo[d]	Benign paroxysmal positional vertigo[d]
31	Acute bronchitis	Acute bronchitis[d]	Acute bronchitis[d]
32	Acute bronchitis	Acute bronchitis[d]	Acute bronchitis[d]
33	Acute conjunctivitis	Viral conjunctivitis[d]	Viral conjunctivitis[d]
34	Acute pharyngitis	Viral upper respiratory tract infection	Upper respiratory tract infection
37	Bee sting without anaphylaxis	Pain of the sting	Localized allergic reaction to a bee sting[d]
38	Canker sore	Recurrent aphthous stomatitis[d]	Recurrent aphthous stomatitis[d]
39	Candida yeast infection	Vaginal candidiasis[d]	Vulvovaginal candidiasis[d]
42	Stye	Hordeolum[d]	Hordeolum[d]
43	Viral upper respiratory tract infection	Acute sinusitis[d]	Acute sinusitis[d]

XSL•FO
RenderX

Case number of the original vignette	Original diagnosis (Answer)	Output from history only (History)[b]	Output from all information (All)[c]
44	Viral upper respiratory tract infection	Common viral illness, such as the common cold or influenza[d]	Viral upper respiratory tract infection[d]

[a]We repeated outputs until a single plausible diagnosis was made, with a maximum of 5 attempts.

[b]Matching answers between Answer and History: 23/30 (76.6%); median trial count 1 (Q1 1, Q2 1, Q3 1).

[c]Matching answers between History and All: 28/30 (93.3%); median trial count 1 (Q1 1, Q2 1, Q3 1).

[d]The output matched with that of "Answer."

[e]N/A: not applicable.

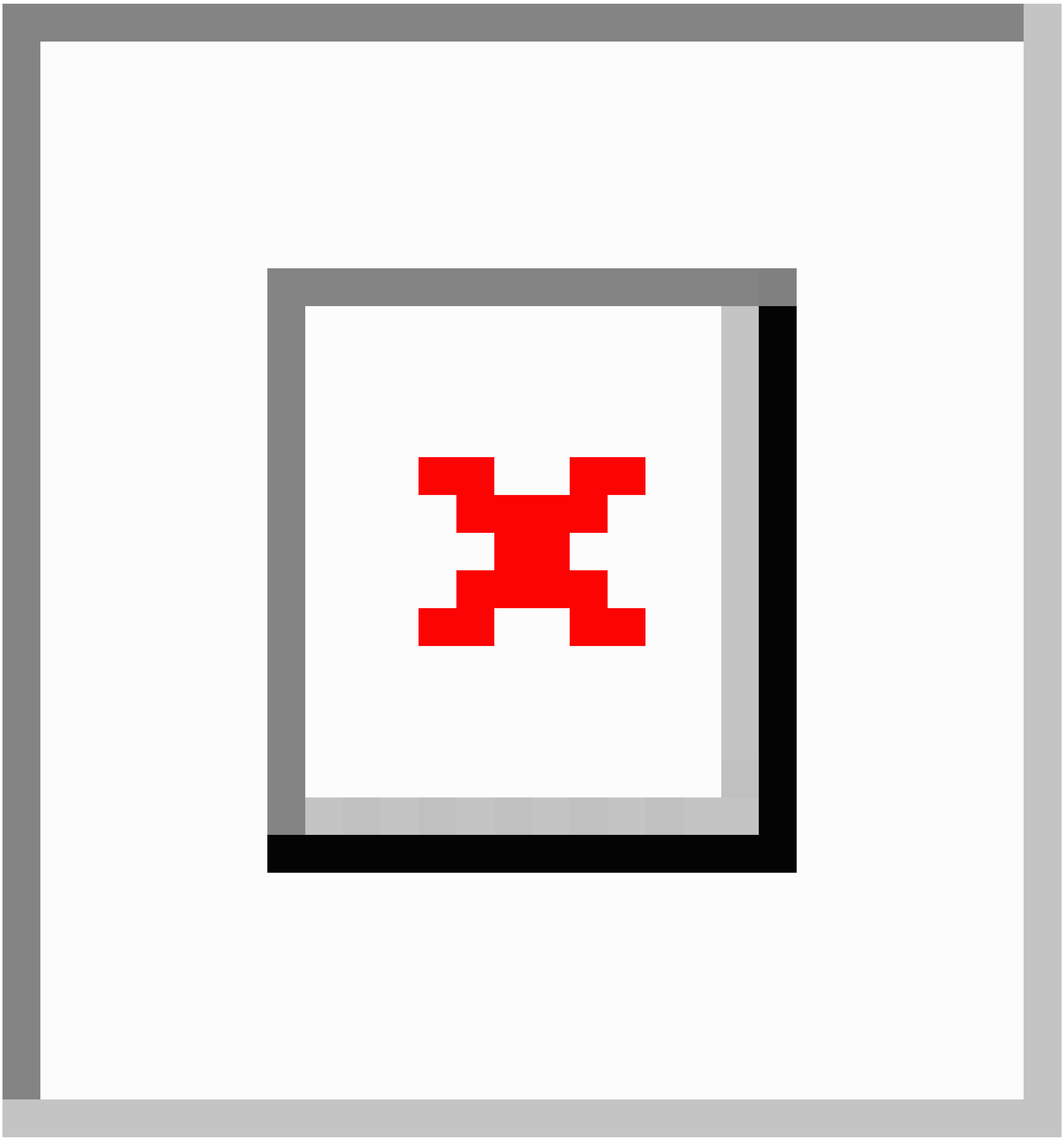
[f]We attempted to obtain a diagnosis 5 times but failed.

[g]We attempted to obtain a diagnosis twice but failed.

Figure 2 presents details regarding the number of attempts required. On average, 1.27 attempts were needed for inputs involving only medical history followed by the question "What is the most likely diagnosis?" When all possible information, including physical examination findings and laboratory data, were inputted, followed by the same question, an average of 1.00 attempt was required. Regarding the 2 cases shown in Figure 2 that required 5 attempts, ChatGPT was unable to narrow down the diagnosis to the single most likely option. Consequently, these cases were counted as mismatches with the correct diagnoses listed in *The BMJ* vignettes.

XSL•FO
RenderX

Figure 2. Data collection and measurements.



Discussion

Principal Findings

Despite the advancements in medical knowledge and diagnostic techniques, misdiagnosis remains a significant issue. AI has shown promise in the diagnosis and treatment of medical conditions; however, there is limited understanding of how AI uses patient history for diagnostic purposes. Our study aimed to investigate the extent to which AI (ChatGPT) can use information from medical history to accurately diagnose common diseases, which are frequently encountered in general outpatient, emergency, and ward management settings. Although some studies have investigated the accuracy of AI-based medical diagnosis, our study is novel because it emphasizes the importance of patient history. We compared the diagnostic accuracy of diagnoses made on the basis of only patient history and those made using all the information; this makes our study unique. To the best of our knowledge, no previous research has been conducted on this topic.

Our study investigated the role of patient history in AI-assisted medical diagnoses using ChatGPT. We analyzed 30 standardized patient vignettes from *The BMJ* to assess the concordance rates between AI-proposed diagnoses based on medical history only and those based on both medical history and additional information. Our results showed high concordance rates of 76.6% between the “Answer” and “History” groups, suggesting the importance of patient history in AI-assisted diagnoses and highlighting the potential of AI in improving diagnostic

XSL•FO
RenderX

accuracy. This result is similar to that of a previous study that involved actual physicians instead of ChatGPT [9,10].

Characteristics of cases that did not lead to appropriate diagnoses based on history alone include, for instance, the following: an appendicitis case (case 2 in Multimedia Appendix 1) for which there was no documentation of pain migration in the medical history, a meningitis case (case 10 in Multimedia Appendix 1) wherein only headache and fever were documented, an otitis media case (case 16 in Multimedia Appendix 1) wherein only upper respiratory symptoms were recorded with no mention of ear-related symptoms, errors in identifying the causative agent in a case of acute gastroenteritis (case 27 in Multimedia Appendix 1), and an acute pharyngitis case (case 34 in Multimedia Appendix 1) that lacked the necessary medical history to determine the Centor score. Such omissions in the medical history could be considered contributing factors to the misdiagnoses. When physical findings and test data were added, an accurate diagnosis was achieved in 28 out of 30 cases (93.3%), showing a 16.7% increase in the accuracy rate. These two cases were of acute pharyngitis diagnosed as acute upper respiratory tract infection and *Salmonella* enteritis diagnosed as acute gastroenteritis. While we considered these incorrect diagnoses for the purpose of this study, they could have been deemed correct under certain criteria. Of the 7 cases that did not match between "Answer" and "History," 6 were of infectious diseases (21 of 30 cases were of infectious diseases). These included cases where appendicitis was mistaken for acute gastroenteritis, acute otitis media and acute pharyngitis were mistaken for upper respiratory infections, and a *Salmonella* infection was mistaken for a *Campylobacter* infection. Physical examinations or tests may help identify the site of infection or pathogen in cases of intra-abdominal or head and neck infections.

There are situations in which physical examination and clinical test information may not be available in clinical settings. For instance, digital patient encounters owing to the impact of the COVID-19 pandemic often preclude physical examinations and clinical tests. The widespread use of telemedicine approaches in COVID-19 management, from screening to follow-up, has demonstrated the community's acceptance and interest in telehealth solutions [32]. Moreover, even in face-to-face consultations, there are scenarios, such as in clinics, where detailed clinical tests may not be feasible depending on the setting. Furthermore, we cannot perform all physical examinations and tests on all patients. Therefore, we should consider potential differential diagnoses and decide which pertinent physical examinations or tests are the most suitable and should be performed. Most importantly, it has been reported that one rarely makes a correct diagnosis when one cannot make a differential diagnosis based on history [11]. In addition, accurately predicting the diagnosis based on medical history is associated with a higher diagnostic accuracy of the physical examination, whereas incorrect prediction of the diagnosis based on medical history is associated with a lower diagnostic accuracy of the physical examination [33]. Based on these findings and suggestions, medical diagnosis using ChatGPT is considered heavily dependent on history.

Using AI for diagnosis can enhance diagnostic accuracy by more efficiently collecting medical histories. For instance, diagnosing acute appendicitis is sometimes challenging. AI may face the same challenge as that observed when, in our study, AI mistakenly identified acute appendicitis as acute gastroenteritis. This misdiagnosis may have occurred because the case lacked specific medical histories characteristic of appendicitis, such as pain migration. By configuring AI systems to verify pain migration in patients with abdominal pain, especially for such common conditions, diagnostic precision may improve.

There are 2 possible limitations in our study. First, it remains unclear whether similar results could be obtained with other vignettes or actual patients. Unlike using preprovided vignettes, among which we included 30 cases, diagnosis can be more challenging in clinical settings because it requires taking a medical history from patients. We included 30 cases from among the vignettes, which include some of the most commonly observed conditions in the outpatient setting. Although covering all the existing conditions is not feasible, we do not know if the case volume in our study is sufficiently high. This study included relatively simple cases in which patients had very few comorbidities, potentially making the diagnosis less challenging. Moreover, patients with psychiatric conditions tend to present with complex and lengthy case histories, and the wording used by mental health clinicians may differ, be inconsistent, be vague, or fail to pinpoint a diagnosis. Our vignettes did not include a diagnosis of any mental illness. Due to the abovementioned reasons, our results may not apply to all clinical settings. Furthermore, when we consider what the patient reports, results may differ if languages other than English are used since ChatGPT does not recognize some languages, and each language may have its unique nuance. This highlights the importance of linguistic diversity and cultural context in AI applications, particularly in medical diagnoses where patient communication and history are critical. Future iterations of AI systems should aim to incorporate a broader range of languages and understand cultural nuances to ensure more accurate and inclusive diagnostic support. This idea is important in the context of health inequality. Furthermore, disparities in technology access may pose some challenges. Future research should address these barriers to ensure equitable access to AI-assisted diagnostic tools.

Second, we encountered cases where the input of medical history followed by the question, "What is the most likely diagnosis?" failed to yield a single most likely diagnosis even after 5 attempts, which could have introduced bias into our results, although we only had 2 such cases.

In the future, studies should focus on training AI by implementing evidence-based medical information, enabling it to present the underlying reasons and guidelines for diagnoses. In the event of a misdiagnosis, analyzing the process that led to the false diagnosis could be challenging in an AI-assisted medical diagnosis. Given the current situation where reflection on misdiagnoses is not always feasible, AI should be used as an auxiliary tool in medical diagnosis. This approach underscores the importance of AI, deeming it a support system rather than a definitive diagnostic solution. This area needs

XSL•FO
RenderX

further investigation. Future studies should also verify our results with certain common conditions or diseases, such as the top 10 diseases identified in the Global Burden of Diseases study [34], potentially leveraging the benefits and limitations of AI-assisted medical diagnosis.

Conclusions

Relevant patient history is essential for AI-assisted diagnosis. The input of relevant patient history or the development of AI systems capable of obtaining comprehensive medical histories is vital for AI-assisted medical diagnosis. Furthermore, even in the modern era of advanced medical knowledge and clinical testing, the significance of patient history in diagnosis remains crucial.

Data Availability

All of our clinical vignettes, results, and prompts used are provided in Multimedia Appendix 1.

Authors' Contributions

FF conceptualized the study, designed the methodology, collected the data, and drafted the manuscript. YY, DY, and SU conceptualized the study, designed the methodology, and reviewed and edited the manuscript. SY, YL, KS, TT, KN, TU, and MI conceptualized the study and reviewed and edited the manuscript. No generative artificial intelligence was used in writing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical Vignettes used in our study.
[PDF File, 159 KB - mededu_v10i1e52674_app1.pdf]

Multimedia Appendix 2
Explanation of the prompts we used in our study.
[PDF File, 48 KB - mededu_v10i1e52674_app2.pdf]

References

1. Omron R, Kotwal S, Garibaldi BT, Newman-Toker DE. The diagnostic performance feedback “calibration gap”: why clinical experience alone is not enough to prevent serious diagnostic errors. AEM Educ Train 2018 Oct;2(4):339-342. [doi: 10.1002/aet2.10119] [Medline: 30386846]
2. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. In: Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care: National Academies Press; 2015.
3. Friberg N, Ljungberg O, Berglund E, et al. Cause of death and significant disease found at autopsy. Virchows Arch 2019 Dec;475(6):781-788. [doi: 10.1007/s00428-019-02672-z] [Medline: 31691009]
4. Shojania KG, Burton EC, McDonald KM, Goldman L. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. JAMA 2003 Jun 4;289(21):2849-2856. [doi: 10.1001/jama.289.21.2849] [Medline: 12783916]
5. Schmitt BP, Kushner MS, Wiener SL. The diagnostic usefulness of the history of the patient with dyspnea. J Gen Intern Med 1986;1(6):386-393. [doi: 10.1007/BF02596424] [Medline: 3794838]
6. Kuijpers C, Fronczek J, van de Goot FRW, Niessen HWM, van Diest PJ, Jiwa M. The value of autopsies in the era of high-tech medicine: discrepant findings persist. J Clin Pathol 2014 Jun;67(6):512-519. [doi: 10.1136/jclinpath-2013-202122] [Medline: 24596140]
7. Ball JR, Balogh E. Improving diagnosis in health care: highlights of a report from the National Academies of Sciences, Engineering, and Medicine. Ann Intern Med 2016 Jan 5;164(1):59-61. [doi: 10.7326/M15-2256] [Medline: 26414299]
8. Winters B, Custer J, Galvagno SM, et al. Diagnostic errors in the intensive care unit: a systematic review of autopsy studies. BMJ Qual Saf 2012 Nov;21(11):894-902. [doi: 10.1136/bmjqs-2012-000803] [Medline: 22822241]
9. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J 1975 May 31;2(5969):486-489. [doi: 10.1136/bmj.2.5969.486] [Medline: 1148666]
10. Peterson MC, Holbrook JM, Hales DV, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. Obstet Gynecol Surv 1992 Oct;47(10):711-712. [doi: 10.1097/00006254-199210000-00013]

XSL•FO
RenderX

11. Gruppen LD, Palchik NS, Wolf FM, Laing TJ, Oh MS, Davis WK. Medical student use of history and physical information in diagnostic reasoning. Arthritis Care Res 1993 Jun;6(2):64-70. [doi: 10.1002/art.1790060204] [Medline: 8399428]
12. Tsukamoto T, Ohira Y, Noda K, Takada T, Ikusaka M. The contribution of the medical history for the diagnosis of simulated cases by medical students. Int J Med Educ 2012 Apr;3:78-82. [doi: 10.5116/ijme.4f8a.e48c]
13. Chen ZH, Lin L, Wu CF, Li CF, Xu RH, Sun Y. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. Cancer Commun (Lond) 2021 Nov;41(11):1100-1115. [doi: 10.1002/cac2.12215] [Medline: 34613667]
14. Mitsala A, Tsalikidis C, Pitiakoudis M, Simopoulos C, Tsaroucha AK. Artificial intelligence in colorectal cancer screening, diagnosis and treatment. A new era. Curr Oncol 2021 Apr 23;28(3):1581-1607. [doi: 10.3390/curroncol28030149] [Medline: 33922402]
15. Ochiai K, Ozawa T, Shibata J, Ishihara S, Tada T. Current status of artificial intelligence-based computer-assisted diagnosis systems for gastric cancer in endoscopy. Diagnostics (Basel) 2022 Dec 13;12(12):3153. [doi: 10.3390/diagnostics12123153] [Medline: 36553160]
16. Calisto FM, Santiago C, Nunes N, Nascimento JC. Breastscreening-AI: evaluating medical intelligent agents for human-AI interactions. Artif Intell Med 2022 May;127:102285. [doi: 10.1016/j.artmed.2022.102285] [Medline: 35430044]
17. Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. World J Gastroenterol 2019 Feb 14;25(6):672-682. [doi: 10.3748/wjg.v25.i6.672] [Medline: 30783371]
18. Peng S, Liu Y, Lv W, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. Lancet Digit Health 2021 Apr;3(4):e250-e259. [doi: 10.1016/S2589-7500(21)00041-8] [Medline: 33766289]
19. Drukker L, Noble JA, Papageorghiou AT. Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. Ultrasound Obstet Gynecol 2020 Oct;56(4):498-505. [doi: 10.1002/uog.22122] [Medline: 32530098]
20. Guermazi A, Tannoury C, Kompel AJ, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology 2022 Mar;302(3):627-636. [doi: 10.1148/radiol.210937] [Medline: 34931859]
21. Zhang K, Liu X, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020 Jun 11;181(6):1423-1433. [doi: 10.1016/j.cell.2020.04.045] [Medline: 32416069]
22. Gore JC. Artificial intelligence in medical imaging. Magn Reson Imaging 2020 May;68:A1-A4. [doi: 10.1016/j.mri.2019.12.006] [Medline: 31857130]
23. Okagawa Y, Abe S, Yamada M, Oda I, Saito Y. Artificial intelligence in endoscopy. Dig Dis Sci 2022 May;67(5):1553-1572. [doi: 10.1007/s10620-021-07086-z] [Medline: 34155567]
24. Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl 2004 Sep;86(5):334-338. [doi: 10.1308/147870804290] [Medline: 15333167]
25. Revilla-León M, Gómez-Polo M, Barmak AB, et al. Artificial intelligence models for diagnosing gingivitis and periodontal disease: a systematic review. J Prosthet Dent 2023 Dec;130(6):816-824. [doi: 10.1016/j.prosdent.2022.01.026] [Medline: 35300850]
26. Chung H, Jo Y, Ryu D, Jeong C, Choe SK, Lee J. Artificial-intelligence-driven discovery of prognostic biomarker for sarcopenia. J Cachexia Sarcopenia Muscle 2021 Dec;12(6):2220-2230. [doi: 10.1002/jcsm.12840] [Medline: 34704369]
27. Uzun Ozsahin D, Ozgocmen C, Balcioglu O, Ozsahin I, Uzun B. Diagnostic AI and cardiac diseases. Diagnostics (Basel) 2022 Nov 22;12(12):2901. [doi: 10.3390/diagnostics12122901] [Medline: 36552908]
28. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 8;351:h3480. [doi: 10.1136/bmj.h3480] [Medline: 26157077]
29. North F, Jensen TB, Stroebel RJ, et al. Self-triage use, subsequent healthcare utilization, and diagnoses: a retrospective study of process and clinical outcomes following self-triage and self-scheduling for ear or hearing symptoms. Health Serv Res Manag Epidemiol 2023;10:23333928231168121. [doi: 10.1177/23333928231168121] [Medline: 37101803]
30. Riboli-Sasco E, El-Osta A, Alaa A, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res 2023 Jun 2;25:e43803. [doi: 10.2196/43803] [Medline: 37266983]
31. Radionova N, Ög E, Wetzel AJ, Rieger MA, Preiser C. Impacts of symptom checkers for laypersons' self-diagnosis on physicians in primary care: scoping review. J Med Internet Res 2023 May 29;25:e39219. [doi: 10.2196/39219] [Medline: 37247214]
32. Khoshrounejad F, Hamednia M, Mehrjerd A, et al. Telehealth-based services during the COVID-19 pandemic: a systematic review of features and challenges. Front Public Health 2021;9:711762. [doi: 10.3389/fpubh.2021.711762] [Medline: 34350154]
33. Shikino K, Ikusaka M, Ohira Y, et al. Influence of predicting the diagnosis from history on the accuracy of physical examination. Adv Med Educ Pract 2015;6:143-148. [doi: 10.2147/AMEP.S77315] [Medline: 25759604]
34. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence

Edited by D Chartash, G Eysenbach, TDA Cardoso; submitted 12.09.23; peer-reviewed by C Baxter, H Sun; revised version received 31.01.24; accepted 15.02.24; published 08.04.24.

Please cite as:
Fukuzawa F, Yanagita Y, Yokokawa D, Uchida S, Yamashita S, Li Y, Shikino K, Tsukamoto T, Noda K, Uehara T, Ikusaka M
Importance of Patient History in Artificial Intelligence–Assisted Medical Diagnosis: Comparison Study
JMIR Med Educ 2024;10:e52674
URL: https://mededu.jmir.org/2024/1/e52674
doi:10.2196/52674

© Fumitoshi Fukuzawa, Yasutaka Yanagita, Daiki Yokokawa, Shun Uchida, Shiho Yamashita, Yu Li, Kiyoshi Shikino, Tomoko Tsukamoto, Kazutaka Noda, Takanori Uehara, Masatomi Ikusaka. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.4.2024. 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.

XSL•FO
RenderX

Original Paper

AI Education for Fourth-Year Medical Students: Two-Year Experience of a Web-Based, Self-Guided Curriculum and Mixed Methods Study

Areeba Abid[1], BS; Avinash Murugan[2], MD, MBA; Imon Banerjee[3], PhD; Saptarshi Purkayastha[4], PhD; Hari Trivedi[5], MD; Judy Gichoya[5], MD

[1]Emory University School of Medicine, Atlanta, GA, United States
[2]Yale New Haven Hospital, New Haven, CT, United States
[3]Mayo Clinic, Phoenix, GA, United States
[4]Indiana University-Purdue University, Indianapolis, IN, United States
[5]Department of Radiology, Emory University, Atlanta, GA, United States

Corresponding Author:
Areeba Abid, BS
Emory University School of Medicine
2015 Uppergate Dr
Atlanta, GA, 30307
United States
Phone: 1 (404) 727 4018
Email: areeba.abid@emory.edu

Abstract

Background: Artificial intelligence (AI) and machine learning (ML) are poised to have a substantial impact in the health care space. While a plethora of web-based resources exist to teach programming skills and ML model development, there are few introductory curricula specifically tailored to medical students without a background in data science or programming. Programs that do exist are often restricted to a specific specialty.

Objective: We hypothesized that a 1-month elective for fourth-year medical students, composed of high-quality existing web-based resources and a project-based structure, would empower students to learn about the impact of AI and ML in their chosen specialty and begin contributing to innovation in their field of interest. This study aims to evaluate the success of this elective in improving self-reported confidence scores in AI and ML. The authors also share our curriculum with other educators who may be interested in its adoption.

Methods: This elective was offered in 2 tracks: technical (for students who were already competent programmers) and nontechnical (with no technical prerequisites, focusing on building a conceptual understanding of AI and ML). Students established a conceptual foundation of knowledge using curated web-based resources and relevant research papers, and were then tasked with completing 3 projects in their chosen specialty: a data set analysis, a literature review, and an AI project proposal. The project-based nature of the elective was designed to be self-guided and flexible to each student's interest area and career goals. Students' success was measured by self-reported confidence in AI and ML skills in pre and postsurveys. Qualitative feedback on students' experiences was also collected.

Results: This web-based, self-directed elective was offered on a pass-or-fail basis each month to fourth-year students at Emory University School of Medicine beginning in May 2021. As of June 2022, a total of 19 students had successfully completed the elective, representing a wide range of chosen specialties: diagnostic radiology (n=3), general surgery (n=1), internal medicine (n=5), neurology (n=2), obstetrics and gynecology (n=1), ophthalmology (n=1), orthopedic surgery (n=1), otolaryngology (n=2), pathology (n=2), and pediatrics (n=1). Students' self-reported confidence scores for AI and ML rose by 66% after this 1-month elective. In qualitative surveys, students overwhelmingly reported enthusiasm and satisfaction with the course and commented that the self-direction and flexibility and the project-based design of the course were essential.

Conclusions: Course participants were successful in diving deep into applications of AI in their widely-ranging specialties, produced substantial project deliverables, and generally reported satisfaction with their elective experience. The authors are

XSL•FO
RenderX

hopeful that a brief, 1-month investment in AI and ML education during medical school will empower this next generation of physicians to pave the way for AI and ML innovation in health care.

(JMIR Med Educ 2024;10:e46500) doi:10.2196/46500

KEYWORDS

medical education; machine learning; artificial intelligence; elective curriculum; medical student; student; students; elective; electives; curricula; curriculum; lesson plan; lesson plans; educators; educator; teacher; teachers; teaching; computer programming; programming; coding; programmer; programmers; self guided; self directed

Introduction

Artificial intelligence (AI) and machine learning (ML) are poised to have a substantial impact in the health care space with many disruptive technologies on the horizon. Innovations in clinical care are increasingly impacted by the development and implementation of AI and ML, and as future clinicians, medical students need to become innovators and active participants in technological changes that will affect how they provide care for their patients. There is much excitement and curiosity among medical students about these technologies [1]. However, few programs exist to deliberately expose future physicians to their role in medicine, let alone to empower students to actively participate in AI and ML innovation [2]. While a plethora of high-quality web-based resources exist to teach programming skills and ML model development, there are few introductory curricula specifically tailored to medical students without a background in data science or programming. Additionally, there is little guidance provided to medical students on where to begin. Some medical societies do have AI outreach activities, but these are limited to trainees within their specialty [3-5].

The authors theorized that a 1-month elective for fourth-year students, composed of existing web-based resources and a project-based structure, would empower students to learn about the impact of AI and ML in their chosen specialty and begin contributing to innovation in their field of interest. The authors also aimed for the elective to be specialty-agnostic and customizable to each student's career goals. The goal of this senior elective is to demystify AI and ML in health care, enabling students to have informed conversations about these technologies and participate in their clinical advancement. The target participant in the elective is any senior medical student with an interest in AI, with no prerequisites for technical, mathematical, or engineering skills.

In this paper, we evaluate the success of this elective over a 2-year period based on self-reported confidence scores in AI and ML. We also publish our curriculum for other educators who may be interested in its adoption.

Methods

Design

We built our elective following advice on designing medical electives with the principles articulated by Ramalho et al [6], which emphasize that a one-size-fits-all approach is often inadequate and that electives benefit from allowing students to carve their own paths. Creating a medical elective in an overloaded, overworked environment is nontrivial, but prior studies on peer-organized coursework gave us insights into the effectiveness of peer-organized research in building academic confidence, as well as the importance of clearly defined learning objectives [7,8].

Technical and Nontechnical Tracks

Given the wide-ranging skill sets that medical students are equipped with before coming to medical school, this elective was offered in 2 tracks: Technical and Nontechnical. The Technical track was intended for the subset of students who were already competent computer programmers. This course did not aim to teach noncoding students how to code because it was expected that 1 month would not be sufficient time for students to make meaningful progress. Therefore, the Nontechnical track was offered to students with no technical background and focused on building a conceptual understanding of AI. Our goal for the Nontechnical track was to help students without a technical background develop a skill set and vocabulary that would enable them to participate in AI and ML evaluation and implementation processes in future collaborations with technical colleagues.

For both the Technical and Nontechnical tracks, the course was designed to address the following learning objectives:

1. Compare and contrast AI and ML.
2. State and differentiate various ML techniques (supervised/unsupervised, classification/regression, etc).
3. Appreciate the growing impact of ML in medicine, broadly and in the student's chosen specialty.
4. Develop an intuition of how machines "learn." Describe how neural networks are structured, trained, and evaluated. Learn vocabulary and concepts used to describe model training (loss functions, gradient descent, and backpropagation).
5. Understand the limitations and pitfalls of ML (reproducibility, interpretability, and bias).
6. Understand what kinds of medical problems can and cannot be solved by ML.
7. Describe issues that may arise in the implementation of an ML algorithm in clinical practice.
8. Discuss ethical issues that concern the use of ML in health care.

Didactic and Project-Based Components

In this self-guided, web-based course, students were referred to existing web-based courses and relevant research papers to supplement these learning objectives (Multimedia Appendix 1 [9-22]) but were expected to guide their own learning beyond this. Students were asked to share and write down their personal

XSL•FO
RenderX

goals at the beginning of the elective to guide their learning. They were also encouraged to spend time after each section on independent research to address lingering questions. The learning objectives and course resources were provided to students on a central document and students were able to follow along at their own pace. Because the course aimed to empower an individual student's interests and career goals, the elective was designed to establish a baseline level of understanding for all students, while also allowing students the freedom to dive deeper into the areas they were drawn to. Students were supported by the course's faculty advisor, a physician with substantial leadership and experience in AI and ML research.

Project Deliverables

Students were then tasked with completing at least 1 of the following project-based deliverables, and encouraged to complete others as their interests dictated:

1. Complete a literature review on the state of AI and ML in the student's chosen specialty.
2. Find and analyze 3 open-source health care data sets, considering strengths, weaknesses, and sources of error and bias.
3. Write a Project Proposal addressing a problem in the student's chosen specialty that can be solved with AI, with a discussion surrounding the implementation complexities.
4. Technical track only: Train and evaluate a clinical ML algorithm.

Details on these projects are provided in Multimedia Appendix 2 [23].

The full curriculum is hosted on the Emory Health Care Innovations and Translational Informatics Lab GitHub repository [24].

This course was initially designed during the COVID-19 pandemic, and maintained a web-based format throughout the 2 years it has been offered. All recommended resources were freely available to students on the web, although some required institutional access. The students attended weekly web-based laboratory meetings to discuss their progress and to be exposed to more advanced research in AI and ML. Students were also encouraged to identify an additional advisor (beyond the elective director, who they met with once a week) within their chosen specialty, who could provide domain expertise for their projects.

Qualitative Survey Data

Initially, the authors collected feedback from students qualitatively through one-on-one meetings; this feedback was used to improve the format and support structure of the elective. Beginning in October 2021, students were also asked for open-ended feedback on the strengths and weaknesses of the elective through anonymous surveys. They were asked:

- What was the most meaningful project or experience you completed during the elective? Do you intend to continue work on it past the end of the elective?
- Did you gain what you hoped to get out of this elective? Please explain.
- What resources were most useful to you during the elective?
- What could be most improved in the curriculum design of this elective?

Quantitative Survey Data

Beginning in October 2021, quantitative pre and postelective surveys were implemented using Google Forms to assess the effectiveness of the elective format and resources provided. Students were asked to fill out formal surveys to rate their confidence in AI and ML concepts and in technical data science and coding skills.

Before starting the elective, students were asked:

- How familiar are you with AI or ML concepts? (Likert scale, 1-5)
- How would you rate your technical data science or coding experience? (Likert scale, 1-5)

After completing the elective, students were asked:

- Did you choose the Technical or Nontechnical Track?
- After completing this elective, how familiar are you with AI or ML concepts? (Likert scale, 1-5)
- After completing this elective, how would you rate your technical data science or coding experience? (Likert scale, 1-5)

Statistical Analysis

Quantitative and discrete data from self-reported confidence scores was analyzed using the Wilcoxon rank sum test. Qualitative survey responses were reviewed in a descriptive manner rather than undergoing a formal analysis. Responses were manually examined for common themes, trends, and noteworthy insights, but no systematic coding framework was used and representative responses are included in the "Results" section.

Ethical Considerations

This study was deemed exempt from review by Emory University's institutional review board, under the category "Educational Tests, Surveys, Interviews, Observations." This is justified based on anonymity and minimal risk to survey participants. All participants were able to opt out of this educational experience and from data collection. Survey data were collected anonymously. Students were not compensated for participation.

Results

Overview

This web-based, self-directed elective was offered on a pass-or-fail basis each month to fourth-year students at Emory University School of Medicine beginning in May 2021. A maximum of 3 students were allowed to enroll each month. As of June 2022, a total of 19 students had signed up and completed the elective. All students successfully met elective requirements and passed the course. The students represented a diverse range of chosen specialties: diagnostic radiology (n=3), general surgery (n=1), internal medicine (n=5), neurology (n=2), obstetrics and gynecology (n=1), ophthalmology (n=1),

XSL•FO
RenderX

orthopedic surgery (n=1), otolaryngology (n=2), pathology (n=2), and pediatrics (n=1).

Given the limited time and open-ended nature of the course, students elected to spend varying amounts of time on each of the project components based on their interests and were not required to complete all 3 projects as long as they produced at least 1 significant deliverable. The vast majority of students (17 out of 19 students) chose the Nontechnical track. Most students (11/19, 58%) chose to focus their efforts on 2 of the 3 projects; 8 (42%) completed all 3 projects, and 1 (5%) submitted only a project proposal. Since the elective was intended to be flexible to students' interests, students were evaluated on a pass-or-fail basis based on demonstrated effort as determined by the faculty advisor, rather than strict adherence to project deliverables. All students received a passing grade. Project proposals submitted by students were wide-ranging, including AI applications such as "Smartphone Detection of Anterior Uveitis," "Predicting Postpartum Hemorrhage," "Image Enhancement in Video Laryngoscopy," and "Audiometry for Pediatric Heart Murmur Screening." Four (25%) students indicated that they intended to continue working on their projects beyond the end of the elective.

Qualitative Survey Results

Qualitative feedback collected from students before October 2021 (n=4) indicated that students wanted more support and guidance in their field of interest; given this feedback, the authors created more structure for the elective and encouraged students to find an additional specialty-specific mentor who could contribute domain expertise.

Students were asked if they gained what they hoped for from their elective experience. Students who sought a basic conceptual understanding reported satisfaction, but some reported an unmet desire for a deeper technical understanding:

- "I wanted to learn more generally how AI/ML can be used and is being used in medicine. I definitely achieved this goal."
- "I feel that I learned AI/ML fundamentals, am now able to better read and understand AI/ML medical literature, and have thought through the essential design elements of an AI/ML proposal."
- "I learned about the clinical applications of ML and how it is used to help rather than replace radiologists. I also have learned that the technology is advanced, but the application is still early in medicine."
- "I found the course very valuable as an introduction to what ML is and how it is used. However, I had hoped to gain more insight into what research is being conducted in ML from a technical perspective and what these advances may mean from a translational perspective."

Students were also asked what aspects of the course were most beneficial. Four students commented that the self-directed and flexible nature of the course was essential. Two students commented that the project proposal was the most essential element. Five (26%) students reported that they intended to continue working on their projects after the end of the elective month.

When asked for constructive feedback, 2 students commented that they desired more concrete guidance on the projects. Some students felt strained to finish the project proposal within 1 month, with one commenting that students should not expect to finish the proposal in 1 month, and 2 recommending future students pick a project as early as possible, rather than waiting until after the literature review and data set project.

Quantitative Survey Results

After October 2021, students were asked to fill out formal surveys collecting feedback and self-reported confidence in skills gained during the elective. Fifteen students filled out the preintervention survey, and 12 students completed the postintervention survey. These results are shown in Table 1.

Table 1. Pre- and postintervention confidence scores in AI[a] or ML[b] concepts and technical skills.

	Mean (SD)	Median (IQR)
"On a scale of 1-5, how well do you understand AI or ML concepts?"[c]		
Preintervention (n=15)	2.5 (1.3)	2 (3)
Postintervention (n=12)	4.1 (0.7)	4 (3)
"On a scale of 1-5, rate your technical data science skills"[d]		
Preintervention (n=15)	2.6 (1.4)	3 (0.25)
Postintervention (n=12)	1.9 (1.3)	1 (2)

[a]AI: artificial intelligence.

[b]ML: machine learning.

[c]Relative difference is 66% and Wilcoxon rank sum *P* value is .003.

[d]Relative difference is –26% and Wilcoxon rank sum *P* value is .20.

XSL•FO
RenderX

Discussion

Principal Results

Students who participated in this elective were successful in diving deep into the potential of AI and ML in their area of interest and generally reported satisfaction with their elective experience. Students were asked to quantitatively rate their familiarity with both AI and ML concepts and coding or data science; the self-reported confidence scores for AI and ML rose by 66%, and these results were found to be statistically significant when analyzed by the Wilcoxon rank sum test. This exposure to AI and ML is a substantial improvement from the status quo, in which most medical students receive little to no exposure during the course of their training; in 1 study from 2022, 66.5% of students reported 0 hours of AI or ML teaching, and 43.4% had never heard the term "machine learning" [25]. Previous literature includes effective AI curricula developed for other types of health care trainees, such as radiology residents, but there is little to no literature on curricula evaluated for a fourth-year medical student audience as described in this paper [26,27].

Self-reported confidence in technical skills (coding and data science) fell by 26%, although this result was not found to be statistically significant. The authors attribute these results to an initial overconfidence prior to the elective, followed by an increased awareness of the technical complexity of model development after the elective.

Because this was a self-guided elective, student output varied with each student's level of motivation and goals prior to entering the elective. Students who had defined a specific area of interest tended to benefit more from their experience than students who came in with no clear goals set. This course could be improved by providing further assistance early on in helping students to finalize a project area early so that they feel less strained by time toward the end of the month.

Students produced a wide range of deliverables in their chosen specialty. Since most fourth-year students have chosen their specialty and have established connections with faculty in their field, the self-guided nature of the course allowed flexibility for students to seek out appropriate mentors and propose reasonable projects in their areas of interest.

Limitations and Future Directions

Limitations of this study include the small number of participants, especially in the Technical track, restricting the generalizability of this study. Only 2 (11%) students chose the Technical track, so there is insufficient data to evaluate this curriculum; this was likely due to the requirement that students interested in the Technical track have in-depth coding experience and receive approval from the course director to ensure a high likelihood of success. However, the authors recommend screening applicants to make sure that they do in fact possess the required level of comfort in coding before attempting to develop an ML model, as we observed a tendency for students to underestimate the complexity of this task. Based on qualitative observations that students spent more time than expected preparing data for training, the authors suggest providing select, cleaned data sets for students in the Technical track, allowing them to focus on model building, training, and testing.

Another substantial limitation is that assessments relied only on students' self-reported confidence, which has been shown to be a flawed metric [28]. Further studies would benefit from a refined objective assessment tool of students' competencies, as well as replication of this study at other medical schools.

Since launching this fourth-year elective, we have also adapted this curriculum to a shorter elective targeting second-year medical students and were invited to participate in a National Academies forum on AI for Health Profession Education to disseminate this curriculum to other learners [29].

Conclusions

Overall, in the 2 years since launching the elective at Emory University School of Medicine, the authors have already seen substantial excitement and appreciation from senior medical students, with continued excitement in the elective's third year. Most students entered the elective with minimal previous experience in AI and ML and were successful in completing self-guided research and proposing creative and realistic AI and ML projects. The authors are hopeful that a brief, 1-month investment in AI and ML education during medical school can lay the groundwork for these future physicians to continue to engage with AI and ML research and empower this next generation of physicians to pave the way for AI and ML innovation in health care.

Acknowledgments

This study would not have been possible without the support of Emory University School of Medicine. The authors are grateful to Meredith Greer for her guidance in curricular development.

Data Availability

The data sets generated or analyzed during this study are not publicly available due to ensure participant confidentiality and privacy in compliance with the institutional review board exemption status, but are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

AA and JG contributed to the conceptualization, investigation, and methodology; analysis of results; and the writing of the manuscript. AM contributed to the conceptualization and design of the course, along with the review and editing of the manuscript. IB, SP, and HT contributed to the administration of the elective and review and editing of the manuscript.

Conflicts of Interest

JG is a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program and declares support from Radiological Society of North America Health Disparities grant (#EIHD2204), Lacuna Fund (#67), Gordon and Betty Moore Foundation, and National Institutes of Health (National Institute of Biomedical Imaging and Bioengineering) Medical Imaging and Data Resource Center grant (contracts 75N92020C00008 and 75N92020C00021). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Multimedia Appendix 1
Learning objectives and corresponding curated resources.
[DOCX File , 17 KB - mededu_v10i1e46500_app1.docx]

Multimedia Appendix 2
Project components and deliverables.
[DOCX File , 16 KB - mededu_v10i1e46500_app2.docx]

References

1. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86 [FREE Full text] [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
2. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
3. Balthazar P, Tajmir SH, Ortiz DA, Herse CC, Shea LAG, Seals KF, et al. The artificial intelligence journal club (#RADAIJC): a multi-institutional resident-driven web-based educational initiative. Acad Radiol 2020;27(1):136-139 [FREE Full text] [doi: 10.1016/j.acra.2019.10.005] [Medline: 31685386]
4. Staziaki PV, Yi PH, Li MD, Daye D, Kahn CE, Gichoya JW. The radiology: artificial intelligence trainee editorial board: initial experience and future directions. Acad Radiol 2022;29(12):1899-1902 [FREE Full text] [doi: 10.1016/j.acra.2022.04.010] [Medline: 35606258]
5. Perchik JD, Smith AD, Elkassem AA, Park JM, Rothenberg SA, Tanwar M, et al. Artificial intelligence literacy: developing a multi-institutional infrastructure for AI education. Acad Radiol 2023;30(7):1472-1480 [FREE Full text] [doi: 10.1016/j.acra.2022.10.002] [Medline: 36323613]
6. Ramalho AR, Vieira-Marques PM, Magalhães-Alves C, Severo M, Ferreira MA, Falcão-Pires I. Electives in the medical curriculum - an opportunity to achieve students' satisfaction? BMC Med Educ 2020;20(1):449 [FREE Full text] [doi: 10.1186/s12909-020-02269-0] [Medline: 33225951]
7. Walling A, Merando A. The fourth year of medical education: a literature review. Acad Med 2010;85(11):1698-1704 [FREE Full text] [doi: 10.1097/ACM.0b013e3181f52dc6] [Medline: 20881826]
8. Abid A. Artificial intelligence & machine learning in medicine. GitHub. 2023. URL: https://github.com/Emory-HITI/AI-ML-Elective [accessed 2023-02-13]
9. Lungren M, Yeung S. Fundamentals of machine learning for healthcare. Coursera. URL: https://www.coursera.org/learn/fundamental-machine-learning-healthcare [accessed 2024-01-16]
10. Géron A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 3rd Edition. Sebastapol, CA: O'Reilly; 2022.
11. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347-1358 [FREE Full text] [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
12. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126 [FREE Full text] [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
13. 3Blue1Brown. Neural networks video series. YouTube. 2018. URL: https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi [accessed 2024-01-16]
14. Abid A. Intro to machine learning. MedAI. URL: https://med-ai.weebly.com/workshops.html [accessed 2024-01-16]
15. Teachable machine. Google. URL: https://teachablemachine.withgoogle.com/ [accessed 2024-01-16]
16. Jacobsen JH, Geirhos R, Michaelis C. Shortcuts: how neural networks love to cheat. The Gradient. 2020. URL: https://thegradient.pub/shortcuts-neural-networks-love-to-cheat/ [accessed 2024-01-16]
17. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]

XSL•FO
RenderX

18. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med 2020;3:107 [FREE Full text] [doi: 10.1038/s41746-020-00318-y] [Medline: 32885053]
19. Bias in predictive algorithms. Khan Academy. URL: https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101/x2d2f703b37b450a3:machine-learning-and-bias/a/bias-in-predictive-algorithms [accessed 2024-01-16]
20. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4(6):e406-e414 [FREE Full text] [doi: 10.1016/S2589-7500(22)00063-2] [Medline: 35568690]
21. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
22. Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng 2021;14:156-180 [FREE Full text] [doi: 10.1109/RBME.2020.3013489] [Medline: 32746371]
23. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322(18):1806-1816 [FREE Full text] [doi: 10.1001/jama.2019.16489] [Medline: 31714992]
24. Nazha B, Salloum RH, Fahed AC, Nabulsi M. Students' perceptions of peer-organized extra-curricular research course during medical school: a qualitative study. PLoS One 2015;10(3):e0119375 [FREE Full text] [doi: 10.1371/journal.pone.0119375] [Medline: 25764441]
25. Blease C, Kharko A, Bernstein M, Bradley C, Houston M, Walsh I, et al. Machine learning in medical education: a survey of the experiences and opinions of medical students in Ireland. BMJ Health Care Inform 2022;29(1):e100480 [FREE Full text] [doi: 10.1136/bmjhci-2021-100480] [Medline: 35105606]
26. Lindqwister AL, Hassanpour S, Lewis PJ, Sin JM. AI-RADS: an artificial intelligence curriculum for residents. Acad Radiol 2021;28(12):1810-1816 [FREE Full text] [doi: 10.1016/j.acra.2020.09.017] [Medline: 33071185]
27. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. Acad Med 2021;96(11S):S62-S70 [FREE Full text] [doi: 10.1097/ACM.0000000000004291] [Medline: 34348374]
28. Blanch-Hartigan D. Medical students' self-assessment of performance: results from three meta-analyses. Patient Educ Couns 2011;84(1):3-9 [FREE Full text] [doi: 10.1016/j.pec.2010.06.037] [Medline: 20708898]
29. Artificial intelligence in health professions education: a workshop series. National Academies of Sciences. URL: https://www.nationalacademies.org/our-work/maximizing-the-promise-and-mitigating-the-peril-of-artificial-intelligence-in-health-professions-education-a-workshop [accessed 2024-01-16]

Abbreviations

AI: artificial intelligence
ML: machine learning

Edited by T de Azevedo Cardoso; submitted 14.02.23; peer-reviewed by H Cho, B Meskó, E Greene MD Ast Professor USUHS; comments to author 21.09.23; revised version received 07.11.23; accepted 21.12.23; published 20.02.24.

Please cite as:
Abid A, Murugan A, Banerjee I, Purkayastha S, Trivedi H, Gichoya J
AI Education for Fourth-Year Medical Students: Two-Year Experience of a Web-Based, Self-Guided Curriculum and Mixed Methods Study
JMIR Med Educ 2024;10:e46500
URL: https://mededu.jmir.org/2024/1/e46500
doi:10.2196/46500
PMID:38376896

©Areeba Abid, Avinash Murugan, Imon Banerjee, Saptarshi Purkayastha, Hari Trivedi, Judy Gichoya. Originally published in JMIR Medical Education (https://mededu.jmir.org), 20.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard

Faiza Farhat[1], PhD; Beenish Moalla Chaudhry[2], PhD; Mohammad Nadeem[3], PhD; Shahab Saquib Sohail[4], PhD; Dag Øivind Madsen[5], PhD

[1]Department of Zoology, Aligarh Muslim University, Aligarh, India
[2]School of Computing and Informatics, The University of Louisiana, Lafayette, LA, United States
[3]Department of Computer Science, Aligarh Muslim University, Aligarh, India
[4]School of Computing Science and Engineering, VIT Bhopal University, Sehore, India
[5]School of Business, University of South-Eastern Norway, Hønefoss, Norway

Corresponding Author:
Dag Øivind Madsen, PhD
School of Business
University of South-Eastern Norway
Bredalsveien 14
Hønefoss, 3511
Norway
Phone: 47 31008732
Email: dag.oivind.madsen@usn.no

Abstract

Background: Large language models (LLMs) have revolutionized natural language processing with their ability to generate human-like text through extensive training on large data sets. These models, including Generative Pre-trained Transformers (GPT)-3.5 (OpenAI), GPT-4 (OpenAI), and Bard (Google LLC), find applications beyond natural language processing, attracting interest from academia and industry. Students are actively leveraging LLMs to enhance learning experiences and prepare for high-stakes exams, such as the National Eligibility cum Entrance Test (NEET) in India.

Objective: This comparative analysis aims to evaluate the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions.

Methods: In this paper, we evaluated the performance of the 3 mainstream LLMs, namely GPT-3.5, GPT-4, and Google Bard, in answering questions related to the NEET-2023 exam. The questions of the NEET were provided to these artificial intelligence models, and the responses were recorded and compared against the correct answers from the official answer key. Consensus was used to evaluate the performance of all 3 models.

Results: It was evident that GPT-4 passed the entrance test with flying colors (300/700, 42.9%), showcasing exceptional performance. On the other hand, GPT-3.5 managed to meet the qualifying criteria, but with a substantially lower score (145/700, 20.7%). However, Bard (115/700, 16.4%) failed to meet the qualifying criteria and did not pass the test. GPT-4 demonstrated consistent superiority over Bard and GPT-3.5 in all 3 subjects. Specifically, GPT-4 achieved accuracy rates of 73% (29/40) in physics, 44% (16/36) in chemistry, and 51% (50/99) in biology. Conversely, GPT-3.5 attained an accuracy rate of 45% (18/40) in physics, 33% (13/26) in chemistry, and 34% (34/99) in biology. The accuracy consensus metric showed that the matching responses between GPT-4 and Bard, as well as GPT-4 and GPT-3.5, had higher incidences of being correct, at 0.56 and 0.57, respectively, compared to the matching responses between Bard and GPT-3.5, which stood at 0.42. When all 3 models were considered together, their matching responses reached the highest accuracy consensus of 0.59.

Conclusions: The study's findings provide valuable insights into the performance of GPT-3.5, GPT-4, and Bard in answering NEET-2023 questions. GPT-4 emerged as the most accurate model, highlighting its potential for educational applications. Cross-checking responses across models may result in confusion as the compared models (as duos or a trio) tend to agree on only a little over half of the correct responses. Using GPT-4 as one of the compared models will result in higher accuracy consensus. The results underscore the suitability of LLMs for high-stakes exams and their positive impact on education. Additionally, the

XSL•FO
RenderX

study establishes a benchmark for evaluating and enhancing LLMs' performance in educational tasks, promoting responsible and informed use of these models in diverse learning environments.

(JMIR Med Educ 2024;10:e51523) doi:10.2196/51523

KEYWORDS

accuracy; AI model; artificial intelligence; Bard; ChatGPT; educational task; GPT-4; Generative Pre-trained Transformers; large language models; medical education, medical exam; natural language processing; performance; premedical exams; suitability

Introduction

Large language models (LLMs) are potent natural language processing tools, excelling in a range of artificial intelligence (AI) tasks, from news writing to product descriptions. They have garnered widespread attention across academia and industry [1,2], going beyond the scope of natural language processing into tasks related to health care [3], neuroscience [4], philosophy [5], marketing and finance [6,7], sociology [8], education, and others [9,10]. The development of LLMs and chatbots is experiencing an upsurge, with established companies and emerging start-ups actively engaged in their creation [11], catering to general or specific purposes [12]. Prominent examples include Generative Pre-trained Transformers (GPT)-3.5 (OpenAI), GPT-4 (OpenAI), and Bard (Google LLC) [13,14]. Other notable examples are BlenderBot, Galactica, LLaMA (FAIR) [15], Alpaca (Stanford), BloombergGPT [16], Chinchilla (DeepMind), and PaLM [17], heralding the emergence of even more chatbots in the future [12].

The public release of ChatGPT in November 2022 and Bard in March 2023 has garnered significant attention due to their general purpose and flexible nature. ChatGPT [18], built on the GPT-3.5 architecture, has become popular for its remarkable ability to generate coherent and human-like responses. GPT-4.0 represents the latest iteration, incorporating enhanced language generation and improved multiturn conversation handling. Both GPT-3.5 and GPT-4.0 have been specifically trained to interact with users in a conversational manner, maintaining context, handling follow-up questions, and even correcting themselves. Bard, on the other hand, leverages Google's LaMDA [19], enabling it to handle a diverse range of language-related tasks and provide in-depth information.

In educational settings, students are using LLMs such as Bard, GPT-3.5, and GPT-4 to enrich their daily learning experiences [20,21]. They aid students in test preparation, offer research assistance, and contribute to their overall performance improvement and knowledge acquisition [22]. It has been observed that LLMs, despite their impressive performance, can sometimes generate text that includes fabricated or incorrect information [13,23]. Consequently, researchers have directed their attention toward investigating the test-taking capabilities of different LLMs. Numerous research studies have delved into the assessment of GPT-3.5's efficacy in multiple-choice exams in higher education domains [24]. Some investigations have specifically focused on ChatGPT's test-taking performance in diverse professional fields, including business [25], accounting [26], law [27], and medicine [28]. In the medical realm, authors in Bommineni et al [29] examined its competence in tackling the Medical College Admissions Test, which serves as a prerequisite for admission to most medical schools in the United States. In Gilson et al [30] and Kung et al [31], authors have scrutinized ChatGPT's aptitude in the United States Medical Licensing Examination (USMLE), while Teebagy et al [32] conducted a comparative study of GPT-3.5 and GPT-4's performance in the Ophthalmic Knowledge Assessment Program exam. Additionally, Ali et al [33] undertook a comparison of GPT-3.5, GPT-4, and Google Bard, using questions specifically prepared for neurosurgery oral board examinations. Similarly, Zhu et al [28] investigated ChatGPT's performance in several medical topics, namely, the American Heart Association, advanced cardiovascular life support, and basic life support exams.

Despite the successful integration of LLMs in educational environments, a crucial question remains: can LLMs provide the necessary accuracy and reliability required for critical assessments? The published studies predominantly focus on specialized fields within medicine, with few investigations addressing the effectiveness of AI tools for medical school entrance examinations [29]. Additionally, such comparisons made in the literature typically revolve around the performance of a solitary LLM against human abilities [24,34], with limited exploration of how they compare against other LLMs or baseline models, which could provide valuable insights into the strengths and weaknesses of different LLMs. Our primary objective is to bridge this knowledge gap by undertaking a comparative analysis of 3 notable chatbots: GPT-3.5, GPT-4, and Bard, for a standardized medical school exam known as the National Eligibility cum Entrance Test (NEET).

NEET [35] is a competitive entrance exam in India for Bachelor of Medicine and Bachelor of Dental Surgery programs in both government and private colleges. Introduced in 2013 by the Medical Council of India, NEET replaced various state-level and institution-specific tests to standardize medical admissions. Since 2019, the National Testing Agency (NTA) has been responsible for conducting and supervising the NEET. The exam comprises a total of 200 multiple-choice questions aimed at testing knowledge, understanding, and aptitude in 4 subjects: physics, chemistry, botany, and zoology. Candidates can only attempt a maximum of 45 questions per subject, for a total of 180 out of 200 questions. Correct answers are awarded 4 points, while each incorrect response leads to a 1-point deduction. Candidates are allotted 3 hours to complete the examination. To qualify for admission to a medical school, candidates must obtain a minimum or cutoff score, which can change year by year. The cutoff score for NEET-2023 was 137 out of 720. In 2023, over 2.03 million students took the NEET exam [24], a number that has been rising annually by 10% to 16.5%, highlighting the exam's widespread popularity and importance.

XSL•FO
RenderX

Among the 1.15 million candidates who qualified in 2023, only 2 scored full marks (720/720), only 1 scored 716 out of 720, a total of 17 scored 715 out of 720, and 6 scored 711 out of 720 [36]. NEET's rigorous nature, coupled with its widespread adoption, underscores its importance as the primary evaluation tool for determining students' knowledge, aptitude, and readiness for pursuing medical and dental education at the undergraduate level [35].

In this investigation, to evaluate the performance of the 3 mainstream LLMs, namely GPT-3.5, GPT-4, and Google Bard, in answering questions related to the NEET 2023 exam, we used rigorous statistical analyses. We scrutinized each model's performance across 3 pivotal frameworks: overall comparison, subject-level comparison, and topic-level comparison. The outcomes of this study can help premed students make informed decisions about incorporating LLMs into their test preparation strategies. To the best of our knowledge, this marks the first endeavor to undertake such a study.

Methods

Question Set Selection and Preparation

In this paper, we tested the performance of the 3 LLMs on NEET-2023, which was obtained as a portable document file. Although the exam consists of 200 questions, due to the presence of illustrations and diagrams, it was not possible to process all the questions. As a result, we excluded questions with illustrations, resulting in a set of 175 questions for this study. This sample size is large enough to statistically justify each model's performance on the entire exam, with a 95% CI and a 5% margin of error. The selected questions were then manually presented to Bard, GPT-3.5, and GPT-4, and the responses were documented in Excel (Microsoft Corporation).

Data Analysis

We compared responses generated by each model against the correct answers from the official answer key on the NEET website. Based on this comparison, the responses were either marked as correct (1) or incorrect (0).

Prediction Performance

Excel's built-in functionalities were then used to generate the following comparison metrics to assess predictive performance of the LLMs:

1. Accuracy is defined as the percentage of correct responses obtained by a model. In the context of this research, accuracy was obtained using the formula:

   Accuracy = Correct Responses / Total Responses

2. Accuracy consensus is defined as the ratio between correct answers upon which the compared models agree to all the answers (correct and incorrect) upon which the compared models agree. The formula is

   Accuracy consensus = Correct Responses / Total Consensus

Scoring Performance

Next, we calculated the overall, subject-level, and topic-level percentage scores for each LLM following the NTA's scoring rules. Each correct answer was awarded 4 points, while each incorrect answer resulted in a deduction of 1 point. We merged zoology and botany into a single biology category, as the topic-level analysis included questions from both fields. The overall score percentage for each model was determined by dividing the total points scored by the maximum possible points, which was 700. Subject-level percentages were derived by dividing each model's total points by the maximum points available in that subject. Similarly, topic-level percentages were calculated by dividing the total points scored in each topic by the maximum points available for that topic, which varied across different topics.

Results

Prediction Performance

The results demonstrated that GPT-4 had higher accuracy and consensus compared to GPT-3.5 and Bard. It also consistently outperformed the other models across subjects and topics. GPT-3.5 and Bard showed variations in their performances, with specific strengths in certain subjects and topics.

Overall Accuracy

The overall accuracy rates of the models were as follows:

1. GPT-4 achieved the highest accuracy rate of approximately 54.3% by correctly identifying 95 out of 175 responses.
2. GPT-3.5 demonstrated an accuracy of 36.7%, with 64 out of 175 correct responses.
3. Bard achieved the lowest accuracy of approximately 33.1%, based on 58 out of 175 correct answers.

Subject-Level Accuracy

Table 1 presents the number of correct responses obtained by each model in each of the 3 subject areas covered by NEET. It was evident that GPT-4 is consistently more accurate than both Bard and GPT-3.5 in all 3 subjects. For each subject, the number of correct responses obtained by GPT-3.5 and Bard differed by ±3, indicating relatively similar subject-level accuracy rates. On the other hand, GPT-4 was substantially more accurate than the other models, generating 4 to 16 more correct answers per subject. In physics, GPT-4 achieved 73% (29/40) accuracy, followed by GPT-3.5 with 45% (18/40), and Bard with 38% (15/40). Similarly, in chemistry, GPT-4's accuracy rate was 44% (16/36), while GPT-3.5 and Bard achieved an accuracy rate of 33% (12/36). Shifting to biology, GPT-4 maintained its lead with 51% (50/99) accuracy, followed by GPT-3.5 with 34% (34/99), and then Bard with 31% (31/99).

XSL•FO
RenderX

Table 1. Number of correct responses (n) and accuracy rates in each subject per model.

Subject	GPT[a]-4, n (%)	GPT-3.5, n (%)	Bard, n (%)
Biology (n=99)	50 (51)	34 (34)	31 (31)
Chemistry (n=36)	16 (44)	12 (33)	12 (33)
Physics (n=40)	29 (73)	18 (45)	15 (38)

[a]GPT: Generative Pre-trained Transformers.

Topic-Level Accuracy

Table 2 displays the number of correct responses obtained from each model on various topics. GPT-4 was the most accurate in 9 (50%) out of 18 topics. Moreover, for at least half (2-4) of the topics in each subject, GPT-4 demonstrated the highest accuracy. GPT-3.5 was the most accurate (8/15, 53%) in inorganic chemistry. In addition, it was more accurate than Bard in 7 topics across the 3 subjects. However, it had a 0% accuracy in population and ecology (biology) and simple harmonic motion and waves (physics). Bard was the most accurate in the topics on plant kingdom and ecosystem and environment issues. Furthermore, it was more accurate than GPT-3.5 in 5 topics across all 3 subjects. However, it has a 0% accuracy for 2 physics topics, namely modern physics and electronics and optics. GPT-4 and GPT-3.5 had similar accuracies in 1 physics topic (modern physics and electronics: 2/4, 50%) and 2 biology topics (cell biology and genetics: 7/16, 44%; and ecosystem and environmental issues: 2/5, 40%). GPT-4 and Bard are 100% accurate in the topics on simple harmonic motion and waves. All 3 models were at the same level of accuracy in the topics on biomolecules and heat and thermodynamics.

In a nutshell, GPT-4 had a higher accuracy across a wide range of topics (15/18, 83%), while GPT-3.5's and Bard's accuracies were well below GPT-4's. Moreover, they showed variations in their accuracies across topics.

Table 2. Number of correct responses for each topic per model.

Topic	GPT[a]-4, n (%)	GPT-3.5, n (%)	Bard, n (%)
Biotechnology (n=11)	7 (64)[b]	6 (55)	4 (36)
Evolution and health (n=9)	7 (78)[b]	4 (44)	2 (22)
Population and ecology (n=6)	1 (17)[b]	0 (0)	1 (17)[b]
Biomolecules (n=3)	1 (33)	1 (33)	1 (33)
Cell biology and genetics (n=16)	7 (44)[b]	7 (44)[b]	3 (19)
Ecosystem and environmental issues (n=5)	2 (40)	2 (40)	3 (60)[b]
Plant kingdom (n=25)	8 (32)	6 (24)	11 (44)[b]
Animal kingdom (n=24)	17 (71)[b]	8 (33)	6 (25)
Physical chemistry (n=12)	6 (50)[b]	3 (25)	4 (33)
Organic chemistry (n=9)	3 (33)[b]	1 (11)	2 (22)
Inorganic chemistry (n=15)	7 (47)	8 (53)[b]	6 (40)
Mechanics (n=12)	8 (67)[b]	6 (50)	6 (50)
Heat and thermodynamics (n=3)	1 (33)	1 (33)	1 (33)
Electrostatics and electricity (n=11)	10 (91)[b]	5 (45)	6 (55)
Optics (n=3)	3 (100)[b]	2 (67)	0 (0)
Simple harmonic motion and waves (n=1)	1 (100)[b]	0 (0)	1 (100)[b]
Magnetism (n=6)	4 (67)[b]	2 (33)	1 (17)
Modern physics and electronics (n=4)	2 (50)[b]	2 (50)[b]	0 (0)

[a] GPT: Generative Pre-trained Transformers.

[b]Highest accuracy within a topic.

XSL•FO
RenderX

Accuracy Consensus

Overall Accuracy Consensus

The accuracy consensus for the pairs were approximately as follows:

1. Bard and GPT-3.5 were correct on 29 out of 69 matching responses, giving the pair an accuracy consensus of 0.42 and an accuracy of 29 (16.6%) out of 175.
2. Bard and GPT-4 were correct on 42 out of 75 matching responses, resulting in an accuracy consensus of 0.56 and an accuracy of 42 (24%) out of 175.
3. GPT-3.5 and GPT-4 were correct on 45 out of 79 matching responses, giving the pair an accuracy consensus of 0.57 and an accuracy of 45 (25.7%) out of 175.
4. All 3 models were correct on 29 out of 49 matched responses. The accuracy consensus of the trio was approximately 0.59 and an accuracy of 29 (16.6%) out of 175.

This ascending trend in accuracy consensus indicated that GPT-4 enhanced the agreement on correct responses, especially when used in conjunction with either Bard or GPT-3.5. The best accuracy consensus and accuracy were obtained when GPT-3.5 and GPT-4 were considered together. Moreover, the collective intelligence of these models was as good as the weakest duo, that is, Bard and GPT-3.5 combined.

Subject-Level Accuracy Consensus

Table 3 shows the total number of correct matching responses and accuracy consensus at the subject level for each model.

Table 3. Subject-level total correct matching responses and accuracy consensus across compared models.

Subject	GPT[a]-3.5 vs Bard		Bard vs GPT-4		GPT-3.5 vs GPT-4		Bard, GPT-3.5, and GPT-4	
	Total correct matching responses, n	Accuracy consensus	Total correct matching responses, n	Accuracy consensus	Total correct matching responses, n	Accuracy consensus	Total correct matching responses, n	Accuracy consensus
Biology	17	0.4	22	0.46	23	0.48[b]	17	0.52
Chemistry	4	0.31	7	0.50	8	0.50[b]	4	0.50
Physics	8	0.58	13	1.00	14	0.93[b]	8	1.00

[a]GPT: Generative Pre-trained Transformers.

[b]Highest accuracy within a subject.

The subject-level accuracy consensus revealed following insights.

For biology, the highest accuracy consensus was observed between GPT-3.5 and GPT-4 (n=23, ratio of 0.48), indicating GPT-4's superior performance. This duo also produced the highest accuracy, that is, 23 (23%) out of 99. Even though the accuracy consensus of the trio was the highest, it did not correspond to the highest accuracy (17/99, 17%).

For chemistry, both comparisons involving GPT-4 (Bard vs GPT-4 and GPT-3.5 vs GPT-4) yielded a higher accuracy consensus ratio of 0.50. However, the duo of GPT-3.5 and GPT-4 resulted in highest accuracy, that is, 8 (22%) out of 36.

For physics, Bard versus GPT-4 and the collective comparison of all models achieved a perfect accuracy consensus of 1.00 and an accuracy of 13 (32%) out of 40. However, the highest accuracy (14/40, 35%) was shown by GPT-3.5 versus GPT-4, with comparable accuracy consensus of 0.93.

These points demonstrate GPT-4's dominance across subjects, with physics showcasing the highest consensus scores. This suggests that when GPT-4 is used in tandem with any other model, the duo or trio will corroborate each other's responses more than when Bard and GPT-3.5 are considered together.

Topic-Level Accuracy Consensus

Table 4 shows the total number of correct matching responses and accuracy consensus at the topic level for each model.

The following observations can be made about data presented in Table 4.

GPT-3.5 versus GPT-4 demonstrated the highest accuracy consensus and number of correct matching responses in 11 (61%) out of 18 topics. This trend was followed by the Bard versus GPT-4 duo, which showed the highest number of accurate responses and accuracy consensus in 7 (39%) out of 18 topics.

“Biomolecules,” “heat and thermodynamics,” “optics,” and “simple harmonic motion and waves” had low or zero accuracy consensus for all or most comparisons.

Hence, the combined intelligence of the models cannot help with the preparation of all the topics, if the goal is to seek consensus or confirmation of responses across models.

XSL•FO
RenderX

Table 4. Topic-level correct matching responses and accuracy consensus across compared models.

Topic	GPT[a]-3.5 vs Bard		Bard vs GPT-4		GPT-3.5 vs GPT-4		Bard, GPT-3.5, and GPT-4	
	Total correct matching responses, n	Accuracy consensus	Total correct matching responses, n	Accuracy consensus	Total correct matching responses, n	Accuracy consensus	Total correct matching responses, n	Accuracy consensus
Biotechnology	3	0.75	3	0.60	4	0.80[b]	3	0.75
Evolution and health	3	0.75[b]	3	0.50	3	0.75[b]	3	0.75[b]
Population and ecology	2	0.67	2	0.67	3	1.00[b]	2	1.00
Biomolecules	0	N/A[c]	0	N/A	0	N/A	0	N/A
Cell biology and genetics	3	0.30	3	0.43[b]	4	0.36[b]	3	0.43[b]
Ecosystem and environmental issues	1	0.33	2	0.67[b]	1	0.33	1	0.50
Plant kingdom	2	0.22	4	0.31[b]	3	0.30	2	0.33
Animal kingdom	3	0.38	5	0.50[b]	5	0.50[b]	3	0.43
Physical chemistry	2	0.67	3	0.50	4	0.67[b]	2	0.67
Organic chemistry	1	0.50	3	0.75[b]	1	0.33	1	1.00
Inorganic chemistry	1	0.13	1	0.25	3	0.43[b]	1	0.25
Mechanics	2	0.50	3	1.00	4	1.00[b]	2	1.00
Heat and thermodynamics	0	N/A	0	N/A	0	N/A	0	N/A
Electrostatics and electricity	3	0.60	5	1.00	6	1.00[b]	3	1.00
Optics	0	N/A	0	N/A	1	1.00[b]	0	N/A
Simple harmonic motion and waves	0	N/A	0	N/A	0	N/A	0	N/A
Magnetism	2	0.50	2	1.00[b]	2	1.00[b]	2	1.00[b]
Modern physics and electronics	1	1.00	3	1.00[b]	1	1.00	1	1.00

[a]GPT: Generative Pre-trained Transformers.

[b]Highest combination of accurate responses and accuracy consensus in a topic.

[c]N/A: not applicable.

Scoring Performance

Overall Scores

GPT-4 achieved the highest score with 300 (42.9%) out of 700 points, outperforming GPT-3.5, which scored 145 (20.7%) out of 700 points, and Bard, which obtained 115 (16.4%) out of 700 points. To qualify for the NEET-2023 entrance test, candidates needed to secure at least 137 out of 720 points, which represents 19.6% of the total points. It was evident that GPT-4 passed the entrance test with flying colors, showcasing exceptional performance. On the other hand, GPT-3.5 managed to meet the qualifying criteria, but with a substantially lower score. However, Bard failed to meet the qualifying criteria and, hence, did not pass the test.

Subject-Level Scores

The subject-level scores, as per NEET's grading rubric, are detailed in Table 5. GPT-4 achieved the highest overall score of 42.9% (300/700), outperforming both GPT-3.5 (145/700, 20.7%) and Bard (115/700, 16.4%). In all 3 subjects, GPT-4 obtained the highest scores. GPT-3.5 scored higher than Bard in biology and physics but tied with Bard in chemistry.

XSL•FO
RenderX

Table 5. Subject and topic level scores for Bard, Generative Pre-trained Transformers (GPT)-3.5, and GPT-4.

Subject and topic	Scores obtained		
	Bard	GPT-3.5	GPT-4
Overall (n=700), n (%)	115 (16.4%)[a]	145 (20.7%)	300 (42.9%)[b]
Biology (n=396)			
Overall	56[a]	71	151[b]
Animal kingdom	6[a]	16	61[b]
Plant kingdom	30[b]	5[a]	15
Ecosystem and environmental issues	10[b]	5	5
Cell biology and genetics	–1[a]	19[b]	19[b]
Biomolecules	2[b]	2[b]	2[b]
Population and ecology	–1[b]	–6[a]	–1[b]
Evolution and health	1[a]	11	26[b]
Biotechnology	9[a]	19	24[b]
Chemistry (n=160)			
Overall	24	24	44[b]
Inorganic chemistry	15[a]	25[b]	20
Organic chemistry	1	–4[a]	6[b]
Physical chemistry	8	3[a]	18[b]
Physics (n=144)			
Overall	35[a]	50	105[b]
Modern physics and electronics	–4[a]	6[b]	6[b]
Magnetism	–1[a]	4	14[b]
Simple harmonic motion and waves	4[b]	–1[a]	4[b]
Optics	–3[a]	7	12[b]
Electrostatics and electricity	19	14[a]	39[b]
Heat and thermodynamics	2[b]	2[b]	2[b]
Mechanics	18	18	28[b]

[a]Lowest scorer within the topic.

[b]Top scorer within the topic.

We then analyzed the breakdown of the total scores obtained by Bard, GPT-3.5, and GPT-4, categorized by subject. Of the total GPT-4 score, 50.3% (151/300) came from biology, 35% (105/300) came from physics, and 14.7% (44/300) came from chemistry. For GPT-3.5, biology contributed 49% (71/145) of the score, physics contributed 34.5% (50/145), and chemistry contributed 16.6% (24/145). Lastly, Bard's score breakdown showed that 48.7% (56/115) from biology, 30.4% (35/115) came from physics, and 20.9% (24/115) came from chemistry.

These results show that GPT-4 outperformed both GPT-3.5 and Bard in the NEET grading rubric, achieving the highest overall score and the top scores in each individual subject. While GPT-3.5 demonstrated better performance than Bard in biology and physics, it tied with Bard in chemistry. The breakdown of scores by subject revealed that for all 3 models, the largest portion of their scores came from biology (understandably, because there were twice as many questions in this category), followed by physics, and then chemistry, indicating a consistent pattern in their relative strengths across these subjects.

Topic-Level Scores

The results in Table 5 shows that GPT-4 exhibited strong performance across all topics in physics but showed a relative weakness in inorganic chemistry within the chemistry subject. Bard, compared to the GPT versions, excelled specifically in the biology topics of the plant kingdom and ecosystem and

XSL•FO
RenderX

environmental issues. Both GPT models performed equally well in cell biology and genetics (biology) and in modern physics and electronics (physics). Additionally, GPT-3.5 stood out for its excellent performance in inorganic chemistry, highlighting its strength in this area of the chemistry subject.

Discussion

Overview

We evaluated the decision-making performance of 3 models—Bard, GPT, and GPT-4—using accuracy, accuracy consensus, and test scores for the NEET-2023 entrance test. Subject-wise and topic-wise analyses were also conducted. GPT-4 consistently outperformed Bard and GPT across all subjects, achieving the highest accuracy rates: 73% (29/40) in physics, 44% (16/36) in chemistry, and 51% (50/99) in biology. Topic-wise comparisons also demonstrated GPT-4's excellence in 15 (79%) out of 19 topics, with Bard and GPT excelling in certain topics. Particularly, Bard excelled in simple harmonic motion and waves, while GPT showed strength in inorganic chemistry. Overall, GPT-4 emerged as the top performer, excelling in both subjects and specific topics. Our findings are in line with previous studies that have also examined how LLMs perform on exams related to medical education. Bommineni et al [29] found that GPT-3.5 performs at or above the median performance of the Medical College Admissions Test takers. Ali et al [33] reported that GPT-4 outperformed both GPT-3.5 and Bard by achieving the highest score of 82.6% in specialized questions prepared for neurosurgery oral board examinations. Friederichs et al [34] found that GPT-3.5 answered about two-thirds of the multiple-choice questions correctly and outperformed nearly all medical students in years 1-3 of their studies. Gilson et al [30] reported that GPT-3.5's performance on the USMLE was either at or near the minimum passing threshold, even without domain-specific fine-tuning. Below, we present both practical and research implications of our findings to enrich the existing literature.

Implications

Practical Implications

The findings have important implications for users who need to select a model based on specific requirements and their desired score. The subject- and topic-level scores highlight the suitability of different models for different domains. GPT-4 appears to have the highest score (300/700, 42.9%), followed by GPT-3.5 (145/700, 20.7%), and then Bard (115/700, 16.4%). This demonstrates that Bard was not able to pass the NEET-2023 admission exam, and GPT-3.5 was only 2% (14/700) away from the cutoff score, which is 19% (133/700).

Although GPT-4 appears to be the preferred choice for NEET preparation, it is important to note that GPT-4 is a subscription-based service and the pricing model is uniform across the globe, which makes this model less accessible to the general audience in some parts of the world, particularly low-income countries. When cost is an issue, prospective medical school students might consider using GPT-3.5 and Bard in tandem to develop specialized knowledge and expertise in specific subject topics. The accuracy consensus metric demonstrates that the duo was correct on 29 (42%) out of 69 matching responses, reaching 16.6% (29/175) overall accuracy. However, this duo did not excel in any of the subjects, compared to the other duos. Moreover, at the topic level, it only excelled in "evolution and health." These results suggest that, in the absence of GPT-4, while students may consider both GPT-3.5 and Bard together for exam preparation, due to the low level of consensus between these models, the total score would still fall below the cutoff score. Moreover, students would be more often confused about the correct responses while cross-checking answers with these models. Therefore, it is recommended that, for exam preparation, students do not solely rely on these models or model duos; instead, they should consult primary sources in conjunction with these models.

Research Implications

While scoring performance comparisons help us evaluate whether these models are able to ace the NEET-2023 exam or not, prediction performance comparisons help us evaluate their long-term performance beyond NEET 2023. The models' predictive accuracy rates match their scoring performance. GPT-4 demonstrated the highest accuracy rate among the 3 models, indicating its superior capability to provide correct responses and its reliability as an accurate study partner. However, there is still plenty of room for improvement since its accuracy was only at 54.3% (95/175), suggesting that anyone using this model for exam preparation would be exposed to a little over 50% (100/200) of accurate information. GPT-3.5 (64/175, 37.6%) and Bard (58/175, 33.1%) had similar overall accuracy rates that are much lower compared to GPT-4's, suggesting that these 2 models would require significant fine-tuning to qualify as reliable study aids for NEET.

The subject- and topic-level accuracy comparisons highlight specific areas where these models could benefit from domain-specific enhancements. GPT-4 demonstrated superior accuracy across all 3 subjects and 15 topics but required further improvements in 3 topics, that is, ecosystem and environmental issues, plant kingdom, and inorganic chemistry. GPT-4 excelled in at least 1 topic from each subject category, including simple harmonic motion and waves and optics in physics, physical chemistry in chemistry, and evolution and biotechnology in biology. Bard excelled in simple harmonic motion and waves, and GPT-3.5 notably excelled in inorganic chemistry. GPT-3.5, besides requiring improvements in its overall prediction capabilities, needs to develop predictive expertise in population and ecology (biology) and simple harmonic motion and waves (physics). Similarly, Bard needs to develop predictive capabilities in modern physics and electronics and optics, in addition to requiring substantial enhancements in its overall predictive capabilities.

In summary, the implications and applications of this study on LLM and education are far-reaching. First, it could serve as a benchmark for evaluating and improving LLMs' performance in exams and other educational tasks, enhancing the overall effectiveness of these models in educational settings. Second, the use of LLMs as tutors, mentors, or peers has the potential to significantly enhance students' learning outcomes and motivation, particularly in a country such as India with a vast

XSL•FO
RenderX

student population and diverse learning needs. Last, this approach could serve as a platform to explore and address ethical and social concerns related to LLMs in education, such as issues of fairness, bias, privacy, and accountability, ensuring responsible and informed use of these models in educational contexts.

Limitations and Further Research

Similar to any other research, this study has certain limitations that should be considered carefully. It is important to note that this study did not involve direct input from actual students, teachers, or medical school boards to understand their perspectives on these mainstream LLMs' capability to answer questions on basic science concepts. Moreover, we do not know how prospective examinees are using these models for exam preparation or whether they trust them for critical issues such as exam preparation.

LLMs have evolved considerably just in the last 6 months. Therefore, the results of this study will have to be revisited at a later stage. For example, it is possible (and likely) that the relative performance of the different models will change over time. While Bard is currently lagging GPT-3.5 in this area, improvements to the model could mean that it might catch up to GPT-3.5 in the future. Since there is currently an "AI race" among many technology firms, it is only a matter of time before new models are introduced that could perform better on these types of questions.

Conclusion

In this study, we conducted a comparative analysis of 3 notable chatbots, Bard, GPT-3.5, and GPT-4, to evaluate their performance on NEET-2023, a highly competitive medical school entrance examination in India. The study involved the preparation of NEET-2023 questions for the chatbots, data collection, data analysis, and scoring performance assessments.

Our results indicate that GPT-4 not only passed the NEET-2023 entrance test with a score of 42.9% (300/700) but also demonstrated higher accuracy and consensus compared to both GPT-3.5 and Bard. Particularly, GPT-4 consistently outperformed the other models across subjects and topics, achieving an overall accuracy of approximately 54.3% (95/175). GPT-3.5 and Bard, on the other hand, showed variations in their performances, with specific strengths in certain subjects and topics. Regarding subject-wise scoring, GPT-4 excelled in physics and biology while Bard performed well in chemistry.

These findings shed light on the proficiency of LLMs in answering high-stakes examination questions, particularly in the context of medical entrance exams such as the NEET. GPT-4's superior performance and accuracy suggest its potential utility as a valuable resource for medical students seeking assistance in test preparation and knowledge acquisition. However, it is essential to note that despite their impressive performance, LLMs such as Bard, GPT-3.5 and GPT-4 can sometimes generate text containing fabricated or incorrect information. This raises concerns about the credibility of information produced by LLMs, especially in educational settings where accuracy is crucial.

It is also important to acknowledge that LLMs, including GPT, come with both positive and negative consequences [37,38]. Friederichs et al [34] argue that the ability to acquire knowledge is a basic determinant of a physician's performance, and GPT-3.5 should be looked upon as a tool that provides easy access to a lot of relevant information, eventually aiding in clinical decision-making processes. On the other hand, Mbakwe et al [39] have commented that GPT-3.5's success on exams such as the USMLE demonstrates the flaws of medical education, which is "mostly focused on the rote memorization of mechanistic models of health and disease" and does not reward critical thinking to the same extent.

Further research and development are warranted to address the limitations and challenges posed by LLMs and ensure their reliable and accurate use in education and other domains. Moreover, future investigations can explore the suitability of LLMs for addressing the needs of diverse professional fields beyond medical entrance exams.

In conclusion, this study contributes valuable insights into the capabilities of Bard, GPT-3.5, and GPT-4 in handling high-stakes examination questions. As LLMs continue to evolve, their potential to revolutionize education and other industries remains promising, albeit with the need for continuous improvements and validation of their accuracy and reliability.

Data Availability

Data can be obtained through a reasonable request to the corresponding author.

Authors' Contributions

FF and SSS contributed to conceptualization. FF and DØM performed the data acquisition. FF and BMC performed the data analysis. FF, BMC, MN, and SSS contributed to writing and drafting. BMC, MN and DØM contributed to reviewing and proofreading. SSS was the collaborative lead.

Conflicts of Interest

None declared.

References

XSL•FO
RenderX

1. Farhat F, Sohail SS, Madsen D. How trustworthy is ChatGPT? the case of bibliometric analyses. Cogent Eng 2023;10(1):2222988 [FREE Full text] [doi: 10.1080/23311916.2023.2222988]
2. Dwivedi YK, Pandey N, Currie W, Micu A. Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: practices, challenges and research agenda. Int J Contemp Hosp Manag 2023 Jun 07;36(1):1-12 [FREE Full text] [doi: 10.1108/ijchm-05-2023-0686]
3. Sohail SS, Madsen D, Farhat F, Alam MA. ChatGPT and vaccines: can AI chatbots boost awareness and uptake? Ann Biomed Eng 2023:1-5 [FREE Full text] [doi: 10.1007/s10439-023-03305-y] [Medline: 37428336]
4. Liu JL, Zheng J, Cai X, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience.: CellPress; 2023 Aug 09. URL: https://www.cell.com/iscience/fulltext/S2589-0042(23)01667-X [accessed 2024-01-27]
5. Floridi L. AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philos Technol 2023;36:15 [FREE Full text] [doi: 10.1007/s13347-023-00621-y]
6. Beerbaum D. Generative artificial intelligence (GAI) with ChatGPT for accounting: a business case. Social Science Research Network. 2023. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4385651 [accessed 2023-12-12]
7. Rane N. Role and challenges of ChatGPT and similar generative artificial intelligence in human resource management. Social Science Research Network. 2023. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4603230 [accessed 2024-01-24]
8. McGee RW. What are the top 20 questions in sociology? a ChatGPT reply. ResearchGate. 2023. URL: https://www.researchgate.net/profile/Robert-Mcgee-5/publication/369972268_What_Are_the_Top_20_Questions_in_Sociology_A_ChatGPT_Reply/links/643727154e83cd0e2fab3dc1/What-Are-the-Top-20-Questions-in-Sociology-A-ChatGPT-Reply.pdf [accessed 2023-12-13]
9. Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. ArXiv. Preprint posted online on March 7, 2023 [FREE Full text]
10. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. ArXiv. Preprint posted online on November 24, 2023 [FREE Full text]
11. Aljanabi M, Ghazi M, Ali AH, Abed SA, ChatGPT. ChatGPT: open possibilities. Iraqi J Comput Sci Math 2023;4(1):62-64 [FREE Full text] [doi: 10.52866/20ijcsm.2023.01.01.0018]
12. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Sys 2023;3:121-154 [FREE Full text] [doi: 10.1016/j.iotcps.2023.04.003]
13. Google's AI -- Bard. Bard. 2023. URL: https://bard.google.com/chat [accessed 2024-01-27]
14. Borji A, Mohammadian M. Battle of the Wordsmiths: comparing ChatGPT, GPT-4, Claude, and Bard. Social Science Research Network. 2023. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4476855 [accessed 2023-12-13]
15. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. ArXiv. Preprint posted online on February 27, 2023 .
16. Wu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, et al. BloombergGPT: a large language model for finance. ArXiv. Preprint posted online on May 9, 2023 .
17. Chowdhery A, Narang A, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res 2023;24(240):1-113 [FREE Full text]
18. OpenAI's ChatGPT. OpenAI. 2022. URL: https://openai.com/ChatGPT [accessed 2024-01-27]
19. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. LaMDA: language models for dialog applications. ArXiv. Preprint posted online on February 10, 2022 .
20. Tlili A, Shehata B, Adarkwah MA, Bozkurt A, Hickey DT, Huang R, et al. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learn Environ 2023;10(1):15 [FREE Full text] [doi: 10.1186/s40561-023-00237-x]
21. Hong WCH. The impact of ChatGPT on foreign language teaching and learning: opportunities in education and research. J Educ Technol Inov 2023;5(1).
22. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
23. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1-38 [FREE Full text] [doi: 10.1145/3571730]
24. Newton P, Xiromeriti M. ChatGPT performance on MCQ exams in higher education: a pragmatic scoping review. EdArXiv. Preprint posted online on December 13, 2023 [FREE Full text] [doi: 10.35542/osf.io/sytu3]
25. Terwiesch C. Would ChatGPT get a Wharton MBA? a prediction based on its performance in the operations management course. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania. 2023. URL: https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Would-ChatGPT-get-a-Wharton-MBA.pdf [accessed 2023-12-13]

XSL•FO
RenderX

26. Wood DA, Achhpilia MP, Adams MT, Aghazadeh S, Akinyele K, Akpan M, et al. The ChatGPT artificial intelligence chatbot: how well does it answer accounting assessment questions? Issues Account Educ 2023;38(4):81-108 [FREE Full text] [doi: 10.2308/ISSUES-2023-013]
27. Choi JH, Hickman KE, Monahan A, Schwarcz D. ChatGPT goes to law school. SSRN Journal 2023 Jan 25;71(3):1-16 [FREE Full text] [doi: 10.2139/ssrn.4335905]
28. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023;188:109783 [FREE Full text] [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
29. Bommineni, VL, Bhagwagar S, Balcarcel D, Davatzikos C, Boyer D. Performance of ChatGPT on the MCAT: the road to personalized and equitable premedical learning. Medrxiv. 2023 Jun 23. URL: https://www.medrxiv.org/content/10.1101/2023.03.05.23286533v3 [accessed 2024-01-26]
30. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor R, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9(2):e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
31. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
32. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP exam: a comparative study with ChatGPT-3.5. medRxiv. Preprint posted online on April 03, 2023 [FREE Full text]
33. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PLZ, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023:1090-1098 [FREE Full text] [doi: 10.1227/neu.0000000000002551] [Medline: 37306460]
34. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023 Dec;28(1):2220920 [FREE Full text] [doi: 10.1080/10872981.2023.2220920] [Medline: 37307503]
35. Arumugam V, Mamilla R, Anil C. NEET for medics: a guarantee of quality? an exploratory study. Qual Assur Educ 2019 Apr 01;27(2):197-222. [doi: 10.1108/QAE-07-2018-0080]
36. Liu X, Fang C, Wang J. Performance of ChatGPT on clinical medicine entrance examination for Chinese Postgraduate in Chinese. medRxiv. Preprint posted online on April 18, 2023 [FREE Full text]
37. Sohail SS. A promising start and not a panacea: ChatGPT's early impact and potential in medical science and biomedical engineering research. Ann Biomed Eng 2023 Aug 04. [doi: 10.1007/s10439-023-03335-6] [Medline: 37540292]
38. Farhat F. ChatGPT as a complementary mental health resource: a boon or a bane. Ann Biomed Eng 2023 Jul 21. [doi: 10.1007/s10439-023-03326-7] [Medline: 37477707]
39. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb;2(2):e0000205 [FREE Full text] [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]

Abbreviations

AI: artificial intelligence
FP: false positive
GPT: Generative Pre-trained Transformers
LLM: large language model
NEET: National Eligibility cum Entrance Test
NTA: National Testing Agency
TP: true positive
USMLE: United States Medical Licensing Examination

Edited by T Leung, T de Azevedo Cardoso; submitted 02.08.23; peer-reviewed by R Odabashian, M Májovský; comments to author 05.09.23; revised version received 22.09.23; accepted 30.10.23; published 21.02.24.

Please cite as:
Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen DØ
Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard
JMIR Med Educ 2024;10:e51523
URL: https://mededu.jmir.org/2024/1/e51523
doi:10.2196/51523
PMID:38381486

XSL•FO
RenderX

©Faiza Farhat, Beenish Moalla Chaudhry, Mohammad Nadeem, Shahab Saquib Sohail, Dag Øivind Madsen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Original Paper

Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study

Takahiro Nakao[1], MD, PhD; Soichiro Miki[1], MD, PhD; Yuta Nakamura[1], MD, PhD; Tomohiro Kikuchi[1,2], MD, PhD; Yukihiro Nomura[1,3], PhD; Shouhei Hanaoka[4], MD, PhD; Takeharu Yoshikawa[1], MD, PhD; Osamu Abe[4], MD, PhD

[1]Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, Bunkyo-ku, Tokyo, Japan

[2]Department of Radiology, School of Medicine, Jichi Medical University, Shimotsuke, Tochigi, Japan

[3]Center for Frontier Medical Engineering, Chiba University, Inage-ku, Chiba, Japan

[4]Department of Radiology, The University of Tokyo Hospital, Bunkyo-ku, Tokyo, Japan

Corresponding Author:
Takahiro Nakao, MD, PhD
Department of Computational Diagnostic Radiology and Preventive Medicine
The University of Tokyo Hospital
7-3-1 Hongo
Bunkyo-ku, Tokyo, 113-8655
Japan
Phone: 81 358008666
Email: tanakao-tky@umin.ac.jp

Abstract

Background: Previous research applying large language models (LLMs) to medicine was focused on text-based information. Recently, multimodal variants of LLMs acquired the capability of recognizing images.

Objective: We aim to evaluate the image recognition capability of generative pretrained transformer (GPT)-4V, a recent multimodal LLM developed by OpenAI, in the medical field by testing how visual information affects its performance to answer questions in the 117th Japanese National Medical Licensing Examination.

Methods: We focused on 108 questions that had 1 or more images as part of a question and presented GPT-4V with the same questions under two conditions: (1) with both the question text and associated images and (2) with the question text only. We then compared the difference in accuracy between the 2 conditions using the exact McNemar test.

Results: Among the 108 questions with images, GPT-4V's accuracy was 68% (73/108) when presented with images and 72% (78/108) when presented without images (*P*=.36). For the 2 question categories, clinical and general, the accuracies with and those without images were 71% (70/98) versus 78% (76/98; *P*=.21) and 30% (3/10) versus 20% (2/10; *P*≥.99), respectively.

Conclusions: The additional information from the images did not significantly improve the performance of GPT-4V in the Japanese National Medical Licensing Examination.

(JMIR Med Educ 2024;10:e54393) doi:10.2196/54393

KEYWORDS

AI; artificial intelligence; LLM; large language model; language model; language models; ChatGPT; GPT-4; GPT-4V; generative pretrained transformer; image; images; imaging; response; responses; exam; examination; exams; examinations; answer; answers; NLP; natural language processing; chatbot; chatbots; conversational agent; conversational agents; medical education

Introduction

The field of natural language processing is rapidly developing with the advent of large language models (LLMs). LLMs are models trained with massive text data sets and achieve the capability to understand and generate text in natural languages. With the introduction of ChatGPT (OpenAI) [1] and other LLM-based chatbot services, many people have started to benefit from the use of LLMs. Although ChatGPT and its underlying model, generative pretrained transformer (GPT) [2,3], were not specifically developed for medical purposes, they possess a considerable amount of medical knowledge. They have achieved good scores in the United States Medical Licensing Examination [4] and are being explored for various applications for clinical and educational purposes [5-7]. GPT can also understand languages other than English. The latest model, GPT-4, has

XSL•FO
RenderX

been reported to achieve passing scores in medical licensing examinations in non–English speaking countries such as Japan, China, Poland, and Peru [8-13].

Despite these successes, there is still a significant challenge in applying LLMs to real-world problems with non–text-based information. Radiological, pathological, and many other types of visual information play a crucial role in determining a patient's management. Very recently, researchers have proposed multimodal variants of LLMs that can handle not only text but various types of input including images [14]. Providing medical images to multimodal LLMs may realize an even higher accuracy in solving medical-related problems. However, in previous studies on the accuracy rate of medical licensing examinations, questions with images were either not mentioned at all or explicitly excluded from the studies. To the best of our knowledge, no study directly evaluated the performance in solving questions with images. Therefore, in this study, we investigated the image recognition capabilities and limitations of GPT-4V [3,15], one of the most potent publicly available multimodal (vision and language) models, in solving medical questions. We focused on the Japanese National Medical Licensing Examination to examine how the visual information affects GPT-4V's performance.

Methods

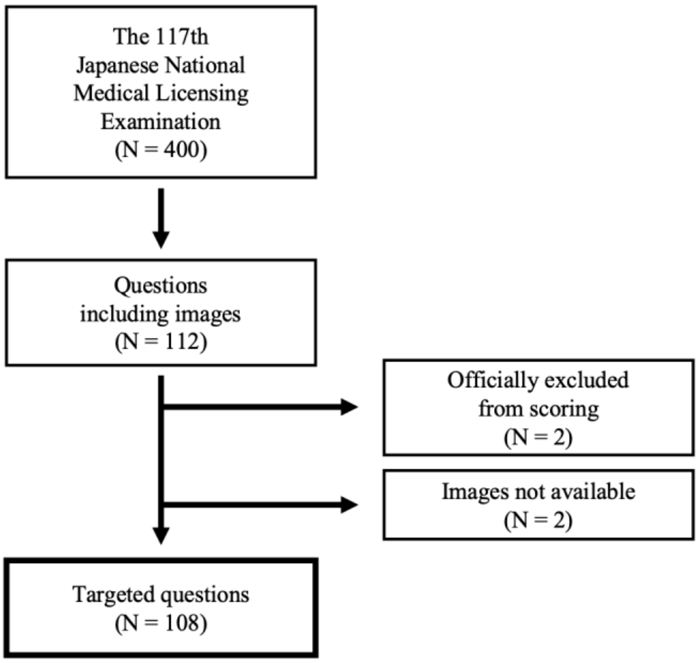
Overview

From the questions of the 117th Japanese National Medical Licensing Examination, held in February 2023, we focused on those that included images as part of a question. Since some of these questions can be answered correctly without interpreting images, we measured the benefit of adding image information by comparing the accuracy rates of ChatGPT under two different conditions: (1) with both the question text and associated images and (2) with the question text only.

Data Set Details

Figure 1 shows the summary of our data set. The questions and correct answers of the 117th Japanese National Medical Licensing Examination are publicly available for download on the official website of the Ministry of Health, Labour and Welfare [16]. All the questions are in a format in which a specified number of choices, typically 1, are to be selected from 5 options. Of the questions that had images, 2 were officially excluded from scoring because they were either too difficult or inappropriate. Additionally, for 2 questions, images of female genitals were not made public on the aforementioned website. These 4 questions were excluded from our study.

Figure 1. Summary of the questions included in this study.



The questions in the Japanese National Medical Licensing Examination were divided into 2 categories: clinical questions and general questions. In clinical questions, clinical information about a specific case is first presented, such as medical history and test results, and answers to questions about the case are required. General questions are about basic medical knowledge, and one is required to choose the correct answer among options for a short question text (typically of 1 or 2 sentences) with an image.

Some clinical questions consisted of multiple subquestions, in which case the background common to all the subquestions was first described, followed by the subquestions. In such cases, each subquestion was individually included in the following analysis if either the subquestion itself or the background part contained an image.

As a result, counting subquestions individually, out of 400 questions, we collected 108 questions that had images, such as photographs of lesions, radiographic images, histopathological images, electrocardiograms, and graphs representing statistical data. Among them, 98 were clinical questions and 10 were general questions.

XSL•FO
RenderX

Experimental Details

We used ChatGPT (September 25, 2023, version) enabled with GPT-4V, which is a multimodal model capable of processing both text and images. This version of ChatGPT asserts it was trained with information up to January 2022, meaning that it had no direct prior knowledge about our target examination. All the question statements and images were manually entered through ChatGPT's web interface. One of the authors, TN, who has 10 years of experience as a medical doctor, reviewed the outputs to interpret the response output by ChatGPT.

A new chat session was created for each question and each condition (ie, with or without images). For questions that comprised multiple subquestions, the background information part and each subquestion were entered into ChatGPT in this order within the same chat session. Subquestions without images were also input to provide ChatGPT with enough context, but they were excluded from the accuracy calculations and the subsequent statistical analysis described below.

The questions were presented to ChatGPT without any preceding or custom instructions. Sometimes, ChatGPT did not respond with the specified number of choices, in which case an additional instruction, such as "select only one option" or "select two options," was provided in Japanese. This additional instruction produced the correct number of options for all the questions.

Statistical Analysis

The difference in ChatGPT's performance between the 2 conditions (ie, with or without images) was analyzed using the exact McNemar test. A *P* value of less than .05 was considered statistically significant. The analysis was conducted using R (version 4.3.1; R Foundation for Statistical Computing).

Ethical Considerations

This study was conducted solely using publicly available resources, therefore, approval from the institutional review board of our institution was not required.

Results

Table 1 shows the results of our experiment. ChatGPT correctly answered 68% (73/108) of image-based questions when provided with both the question text and images, whereas it correctly answered 72% (78/108) of image-based questions when only the question text was provided. There was no significant difference in accuracy between these 2 conditions (*P*=.36). For the clinical questions, the accuracies when presented with and without images were 71% (70/98) and 78% (76/98), respectively. For the general questions, the accuracies were 30% (3/10) when presented with images and 20% (2/10) without images. We have included examples of the input and output along with their English translations in Multimedia Appendix 1, and we have also provided a summary of image interpretation for each question where the results differed depending on the presence of image input (N=7+12) in Multimedia Appendix 2.

Table 1. Performance of ChatGPT in answering questions from the 117th Japanese National Medical Licensing Examination, when presented with or without associated images for each question.

	With images		
	Correct	Incorrect	Total
Overall (*P*=.36)			
Without images, n (%)			
Correct	66 (61)	12 (11)	78 (72)
Incorrect	7 (6)	23 (21)	30 (28)
Total	73 (68)	35 (32)	108 (100)
Clinical (*P*=.21)			
Without images , n (%)			
Correct	65 (66)	11 (11)	76 (78)
Incorrect	5 (5)	17 (17)	22 (22)
Total	70 (71)	28 (29)	98 (100)
General (*P*≥.99)			
Without images, n (%)			
Correct	1 (10)	1 (10)	2 (20)
Incorrect	2 (20)	6 (60)	8 (80)
Total	3 (30)	7 (70)	10 (100)

XSL•FO
RenderX

Discussion

Principal Results

In this study, we examined the image recognition capabilities of GPT-4V using questions associated with images from the Japanese National Medical Licensing Examination. To the best of our knowledge, this is the first study in which the capability of multimodal LLM for the Japanese National Medical Licensing Examination was investigated. Contrary to our initial expectations, the inclusion of image information did not result in any improvement in accuracy. Instead, we even observed a slight decrease, albeit not significant. This indicates that, at the moment, GPT-4V cannot effectively interpret images related to medicine. The passing score rate for the 117th Japanese National Medical Licensing Examination is approximately 75% (and 80% for some questions marked as "essential") [16]. In this study, GPT-4V failed to reach this passing score rate for the questions it was tested on. Considering that 92% of human candidates passed, it implies that the image interpretation skills of GPT-4V will fall short of those possessed by many medical students.

For the clinical questions, in which sufficient clinical information including patient history was available in the text form, GPT-4V was able to choose the correct answers solely from the textual information in the majority (76/98, 78%) of questions, but the addition of images did not improve the accuracy. On the other hand, for the general questions, there was little information in the question text, and GPT-4V had to determine the correct answer by interpreting the images. For these, GPT-4V yielded an accuracy rate that was hardly any better than random guessing even when presented with images. Our results suggest that, for both categories of questions, GPT-4V failed to use visual information to improve its accuracy. We observed that GPT-4V often either explicitly stated that it was unable to interpret the images or failed to provide information beyond what was evident from the question text. In our retrospective review, even in questions where GPT-4V gave correct answers only when presented with images, there were only 2 out of 7 questions where it provided a correct interpretation of the image and used that as a critical clue. Conversely, in questions where GPT-4V provided incorrect answers only when presented with images, it sometimes made incorrect or insufficient interpretations of the images, leading to incorrect answers (4 out of 12).

ChatGPT may serve as a valuable teaching assistant in medical education; however, the inaccuracies in its responses are a significant concern [5,7]. Our current findings suggest that, especially with medical-related images, GPT-4V should not be relied upon as a primary source of information for medical education or practice. If used, extreme caution should be exercised regarding the accuracy of its responses. OpenAI officially states [15] that they "do not consider the current version of GPT-4V to be fit for performing any medical function or substituting professional medical advice, diagnosis, or treatment, or judgment" due to its imperfect performance in the medical domain. Yang et al [17] have comprehensively examined the capabilities of GPT-4V in various tasks including medical image understanding and radiology report generation, and they stated that GPT-4V could correctly diagnose some medical images. However, as they acknowledge, their results contained a considerable number of errors, such as overlooking obvious lesions and errors in laterality. According to the case studies by Wu et al [18], GPT-4V could recognize the modality and anatomy of medical images, but it could hardly make accurate diagnoses and its prediction relied heavily on the patient's medical history. The results of our experiment supported these previous reports.

Considering the well-known high performance of GPT-4V in more generic image recognition tasks [3,17], the most probable reason for its limited image recognition performance in the medical field is that it may simply not have been trained with a sufficient number of medical-related images. LLMs are trained with a vast data set available on the internet, but medical images are not as readily accessible, partly due to privacy concerns. Some researchers are now working on developing multimodal LLMs specialized for medicine based on open-source LLMs [19,20]. These models use publicly available data sets that combine medical images and text, including MIMIC-CXR [21], which contains chest x-ray images with their associated reports, and PMC-OA [22], a compilation of the figures and captions from open-access medical journal papers. The rise of multimodal LLMs is expected to stimulate the publication of more such data sets, thereby advancing the development of multimodal LLMs in the medical field. Moreover, although there are limited medical-related images publicly available on the internet, hospitals have a vast amount of image data. A large part of this is accompanied by textual interpretations in the form of reports or medical records, which may serve as an ideal data set for training multimodal LLMs. In highly specialized domains such as medicine, there remains a significant value in developing domain-specific models using such medical data sets.

Limitations

This study had several limitations. First, ChatGPT was not provided any prior instructions and was directly presented with only the questions themselves. This might have negatively affected its capability to interpret images as the capabilities of LLMs are known to be affected by such "prompt engineering." This will be a subject for future investigation. Second, this study specifically targeted the Japanese National Medical Licensing Examination, and thus, further analysis is necessary to determine whether its conclusions can be generalized to questions in other languages or of different types. However, as mentioned earlier, the limited capability of GPT-4V to interpret medical images has also been demonstrated in other studies focusing on English [17,18], and our results are consistent with those findings. Since ChatGPT's proficiency in non-English interpretation is known to be inferior to that in English interpretation, translating the question text into English before inputting it to ChatGPT might have improved the model's image interpretation capability. However, in a previous study by Yanagita et al [10], in which nonimage questions from the Japanese National Medical Licensing Examination were the target, satisfactory results were achieved even when the questions were input in Japanese. Thus, we adopted the same approach in our study. Third, although our results were based on the same version of ChatGPT and the

XSL•FO
RenderX

same question was evaluated with and without images on the same day, we cannot exclude the possibility that different models were used internally. Lastly, only a single evaluation was conducted for each condition and question. ChatGPT's outputs have some randomness, and responses may differ across multiple evaluations. With ChatGPT's application programming interface, users can programmatically control the degree of randomness by specifying a parameter called *temperature* and obtain mostly deterministic responses. However, during the time of this study, the application programming interface for GPT-4V was not available.

Conclusions

At present, GPT-4V's capability to interpret medical images may be insufficient. In highly specialized fields such as medicine, it is considered meaningful to develop field-specific multimodal models.

Acknowledgments

The Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, is sponsored by HIMEDIC Inc and Siemens Healthcare K.K.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of inputs and outputs from GPT-4V.
[PDF File (Adobe PDF File), 997 KB - mededu_v10i1e54393_app1.pdf]

Multimedia Appendix 2
Summary of image interpretation by GPT-4V.
[DOC File , 50 KB - mededu_v10i1e54393_app2.doc]

References

1. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-10-23]
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. ArXiv Preprint posted online July 22, 2020. [FREE Full text] [doi: 10.48550/arXiv.2005.14165]
3. OpenAI. GPT-4 technical report. ArXiv. Preprint posted online December 19, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]
4. Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. How large language models perform on the United States medical licensing examination: a systematic review. medRxiv Preprint posted online September 07, 2023. [FREE Full text] [doi: 10.1101/2023.09.03.23294842]
5. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2023 (forthcoming) [FREE Full text] [doi: 10.1002/ase.2270] [Medline: 36916887]
6. Cooper A, Rodman A. AI and medical education - a 21st-century Pandora's Box. N Engl J Med 2023;389(5):385-387. [doi: 10.1056/NEJMp2304993] [Medline: 37522417]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
8. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health 2024 Jan;3(1):e0000433 [FREE Full text] [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]
9. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
10. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. JMIR Form Res 2023;7:e48023 [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
11. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-english national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2023;2(12):e0000397 [FREE Full text] [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
12. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Torres BGQ, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ 2023;9:e48039 [FREE Full text] [doi: 10.2196/48039] [Medline: 37768724]

XSL•FO
RenderX

13. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
14. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. ArXiv Preprint posted online February 16, 2024. [FREE Full text] [doi: 10.48550/arXiv.2304.00685]
15. GPT-4V(ision) System Card. OpenAI. 2023 Sep 25. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf [accessed 2023-10-23]
16. 第１１７回医師国家試験問題および正答について [The 117th national medical licensing examination questions and correct answers]. Ministry of Health, Labour and Welfare. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230502-01.html [accessed 2023-10-24]
17. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). ArXiv Preprint posted online October 11, 2023. [FREE Full text] [doi: 10.48550/arXiv.2309.17421]
18. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, et al. Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. ArXiv Preprint posted online December 04, 2023. [FREE Full text] [doi: 10.48550/arXiv.2310.09909]
19. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-flamingo: a multimodal medical few-shot learner. ArXiv Preprint posted online July 27, 2023. [FREE Full text] [doi: 10.48550/arXiv.2307.15189]
20. Xu S, Yang L, Kelly C, Sieniek M, Kohlberger T, Ma M, et al. ELIXR: Towards a general purpose X-Ray artificial intelligence system through alignment of large language models and radiology vision encoders. ArXiv Preprint posted online September 07, 2023. [FREE Full text] [doi: 10.48550/arXiv.2308.01317]
21. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317 [FREE Full text] [doi: 10.1038/s41597-019-0322-0] [Medline: 31831740]
22. Lin W, Zhao Z, Zhang X, Wu C, Zhang Y, Wang Y, et al. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. ArXiv Preprint posted online March 13, 2023. [FREE Full text] [doi: 10.48550/arXiv.2303.07240]

Abbreviations

GPT: generative pretrained transformer
LLM: large language model

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 08.11.23; peer-reviewed by D Hu, M Chatzimina; comments to author 07.12.23; revised version received 26.12.23; accepted 16.02.24; published 12.03.24.

Please cite as:
Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, Yoshikawa T, Abe O
Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study
JMIR Med Educ 2024;10:e54393
URL: https://mededu.jmir.org/2024/1/e54393
doi:10.2196/54393
PMID:38470459

©Takahiro Nakao, Soichiro Miki, Yuta Nakamura, Tomohiro Kikuchi, Yukihiro Nomura, Shouhei Hanaoka, Takeharu Yoshikawa, Osamu Abe. Originally published in JMIR Medical Education (https://mededu.jmir.org), 12.03.2024. 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.

XSL•FO
RenderX

Original Paper

Performance of GPT-4V in Answering the Japanese Otolaryngology Board Certification Examination Questions: Evaluation Study

Masao Noda[1,2], MD, MBA, PhD; Takayoshi Ueno[2], MD, PhD; Ryota Koshu[1], MD; Yuji Takaso[1,2], MD, PhD; Mari Dias Shimada[1], MD; Chizu Saito[1], MD; Hisashi Sugimoto[2], MD, PhD; Hiroaki Fushiki[3], MD, PhD; Makoto Ito[1], MD, PhD; Akihiro Nomura[4], MD, PhD; Tomokazu Yoshizaki[2], MD, PhD

[1]Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, Shimotsuke, Japan
[2]Department of Otolaryngology and Head and Neck Surgery, Kanazawa University, Kanazawa, Japan
[3]Department of Otolaryngology, Mejiro University Ear Institute Clinic, Saitama, Japan
[4]College of Transdisciplinary Sciences for Innovation, Kanazawa University, Kanazawa, Japan

Corresponding Author:
Masao Noda, MD, MBA, PhD
Department of Otolaryngology and Head and Neck Surgery
Jichi Medical University
Yakushiji 3311-1
Shimotsuke, 329-0498
Japan
Phone: 1 0285442111
Email: doforanabdosuc@gmail.com

Abstract

Background: Artificial intelligence models can learn from medical literature and clinical cases and generate answers that rival human experts. However, challenges remain in the analysis of complex data containing images and diagrams.

Objective: This study aims to assess the answering capabilities and accuracy of ChatGPT-4 Vision (GPT-4V) for a set of 100 questions, including image-based questions, from the 2023 otolaryngology board certification examination.

Methods: Answers to 100 questions from the 2023 otolaryngology board certification examination, including image-based questions, were generated using GPT-4V. The accuracy rate was evaluated using different prompts, and the presence of images, clinical area of the questions, and variations in the answer content were examined.

Results: The accuracy rate for text-only input was, on average, 24.7% but improved to 47.3% with the addition of English translation and prompts (*P*<.001). The average nonresponse rate for text-only input was 46.3%; this decreased to 2.7% with the addition of English translation and prompts (*P*<.001). The accuracy rate was lower for image-based questions than for text-only questions across all types of input, with a relatively high nonresponse rate. General questions and questions from the fields of head and neck allergies and nasal allergies had relatively high accuracy rates, which increased with the addition of translation and prompts. In terms of content, questions related to anatomy had the highest accuracy rate. For all content types, the addition of translation and prompts increased the accuracy rate. As for the performance based on image-based questions, the average of correct answer rate with text-only input was 30.4%, and that with text-plus-image input was 41.3% (*P*=.02).

Conclusions: Examination of artificial intelligence's answering capabilities for the otolaryngology board certification examination improves our understanding of its potential and limitations in this field. Although the improvement was noted with the addition of translation and prompts, the accuracy rate for image-based questions was lower than that for text-based questions, suggesting room for improvement in GPT-4V at this stage. Furthermore, text-plus-image input answers a higher rate in image-based questions. Our findings imply the usefulness and potential of GPT-4V in medicine; however, future consideration of safe use methods is needed.

(JMIR Med Educ 2024;10:e57054) doi:10.2196/57054

XSL•FO
RenderX

KEYWORDS

artificial intelligence; GPT-4v; large language model; otolaryngology; GPT; ChatGPT; LLM; LLMs; language model; language models; head; respiratory; ENT: ear; nose; throat; neck; NLP; natural language processing; image; images; exam; exams; examination; examinations; answer; answers; answering; response; responses

Introduction

Advancements in artificial intelligence (AI) in the field of medicine have led to revolutionary changes in diagnosis, treatment, and education. The evolution of natural language processing technologies has significantly affected medical education and evaluation methods [1,2]. The use of large-scale language models contributes to the optimization of complex problem-solving and learning processes, and the effectiveness of these models has been reported in Japanese medicine [3-5]. These AI models can learn from medical literature and clinical cases and generate answers that rival those of human experts.

We have verified the effectiveness of large-scale language-processing models in medical licensing and otolaryngology board certification examinations [6]. Although a certain level of accuracy has been achieved through prompt engineering, these validations have been primarily limited to text-based information processing, and challenges remain in the analysis of complex medical data containing images and diagrams.

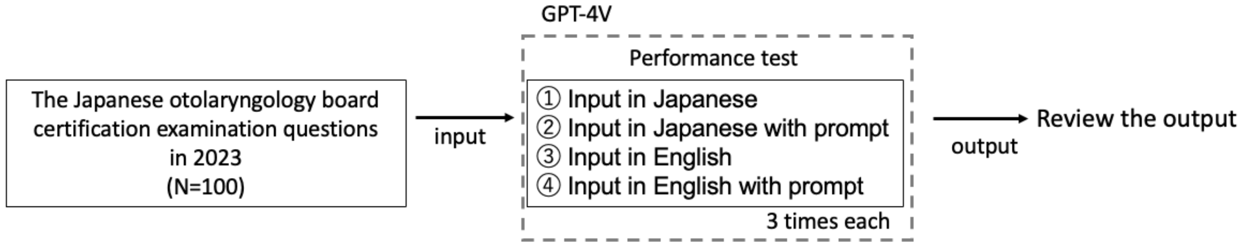
ChatGPT-4 Vision (GPT-4V), announced on September 25, 2023, includes the addition of image input capabilities, potentially expanding its application in the medical field [7]. The current version of the model includes information up to April 2023; it does not encompass the 2023 board examination.

In this study, we aimed to assess the answering capabilities and accuracy of GPT-4V using 100 questions, including image-based questions, from the 2023 otolaryngology board certification examination.

Methods

We evaluated the performance of GPT-4V (Open AI), the latest version of the generative pretrained transformer (GPT) model, using 100 questions from the 2023 otolaryngology specialist examination, which was held on August 5, 2023 (54 text-only and 46 image-based questions; Figure 1).

Figure 1. Study overview. GPT-4V: ChatGPT-4 Vision.



The study design was based on previously reported methods and compared the effectiveness of the following four GPT-4V input approaches: (1) direct input of the question text and images, (2) input of the question text with Japanese prompts added, (3) input of the question text after translation to English, and (4) input of the translated question text with English prompts added [5,6,8] (examples images of prompts for English translation and answering medical questions; Figure S1 in Multimedia Appendix 1).

Each approach was implemented 3 times to evaluate its accuracy. All inputs were entered manually, and both questions and answers were independently scrutinized by otolaryngology specialists (MN and TU) to ensure medical validity [9].

We compiled the correct answer rate and the number of answered and unanswered questions, then conducted an analysis based on the presence of images, the different prompts, the content of the questions, and the associated fields. In addition, the case in which the respondent with no options, and refrained from giving a medical answer was counted as “Output errors.”

Questions were categorized into fields, such as ear; nasal allergy; speech, swallowing, and larynx; oropharynx; head and neck; general; and infectious disease. Question content was classified as treatment, details of the disease and diagnosis, examination, anatomy, systems, and others. Image-based questions were classified as photographs (endoscopic images, microscopic images, and gross photographs), radiological images (computed tomography, magnetic resonance imaging, and positron emission tomography), graphs (audiogram, olfactometry, polysomnography, electronystagmography, etc), and histopathological images.

Finally, to examine the impact of image-based questions on the program’s ability to respond, we compared the responses to text-only questions with those to questions that included figures. We then added an English translation of the text (including the text provided along with figures) and analyzed the difference.

Regarding statistical methods, comparisons among 3 or more groups were performed using 1-way ANOVA. Subsequently, multiple comparison tests (Bonferroni method) were used to compare each group, while comparisons between 2 groups were conducted using the 2-tailed Student *t* test. A significance level of .05 was set for determination.

XSL•FO
RenderX

Results

Performance Evaluation Based on Prompt Type

Input of only the question text resulted in an average correct answer rate of 24.7% (23%, 26%, and 25% in the first, second, and third rounds, respectively). When Japanese prompts were added, the average increased to 36.7% (38%, 33%, and 39%, respectively; *P*=.002); with translation to English, the average rate was 31.3% (33%, 31%, and 30%, respectively; *P*=.06); and with the addition of English translation and English prompts, the average increased to 47.3% (44%, 49%, and 49%, respectively; *P*<.001). The results of all input methods are shown in Table 1.

Table 1. Results of each input method.

Results	Japanese			Japanese with prompt			English			English with prompt		
	Text-only	Image-based	Total	Text-only	Image-based	Total	Text-only	Image-based	Total	Text-only	Image-based	Total
Questions, n	54	46	100	54	46	100	54	46	100	54	46	100
Correct answers, n (%)	23.0 (41.5)	1.7 (3.7)	24.7 (24.7)	25.0 (46.3)	11.7 (25.4)	36.7 (36.7)	23.7 (43.8)	7.7 (16.7)	31.3 (31.3)	28.3 (52.5)	19.0 (41.3)	47.3 (47.3)
Incorrect answers, n	26.0	3.0	29.0	26.0	15.7	41.7	26.3	14.7	41.0	25.3	24.3	50.0
Output errors, n (%)	6.3 (11.4)	40.0 (89.6)	46.3 (46.3)	3.0 (5.6)	18.7 (39.1)	21.7 (21.7)	4.0 (7.4)	23.7 (51.5)	27.7 (27.7)	0.3 (0.6)	2.7 (5.8)	2.7 (2.7)

The nonresponse rate after input of only the question text was, on average, 46.3%. With Japanese prompts, it was 21.7% (*P*<.001). After translation to English, the average was 27.7% (*P*=.002), and with English prompts, it decreased to an average of 2.7% (*P*<.001).

Performance Based on the Presence of Images

There were 46 questions with images, and 54 were text-only. Text-only questions had a higher correct answer rate than that for image-based questions. However, the addition of English translation and prompts significantly increased the correct answer rate, even for questions with images.

The nonresponse rate for image-based questions was higher than that for text-only questions (11.4% vs 89.6%, respectively; Table 1). With Japanese prompts, the nonresponse rates were 5.6% and 39.1%, respectively. With English translation, they were 7.4% and 51.5%, respectively. With the addition of English translation and prompts, they significantly decreased to 0.6% and 5.8%, respectively.

Correct Answer Rates Based on the Question's Field

As shown in Table 2, general questions and those from the fields of head and neck and nasal allergies had relatively high correct answer rates.

XSL•FO
RenderX

Table 2. Results based on the question's field.

Results	Ear	Nasal allergy	Speech, swallowing, and larynx	Oropharynx	Head and neck	General	Infectious disease
Questions, n	29	18	18	11	10	11	3
Japanese							
Correct answers, n (%)	5.0 (17.2)	6.0 (33.3)	1.7 (9.3)	2.0 (18.2)	3.0 (30)	8.0 (72.7)	0.0 (0)
Incorrect answers, n	9.7	7.0	5.0	2.3	2.0	2.0	0.0
Japanese with prompt							
Correct answers, n (%)	7.7 (26.4)	10.3 (57.4)	3.7 (20.4)	3.0 (27.3)	4.3 (43.3)	6.3 (57.6)	1.3 (44.4)
Incorrect answers, n	13.7	4.0	9.7	6.0	3.7	3.3	1.3
English							
Correct answers, n (%)	8.3 (28.7)	9.0 (50)	1.3 (7.4)	4.0 (36.4)	4.7 (46.7)	6.7 (60.6)	0.3 (11.1)
Incorrect answers, n	14.0	1.7	8.7	4.7	3.7	3.3	2.0
English with prompt							
Correct answers, n (%)	9.7 (33.3)	11.3 (63)	7.0 (38.9)	4.7 (42.4)	7.3 (73.3)	6.3 (57.6)	1.0 (33.3)
Incorrect answers, n	18.3	6.0	11.0	6.3	2.7	3.7	2.0

For the fields of head and neck and nasal allergies, respectively, with text-only input, the rates were 72.7%, 30%, and 33.3%, respectively. With Japanese prompts, they were 57.6%, 43.3%, and 57.4%, respectively. With English translation, they were 60.6%, 46.7%, and 50%, respectively. With English translation and prompts, they were 57.6%, 73.3%, and 63%, respectively. Furthermore, in all fields, the correct answer rate improved with the addition of English translation and prompts.

Correct Answer Rates Based on Question Content

As shown in Table 3, questions related to anatomy had the highest correct answer rates: 44.4% for question text only, 55.6% with Japanese prompts, 51.9% with English translation, and 66.7% with English translation and prompts. The correct answer rates for all question content categories improved with the addition of English translation and prompts.

XSL•FO
RenderX

Table 3. Results based on question content.

Results	Treatment	Details of the disease and diagnosis	Examination	Anatomy	Systems	Others
Questions, n	37	32	13	9	7	2
Japanese						
Correct answers, n (%)	6.7 (18)	7.0 (21.9)	3.0 (23.1)	4.0 (44.4)	3.0 (42.9)	2.0 (100)
Incorrect answers, n	13.7	9.3	7.0	2.0	1.0	0.0
Japanese with prompt						
Correct answers, n (%)	12.7 (34.2)	11.0 (34.4)	4.0 (30.8)	5.0 (55.6)	2.0 (28.6)	2.0 (100)
Incorrect answers, n	14.7	15.7	8.0	1.7	1.7	0.0
English						
Correct answers, n (%)	10.0 (27)	10.7 (33.3)	4.0 (30.8)	4.7 (51.9)	3.0 (42.9)	2.0 (100)
Incorrect answers, n	14.7	13.3	7.0	2.0	1.0	0.0
English with prompt						
Correct answers, n (%)	16.7 (45)	14.3 (44.8)	4.7 (35.9)	6.0 (66.7)	3.7 (52.4)	2.0 (100)
Incorrect answers, n	19.7	17.0	8.3	2.7	2.3	0.0

Correct Answer Rates of Image-Based Questions According to the Type of Image

Table 4 shows the results for each type of figure among the 46 image-based questions. There were 23 questions based on photographs, 11 questions based on radiological images, 8 questions based on graphs, and 4 questions based on histopathological images. While the percentage of correct answers for questions based on radiological images was relatively high, this percentage was low for questions based on graphs, such as physiological tests. In the English translation and prompts, the percentage of correct answers for questions based on radiological images was 51.5%, while that for questions based on graphs was 29.2%.

Table 4. Results for image-based questions discriminated according to the type of image.

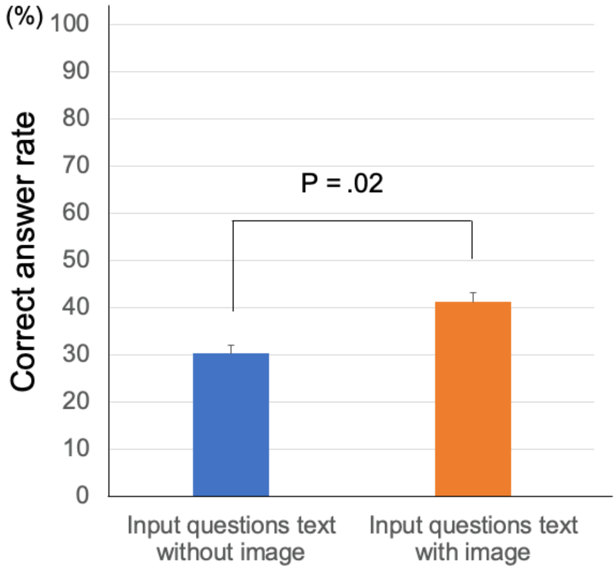
Results	Photograph	Radiological image	Graph	Histopathological image
Questions, n	23	11	8	4
Japanese				
Correct answers, n (%)	0.7 (2.9)	0.0 (0)	1.0 (12.5)	0.0 (0)
Incorrect answers, n	0.0	2.0	1.0	0.0
Japanese with prompt				
Correct answers, n (%)	5.7 (24.6)	3.3 (30.3)	2.0 (25)	0.7 (16.7)
Incorrect answers, n	7.7	4.0	1.3	1.7
English				
Correct answers, n (%)	2.7 (11.6)	2.7 (24.2)	2.0 (25)	0.3 (8.3)
Incorrect answers, n	8.3	4.0	1.7	1.0
English with prompt				
Correct answers, n (%)	9.3 (40.6)	5.7 (51.5)	2.3 (29.2)	1.7 (41.7)
Incorrect answers, n	12.0	5.0	5.0	2.3

Performance Based on Image-Based Questions Text-Only Input Versus Text-Plus-Image Input

Figure 2 shows the performance of GPT-4V based on imaged-based questions with text-only input and with text-plus image input. On image-based questions with text-only input, the average correct answer rate was 30.4%; and with text-plus-image input, the average correct answer rate was 41.3% (*P*=.02; Figure 2).

XSL•FO
RenderX

Figure 2. Performance of ChatGPT-4 Vision on image-based questions.



Discussion

Principal Results

In this study, we evaluated the accuracy of GPT-4V in answering 100 questions, including 46 image-based and 54 text-only questions, from the 2023 otolaryngology board certification examination. The results confirmed that the accuracy was higher for text-only questions than for image-based questions. As for the performance of figure recognition, the correct answer rate with text-plus-image input was higher than that with text-only-input. Moreover, we found that the accuracy improved with the addition of English translations and prompts, but responses were often avoided for simple question inputs, suggesting limitations in medical responses. Variability in accuracy was also evident depending on the field and content of the questions.

Our findings showed that the accuracy of GPT-4V for image-based questions was lower than that for text-only questions. This suggests that, although AI excels at analyzing textual information, it still has limitations in analyzing image-based data [10]. Medical images contain complex and diverse information that requires specialized knowledge for interpretation. Therefore, AI remains inferior to human experts. To improve the accuracy of AI for image analysis, further studies on specialized prompts, the development of more advanced image-recognition technologies, and training focused on medical images are necessary.

Comparison With Prior Work

In relation to medical education, the performance of GPT on licensing examinations and specialist-level medical examinations has been verified and reported [1,11-14]. In English-speaking regions, relatively high accuracy rates have been reported [1,14], whereas in non–English-speaking regions, there is variability [11-13,15]. In addition, accuracy rates differ not only by language but also by the type of examination. Generally, there are more favorable reports for national medical licensing examinations, while there are comparatively poorer reports for specialist-level exams [16,17]. Even when looking at Japanese language reports, while national examinations and general practice examinations have shown good results [3-5,18], ophthalmology, pharmacist, nursing, and dentistry examinations have around a 50%-70% accuracy rate [19-22], with the otolaryngology field in this study showing comparable results [6]. In our previous study, the otolaryngology field tended to have a higher frequency of wrong answers for questions about the ear, larynx, and voice, as well as for questions about examination and treatment. This trend has not changed, suggesting that there are strengths and weaknesses within the specialty. Although the percentage of correct answers was lower for image-based questions than for text-only questions, the percentage of correct answers for text-only questions was higher for general and nasal allergy questions compared with those associated with other question areas, which may have affected the difference in the percentage of correct answers according to the specific field. It is believed that there is room for improvement in GPT's performance, especially in highly specialized fields.

Regarding the effectiveness of prompts for image-based questions, there are reports that the additional input of figures is no different from the input of text only in the Japanese National Medical Practitioners' Examination [23]. On the other hand, in our study, the percentage of correct answers was approximately 10% higher when figures and text were added compared with text-only input. In addition, among the imaged-based questions, the percentage of correct responses was lower for questions related to physiological tests such as hearing tests and polysomnography than for questions related to radiography and microscopy images.

Although there are likely to be differences in the ability to recognize diagrams depending on the field and specialization, it is thought that the search for dedicated prompts, the development of more advanced image recognition techniques, and training specific to medical images will be necessary to further improve the accuracy of image analysis. Converting the physiological tests so that they can be recognized as numerical

XSL•FO
RenderX

values rather than image recognition could further increase the percentage of correct responses.

The fact that accuracy improved with the addition of English translations and prompts suggests AI is optimized for specific formats and languages. The processing capabilities of GPT-4 for text are specialized in English, and the addition of English prompts was believed to increase the likelihood of generating more accurate answers. Our findings further showed that prompts can enhance the quality of AI answers. This effect was valid for image-based as well as text-only questions, emphasizing the need for effective prompts for medical images.

Limitations

The frequent avoidance of generating answers for simple inputs indicates the limitations of AI in terms of complex medical concepts and specialized knowledge. In the medical field, many problems require specific expertise and contexts, making it challenging for AI to provide adequate answers. Furthermore, the issue of hallucinations, where incorrect answers are presented as if they were correct, has become problematic. This includes instances where AI ignores specific facts, engages in illogical reasoning, or fails to apply concepts to new situations [14,24,25]. There is also concern that such inaccuracies could present barriers to direct comprehension by patients, necessitating careful consideration of how AI is used in practice [26].

In addition, the correlation between the difficulty level for specialists and the difficulty level for GPT-4V is not clear, since neither the percentage of correct answers per question nor the minimum number of correct answers required to pass the examination have been reported. Understanding the difference would allow for further consideration of the situations in which the GPT-4V is used. This highlights the importance of understanding these limitations and appropriately using AI in medical education and clinical diagnoses within the otolaryngology field. Though AI suggestions should be considered when making medical judgments, medical professionals need to make the final decisions.

Conclusions

GPT-4V demonstrated a certain level of accuracy for the 2023 otolaryngology board certification examination, and text-plus-image input increased the accuracy of image-based questions. However, the capabilities of AI for image-based questions were limited. Our findings can form the basis for further research and development of the application of AI in the medical field. Future studies should focus on improving the capabilities of AI in image analysis, designing more effective prompts, and developing multilingual support.

Acknowledgments

The authors would like to thank Ryosei Moto for his technical assistance.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Explanations of this study.
[DOCX File , 261 KB - mededu_v10i1e57054_app1.docx]

References

1. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Red Hook, NY, United States: Curran Associates Inc; 2020 Presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; December 6-12, 2020; Vancouver, BC, Canada p. 1877-1901 URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883
3. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
4. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. JMIR Form Res 2023;7:e48023 [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
5. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained transformer on the national medical licensing examination in Japan. PLOS Digit Health 2024;3(1):e0000433 [FREE Full text] [doi: 10.1371/journal.pdig.0000433] [Medline: 38261580]
6. Noda M, Ueno T, Koshu R. A study of the performance of the generative pretrained transformer in the Japanese otorhinolaryngology specialty examination. Nippon Jibiinkoka Tokeibugeka Gakkai Kaiho (Tokyo) 2023;126(11):1217-1223 [FREE Full text] [doi: 10.3950/jibiinkotokeibu.126.11_1217]

XSL•FO
RenderX

7. GPT-4V(ision) system card. OpenAI. 2023. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf [accessed 2024-03-19]
8. Bsharat SM, Myrzakhan A, Shen Z. Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. ArXiv. Preprint posted online on December 26, 2023 [FREE Full text] [doi: 10.48550/arXiv.2312.16171]
9. Answers to multiple-choice questions for the 35rd Specialist Certification Examination. URL: https://www.jstage.jst.go.jp/article/jibiinkotokeibu/126/11/126_1249/_pdf/-char/ja [accessed 2024-03-19]
10. OpenAI. GPT-4 technical report. ArXiv. Preprint posted online on March 4, 2024 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]
11. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform 2023;30(1):e100815 [FREE Full text] [doi: 10.1136/bmjhci-2023-100815] [Medline: 38081765]
12. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian national licensing medical examination: a cross-sectional study. J Educ Eval Health Prof 2023;20:30 [FREE Full text] [doi: 10.3352/jeehp.2023.20.30] [Medline: 37981579]
13. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
14. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024;26:e52113 [FREE Full text] [doi: 10.2196/52113] [Medline: 38261378]
15. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023;13(1):18562 [FREE Full text] [doi: 10.1038/s41598-023-45837-2] [Medline: 37899405]
16. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: observational study. JMIR Med Educ 2024;10:e50842 [FREE Full text] [doi: 10.2196/50842] [Medline: 38236632]
17. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. J Orthop 2024;50:70-75. [doi: 10.1016/j.jor.2023.11.056] [Medline: 38173829]
18. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ 2023;9:e52202 [FREE Full text] [doi: 10.2196/52202] [Medline: 38055323]
19. Sakai D, Maeda T, Ozaki A, Kanda GN, Kurimoto Y, Takahashi M. Performance of ChatGPT in board examinations for specialists in the Japanese ophthalmology society. Cureus 2023;15(12):e49903 [FREE Full text] [doi: 10.7759/cureus.49903] [Medline: 38174202]
20. Ohta K, Ohta S. The performance of GPT-3.5, GPT-4, and Bard on the Japanese national dentist examination: a comparison study. Cureus 2023;15(12):e50369 [FREE Full text] [doi: 10.7759/cureus.50369] [Medline: 38213361]
21. Kunitsu Y. The potential of GPT-4 as a support tool for pharmacists: analytical study using the Japanese national examination for pharmacists. JMIR Med Educ 2023;9:e48452 [FREE Full text] [doi: 10.2196/48452] [Medline: 37837968]
22. Kaneda Y, Takahashi R, Kaneda U, Akashima S, Okita H, Misaki S, et al. Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese nursing examination. Cureus 2023;15(8):e42924 [FREE Full text] [doi: 10.7759/cureus.42924] [Medline: 37667724]
23. Nakao T, Miki S, Nakamra Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in Japanese national medical licensing examination. medRxiv. Preprint posted online on November 8, 2023 [FREE Full text] [doi: 10.1101/2023.11.07.23298133]
24. Knopp MI, Warm EJ, Weber D, Kelleher M, Kinnear B, Schumacher DJ, et al. AI-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds. JMIR Med Educ 2023;9:e50373 [FREE Full text] [doi: 10.2196/50373] [Medline: 38145471]
25. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
26. Razdan S, Siegal AR, Brewer Y, Sljivich M, Valenzuela RJ. Assessing ChatGPT's ability to answer questions pertaining to erectile dysfunction: can our patients trust it? Int J Impot Res 2023. [doi: 10.1038/s41443-023-00797-z] [Medline: 37985815]

Abbreviations

AI: artificial intelligence
GPT: generative pretrained transformer
GPT-4V: ChatGPT-4 Vision

XSL•FO
RenderX

Edited by G Eysenbach; submitted 03.02.24; peer-reviewed by H Yoshimura, Y Kunitsu, S Murono; comments to author 16.02.24; revised version received 22.02.24; accepted 09.03.24; published 28.03.24.

Please cite as:
Noda M, Ueno T, Koshu R, Takaso Y, Shimada MD, Saito C, Sugimoto H, Fushiki H, Ito M, Nomura A, Yoshizaki T
Performance of GPT-4V in Answering the Japanese Otolaryngology Board Certification Examination Questions: Evaluation Study
JMIR Med Educ 2024;10:e57054
URL: https://mededu.jmir.org/2024/1/e57054
doi:10.2196/57054
PMID:38546736

©Masao Noda, Takayoshi Ueno, Ryota Koshu, Yuji Takaso, Mari Dias Shimada, Chizu Saito, Hisashi Sugimoto, Hiroaki Fushiki, Makoto Ito, Akihiro Nomura, Tomokazu Yoshizaki. Originally published in JMIR Medical Education (https://mededu.jmir.org), 28.03.2024. 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.

XSL•FO
RenderX

Viewpoint

Embracing ChatGPT for Medical Education: Exploring Its Impact on Doctors and Medical Students

Yijun Wu[1,2], MD; Yue Zheng[1,2], MD; Baijie Feng[3], MBBS; Yuqi Yang[3], MBBS; Kai Kang[1,2], MD; Ailin Zhao[4], MD

[1]Cancer Center, West China Hospital, Sichuan University, Chengdu, China

[2]Laboratory of Clinical Cell Therapy, West China Hospital, Sichuan University, Chengdu, China

[3]West China School of Medicine, Sichuan University, Chengdu, China

[4]Department of Hematology, West China Hospital, Sichuan University, Chengdu, China

Corresponding Author:
Ailin Zhao, MD
Department of Hematology
West China Hospital
Sichuan University
37 Guoxue Street
Chengdu
China
Phone: 86 17888841669
Email: irenez20@outlook.com

Abstract

ChatGPT (OpenAI), a cutting-edge natural language processing model, holds immense promise for revolutionizing medical education. With its remarkable performance in language-related tasks, ChatGPT offers personalized and efficient learning experiences for medical students and doctors. Through training, it enhances clinical reasoning and decision-making skills, leading to improved case analysis and diagnosis. The model facilitates simulated dialogues, intelligent tutoring, and automated question-answering, enabling the practical application of medical knowledge. However, integrating ChatGPT into medical education raises ethical and legal concerns. Safeguarding patient data and adhering to data protection regulations are critical. Transparent communication with students, physicians, and patients is essential to ensure their understanding of the technology's purpose and implications, as well as the potential risks and benefits. Maintaining a balance between personalized learning and face-to-face interactions is crucial to avoid hindering critical thinking and communication skills. Despite challenges, ChatGPT offers transformative opportunities. Integrating it with problem-based learning, team-based learning, and case-based learning methodologies can further enhance medical education. With proper regulation and supervision, ChatGPT can contribute to a well-rounded learning environment, nurturing skilled and knowledgeable medical professionals ready to tackle health care challenges. By emphasizing ethical considerations and human-centric approaches, ChatGPT's potential can be fully harnessed in medical education, benefiting both students and patients alike.

(JMIR Med Educ 2024;10:e52483) doi:10.2196/52483

KEYWORDS

artificial intelligence; AI; ChatGPT; medical education; doctors; medical students

Introduction

ChatGPT, whose name is derived from "generative pre-trained transformer," is a large natural language processing model grounded in artificial intelligence (AI) technology, demonstrating remarkable performance across various language-related tasks [1]. Within the realm of medical education, ChatGPT emerges as a highly promising tool with considerable potential [2]. Through training in the ChatGPT model, medical students and doctors can enhance their clinical reasoning and decision-making capabilities, consequently leading to improved performance in case analysis and diagnosis. Moreover, ChatGPT offers personalized and efficient learning experiences for medical learners by facilitating simulated dialogues, providing intelligent tutoring, and offering automated question-answering, thereby deepening students' comprehension of medical knowledge [3].

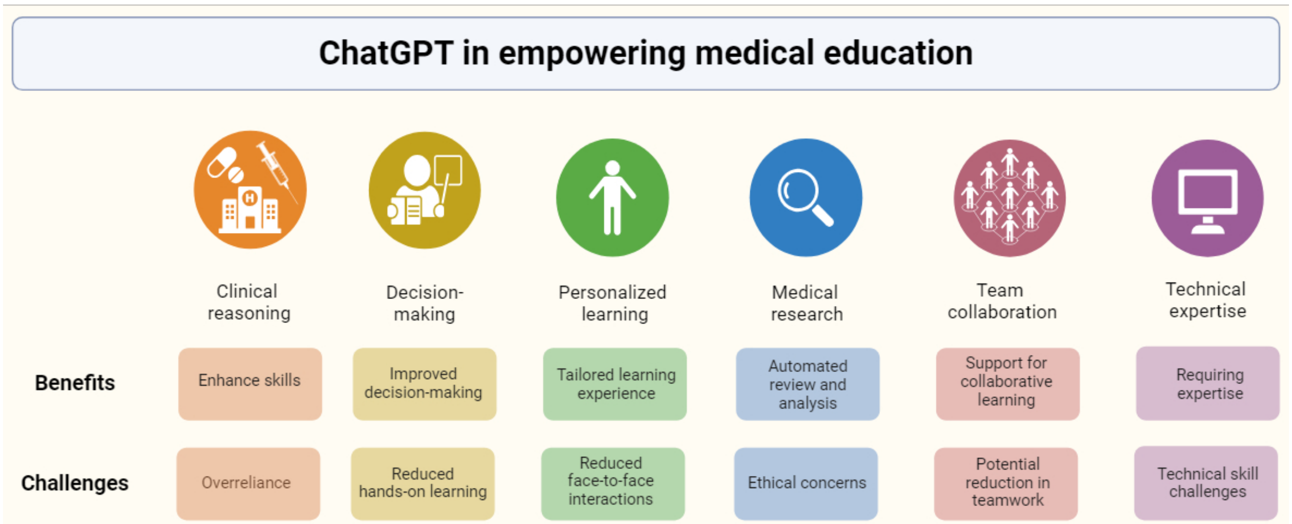
In the realm of transformative technologies in medical education, ChatGPT prominently distinguishes itself, standing out from other large language models by virtue of its unique architecture

XSL•FO
RenderX

and comprehensive training data [4,5]. A pivotal factor setting ChatGPT apart is its monumental scale, boasting an impressive 175 billion parameters. This scale-driven proficiency contrasts starkly with smaller models that may struggle when confronted with complicated queries or tasked with producing coherent replies. With its intricate architectural foundation, ChatGPT possesses the capability to comprehend and generate human-like text across a diverse spectrum of topics, showcasing remarkable coherence and context awareness. What renders the ChatGPT truly distinctive is its specialized focus on fostering dynamic and coherent conversations, thereby excelling in maintaining context over extended interactions. This stands in stark contrast to models primarily designed for single-turn tasks. In educational contexts such as problem-based learning (PBL), team-based learning (TBL), case-based learning (CBL), and precision medical education, ChatGPT takes center stage as a focal point, primarily due to its potential to elevate dynamic learning experiences.

Nevertheless, obstacles occur in the implementation of ChatGPT [6]. On the one hand, the effective training and use of the model requires a high level of technical expertise and skill. On the other hand, concerns related to data security and ethical considerations demand careful attention. To fully harness the potential of ChatGPT in medical education, these challenges must be overcome and concerted efforts should be directed toward integrating AI technology with medical education. By leveraging the capabilities of ChatGPT alongside these innovative teaching approaches, medical education can achieve new heights, fostering a generation of skilled and knowledgeable medical professionals ready to tackle the challenges of the health care field. This paper aims to illuminate both the benefits and the challenges of ChatGPT in medical education (Figure 1).

Figure 1. Benefits and challenges of ChatGPT in medical education.



Potential Benefits of ChatGPT in Medical Education

Overview

In the context of medical education, ChatGPT holds immense promise for bolstering the clinical reasoning and decision-making abilities of medical students and physicians [7]. By training the ChatGPT model, medical learners can tap into its powerful natural language generation and understanding capabilities to master the methods and skills of clinical reasoning and decision-making [8,9]. These competencies are critical components of medical education and fundamental skills that medical students and physicians must possess.

Educational Paradigms: Traditional Vs Enhanced by ChatGPT

Traditional medical education typically follows a teacher-centric approach, where the content and pace of learning are determined mainly by instructors. This often leads to passive student engagement and a lack of personalized education that caters to individual differences. However, the introduction of ChatGPT allows for more personalized and efficient medical education (Table 1). By generating learning materials based on each student's learning status and needs, ChatGPT empowers students to take a more autonomous approach to learning and gain a customized educational experience aligned with their preferences [10,11]. For instance, students can engage in simulated dialogues with ChatGPT, discussing medical cases and diagnostic approaches. Additionally, ChatGPT can adapt based on students' feedback and performance, providing personalized intelligent tutoring and answering questions. This personalized dialogue approach can be tailored to each student's unique needs and interests, thereby enhancing their grasp of medical knowledge and skills.

XSL•FO
RenderX

Table 1. Comparison between traditional medical education and medical education with ChatGPT.

Aspect	Traditional medical education	Medical education with ChatGPT
Clinical reasoning	Instructor-led lectures and traditional case discussions	Enhanced clinical reasoning, personalized dialogues, and simulated case analyses
Decision-making	Limited case exposure	Diverse cases and diagnostic approaches
Personalized learning	One-size-fits-all learning materials and standardized assessments	Tailored learning materials, intelligent tutoring, and automated question-answering based on individual progress
Interaction with educators	Limited face-to-face interactions	Continuous personalized feedback
Medical research support	Manual review and analysis	Automated literature review, study design proposals, and statistical analysis
Team collaboration	Emphasizes group discussions and teamwork	Balancing personalized learning and team-based activities
Technical expertise and challenges	Less reliance on technology	Skill in using ChatGPT
Ethical considerations	Data privacy and consent	Addressing ethical implications

ChatGPT Intelligent Tutoring in PBL Integration

The integration of ChatGPT holds promising implications for PBL in medical education. ChatGPT's capacity to offer personalized guidance and stimulate critical thinking aligns seamlessly with the core principles of PBL [12]. In this context, ChatGPT functions as an intelligent tutor, adept at steering students through intricate problems by furnishing pertinent information, detailed explanations, and insightful suggestions. The model's ability to dynamically adjust responses to student queries contributes to creating a vibrant and responsive learning environment. Students can leverage ChatGPT to brainstorm potential solutions, collect relevant research, or validate hypotheses during the problem-solving process [13]. Furthermore, the model can generate patient cases or clinical scenarios based on real-world data, enabling students to apply their knowledge to practical situations. It is essential to design PBL activities that seamlessly incorporate both the advantages offered by ChatGPT and the indispensable experience derived from clinical practice. By maintaining a focus on group discussions and collaborative problem-solving based on actual patient cases, educators ensure that students reap the benefits of ChatGPT's enhancements while retaining the essential skills cultivated through hands-on clinical interactions and in-depth case analyses. As technology continues to advance, it remains imperative to uphold patient-based learning as the cornerstone of medical education. Recognizing that, at its current stage, ChatGPT cannot entirely replace the critical skills honed through genuine patient interactions and the nuanced analysis of complex cases is vital for preserving the integrity and effectiveness of medical education.

Synergizing ChatGPT With Other Collaborative Teaching Methods

ChatGPT's application in medical education should be complemented by other teaching methods, such as CBL, TBL, and small-group sessions. The model's ability to generate diverse perspectives and solutions enhances the overall TBL experience [14]. In CBL scenarios, ChatGPT can function as a case facilitator, generating realistic scenarios, asking probing questions, and providing nuanced feedback. It can simulate authentic patient interactions or complex business dilemmas, allowing learners to apply theoretical knowledge to practical situations. The model's adaptability ensures that the cases presented are tailored to the evolving needs and understanding of the learners. Within the TBL framework, ChatGPT can facilitate collaboration among team members by offering real-time assistance and promoting knowledge sharing. It can contribute to group discussions, help clarify concepts, and prompt critical thinking among team members. ChatGPT can also facilitate preclass preparation by providing students with foundational knowledge and resources related to the upcoming TBL session. By integrating ChatGPT with these methods, medical educators can create a well-rounded learning experience that maximizes the benefits of both individualized learning and TBL. To enhance team collaboration abilities, medical institutions should prioritize the development of medical students through interprofessional education, where students from different health care disciplines collaborate. Encouraging student-led initiatives and group projects also fosters collaboration, leadership, and effective communication among future medical professionals. This multifaceted approach ensures a well-rounded learning experience, maximizing the benefits of both individualized and collaborative learning while preparing students for the complex challenges of the health care field.

ChatGPT in Precision Medical Education

In the evolving landscape of medical education, the concept of precision medical education has gained prominence [15]. This approach aligns with current trends, notably competency-based medical education (CBME) and pedagogical approaches such as PBL, CBL, and TBL [16]. Precision medical education emphasizes tailoring learning experiences to individual student needs, aligning with the principles of personalized and adaptive learning championed by ChatGPT. CBME focuses on learners progressing at their own pace, demonstrating proficiency in specific competencies. ChatGPT's intelligent tutoring and adaptability make it a valuable tool in supporting this competency-based model. By providing personalized guidance, generating relevant content, and fostering critical thinking, ChatGPT contributes to a more precise and effective medical education tailored to each learner's requirements [17]. Furthermore, the integration of ChatGPT with collaborative teaching methods enhances the multifaceted nature of precision

XSL•FO
RenderX

medical education. In scenarios like CBL and TBL, ChatGPT assists learners in navigating complex medical cases, fostering collaborative problem-solving skills essential for modern health care practice. This approach ensures that students not only acquire essential competencies but also develop the ability to collaborate across health care disciplines, aligning with the interprofessional education framework.

As medical education continues to advance, the incorporation of precision medical education, supported by technologies like ChatGPT, becomes imperative. This tailored approach ensures that medical professionals are equipped with the diverse skills needed to address the complexities of contemporary health care, providing a comprehensive and forward-thinking educational experience.

Empowering Medical Research With ChatGPT

ChatGPT proves to be a valuable asset in medical research [18]. The intricate relationship between medical research and education, as aligned with the standards and roles outlined by the World Federation for Medical Education (WFME) [19] and Canadian Medical Education Direction System (CanMEDS) [20], not only provides a profound and practical foundation for medical education but also aligns with the comprehensive development requirements for medical professionals. This close connection ensures that medical education remains consistent with the latest advancements in medical science, fostering the cultivation of well-rounded medical practitioners. Medical research relies heavily on extensive literature to support its content and conclusions. However, reading and analyzing vast amounts of literature can be time-consuming and labor-intensive. ChatGPT streamlines research by automating literature review and analysis. Additionally, ChatGPT aids medical researchers in study design and data analysis [21]. By expediting data processing, extracting data features and patterns, generating research design proposals, and offering statistical analysis methods and data visualization tools, ChatGPT facilitates improved experiment design and data analysis.

Challenges of ChatGPT in Medical Education

Overview

While ChatGPT offers substantial benefits to medical education, it faces a spectrum of challenges [22]. The rapid pace of knowledge evolution within the medical field presents a significant hurdle. New research and clinical guidelines continually emerge, demanding constant updates to ChatGPT to ensure that students are provided with the most current and accurate medical information. This necessitates not only the ability to keep up with knowledge updates but also to ensure their accuracy and credibility.

Potential Devaluation of Collaboration

A notable concern emerges regarding the potential devaluation of the collaborative aspect of learning in medical education, particularly in traditional methodologies such as PBL, CBL, and TBL. Collaboration and teamwork are pivotal in these approaches [23], and ChatGPT may inadvertently diminish the importance of human-to-human interaction. Maintaining a balance between technology and interpersonal relationships is vital for effective learning. While ChatGPT enhances PBL through personalized guidance, educators must underscore the enduring importance of patient-based learning and teamwork. Despite its simulation capabilities and theoretical insights, ChatGPT cannot replace practical experiences gained through real-world interactions, especially in medical education. Acknowledging the model's limitations is crucial to prevent an overreliance on simulated learning. Embedding ChatGPT seamlessly into existing curricula presents a challenge, requiring educators to invest time in designing and integrating AI-driven components aligned with overall learning goals.

Overreliance

Importantly, overreliance on technology may hinder critical thinking and hands-on learning, potentially lowering the quality of education. ChatGPT's answers can vary or even contradict themselves with each query, further impacting student learning [24]. Learning through ChatGPT might inadvertently reduce face-to-face interactions with educators and peers, impacting effective communication skills in clinical practice. ChatGPT may occasionally disseminate inaccurate medical information, making the prompt recognition and correction of such errors critical [25,26]. The establishment of supervision and feedback mechanisms to enhance ChatGPT's accuracy is imperative.

Challenge of Personalized Learning

The challenge of personalized learning is a crucial consideration. Every student has distinct needs and academic levels, requiring ChatGPT to offer tailored education that aligns with individual requirements and progress. Achieving this may necessitate the development of more sophisticated algorithms and technologies. Cultural diversity and inclusivity should also be addressed. Medical education needs to accommodate students from different cultural backgrounds. ChatGPT should be capable of delivering information and using teaching methods that ensure effective comprehension and benefits for all students.

Ethical Considerations

The ethical and privacy dimensions of using ChatGPT in medical education are paramount [27,28]. Handling patient data in an educational context and safeguarding patient privacy are complex and vital concerns. This entails strict adherence to regulatory and ethical guidelines. Identifying and rectifying errors is another noteworthy challenge.

Technological Accessibility

Technological accessibility poses a challenge. The effective use of ChatGPT depends on network connectivity and device availability, which can be problematic in various regions and among specific student populations [29]. Strategies must be devised to use ChatGPT in diverse technological environments.

XSL•FO
RenderX

Future Directions of ChatGPT in Medical Education

Overview

To mitigate these issues, appropriate regulation and supervision are essential. Students should receive training in interpersonal interactions to engage effectively with patients and efforts should be made to provide equal access to technology and learning resources, promoting fair and inclusive medical education. Moving forward, research in this field should explore various promising avenues to enhance our comprehension and application of ChatGPT.

Strategies to Tackle Present Challenges

To specifically address the challenges of ChatGPT on PBL, TBL, and CBL, measures should be taken to mitigate potential drawbacks on collective capabilities. Introducing targeted interventions, such as incorporating collaborative exercises and feedback mechanisms, can help balance individual contributions within a team setting. Emphasizing the importance of teamwork in medical education [30], alongside the integration of ChatGPT, can foster a collaborative learning environment.

There is a pressing need to investigate methods that can augment ChatGPT's capacity to deliver contextually relevant and up-to-date medical information. This involves developing mechanisms for real-time knowledge updates and refining the curation of medical data. Besides, it is crucial to address the ethical and privacy challenges associated with ChatGPT [31]. Future research can focus on devising robust protocols and AI-driven solutions to protect patient data while seamlessly integrating ChatGPT into medical education. Furthermore, exploring innovative approaches for personalizing medical education with ChatGPT presents an exciting opportunity. Research can delve into adaptive learning algorithms and inventive teaching strategies tailored to individual student needs and learning styles. Additionally, there is a need for research on improving ChatGPT's error identification and correction mechanisms, ensuring the highest level of accuracy and reliability in medical content. Finally, we should examine ways to enhance ChatGPT's cultural sensitivity and inclusivity in medical education and acknowledge the diversity of student backgrounds and learning requirements. This holistic approach ensures that ChatGPT not only provides accurate medical information but also aligns with the broader goals of medical education in promoting collaboration, ethical considerations, and cultural competence.

Conclusions

In conclusion, ChatGPT enhances medical education by improving clinical reasoning, personalizing learning, promoting precision medical education, and supporting medical research. However, a balanced and responsible integration requires a focus on ethics and human-centered approaches. Medical educators can achieve this balance by customizing learning paths, blending personalization with group activities, assigning team projects, guiding ChatGPT use, and emphasizing ethics and critical thinking training. These steps create a holistic learning environment that prepares students to excel as independent thinkers and team players in health care, optimizing ChatGPT's role in medical education while maintaining its integrity.

Acknowledgments

This work was supported by Postdoctoral Fellowship Program of CPSF (GZB20230481), National Natural Science Foundation of China (82303773, 82303772, and 82204490), Natural Science Foundation of Sichuan Province (2023NSFSC1885), Key Research and Development Program of Sichuan Province (2023YFS0306), and the 2024 College Students' Innovative Entrepreneurial Training Plan Program (C2024130171 and C2024128798). We express our thanks to BioRender.com for creating Figure 1.

Authors' Contributions

AZ provided the idea and designed the study. YW, YZ, BF, YY, and KK contributed to the conceptualization, writing original draft, and writing—review and editing. All authors contributed to the paper and approved the submitted version.

Conflicts of Interest

None declared.

References

1. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]
2. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med 2023;98(8):867-868 [FREE Full text] [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
3. 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;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
4. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307(2):e230163 [FREE Full text] [doi: 10.1148/radiol.230163] [Medline: 36700838]

XSL•FO
RenderX

5. Arachchige ASPM. Large Language Models (LLM) and ChatGPT: a medical student perspective. Eur J Nucl Med Mol Imaging 2023;50(8):2248-2249. [doi: 10.1007/s00259-023-06227-y] [Medline: 37046082]
6. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
7. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
8. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
9. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023;9:e48163 [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
10. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 2023;104(6):269-274. [doi: 10.1016/j.diii.2023.02.003] [Medline: 36858933]
11. Wang LKP, Paidisetty PS, Cano AM. The next paradigm shift? ChatGPT, artificial intelligence, and medical education. Med Teach 2023;45(8):925. [doi: 10.1080/0142159X.2023.2198663] [Medline: 37036176]
12. Hamid H, Zulkifli K, Naimat F, Yaacob NLC, Ng KW. Exploratory study on student perception on the use of chat AI in process-driven problem-based learning. Curr Pharm Teach Learn 2023;15(12):1017-1025. [doi: 10.1016/j.cptl.2023.10.001] [Medline: 37923639]
13. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2023:1-6 [FREE Full text] [doi: 10.1002/ase.2270] [Medline: 36916887]
14. Park J. Medical students' patterns of using ChatGPT as a feedback tool and perceptions of ChatGPT in a leadership and communication course in Korea: a cross-sectional study. J Educ Eval Health Prof 2023;20:29 [FREE Full text] [doi: 10.3352/jeehp.2023.20.29] [Medline: 38096895]
15. Triola MM, Burk-Rafel J. Precision medical education. Acad Med 2023;98(7):775-781. [doi: 10.1097/ACM.0000000000005227] [Medline: 37027222]
16. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638-645. [doi: 10.3109/0142159X.2010.501190] [Medline: 20662574]
17. Duong MT, Rauschecker AM, Rudie JD, Chen PH, Cook TS, Bryan RN, et al. Artificial intelligence for precision education in radiology. Br J Radiol 2019;92(1103):20190389 [FREE Full text] [doi: 10.1259/bjr.20190389] [Medline: 31322909]
18. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023;28(1):2181052 [FREE Full text] [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
19. van Niekerk JPDV. WFME global standards receive ringing endorsement. Med Educ 2003;37(7):585-586. [doi: 10.1046/j.1365-2923.2003.01561.x] [Medline: 12834413]
20. Ellaway R. CanMEDS is a theory. Adv Health Sci Educ Theory Pract 2016;21(5):915-917. [doi: 10.1007/s10459-016-9724-3] [Medline: 27878472]
21. Ashraf H, Ashfaq H. The role of ChatGPT in medical research: progress and limitations. Ann Biomed Eng 2023. [doi: 10.1007/s10439-023-03311-0] [Medline: 37452215]
22. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
23. Zhao W, He L, Deng W, Zhu J, Su A, Zhang Y. The effectiveness of the combined problem-based learning (PBL) and case-based learning (CBL) teaching method in the clinical practical teaching of thyroid disease. BMC Med Educ 2020;20(1):381 [FREE Full text] [doi: 10.1186/s12909-020-02306-y] [Medline: 33092583]
24. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023;188:109783 [FREE Full text] [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
25. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
26. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3(4):100324 [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]
27. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023;613(7945):612 [FREE Full text] [doi: 10.1038/d41586-023-00191-1] [Medline: 36694020]
28. Ferreira AL, Lipoff JB. The complex ethics of applying ChatGPT and language model artificial intelligence in dermatology. J Am Acad Dermatol 2023;89(4):e157-e158 [FREE Full text] [doi: 10.1016/j.jaad.2023.05.054] [Medline: 37263382]
29. Wu JTY, Shenoy ES, Carey EP, Alterovitz G, Kim MJ, Branch-Elliman W. ChatGPT: increasing accessibility for natural language processing in healthcare quality measurement. Infect Control Hosp Epidemiol 2024;45(1):9-10. [doi: 10.1017/ice.2023.236] [Medline: 37946379]

XSL•FO
RenderX

30. Weller J, Boyd M, Cumin D. Teams, tribes and patient safety: overcoming barriers to effective teamwork in healthcare. Postgrad Med J 2014;90(1061):149-154. [doi: 10.1136/postgradmedj-2012-131168] [Medline: 24398594]
31. Beltrami EJ, Grant-Kels JM. Consulting ChatGPT: ethical dilemmas in language model artificial intelligence. J Am Acad Dermatol 2023;11:S0190-9622(23)00364-X [FREE Full text] [doi: 10.1016/j.jaad.2023.02.052] [Medline: 36907556]

Abbreviations

AI: artificial intelligence
CanMEDS: Canadian Medical Education Direction System
CBL: case-based learning
CBME: competency-based medical education
PBL: problem-based learning
TBL: team-based learning
WFME: World Federation for Medical Education

Edited by T de Azevedo Cardoso, D Chartash; submitted 05.09.23; peer-reviewed by F Zhang, M Hasnain, L Zhu, B McGowan; comments to author 23.10.23; revised version received 03.11.23; accepted 17.01.24; published 10.04.24.

Please cite as:
Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A
Embracing ChatGPT for Medical Education: Exploring Its Impact on Doctors and Medical Students
JMIR Med Educ 2024;10:e52483
URL: https://mededu.jmir.org/2024/1/e52483
doi:10.2196/52483
PMID:38598263

©Yijun Wu, Yue Zheng, Baijie Feng, Yuqi Yang, Kai Kang, Ailin Zhao. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.04.2024. 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.

XSL•FO
RenderX

Original Paper

Appraisal of ChatGPT's Aptitude for Medical Education: Comparative Analysis With Third-Year Medical Students in a Pulmonology Examination

Hela Cherif[1*]; Chirine Moussa[1*]; Abdel Mouhaymen Missaoui[1*]; Issam Salouage[1], Prof Dr; Salma Mokaddem[1]; Besma Dhahri[1], Prof Dr

Faculté de Médecine de Tunis, Université de Tunis El Manar, Tunis, Tunisia

[*]these authors contributed equally

Corresponding Author:
Hela Cherif
Faculté de Médecine de Tunis
Université de Tunis El Manar
15, Rue Djebel Lakhdhar – Bab Saadoun
Tunis, 1007
Tunisia
Phone: 216 50424534
Email: hela.cherif@fmt.utm.tn

Abstract

Background: The rapid evolution of ChatGPT has generated substantial interest and led to extensive discussions in both public and academic domains, particularly in the context of medical education.

Objective: This study aimed to evaluate ChatGPT's performance in a pulmonology examination through a comparative analysis with that of third-year medical students.

Methods: In this cross-sectional study, we conducted a comparative analysis with 2 distinct groups. The first group comprised 244 third-year medical students who had previously taken our institution's 2020 pulmonology examination, which was conducted in French. The second group involved ChatGPT-3.5 in 2 separate sets of conversations: without contextualization (V1) and with contextualization (V2). In both V1 and V2, ChatGPT received the same set of questions administered to the students.

Results: V1 demonstrated exceptional proficiency in radiology, microbiology, and thoracic surgery, surpassing the majority of medical students in these domains. However, it faced challenges in pathology, pharmacology, and clinical pneumology. In contrast, V2 consistently delivered more accurate responses across various question categories, regardless of the specialization. ChatGPT exhibited suboptimal performance in multiple choice questions compared to medical students. V2 excelled in responding to structured open-ended questions. Both ChatGPT conversations, particularly V2, outperformed students in addressing questions of low and intermediate difficulty. Interestingly, students showcased enhanced proficiency when confronted with highly challenging questions. V1 fell short of passing the examination. Conversely, V2 successfully achieved examination success, outperforming 139 (62.1%) medical students.

Conclusions: While ChatGPT has access to a comprehensive web-based data set, its performance closely mirrors that of an average medical student. Outcomes are influenced by question format, item complexity, and contextual nuances. The model faces challenges in medical contexts requiring information synthesis, advanced analytical aptitude, and clinical judgment, as well as in non-English language assessments and when confronted with data outside mainstream internet sources.

(JMIR Med Educ 2024;10:e52818) doi:10.2196/52818

KEYWORDS

medical education; ChatGPT; GPT; artificial intelligence; natural language processing; NLP; pulmonary medicine; pulmonary; lung; lungs; respiratory; respiration; pneumology; comparative analysis; large language models; LLMs; LLM; language model; generative AI; generative artificial intelligence; generative; exams; exam; examinations; examination

XSL•FO
RenderX

Introduction

Artificial intelligence (AI) has emerged as a transformative force across various aspects of modern life. Within the realm of AI, natural language processing (NLP) has gained significant attention as it involves the use of devices to replicate human cognitive processes, encompassing learning, problem-solving, and practical application [1,2]. An exemplary NLP model is ChatGPT, developed by OpenAI. This model uses deep learning algorithms trained on extensive data sets to generate responses simulating human-like interactions. This versatile dialogic agent holds promise in diverse applications, including customer service and chatbots [3,4].

Launched on November 30, 2022, ChatGPT quickly gained popularity, attracting a million users within its first week and achieving unprecedented growth. In June 2023 alone, the ChatGPT website received 1.66 billion visits, underscoring its widespread appeal and use [5,6].

While this rapid development of ChatGPT has generated both excitement and concern across various fields, the impact on medical education has been particularly intriguing [7]. This chatbot technology may present opportunities to revolutionize medical education, offering enhanced efficiency, interactivity, and realism in training scenarios [8,9]. However, these benefits come with significant challenges and uncertainties that need to be carefully addressed and navigated [10,11].

A paramount examination in the medical school curriculum is the pneumology examination. This pivotal assessment evaluates the comprehensive understanding of respiratory diseases and their management—a core competency for any medical practitioner.

Our study aims to evaluate the performance of ChatGPT in the context of pneumology examinations through a comparative analysis with that of third-year medical students.

Methods

Study Design and Participants

This research adopts a cross-sectional design and was conducted at the pneumology teaching section of the Faculty of Medicine of Tunis (FMT), Tunisia, in June 2023. The study uses a comparative approach, involving 2 distinct groups: ChatGPT and medical students.

The first group comprises 244 third-year medical students registered at the FMT. These students had previously taken the pulmonology examination in January 2020. The second group consists of ChatGPT-3.5, a freely available version of ChatGPT, which undertook the same pneumology examination in June 2023.

Pneumology Examination

Question Selection

The pneumology examination of FMT of 2020 is a 90-minute test comprising 50 questions, written in French. These questions underwent validation within the pneumology section of FMT to cover a diverse range of knowledge levels, including both fundamental and advanced concepts. The examination assesses candidates' competency in various fields of pneumology, such as clinical pneumology, microbiology, respiratory radiology, pharmacology, pathology, and thoracic surgery.

The administered version of the examination involved only 45 text-based questions to align with ChatGPT's processing capabilities. Thus, 5 questions based on visual components (images, graphs, and illustrations) were excluded since ChatGPT lacks the ability to process this material within its conversational scope.

A comprehensive mapping of assessment parameters for the administered pneumology examination is presented in Table 1. It encompasses a total of 9 multiple choice questions (MCQs), 13 short open-ended questions (SOEQs), and 7 clinical scenarios. Among the clinical scenarios, 2 were structured with MCQs, while the remaining 5 were constructed with SOEQs.

XSL•FO
RenderX

Table 1. Assessment parameters and question distribution in pneumology examination.

Mapping of pulmonology examination	Findings
Parameters	
Academic year	2020
Target examinees	Third-year medical students
Timing	90 minutes
Grading scale	0-100
Questions, n	45
Question topics, n (%)	
Clinical pneumology	27 (60)
Radiology	7 (16)
Pharmacology	5 (11)
Pathology	3 (7)
Microbiology	2 (4)
Thoracic surgery	1 (2)
Question formats, n (%)	
Independent MCQs[a]	9 (20)
Independent SOEQs[b]	13 (29)
MCQ-structured clinical cases	7 (16)
SOEQ-structured clinical cases	16 (35)
Distribution by difficulty index, n (%)	
Low difficulty index items	12 (27)
Intermediate difficulty index items	25 (56)
High difficulty index items	8 (18)
Distribution by discrimination index, n (%)	
Low discrimination index items	21 (47)
Intermediate discrimination index items	13 (29)
High discrimination index items	11 (24)

[a]MCQ: multiple choice question.

[b]SOEQ: short answer open-ended question.

Item Performance Indexes

Item performance indexes are crucial statistical measures used to assess the effectiveness and quality of test questions, ensuring the reliability and validity of the assessment. These indexes provide valuable insights into the performance of each item concerning difficulty level, discrimination, and its ability to differentiate between high- and low-performing students. In this study, we used common item performance indexes, including the difficulty index (D1) and the discrimination index (D2) [12,13].

The D1 represents the proportion of students who answered an item correctly, calculated by dividing the number of correct responses by the total number of students attempting the item. While the optimal item difficulty may vary based on the specific test format and intended learning outcomes, a value within the 0.3 to 0.7 range is generally preferred [14,15].

On the other hand, the D2 measures an item's capability to differentiate between high-performing and low-performing students. It is determined by comparing the performance of students who achieved high scores on the overall test with those who scored low on the same test for a particular item. D2 levels are classified as follows: high discrimination (D2>0.7), intermediate discrimination (D2 values between 0.3 and 0.7), and low discrimination (D2<0.3) [14,15].

Data Collection and Score System

The database, containing the results and scores of medical students who took the pneumology examination in 2020, along with corresponding performance indexes, was accessible in the pneumology section and used in our comparative analysis.

Two authors (HC and CM) conducted separate conversations with ChatGPT-3.5. In the first conversation, CM presented questions to the chatbot without contextualization (V1). In the

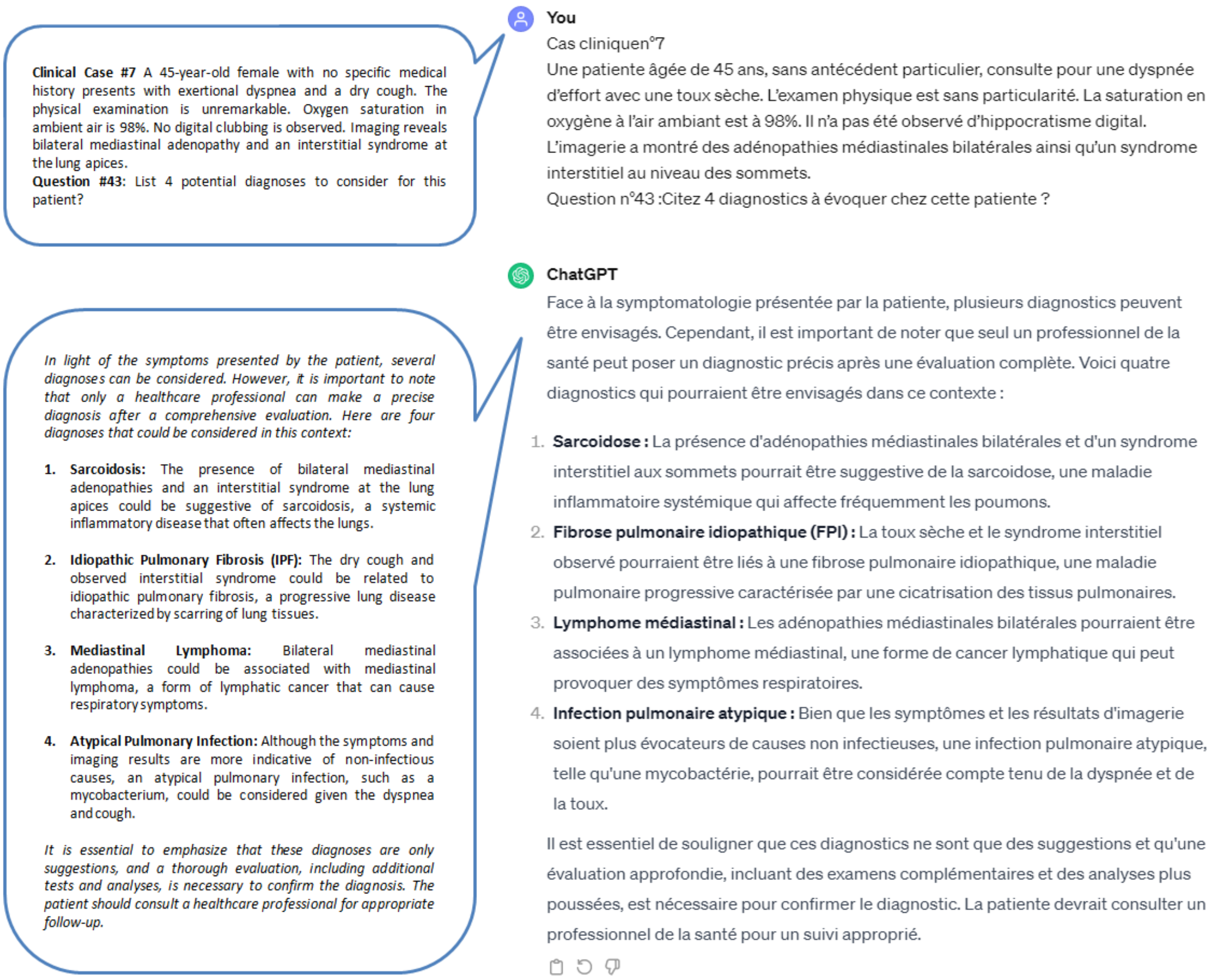
XSL•FO
RenderX

second conversation, conducted by HC, suitable context was provided before posing the questions (V2). The questions were presented in exactly the same order as given to the students. Figures 1 and 2 show illustrations of the dual chat conversations conducted by HC and CM and the respective responses from ChatGPT.

The responses generated by both V1 and V2 were meticulously transcribed and stored in separate files. To ensure objectivity and independence, an impartial pneumology teacher, not involved in this study, conducted the evaluation. This teacher used the same grading scale specifically designed for evaluating student performance in the 2020 examination, ensuring an unbiased and rigorous assessment process.

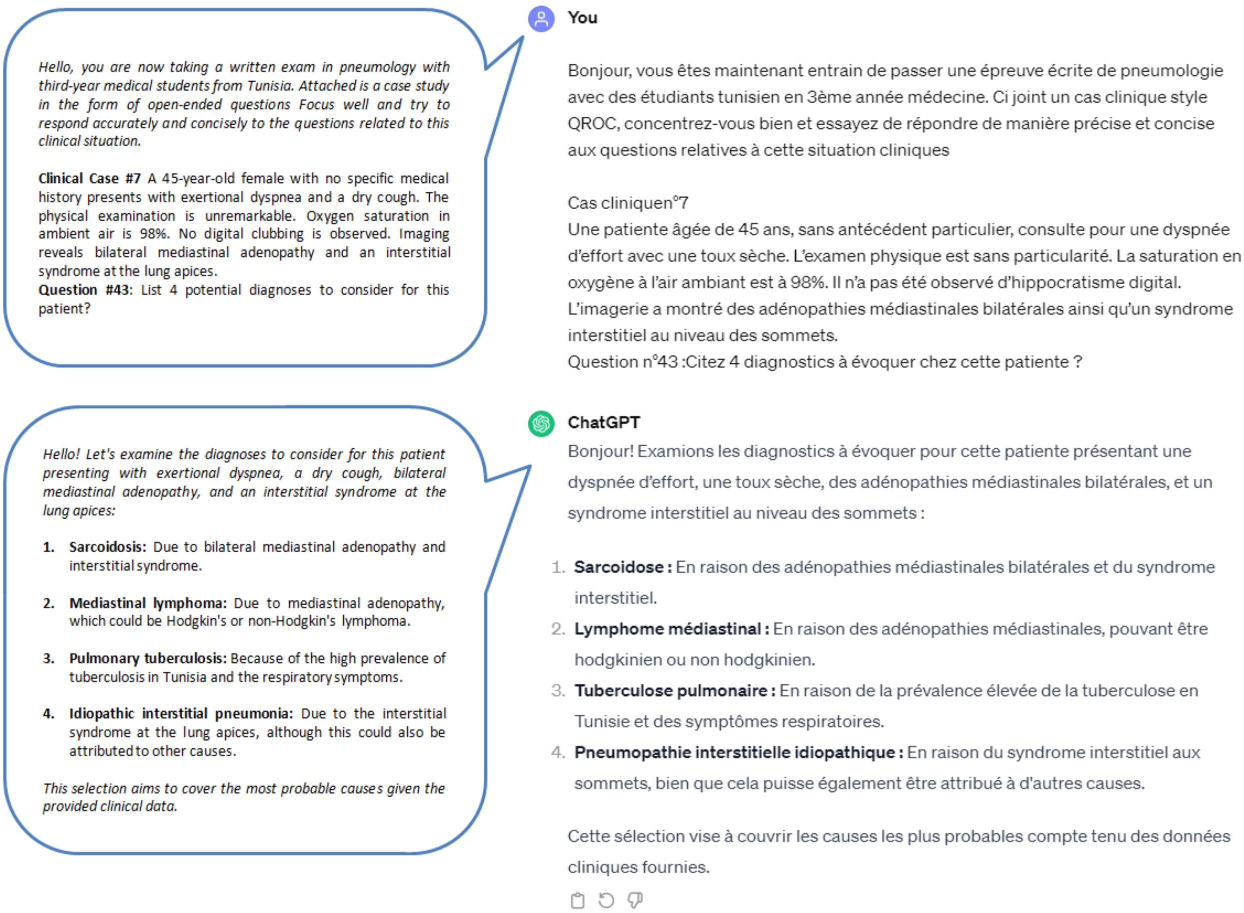
Each question is assigned 1 point. For MCQs, the grading scale was as follows: an incorrect response concealed a correct answer. The assigned grades were 0, 1, or 0.5, based on the nature of the answer provided. SOEQs were assessed as follows: 1 point is awarded for a correct response, 0 points for an incorrect response, and 0.5 points for an omission. For clarity, the global scores achieved by both third-year medical students, and ChatGPT were transformed into a score out of 100 (maximum score). To successfully pass the examination, candidates needed to achieve a global score of ≥50 points.

Figure 1. Illustration of the first version of the conversation with ChatGPT. The question was directly posed to ChatGPT without any context. ChatGPT responded by introducing the answer, emphasizing the necessity for a health care provider to address the case, and provide 4 probable diagnoses as demanded: sarcoidosis, idiopathic pulmonary fibrosis, mediastinal lymphoma, and atypical pulmonary infection. The bubbles in the figure represent the English translation of the conversations conducted in French.



XSL•FO
RenderX

Figure 2. Illustration of the second version of the conversation with ChatGPT. The chatbot was asked the same question (in a separate chat session) but preceded by a briefing about the context of the question, which pertains to a pneumology examination for Tunisian medical students. A clear instruction on how to answer was provided. In light of this context, ChatGPT altered its response, explicitly stating pulmonary tuberculosis, likely due to the endemic nature of *Mycobacterium tuberculosis* in Tunisia and other African countries. The bubbles in the figure represent the English translation of the conversations conducted in French.



Data Analysis

The collected data underwent statistical analysis using SPSS (version 25.0; IBM Corp). Nominal values were expressed as frequencies and percentages, while continuous variables were represented using means and SDs for normally distributed data and medians and quartiles for non-normally distributed data.

A comparative analysis was conducted, evaluating student scores alongside those V1 and V2. This analysis encompassed various factors, including question formats (MCQs, SOEQs, and clinical scenarios), topics (clinical pneumology, microbiology, respiratory radiology, pharmacology, pathology, and thoracic surgery), and item performance indexes. To accurately portray the performance levels of each ChatGPT conversation, we presented results as percentages of the maximum scale attributed to each studied item, along with the ranking of ChatGPT scores among those of third-year medical students.

Ethical Considerations

We have obtained approval from both the Medical Education Committee and the Ethics Committee of the Faculty of Medicine of Tunis to access the data (file number CE-FMT/2024/04/FSI/V2). This approval ensures confidentiality and restricting external use.

Results

Performance of Students in the Pneumology Examination

The median overall score achieved by medical students in the pulmonology examination was 48.9 out of 100 (IQR 40.0-54.7; Table 2). Among the participants (N=244), a modest cohort of 107 students reached the necessary threshold for successful completion of the examination, resulting in an overall success rate of 43.9%.

XSL•FO
RenderX

Table 2. Pneumology examination performance comparison: medical students versus ChatGPT with (V1) and without (V2) contextualization.

Parameters and categories	Maximum category score	Medical students' performance	V1 performance			V2 performance		
		Score, median (IQR)	Score	Percentage score	Rank among students (percentile)	Score	Percentage score	Rank among students (percentile)
Examination topics								
Pathology	3	2.5 (2-3)	2	66.7	133 (40.6)	2.5	83.3	84 (62.5)
Pharmacology	5	3.5 (2.5-4)	3	60	137 (38.8)	3.5	70	96 (57.1)
Microbiology	2	1.5 (1-1.5)	1.5	75	48 (78.6)	1.5	75	48 78.6)
Radiology	7	3.5 (2.1-4.5)	3.5	50	93 (58.5)	4	57.1	64 (71.4)
Thoracic surgery	1	0 (0-0)	1	100	1 (99.6)	0	0	29 (87.1)
Clinical pneumology	27	11 (9-13)	10	37	133 (40.6)	11.5	42.6	97 (56.7)
Question formats								
Independent MCQs[a]	9	4.5 (3.5-5.5)	4	44.4	138 (38.4)	3	33.3	191 (14.7)
Independent SOEQs[b]	13	5 (3.5-6)	4.5	34.6	120 (46.4)	6.5	50	30 (86.6)
MCQ-structured clinical cases	7	2.8 (2-3.5)	1.5	21.4	181 (19.2)	2	28.6	149 (33.5)
SOEQ-structured clinical cases	16	9.5 (7.6-11)	11	68.8	51 (77.2)	11.5	71.9	36 (83.9)
Overall examination score	100	48.9 (40-54.4)	46.7	46.7	133 (40.6)	51.1	51.1	85 (62.1)

[a]MCQ: multiple choice question.

[b]SOEQ: short answer open-ended question.

Significant variations in performance emerged across different question categories. Notably, students (N=244) demonstrated pronounced proficiency in the domains of pathology, pharmacology, and microbiology, with scores exceeding 50% in 88.5% (n=216), 77.5% (n=189), and 74.6% (n=182), respectively. A moderate level of accomplishment was observed in the field of radiology. In contrast, the weakest performances were evident in questions related to thoracic surgery and clinical pneumology, with only 11.5% (n=28) and 22.5% (n=55) of students surpassing the 50% threshold of the maximum score in these areas.

The question format also appeared to significantly influence students' performance. Candidates (N=244) excelled in SOEQ-structured clinical cases and independent MCQs, with 68.9% (n=212) and 56.1% (n=137), respectively, achieving marks exceeding 50% of the maximum achievable. Conversely, the performance in MCQ-structured clinical cases lagged, with only 31.1% (n=76) of candidates reaching scores beyond 50% of the highest attainable mark for this question format. The most challenging performance was observed in independent SOEQs, as only 19.3% (n=47) of students achieved marks surpassing the 50% threshold of the maximum attainable for this particular question format.

Based on these students' outcomes, item performance indexes were computed. A significant proportion of questions (25/45, 56%) exhibited moderate difficulty indexes, while only 18% (8/45) of the questions demonstrated elevated levels of difficulty. Additionally, a substantial fraction of the items (21/45, 47%) showed limited discriminatory power in contrast to 24% (11/45) that displayed a pronounced D2 (Table 1).

Assessment of ChatGPT-3.5 Performance in the Pneumology Examination

V1 performed well, achieving scores exceeding 50% in all question categories except for clinical pneumology. A similar trend emerged with V2, even though it faced challenges in reaching scores above 50% in thoracic surgery and clinical pneumology (Table 2).

The question format significantly impacted ChatGPT's performance. In cases where questions lacked contextualization, V1 fell short of reaching the 50% mark for the maximum score in all question formats, except for SOEQ-structured clinical cases. Similarly, in the responses generated by V2, even when provided with appropriate context, limitations were evident in both independent MCQs and MCQs integrated into clinical cases. Interestingly, V2 demonstrated a higher level of accuracy

XSL•FO
RenderX

in SOEQ-structured clinical cases. Both conversations displayed improved performance in questions with higher D1 and D2 (Table 3).

Considering the overall examination scores, V1 did not meet the passing threshold, achieving a total score of 46.7 out of 100. In contrast, V2 secured a global score of 51.5 out of 100, narrowly achieving success in this examination.

Table 3. Achievement quotient of ChatGPT with (V1) and without (V2) contextualization in the pneumology examination by difficulty and discrimination indexes.

	V1 (%)	V2 (%)
Low difficulty index terms	20.8	16.7
Intermediate difficulty index terms	54	62
High difficulty index terms	62.5	68
Low discrimination index items	31	38.1
Intermediate discrimination index items	61.5	61.5
High discrimination index items	59.1	63.6

Comparative Analysis of ChatGPT Performance and Medical Students' Performance

Question Topic

Comparing the performance of ChatGPT with that of medical students, distinct patterns emerge. V1 demonstrated heightened proficiency in specialized pneumology fields, especially radiology, microbiology, and thoracic surgery. Notably, V1 outperformed 131 (58.5%), 176 (78.6%), and 223 (99.6%) medical students in these respective domains. ChatGPT faced challenges in this conversation when addressing questions related to pathology, pharmacology, and clinical pneumology, achieving lower scores than most medical students. In opposition to that, V2 consistently provided more accurate responses than the majority of medical students across various question categories, regardless of their specialized fields. Noteworthy excellence was observed, particularly in microbiology and thoracic surgery.

Question Format

V1 demonstrated strong proficiency in SOEQ-structured clinical cases, surpassing the performance of 173 (77.2%) medical students. However, its performance weakened in independent MCQs and SOEQs, and it performed less optimally in MCQ-structured clinical cases compared to third-year medical students. In the case of V2, commendable performance was observed in responding to both independent and structured SOEQs within clinical cases. Yet, a notable deficiency emerged in accurately answering all formats of MCQs, ranking only above 33 (14.7%) and 75 (33.5%) students in independent MCQs and MCQ-structured clinical cases, respectively.

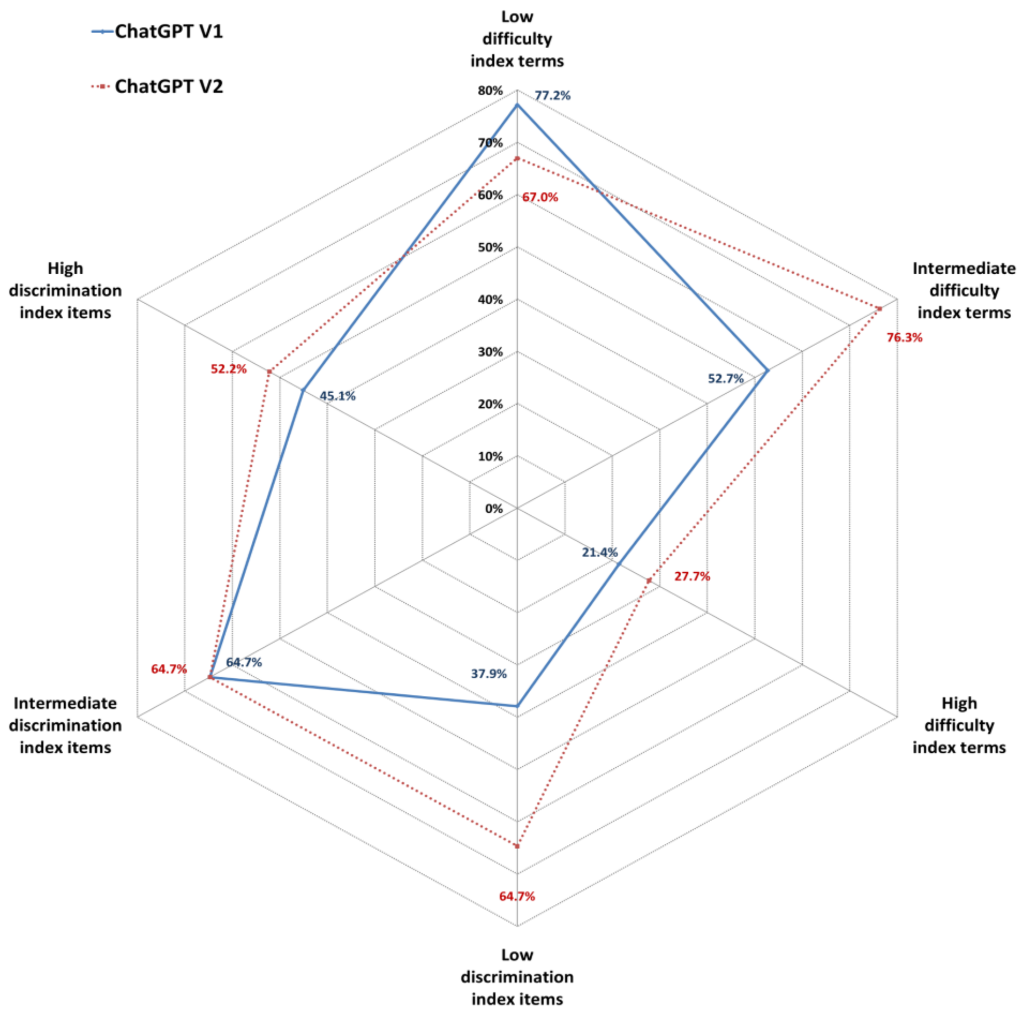
Item Performance Indexes

Both conversations with ChatGPT, particularly V2, performed better than students in handling questions of low and intermediate difficulty. Remarkably, students demonstrated stronger proficiency when tackling highly difficult questions. Regarding the D2, V1 showed similar performance to participants in accurately addressing questions with low and high D2 index values. Additionally, V1 slightly exceeded participants' performance in questions with an intermediate D2 index. V2 consistently outperformed medical students across all question discrimination categories (Figure 3).

In summary, V1 did not pass the examination, but its score surpassed that of 91 (40.6%) students. In contrast, V2 successfully passed the examination, outperforming 139 (62.1%) medical students.

XSL•FO
RenderX

Figure 3. Percentile rank of ChatGPT with (ChatGPT-V1) and without (ChatGPT-V2) contextualization among medical students in the pneumology examination based on difficulty and discrimination indexes. Percentages represent the percentile rank of ChatGPT-V1 and ChatGPT-V2 among medical students.



Discussion

Principal Findings

The cognitive capabilities and knowledge processing of ChatGPT have generated significant discussions in both public and academic circles. This NLP tool has gained attention for its prompt and coherent responses across various subjects, showcasing an impressive capacity to generate essays and offer explanations. However, there is a lack of comprehensive investigations into ChatGPT's performance in medical education and examinations. To address this, this study evaluates ChatGPT using a previously collected data set of pneumology examinations from FMT, enabling direct comparisons between ChatGPT's performance and that of third-year medical students.

Our findings highlight ChatGPT's proficiency in handling diverse biomedical information and clinical data. Powered by a vast corpus of internet text data, ChatGPT demonstrates remarkable expertise in pneumology, particularly excelling in radiology and microbiology. It outperformed a significant proportion of medical students in these paraclinical specialties. Comparable high performance of AI-powered tools in paraclinical sciences has been previously documented before. Rodriguez-Ruiz et al [16], using data from 9 diverse data sets (2652 examinations), including 653 malignancies, found that their AI system exhibited cancer detection accuracy on par with the average breast radiologist, surpassing the performance of 61.4% of the radiologists in their retrospective analysis.

Das et al [17] assessed ChatGPT's accuracy in addressing a test based on the competency-based medical education (CBME) curriculum for microbiology. ChatGPT showcased the ability to answer both first- and second-order knowledge questions related to microbiology. The model exhibited significant potential as an automated question-answering tool in the field of microbiology, achieving an accuracy rate of approximately 80%. In another investigation, ChatGPT demonstrated proficiency in medical biochemistry, another paraclinical specialty. It successfully responded to 200 random medical biochemistry reasoning questions from the CBME curriculum's competency modules [8].

XSL•FO
RenderX

In fields like clinical pneumology that demand careful processing of medical data, ChatGPT shows some limitations when compared to medical students. However, these shortcomings can be improved through adequate contextualization, as seen in the enhanced proficiency of V2. Our findings about clinical pneumology align with previous studies that highlight ChatGPT's challenges in similar medical disciplines requiring advanced judgment and nuanced clinical reasoning, such as neurology and traumatology. For instance, ChatGPT 3.5 achieved an overall accuracy rate of 57%, just below the 58% passing threshold set for the 2022 UK Specialty Certificate Neurology Examination [18].

Moreover, ChatGPT scored 35.8%, which is notably lower than the pass rate for the Fellowship of the Royal College of Surgeons examination in trauma surgery by 30%. This performance was also 8.2% below the average score of participants at all training levels [19]. In a study conducted in India, ChatGPT demonstrated a limited ability to translate basic pharmacology knowledge into clear clinical concepts. It exhibited inconsistency in predicting and explaining common drug interactions [20]. This observation aligns with ChatGPT's modest accuracy in questions related to pharmacology applied to pneumology in our FMT examination.

The way questions are presented greatly affects how well both medical students and AI tools like ChatGPT perform [21,22]. ChatGPT struggled to match the performance of medical students in all question styles, except for SOEQs integrated into clinical scenarios. Even after contextualization, ChatGPT still had a hard time answering MCQs in pulmonology compared to medical students. Zhu et al [23] addressed this concern, suggesting that ChatGPT may be more suitable for responding to open-ended questions than for being presented with a predefined set of options. Considering the ChatGPT's occasional inconsistency in providing identical responses for the same question, the authors recommended posing the question 3 times to ensure response stability.

Other research generally shows good performance by ChatGPT when handling MCQs. For example, a 2023 study by Duong and Solomon [24] revealed ChatGPT's comparable performance to human beings in responding to MCQs on human genetics. ChatGPT also successfully passed the 2022 Italian Residency Admission National Exam, which consists solely of MCQs. Additionally, in the 2022 European Examination in Core Cardiology, ChatGPT answered over 60% of questions correctly, displaying consistency across various MCQs [25]. In this study, the discrepancy in ChatGPT's performance across question formats may be attributed to the high difficulty level of these questions, even for third-year medical students.

ChatGPT clearly outperformed medical students in tasks that required detailed responses, particularly SOEQs integrated into clinical scenarios. This was supported by Qu et al [26], who also emphasized the impressive capability of this NLP software in handling otorhinolaryngology clinical scenarios [26]. Indeed, ChatGPT consistently provided accurate differential diagnoses and well-justified treatment strategies for recognized clinical conditions. It used specialized medical terminology and carefully curated relevant medical history, physical examination, radiological, and laboratory findings. This proficiency can be explained by the similarity between the scenarios in our pneumology examination and the writing style commonly found in textbooks, scientific literature, and other data sources used to train the AI model.

Unlike third-year medical students, ChatGPT surprisingly exhibited limited performance on questions with a high difficulty index. These questions necessitate skills in navigating intricate concepts, synthesizing information, and using strategic analytical abilities. Bhayana et al [27] subjected this chatbot to the Canadian Royal College and American Board of Radiology examinations and their conclusions match our findings. Although ChatGPT successfully passed these examinations, it faced difficulties with questions demanding higher order thinking, such as describing radiological findings, classification, and application of concepts [27]. While certain questions can help tell the difference between students with different levels of ability or knowledge, this D2 might not apply directly to AI-powered models like ChatGPT. A noteworthy observation is ChatGPT's enhanced performance when provided with adequate context, outperforming students irrespective of the theoretical item discrimination.

Ultimately, the findings reveal unexpected limits in ChatGPT's performance during our pneumology examination. It barely passed in the part with contextualized chats, giving an overall modest score of 51.1%. This is different from past research where ChatGPT consistently demonstrates strong performance in English-language medical assessments like the United States Medical Licensing Examination, CBME evaluations, and the European Examination in Core Cardiology [17,25,28]. It appears that its effectiveness diminishes when dealing with evaluations from non-Western institutions and non-English language examinations like our Tunisian examination, written in French. Similarly, this AI chatbot faced challenges in both the Taiwanese pharmacist licensing and Taiwanese family medicine board examinations [29,30]. It also scored below the level of students in a Korean parasitology examination, the Japanese National Medical Licensing Examination, and the Chinese National Medical Licensing Examination [31,32]. This discrepancy likely arises from ChatGPT's limited ability to grasp linguistic nuances in non-English texts, exacerbated by the prevalence of Western-centric internet data. In certain contexts, these data may not fully apply to African and Asian populations, which exhibit slight variations in clinical presentations and disease epidemiology.

Strengths and Limitations of the Study

Our research constitutes the initial exploration of ChatGPT's capabilities in French-language medical examinations, providing a valuable addition to the expanding body of research in medical AI assessment. A notable strength of this study lies in its comparative approach, effectively evaluating ChatGPT's performance alongside that of medical students in a comprehensive pneumology examination. This examination covers various question formats and topics, offering a realistic assessment of the AI's competencies.

However, the study acknowledges several limitations. Conducted at a single institution with a highly homogeneous

XSL•FO
RenderX

population concerning demographics, educational background, and medical curricula, there may be a potential selection bias that affects the external validity of the findings, particularly when extrapolating to more diverse student groups, even from other French-speaking medical universities. Additionally, focusing solely on the pneumology field may limit the generalizability of the findings to a broader academic context.

ChatGPT's inability to process visual elements also introduces an inherent selection bias concerning the administered questions, hindering a comprehensive evaluation of its proficiency in clinical scenarios where visual cues, radiology data, and histological images are significant. It is crucial to recognize that the specific findings related to ChatGPT-3.5 may not necessarily extend to other iterations of ChatGPT or alternative AI models. Furthermore, the absence of cultural adaptation and the scarcity of relevant data for non-Western contexts impeded a thorough exploration of ChatGPT's capabilities, potentially introducing a cultural bias.

Conclusions

In summary, despite its access to a comprehensive web-based data set and quick response generation, ChatGPT performs similarly to an average medical student, with outcomes influenced by question format, item complexity, and contextual factors. Notably, ChatGPT struggles in specific medical contexts requiring information synthesis, advanced analytical skills, and nuanced clinical judgment. Its efficiency also diminishes in non–English language assessments and when confronted with data outside dominant internet sources. These findings suggest the need for further exploration and improvement in the application of AI tools like ChatGPT in medical education, training, and evaluation. It also emphasizes the importance of enhancing its performance across cultural and linguistic contexts.

Acknowledgments

We extend our deep appreciation to Dr Toujani Sonia for his invaluable assistance in impartially and objectively evaluating the responses generated by V1 and V2, thereby ensuring an unbiased and rigorous assessment process. 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.

Data Availability

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

Conflicts of Interest

None declared.

Authors' Contributions

HC and CM conceived the study and designed the methodology. HC and AMM conducted the literature review. HC and CM engaged in conversations with ChatGPT. HC and CM performed data collection and statistical analysis. HC, CM, and AMM jointly drafted the manuscript. SM and BD supervised the research progression. All authors approved the final version of the manuscript.

References

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18(5):544-551 [FREE Full text] [doi: 10.1136/amiajnl-2011-000464] [Medline: 21846786]
2. Hirschberg J, Manning CD. Advances in natural language processing. Science 2015 Jul 17;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]
3. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Syst 2023;2(2):81-83 [FREE Full text] [doi: 10.54097/fcis.v2i2.4465]
4. Introducing ChatGPT. OpenAi. URL: https://openai.com/blog/chatgpt [accessed 2023-08-01]
5. Maheshwari R. Top AI statistics and trends. Forbes Advisor INDIA. 2023. URL: https://www.forbes.com/advisor/in/business/ai-statistics/ [accessed 2023-08-01]
6. Hu K. ChatGPT sets record for fastest-growing user base—analyst note. Reuters. 2023. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2023-08-01]
7. Biswas SS. Role of chat GPT in public health. Ann Biomed Eng 2023;51(5):868-869. [doi: 10.1007/s10439-023-03172-7] [Medline: 36920578]
8. Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus 2023 Apr;15(4):e37023-e37066 [FREE Full text] [doi: 10.7759/cureus.37023]

XSL•FO
RenderX

9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
10. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
11. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
12. Downing SM. Item response theory: applications of modern test theory in medical education. Med Educ 2003 Aug;37(8):739-745. [doi: 10.1046/j.1365-2923.2003.01587.x] [Medline: 12945568]
13. Thomas ML. The value of item response theory in clinical assessment: a review. Assessment 2011 Sep;18(3):291-307. [doi: 10.1177/1073191110374797] [Medline: 20644081]
14. Engelhardt PV. An introduction to classical test theory as applied to conceptual multiple-choice tests. In: Henderson CR, Harper KA, editors. Getting Started in PER. College Park, TX: American Association of Physics Teachers; Apr 2009:1-40.
15. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 2010 Jan;44(1):109-117 [FREE Full text] [doi: 10.1111/j.1365-2923.2009.03425.x] [Medline: 20078762]
16. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019 Sep 01;111(9):916-922 [FREE Full text] [doi: 10.1093/jnci/djy222] [Medline: 30834436]
17. Das D, Kumar N, Longjam L, Sinha R, Deb Roy A, Mondal H, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus 2023 Mar;15(3):e36034 [FREE Full text] [doi: 10.7759/cureus.36034] [Medline: 37056538]
18. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurol Open 2023;5(1):e000451 [FREE Full text] [doi: 10.1136/bmjno-2023-000451] [Medline: 37337531]
19. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma and Orthopaedics) examination? Postgrad Med J 2023 Sep 21;99(1176):1110-1114. [doi: 10.1093/postmj/qgad053] [Medline: 37410674]
20. Juhi A, Pipil N, Santra S, Mondal S, Behera J, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. Cureus 2023 Mar;15(3):e36272 [FREE Full text] [doi: 10.7759/cureus.36272] [Medline: 37073184]
21. Medina MS. Relationship between case question prompt format and the quality of responses. Am J Pharm Educ 2010 Mar 10;74(2):29 [FREE Full text] [doi: 10.5688/aj740229] [Medline: 20414442]
22. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ 2014 Nov 28;14:249 [FREE Full text] [doi: 10.1186/s12909-014-0249-2] [Medline: 25431359]
23. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783 [FREE Full text] [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
24. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. Eur J Hum Genet 2023 May 29:466-468 [FREE Full text] [doi: 10.1038/s41431-023-01396-8] [Medline: 37246194]
25. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023 May;4(3):279-281 [FREE Full text] [doi: 10.1093/ehjdh/ztad029] [Medline: 37265864]
26. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. OTO Open 2023;7(3):e67 [FREE Full text] [doi: 10.1002/oto2.67] [Medline: 37614494]
27. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023 Jun;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
28. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
29. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc 2023 Jul 01;86(7):653-658 [FREE Full text] [doi: 10.1097/JCMA.0000000000000942] [Medline: 37227901]
30. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023 Aug 01;86(8):762-766 [FREE Full text] [doi: 10.1097/JCMA.0000000000000946] [Medline: 37294147]
31. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. J Med Syst 2023 Aug 15;47(1):86. [doi: 10.1007/s10916-023-01961-0] [Medline: 37581690]
32. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]

XSL•FO
RenderX

Abbreviations

AI: artificial intelligence
CBME: competency-based medical education
D1: difficulty index
D2: discrimination index
FMT: Faculty of Medicine of Tunis
MCQ: multiple choice question
NLP: natural language processing
SOEQ: short open-ended question

Edited by AH Sapci, MD; submitted 29.09.23; peer-reviewed by H Mondal, L Zhu; comments to author 20.01.24; revised version received 05.02.24; accepted 26.02.24; published 23.07.24.

Please cite as:
Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B
Appraisal of ChatGPT's Aptitude for Medical Education: Comparative Analysis With Third-Year Medical Students in a Pulmonology Examination
JMIR Med Educ 2024;10:e52818
URL: https://mededu.jmir.org/2024/1/e52818
doi:10.2196/52818
PMID:

©Hela Cherif, Chirine Moussa, Abdel Mouhaymen Missaoui, Issam Salouage, Salma Mokaddem, Besma Dhahri. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.07.2024. 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.

XSL•FO
RenderX

Original Paper

Assessing ChatGPT's Competency in Addressing Interdisciplinary Inquiries on Chatbot Uses in Sports Rehabilitation: Simulation Study

Joseph C McBee[1,2]; Daniel Y Han[1]; Li Liu[3,4], MD, PhD; Leah Ma[5], MSc; Donald A Adjeroh[6], PhD; Dong Xu[7], PhD; Gangqing Hu[1], PhD

[1]Department of Microbiology, Immunology, & Cell Biology, West Virginia University, Morgantown, WV, United States
[2]Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV, United States
[3]College of Health Solutions, Arizona State University, Phoenix, AZ, United States
[4]Biodesign Institute, Arizona State University, Tempe, AZ, United States
[5]College of Health, Education, and Human Services, Wright State University, Dayton, OH, United States
[6]Lane Department of Computer Science & Electrical Engineering, West Virginia University, Morgantown, WV, United States
[7]Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, United States

Corresponding Author:
Gangqing Hu, PhD
Department of Microbiology, Immunology, & Cell Biology
West Virginia University
64 Medical Center Drive
Morgantown, WV, 26506-9177
United States
Phone: 1 304 581 1692
Fax: 1 304 293 7823
Email: gh00001@mix.wvu.edu

Abstract

Background: ChatGPT showcases exceptional conversational capabilities and extensive cross-disciplinary knowledge. In addition, it can perform multiple roles in a single chat session. This unique multirole-playing feature positions ChatGPT as a promising tool for exploring interdisciplinary subjects.

Objective: The aim of this study was to evaluate ChatGPT's competency in addressing interdisciplinary inquiries based on a case study exploring the opportunities and challenges of chatbot uses in sports rehabilitation.

Methods: We developed a model termed PanelGPT to assess ChatGPT's competency in addressing interdisciplinary topics through simulated panel discussions. Taking chatbot uses in sports rehabilitation as an example of an interdisciplinary topic, we prompted ChatGPT through PanelGPT to role-play a physiotherapist, psychologist, nutritionist, artificial intelligence expert, and athlete in a simulated panel discussion. During the simulation, we posed questions to the panel while ChatGPT acted as both the panelists for responses and the moderator for steering the discussion. We performed the simulation using ChatGPT-4 and evaluated the responses by referring to the literature and our human expertise.

Results: By tackling questions related to chatbot uses in sports rehabilitation with respect to patient education, physiotherapy, physiology, nutrition, and ethical considerations, responses from the ChatGPT-simulated panel discussion reasonably pointed to various benefits such as 24/7 support, personalized advice, automated tracking, and reminders. ChatGPT also correctly emphasized the importance of patient education, and identified challenges such as limited interaction modes, inaccuracies in emotion-related advice, assurance of data privacy and security, transparency in data handling, and fairness in model training. It also stressed that chatbots are to assist as a copilot, not to replace human health care professionals in the rehabilitation process.

Conclusions: ChatGPT exhibits strong competency in addressing interdisciplinary inquiry by simulating multiple experts from complementary backgrounds, with significant implications in assisting medical education.

(JMIR Med Educ 2024;10:e51157) doi:10.2196/51157

XSL•FO RenderX

KEYWORDS

ChatGPT; chatbots; multirole-playing; interdisciplinary inquiry; medical education; sports medicine

Introduction

The sports industry is a significant economic contributor in the United States, which was projected to generate US $83.1 billion in revenue in 2023 [1]. Concurrently, sports/recreation-related injuries are prevalent, with an estimated rate of 34 per 1000 individuals, accumulating to an annual total of 8.6 million cases [2]. Sports rehabilitation, aiming to facilitate full recovery, minimize sports downtime, and prevent future injuries, is a process of coordinated efforts between the athlete and health care professionals across various disciplines [3]. However, the rehabilitation process often spans a lengthy period and demands expensive medical and psychological support, making it inaccessible for many patients. In recent years, the integration of artificial intelligence (AI) in sports medicine has shown promise in enhancing both the accessibility to service and the efficacy of treatment outcomes [4,5]. Nevertheless, the use of chatbots in assisting sports rehabilitation is still in its formative stages, with many potential benefits and pitfalls yet to be explored and understood.

ChatGPT, a sophisticated large language model (LLM)–based chatbot, is capable of human-like dialogue [6]. This chatbot exhibits promise as a virtual assistant in medical education by providing real-time personalized feedback and enhancing student engagement [7]. However, controlled assessments in medical education have identified considerable limitations such as the need for precise prompts (also known as prompt engineering), instances of hallucination, and a lack of critical thinking in its responses [8-10]. Another challenge is that many of the topics in health care are interdisciplinary, involving multiple contributors such as physicians, pharmacists, and social workers to ensure better treatment outcomes and patient satisfaction. Unfortunately, current evaluations of ChatGPT are often confined to tasks from a specifical discipline, leaving its competency in addressing interdisciplinary topics largely unexplored [11,12], especially in medical education fields such as sports rehabilitation [5,13].

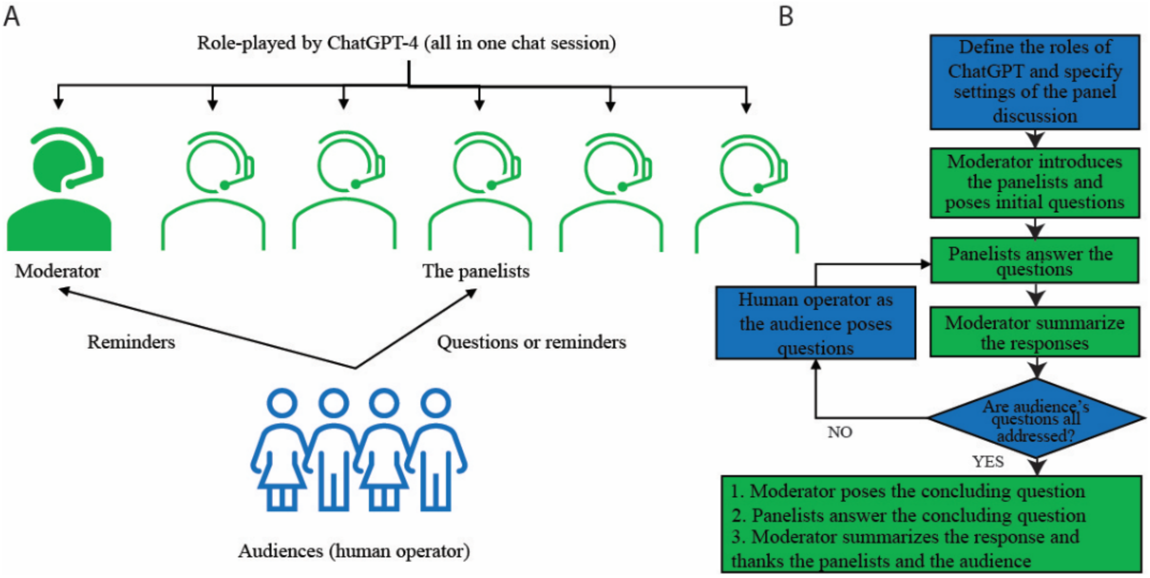
Here, we highlight an attractive feature of ChatGPT in addressing interdisciplinary questions via multirole-playing, which allows the chatbot to assume the roles of several discipline-specific experts simultaneously in one chat session. This unique feature inspired us to propose a model named PanelGPT for exploring interdisciplinary topics through a simulated panel discussion, where ChatGPT assumes the roles of a moderator and various experts on the panel. The aim of the study was to evaluate ChatGPT's competency through PanelGPT in addressing the opportunities and challenges of chatbot uses in sports rehabilitation, an interdisciplinary field that covers topics on patient education, physical therapy, psychological support, nutrition, and ethics.

Methods

The PanelGPT Model

We developed a model named PanelGPT to evaluate ChatGPT's competency in addressing interdisciplinary inquiry (Figure 1A). In this model, ChatGPT assumes the roles of both the moderator and panel experts, while a human operator, representing the audience, poses questions and sends reminders to the moderator or the panelists. Questions from the human operator are directly copied and pasted into the chat session, with ChatGPT determining which panel member(s) should respond. If the discussion stalls, the human operator prompts the moderator or panelists to continue by sending reminders. After each round of discussion, the moderator summarizes the comments before moving to the next question from the audience. Upon conclusion of the discussion, we summarized and evaluated the chatbot's responses based on the literature and our expertise.

Figure 1. Overview of the PanelGPT model for a ChatGPT-simulated panel discussion (A) and a flowchart that delineates the simulation process (B).



XSL•FO
RenderX

Application to Chatbots in Sports Rehabilitation

We applied PanelGPT to explore the pros and cons of chatbot uses in sports rehabilitation. The simulated panelists included 4 experts representing essential disciplines related to the topic: a physiotherapist, psychologist, nutritionist, and AI expert specializing in clinical applications. In addition, a virtual athlete who had successfully recovered from a severe injury participated in the panel. We formulated 4 main questions based on personal experience and/or a literature review. After reviewing the responses from pilot simulations, we added 2 more questions (Multimedia Appendix 1 [14-16]). During one of the pilot simulations, ChatGPT autonomously introduced opening questions, which we subsequently included in the final simulations. This finding also inspired us to instruct the chatbot to ask closing questions at the end of each simulation.

To clarify, our focus is not on using ChatGPT to provide sports rehabilitation advice. Instead, we centered on using ChatGPT to drive a panel discussion titled "Chatbots in sports rehabilitation" in a "self-consistency" manner [17]. The prompts used to steer the final simulations are detailed in Multimedia Appendix 2. A flowchart that outlines the process of the simulation is shown in Figure 1B. At the beginning of the prompts, we instructed ChatGPT to undertake multiple roles and specified other settings in the simulation (Multimedia Appendix 2). Next, the moderator was prompted to introduce the panelists and kick off the discussion with opening questions. Following the responses to these initial questions from the panelists, the moderator was tasked to summarize the responses and open the platform for questions from the audience. In response, the human operator copied each question from the audience directly into the chat session, allowing ChatGPT to select which expert should respond autonomously. After each round of questions and answers, the moderator was prompted to summarize the responses and call for the next question. This process was iterated until all of the audience's questions had been addressed. At the end of the panel discussion, the moderator was asked to propose a closing question and provide a summary of the responses. Additional prompts were introduced as needed to ensure a smooth progression of the panel discussion (Multimedia Appendix 2). We repeated the simulation 3 times using ChatGPT-4 (May 24, 2023, version) with its online web interface [18].

As shown in Multimedia Appendix 1, we initiated the simulation with an opening question and concluded with a closing question. During the simulation, we prompted ChatGPT to simulate a panel discussion on topics from chatbot uses in sports rehabilitation in the order of "patient education," "physical therapy," "psychological support," "nutrition," "tracking & other alternatives," and "ethics." After 3 rounds of simulations, we manually evaluated the panel's response to questions from each topic by referring to the literature and our human expertise.

Ethical Considerations

This work was based on analyzing ChatGPT's response to designed prompts. As the work is classified as not human subjects research, review of the Institutional Review Board of West Virginia University was not required [19].

Results

Overview

The complete chat histories, including prompts and ChatGPT's response from the simulated panel discussion, are accessible in Multimedia Appendices 3-5 (audio versions are available upon request). As expected, 2 or more experts responded to each question (Table 1); the experts generally offered insights from their respective fields of expertise. We evaluated the responses by citing relevant references and according to our own expertise. The most relevant findings are compiled and summarized below for each question.

Table 1. Records of direct responses to questions during the simulation.[a]

Question	Physiotherapist	Psychologist	Nutritionist	Athlete	AI[b] expert
Opening question	1, 2, 3	1, 2, 3	1, 2, 3	1, 2, 3	1, 2, 3
Patient education	1, 2, 3	1, 2, -	-, -, -	1, -, 3	1, 2, 3
Physical therapy	1, 2, 3	-, -, -	-, -, -	-, -, -	1, 2, 3
Psychological support	-, -, -	1, 2, 3	-, -, -	1, -, -	1, 2, 3
Nutrition	-, -, -	-, -, -	1, 2, 3	-, -, -	1, 2, 3
Tracking and other alternatives	1, 2, 3	1, 2, -	-, -, -	1, 2, -	1, 2, 3
Ethics	1, 2, 3	1, 2, -	1, 2, -	1, -, -	1, 2, 3
Closing question	1, 2, 3	1, 2, 3	1, 2, 3	1, 2, 3	1, 2, 3

[a]Numbers 1, 2, and 3 indicate when a response directly targeting the question was made for rounds 1, 2, and 3 of the simulation, respectively, whereas "-" denotes the absence of such a response.

[b]AI: artificial intelligence.

Opening Question

The simulated panel discussion began with introductions and requests for the panelists' perspectives on the role of chatbots in sports rehabilitation, to which all panel members responded (Table 1). The ensuing dialogue identified chatbots as round-the-clock support systems, adept at monitoring, offering reminders, consulting, and nurturing a positive mindset in athletes during their recovery. Similar observations have been reported for orthopedic patients with AI assistance in the

XSL•FO
RenderX

literature [14,20,21]. Looking into the future and consistent with expectations, chatbots might grow increasingly adept at analyzing biomechanical data, emotional indicators, and nutritional needs, thus providing personalized feedback that helps athletes better comprehend their bodies and healing journeys.

Patient Education

The conversation pinpointed several critical factors in educating athletes on using chatbots for rehabilitation. Both the athlete and the psychologist touched on the importance of understanding the benefits of using a chatbot, such as a readily available source of advice and mental support [22]. The AI expert emphasized the education on transparency, including how data are collected, processed, stored, and protected. Effective communication with a chatbot is a nontrivial task [23]. The physiotherapist focused on how to guide users to interact with the chatbot effectively and how to interpret the responses. The discussion also underscored that the chatbot system is designed to enhance recovery, not to replace the human touch. Through education, athletes need to be able to identify situations that call for direct communication with health care professionals.

Physical Therapy

The primary focus of these questions was on the chatbot's potential to facilitate physical therapy by analyzing movements and weight distributions [15]. Relevant responses were from the physiotherapist and the AI expert, who acknowledged that current chatbots primarily interact with users through text and voice, which restricts their direct applicability to the question. However, the AI expert envisioned integrating chatbots, wearables, cameras, and smart devices to analyze an athlete's movement patterns and provide real-time, personalized feedback. A good example, as has also been noted in the literature, is computer vision–based analysis that has been applied to monitor and improve sports performance [24]. The AI expert further highlighted that the accuracy of this application depends on the size and quality of the training data, as well as advances in AI technologies such as machine learning and computer vision.

Psychological Support

This round of discussion explored the role of chatbots in analyzing emotional cues via sentiment analysis, a technique previously shown to enhance patient satisfaction in several medical chatbot applications [16,25,26] and other applications [27]. The panel's responses aligned with the existing literature: by delivering tailored responses to emotions, chatbots offer athletes emotional support and reduce their feelings of isolation. Nevertheless, the panel did not explore the impact of chatbots on psychological outcome measures such as improvements in communication skills, cognitive level, motivation, and abilities in coping with the injury. The psychologist and the AI expert cautioned that sentiment analysis may not always capture human emotions accurately. Thus, the psychological support provided via chatbots should be regarded as a complement to human interventions, which, in our opinion, can extend from health care professionals to coaches, teammates, friends, and family members.

Nutrition

Chatbots have been used for nutrition advice [28-30]. The nutritionist outlined multiple roles for chatbots in nutritional management, such as reminding athletes to stay hydrated, tracking dietary intake, and suggesting meal plans. A personalized dietary plan could use an advanced AI algorithm to analyze factors such as demographics, injury type, recovery stage, allergy history, and signals from wearable devices or health-tracking apps. The AI expert emphasized that building a personalized nutrition model demands a precise understanding of nutritional science, human physiology, and high-quality training data. However, given that chatbots might make mistakes such as recommending diets containing allergens [31] or harmful diet tips that promote eating disorders [32], they should be regarded as supplementary tools to human nutritionists rather than as their replacements.

Tracking and Other Alternatives

Responses from the physiotherapist and the AI expert to this topic largely echoed those provided during the "physical therapy" round. The athlete noted that the automated tracking, recording, and reminding function helps reduce stress, echoing the psychologist's comments. In line with remarks made by other researchers [33], the simulation highlighted several advantages of chatbots over traditional methods in sports medicine. These included reducing the need for manual reporting, offering convenient cloud-based access to records, real-time data collection, instantaneous analysis, and providing immediate advice. Despite these benefits, the simulation lacked a discussion on how chatbots could potentially enhance treatment outcomes over alternative tools such as increasing patient satisfaction or reducing the recovery duration. In addition, the questions were designed to invoke engagement from all panelists. However, the nutritionist unexpectedly did not respond (Table 1).

Ethics

Distinct from other audience-initiated topics, questions regarding ethics prompted responses from all panelists (Table 1). Some comments reiterated points from previous discussions, particularly regarding patient education. The conversation emphasized the need for stringent adherence to medical privacy regulations such as the Health Insurance Portability and Accountability Act in the United States or the General Data Protection Regulation in Europe [34]. This discussion highlighted the necessity of robust protocols for data encryption and storage to ensure security, as well as the need for transparency on data collection, processing, and accessibility. However, the panel did not delve into the merits and drawbacks of open-source, locally deployed chatbots (especially those furnished with domain-specific knowledge) versus commercial and online chatbots about privacy and security [35].

Regarding bias and fairness, it was stressed that chatbot training should use diverse and representative datasets. As users, athletes should retain complete discretion on whether to use chatbots, alternative methods, or a combination of both. The psychologist highlighted the need to implement chatbots in a manner that avoids triggering anxiety or other negative emotions. All the

XSL•FO
RenderX

comments align with the 5 ethical principles proposed by AI4People: beneficence, nonmaleficence, justice, autonomy, and explicability [36].

Closing Question

The moderator was prompted to steer the panel discussion toward its end with a final question. As anticipated, the questions were all forward-thinking (Multimedia Appendix 1). Panelists offered predictions drawing from their respective fields of expertise. Foreseeing rapid advancements in AI and complementary technologies, the panel envisaged a future of precision sports rehabilitation in the chatbot era. In this vision, the rehabilitation program would be tailored to individual needs, bolstered by health care providers, and empowered by chatbots. According to responses from the simulated athlete, this form of personalized support would make rehabilitation feel like a natural part of the recovery process, and the athlete would take charge of the rehabilitation journey.

Discussion

Principal Findings

We evaluated ChatGPT's competency in addressing interdisciplinary inquiry using sports rehabilitation as an example. Using a novel model named PanelGPT, we prompted ChatGPT to explore the pros and cons of chatbot use in sports rehabilitation. ChatGPT answered questions via a simulated panel discussion where it role-played multiple experts, including a physiotherapist, psychologist, nutritionist, AI expert, and athlete. Our analysis of its responses highlighted benefits such as 24/7 support, personalized advice, and automated tracking, as well as challenges such as limited interaction modes, inaccuracies in emotion-related advice, and data privacy concerns. We repeated the experiments with the most recent version, GPT-4o (May 2024), and obtained generally similar results. Thus, our findings highlight the potential of using ChatGPT through PanelGPT to enhance appreciation of any interdisciplinary topic.

The interdisciplinary approach through PanelGPT brings several benefits with significant implications in medical education. First, the responses come from a panelist of experts with complementary expertise, providing different perspectives that are automatically categorized and offering a comprehensive view of the topic in question. For instance, including an athlete on the panel yielded a unique user perspective that could be overlooked in simple prompts. For example, a simple prompt of the questions on "psychological support" to ChatGPT yielded responses rooted in knowledge based on a psychologist (Multimedia Appendix 6). Thus, PanelGPT can offer students a holistic view of a complex interdisciplinary topic and integrates insights that might be missed from traditional educational settings.

Second, as LLMs become increasingly adapted in education, it is important to educate students on alternative, innovative ways of using chatbots. Compared to conventional communication with a chatbot, PanelGPT is novel in that it focuses the chatbot's attention on the question and provides critical contexts for responding to the questions. For instance, when the "physical therapy" questions were simply prompted to ChatGPT, the responses quickly drifted toward other topics such as education and mental health (see Multimedia Appendix 7). With PanelGPT, the response involved a discussion between a physiotherapist and an AI expert, and the topic remained in the context of sports rehabilitation.

Third, the multirole-playing feature of ChatGPT through PanelGPT makes learning more interactive and engaging by encouraging active participation from learners. It also helps learners develop critical thinking skills such as synthesizing information from multiple simulated experts from different backgrounds and evaluating their credibility. This is particularly important when addressing the pros and cons of implementing new technologies in health care settings on topics that are interdisciplinary by nature.

Finally, having a panel of experts enables students to form a balanced view on a specific topic. For example, in addressing the "physical therapy" questions, the physiologist's response highlighted the current limitations of chatbots in text or voice communication, while the AI expert expanded the discussion to the integration of real-time video analysis (Multimedia Appendices 3-5). This balanced view is crucial in medical education, as it allows students to understand both the potential and the limitations of any emerging technology (such as chatbots) that are poised for health care applications.

Limitations

The breadth and depth of a panelist's response depend on the training dataset in the field. In several discussions, such as the "patient education" and "tracking and other alternatives" topics, where we expected feedback from all panelists, there was a noticeable lack of direct responses from the nutritionist. It could be that the dataset used to train ChatGPT for the nutritionist was underrepresented in the rehabilitation field. Indeed, a combined search for "rehabilitation" (or "rehab") and "nutritionist" (or "nutrition") on PubMed yielded 6-8 times fewer hits compared to searches involving the terms "physiotherapist" (or "physiotherapy") or "psychologist" (or "psychology") (as of July 7, 2023). To address this limitation, the human operator could send reminders to the nutritionist to elicit a response. In contrast to the nutritionist, the AI expert responded to questions on all topics. This is expected because of the inherent need for AI expertise in creating such chatbot systems.

The data used to train ChatGPT at the time of our experiments only extended up to September 2021. As such, ChatGPT could not provide comments that would reference more recent developments in chatbots such as ChatGPT itself or BARD (now known as Gemini). The feature to activate Bing within ChatGPT does allow for real-time information browsing from the internet. However, in practice, this disrupted the panel discussion's flow, resulting in a shift back to the regular ChatGPT conversation format and a subsequent loss of the expert identities after several exchanges (as shown in Multimedia Appendices 8-10).

We observed instances where the response to a question from the same expert was vague in one simulation but detailed in

XSL•FO
RenderX

another. This observation suggests that conducting multiple simulations could enhance the efficacy of PanelGPT in providing a well-rounded understanding of the knowledge landscape surrounding an interdisciplinary topic. This practice enables self-consistency checking, which has been shown to improve the reasoning performance of language models [17]. Additionally, summarizing diverse responses from multiple simulations facilitates the identification of contrasting viewpoints and emergent trends in the panel discussion.

Hallucination, the generation of unsupported or false information, is a prevalent issue with LLM-based chatbots. The multiperspective approach of PanelGPT allows the chatbot to draw on the strengths and mitigate the weaknesses of each panelist when responding to specific questions. The current model is constrained by the same chatbot simulating all the panelists in a given chat session. With advances in chatbot development, this model could be extended by integrating responses from other LLM chatbots, especially those possessing domain-specific knowledge. In fact, cross-referencing responses from different experts on the panel powered by distinct models helps mitigate hallucination [37]. Nonetheless, it remains crucial to cross-verify the conclusions drawn from the simulation with literature findings or opinions from human experts to ensure the accuracy of the information.

Throughout the simulation, we noted instances where comments from one expert were acknowledged by another. Intriguingly, contradictory comments between experts were not observed. The richness and depth of the discussion can be further enhanced by using additional prompting strategies. For instance, after each response round, panelists could be prompted to critically evaluate each other's comments to foster consensus or highlight disagreements. Panelists may also be prompted to pose questions to one another, such as seeking clarifications or requesting further details on a given response. Moreover, panelists could prompt the audience to clarify their questions if necessary. These additional prompting tactics make the panel discussion more engaging and mirror a real-life scenario, increasing the likelihood of obtaining a thorough appreciation of the topic.

Conclusions

We presented PanelGPT, an innovative method that capitalizes on the multirole-playing feature of ChatGPT through simulated panel discussions, and applied it to evaluate ChatGPT's competency in addressing interdisciplinary inquiry. In our case study, ChatGPT adequately addressed the opportunities and challenges on chatbot uses in sports rehabilitation. As a generalizable model, we envision PanelGPT as a supplementary tool in the classroom, aiding students in understanding complex interdisciplinary topics in medical education, such as nursing care, sports rehabilitation, stroke rehabilitation, and the management of recurrent pneumonia in long-term care facilities.

Acknowledgments

This study is supported by the National Institutes of Health (NIH)-National Institute of General Medical Sciences (NIGMS) (grants P20 GM103434 and U54 GM-104942 to GH), National Science Foundation (grant 2125872 to GH and DAA), and NIH-National Library of Medicine (NLM) (grant R01LM013438 to LL and grant 5R01LM013392 to DX). JCM is supported by the West Virginia University Cancer Institute summer undergraduate research program. We thank Zien Cheng and Evelyn Shue from GH's lab for proofreading. ChatGPT and Grammarly were used to help polish the writing of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

GH performed the formal analysis and wrote the original draft of the manuscript. All authors contributed to formal analysis and review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questions designed for the simulated panel discussion on “chatbots in sports rehabilitation.”.
[DOCX File , 15 KB - mededu_v10i1e51157_app1.docx]

Multimedia Appendix 2
Prompts used to steer the simulated panel discussion.
[DOCX File , 15 KB - mededu_v10i1e51157_app2.docx]

Multimedia Appendix 3
Prompts and scripts for the 1st round of simulation.
[PDF File (Adobe PDF File), 2050 KB - mededu_v10i1e51157_app3.pdf]

Multimedia Appendix 4

XSL•FO
RenderX

Prompts and scripts for the 2nd round of simulation.
[PDF File (Adobe PDF File), 2194 KB - mededu_v10i1e51157_app4.pdf]

Multimedia Appendix 5
Prompts and scripts for the 3rd round of simulation.
[PDF File (Adobe PDF File), 2145 KB - mededu_v10i1e51157_app5.pdf]

Multimedia Appendix 6
Scripts for a direct prompt on "psychological support.".
[PDF File (Adobe PDF File), 796 KB - mededu_v10i1e51157_app6.pdf]

Multimedia Appendix 7
Scripts for a direct prompt on "physical therapy.".
[PDF File (Adobe PDF File), 760 KB - mededu_v10i1e51157_app7.pdf]

Multimedia Appendix 8
Prompts and scripts for the 1st round of simulation with Bing activated.
[PDF File (Adobe PDF File), 780 KB - mededu_v10i1e51157_app8.pdf]

Multimedia Appendix 9
Prompts and scripts for the 2nd round of simulation with Bing activated.
[PDF File (Adobe PDF File), 678 KB - mededu_v10i1e51157_app9.pdf]

Multimedia Appendix 10
Prompts and scripts for the 3rd round of simulation with Bing activated.
[PDF File (Adobe PDF File), 1742 KB - mededu_v10i1e51157_app10.pdf]

References

1. 2019 PwC Outlook: At the gate and beyond. PricewaterhouseCoopers. URL: https://www.pwc.com/us/en/industries/tmt/assets/pwc-sports-outlook-2019.pdf [accessed 2023-06-22]
2. Sheu Y, Chen LH, Hedegaard H. Sports- and recreation-related injury episodes in the United States, 2011-2014. Natl Health Stat Report 2016 Nov 18(99):1-12 [FREE Full text] [Medline: 27906643]
3. Dhillon H, Dhillon S, Dhillon MS. Current concepts in sports injury rehabilitation. Indian J Orthop 2017;51(5):529-536 [FREE Full text] [doi: 10.4103/ortho.IJOrtho_226_17] [Medline: 28966376]
4. Ramkumar PN, Luu BC, Haeberle HS, Karnuta JM, Nwachukwu BU, Williams RJ. Sports medicine and artificial intelligence: a primer. Am J Sports Med 2022 Mar 26;50(4):1166-1174. [doi: 10.1177/03635465211008648] [Medline: 33900125]
5. Fayed AM, Mansur NSB, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in orthopaedics and sports medicine. J Exp Orthop 2023 Jul 26;10(1):74 [FREE Full text] [doi: 10.1186/s40634-023-00642-8] [Medline: 37493985]
6. Owens B. How Nature readers are using ChatGPT. Nature 2023 Mar 20;615(7950):20. [doi: 10.1038/d41586-023-00500-8] [Medline: 36807343]
7. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2024 Mar 14;17(5):926-931. [doi: 10.1002/ase.2270] [Medline: 36916887]
8. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. JMIR Med Educ 2024 Jan 31;10:e51344 [FREE Full text] [doi: 10.2196/51344] [Medline: 38111256]
9. Magalhães Araujo S, Cruz-Correia R. Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med Educ 2024 Mar 20;10:e51151 [FREE Full text] [doi: 10.2196/51151] [Medline: 38506920]
10. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ 2023 Dec 22;9:e50658 [FREE Full text] [doi: 10.2196/50658] [Medline: 38133908]
11. King RC, Samaan JS, Yeo YH, Peng Y, Kunkel DC, Habib AA, et al. A multidisciplinary assessment of ChatGPT's knowledge of amyloidosis: observational study. JMIR Cardio 2024 Apr 19;8:e53421 [FREE Full text] [doi: 10.2196/53421] [Medline: 38640472]
12. Miao H, Ahn H. Impact of ChatGPT on interdisciplinary nursing education and research. Asian Pac Isl Nurs J 2023 Apr 24;7:e48136 [FREE Full text] [doi: 10.2196/48136] [Medline: 37093625]

XSL•FO
RenderX

13. Zhu W, Geng W, Huang L, Qin X, Chen Z, Yan H. Who could and should give exercise prescription: physicians, exercise and health scientists, fitness trainers, or ChatGPT? J Sport Health Sci 2024 May;13(3):368-372 [FREE Full text] [doi: 10.1016/j.jshs.2024.01.001] [Medline: 38176646]
14. Dwyer T, Hoit G, Burns D, Higgins J, Chang J, Whelan D, et al. Use of an artificial intelligence conversational agent (chatbot) for hip arthroscopy patients following surgery. Arthrosc Sports Med Rehabil 2023 Mar 16;5(2):e495-e505 [FREE Full text] [doi: 10.1016/j.asmr.2023.01.020] [Medline: 37101866]
15. Cheng K, Guo Q, He Y, Lu Y, Gu S, Wu H. Exploring the potential of GPT-4 in biomedical engineering: the dawn of a new era. Ann Biomed Eng 2023 Aug;51(8):1645-1653. [doi: 10.1007/s10439-023-03221-1] [Medline: 37115365]
16. Oh J, Jang S, Kim H, Kim J. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform 2020 Aug;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
17. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv Preprint 2023 Mar 7:1-24. [doi: 10.48550/arXiv.2203.11171]
18. ChatGPT. URL: https://chat.openai.com [accessed 2023-07-08]
19. Not Human Subjects Research and Not Research. West Virginia University Research Compliance Administration. URL: https://human.research.wvu.edu/get-started/determine-protocol-type/nhsr [accessed 2024-07-31]
20. Anthony CA, Rojas EO, Keffala V, Glass NA, Shah AS, Miller BJ, et al. Acceptance and commitment therapy delivered via a mobile phone messaging robot to decrease postoperative opioid use in patients with orthopedic trauma: randomized controlled trial. J Med Internet Res 2020 Jul 29;22(7):e17750 [FREE Full text] [doi: 10.2196/17750] [Medline: 32723723]
21. Bian Y, Xiang Y, Tong B, Feng B, Weng X. Artificial intelligence-assisted system in postoperative follow-up of orthopedic patients: exploratory quantitative and qualitative study. J Med Internet Res 2020 May 26;22(5):e16896 [FREE Full text] [doi: 10.2196/16896] [Medline: 32452807]
22. Haque MDR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth 2023 May 22;11:e44838 [FREE Full text] [doi: 10.2196/44838] [Medline: 37213181]
23. Shue E, Liu L, Li B, Feng Z, Li X, Hu G. Empowering beginners in bioinformatics with ChatGPT. Quant Biol 2023 Jun;11(2):105-108 [FREE Full text] [doi: 10.15302/j-qb-023-0327] [Medline: 37378043]
24. Host K, Ivašić-Kos M. An overview of human action recognition in sports based on computer vision. Heliyon 2022 Jun 5;8(6):e09633 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e09633] [Medline: 35706961]
25. Chaix B, Bibault J, Pienkowski A, Delamon G, Guillemassé A, Nectoux P, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR Cancer 2019 May 02;5(1):e12856 [FREE Full text] [doi: 10.2196/12856] [Medline: 31045505]
26. de Gennaro M, Krumhuber EG, Lucas G. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. Front Psychol 2019 Jan 23;10:3061 [FREE Full text] [doi: 10.3389/fpsyg.2019.03061] [Medline: 32038415]
27. Abbasi A, Li J, Adjeroh D, Abate M, Zheng W. Don't mention it? Analyzing user-generated content signals for early adverse event warnings. Inf Syst Res 2019 Sep;30(3):1007-1028. [doi: 10.1287/isre.2019.0847]
28. Casas J, Mugellini E, Abou Khaled O. Food diary coaching chatbot. 2018 Presented at: 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18; October 8-12, 2018; Singapore p. 1676-1680. [doi: 10.1145/3267305.3274191]
29. Calvaresi D, Eggenschwiler S, Calbimonte JP, Manzo G, Schumacher M. A personalized agent-based chatbot for nutritional coaching. 2021 Presented at: WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence; December 14-17, 2021; Melbourne, Australia p. 682-687. [doi: 10.1145/3486622.3493992]
30. Han R, Todd A, Wardak S, Partridge SR, Raeside R. Feasibility and acceptability of chatbots for nutrition and physical activity health promotion among adolescents: systematic scoping review with adolescent consultation. JMIR Hum Factors 2023 May 05;10:e43227 [FREE Full text] [doi: 10.2196/43227] [Medline: 37145858]
31. Niszczota P, Rybicka I. The credibility of dietary advice formulated by ChatGPT: robo-diets for people with food allergies. Nutrition 2023 Aug;112:112076 [FREE Full text] [doi: 10.1016/j.nut.2023.112076] [Medline: 37269717]
32. Tolentino D. National Eating Disorders Association pulls chatbot after users say it gave harmful dieting tips. NBC News. 2023 Jun 01. URL: https://www.nbcnews.com/tech/neda-pulls-chatbot-eating-advice-rcna87231 [accessed 2023-07-04]
33. Cheng K, Guo Q, He Y, Lu Y, Xie R, Li C, et al. Artificial intelligence in sports medicine: could GPT-4 make human doctors obsolete? Ann Biomed Eng 2023 Aug;51(8):1658-1662. [doi: 10.1007/s10439-023-03213-1] [Medline: 37097528]
34. Priyanshu A, Vijay S, Kumar A, Naidu R, Mireshghallah F. Are chatbots ready for privacy-sensitive applications? An investigation into input regurgitation and prompt-induced sanitization. arXiv Preprint 2024 May 24. [doi: 10.48550/arXiv.2305.15008]
35. Castelvecchi D. Open-source AI chatbots are booming - what does this mean for researchers? Nature 2023 Jun;618(7967):891-892. [doi: 10.1038/d41586-023-01970-6] [Medline: 37340135]
36. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People-An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 2018 Nov 26;28(4):689-707. [doi: 10.1007/s11023-018-9482-5] [Medline: 30930541]

XSL•FO
RenderX

37. Jiang D, Ren X, Lin BY. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. 2023 Presented at: 61st Annual Meeting of the Association for Computational Linguistics; Jul 9-14, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.acl-long.792]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by T de Azevedo Cardoso; submitted 23.07.23; peer-reviewed by K Chen, MN Shalaby; comments to author 15.08.23; revised version received 21.08.23; accepted 23.07.24; published 07.08.24.

Please cite as:
McBee JC, Han DY, Liu L, Ma L, Adjeroh DA, Xu D, Hu G
Assessing ChatGPT's Competency in Addressing Interdisciplinary Inquiries on Chatbot Uses in Sports Rehabilitation: Simulation Study
JMIR Med Educ 2024;10:e51157
URL: https://mededu.jmir.org/2024/1/e51157
doi:10.2196/51157
PMID:39042885

©Joseph C McBee, Daniel Y Han, Li Liu, Leah Ma, Donald A Adjeroh, Dong Xu, Gangqing Hu. Originally published in JMIR Medical Education (https://mededu.jmir.org), 07.08.2024. 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.

XSL•FO
RenderX

Original Paper

Educational Utility of Clinical Vignettes Generated in Japanese by ChatGPT-4: Mixed Methods Study

Hiromizu Takahashi[1], MD, PhD; Kiyoshi Shikino[2], MD, MHPE, PhD; Takeshi Kondo[3], MD, MHPE; Akira Komori[1,4], MD, PhD; Yuji Yamada[5], MD; Mizue Saita[1], MD, PhD; Toshio Naito[1], MD, PhD

[1]Department of General Medicine, Juntendo University Faculty of Medicine, Tokyo, Japan

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

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

[4]Department of Emergency and Critical Care Medicine, Tsukuba Memorial Hospital, Tsukuba, Japan

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

Corresponding Author:
Hiromizu Takahashi, MD, PhD
Department of General Medicine
Juntendo University Faculty of Medicine
Bunkyo
3-1-3 Hongo
Tokyo, 113-0033
Japan
Phone: 81 3 3813 3111
Email: hrtakaha@juntendo.ac.jp

Abstract

Background: Evaluating the accuracy and educational utility of artificial intelligence–generated medical cases, especially those produced by large language models such as ChatGPT-4 (developed by OpenAI), is crucial yet underexplored.

Objective: This study aimed to assess the educational utility of ChatGPT-4–generated clinical vignettes and their applicability in educational settings.

Methods: Using a convergent mixed methods design, a web-based survey was conducted from January 8 to 28, 2024, to evaluate 18 medical cases generated by ChatGPT-4 in Japanese. In the survey, 6 main question items were used to evaluate the quality of the generated clinical vignettes and their educational utility, which are information quality, information accuracy, educational usefulness, clinical match, terminology accuracy (TA), and diagnosis difficulty. Feedback was solicited from physicians specializing in general internal medicine or general medicine and experienced in medical education. Chi-square and Mann-Whitney *U* tests were performed to identify differences among cases, and linear regression was used to examine trends associated with physicians' experience. Thematic analysis of qualitative feedback was performed to identify areas for improvement and confirm the educational utility of the cases.

Results: Of the 73 invited participants, 71 (97%) responded. The respondents, primarily male (64/71, 90%), spanned a broad range of practice years (from 1976 to 2017) and represented diverse hospital sizes throughout Japan. The majority deemed the information quality (mean 0.77, 95% CI 0.75-0.79) and information accuracy (mean 0.68, 95% CI 0.65-0.71) to be satisfactory, with these responses being based on binary data. The average scores assigned were 3.55 (95% CI 3.49-3.60) for educational usefulness, 3.70 (95% CI 3.65-3.75) for clinical match, 3.49 (95% CI 3.44-3.55) for TA, and 2.34 (95% CI 2.28-2.40) for diagnosis difficulty, based on a 5-point Likert scale. Statistical analysis showed significant variability in content quality and relevance across the cases (P<.001 after Bonferroni correction). Participants suggested improvements in generating physical findings, using natural language, and enhancing medical TA. The thematic analysis highlighted the need for clearer documentation, clinical information consistency, content relevance, and patient-centered case presentations.

Conclusions: ChatGPT-4–generated medical cases written in Japanese possess considerable potential as resources in medical education, with recognized adequacy in quality and accuracy. Nevertheless, there is a notable need for enhancements in the precision and realism of case details. This study emphasizes ChatGPT-4's value as an adjunctive educational tool in the medical field, requiring expert oversight for optimal application.

XSL•FO
RenderX

(JMIR Med Educ 2024;10:e59133) doi:10.2196/59133

KEYWORDS

generative AI; ChatGPT-4; medical case generation; medical education; clinical vignettes; AI; artificial intelligence; Japanese; Japan

Introduction

The field of medical artificial intelligence (AI) has seen significant innovations, especially with the development of large language models such as ChatGPT, developed by OpenAI [1,2]. These technologies are being explored for applications across various items, including medical education [3-6], diagnostic assistance [7-9], patient health monitoring [10], and automated document creation [11,12]. However, the use of ChatGPT in health care raises serious concerns about the quality and accuracy of the information generated [13]. Accurate and reliable information is essential in health care, and inaccurate information can have harmful effects on patient health [6].

The importance of case-based learning in medicine has been well established [14]. This teaching approach is vital for medical students and health care professionals to extend their theoretical knowledge and understand the complexity and diversity of the clinical scenarios they will encounter. It fosters essential clinical reasoning and decision-making skills for accurate diagnoses and treatment plans [15,16]. Engaging with real cases helps learners develop the flexibility and adaptability needed in medical practice and encourages a more empathetic and human approach to care for patients and their families [17]. However, in teaching that involves handling actual clinical cases, creating scenarios requires a significant amount of labor, and conducting training using simulated patients can be costly, resulting in limited resources practically usable for education.

ChatGPT can easily create detailed and varied clinical scenarios that mirror actual situations, including disease types, symptom complexity, and patients' backgrounds, without incurring high costs [18]. If it becomes clear that ChatGPT can create clinical vignettes of a level suitable for educational use, it could reduce the labor and costs for educators, allowing learners to engage with a variety of cases. Exposure to a wide range of cases allows learners to deepen their understanding of specific pathologies and treatments, strengthen their clinical judgment and problem-solving skills, and enhance their overall clinical competence, from diagnosis to treatment planning.

Moreover, in Japan, recent reforms in physicians' work styles have mandated significant reductions in physicians' working hours, as part of a national effort to improve work-life balance and reduce instances of overwork. These changes, although beneficial for physician well-being, have created a pressing challenge for medical education, because less time is now available for traditional in-person training and supervision. This situation underscores the urgency of using AI technologies such as generative models to efficiently supplement and enhance the training of medical professionals. However, the extent to which AI can accurately replicate clinical information and scenarios remains a critical question.

To harness the potential of AI in enhancing medical education, this study investigated the educational utility of AI-generated clinical vignettes. These clinical vignettes have the potential to be used in educational scenarios, such as simulated patient interactions, clinical reasoning, and problem-solving exercises. The integration of such AI-generated materials into medical training programs poses significant questions regarding their quality and applicability in real-world educational settings. Thus, several evaluation items were set, and a questionnaire survey of physicians specialized in general internal medicine (GIM) or general medicine (GM) was conducted, to ensure these materials meet the requisite educational standards.

First, information quality (IQ) and information accuracy (IA) were assessed to determine if the clinical vignettes adhered to fundamental quality standards, which are crucial for maintaining educational integrity. Furthermore, the metric of education usefulness (EU) was introduced to explore the actual educational value of these AI-generated clinical vignettes in medical education. Recognizing the potential disparity between AI-generated clinical vignettes and actual clinical scenarios, clinical match (CM) was evaluated to confirm the relevance and applicability of these clinical vignettes in a realistic educational framework.

Furthermore, despite the potential accuracy of the content, the precision of medical terminology and the use of the Japanese language in AI-generated cases raised concerns, necessitating a separate evaluation of terminology accuracy (TA). In addition, diagnosis difficulty (DD) was assessed to understand how variations in the complexity of presented diagnoses might affect both the accuracy of the information and its overall educational value.

This study aimed to contribute to a fundamental understanding of clinical educational content created using ChatGPT-4 by evaluating the quality and EU of clinical vignettes generated in Japanese. The objective was to determine whether the ChatGPT-4–generated vignettes effectively simulate real-life clinical scenarios, thus potentially serving as a valuable resource for medical education and training.

Methods

Study Design

This was an exploratory, web-based, prospective, questionnaire-based survey conducted from January 8 to 28, 2024, using a convergent mixed methods design [19].

Selection of a Generative AI Model

ChatGPT-4 was selected for this study primarily because of its extensive use in previous medical AI research, unlike newer models such as Claude by Anthropic or Llama by Meta. This decision was driven by the availability of a robust body of

XSL•FO
RenderX

literature, enabling us to directly compare our findings with well-established studies in the field.

Medical Case Selection and Case Generation by ChatGPT-4

A total of 18 medical cases were created in Japanese using ChatGPT-4. The selection of cases was based on the 191 fundamental diseases listed in the 2022 revised *Model Core Curriculum for Medical Education* drafted by Japan's Ministry of Education, Culture, Sports, Science, and Technology [20]. These diseases were categorized into 19 areas by organ system, and 1 disease per area was selected for this study. If an area had multiple foundational diseases, the research team chose diseases that, based on patient history and physical findings, seemed likely to suggest a diagnosis. Since 1 area (breast diseases) did not have a foundational disease listed, a total of 18 cases were included (Table 1). Each case was created through a 4-step process. Initially, the output format was set, and ChatGPT-4 was instructed to generate patient histories and physical findings based on the diagnoses. Next, whether the generated cases were typical for the diagnoses based on patient history and physical findings was verified with ChatGPT-4. The third step involved checking for the inclusion of any nonexistent information. Finally, the accuracy of the terminology used in the patient history was assessed. At no point was the generation of specific findings or histories beyond the diagnosis directed (Multimedia Appendix 1).

Table 1. A total of 18 cases selected from the Model Core Curriculum for Medical Education in Japan.

Case	System	Disease name
Case 1	Blood, hematopoietic, and lymphatic system	Vitamin B_{12} deficiency anemia
Case 2	Nervous system	Parkinson disease
Case 3	Skin system	Cellulitis
Case 4	Musculoskeletal system	Spinal disc herniation
Case 5	Circulatory system	Acute aortic dissection
Case 6	Respiratory system	Pulmonary thromboembolism
Case 7	Digestive system	Acute appendicitis
Case 8	Renal-urinary system	Urinary stone disease
Case 9	Reproductive system	Benign prostatic hyperplasia
Case 10	Pregnancy and childbirth	Pregnancy-induced hypertension
Case 11	Pediatrics	Febrile seizures
Case 12	Endocrine, nutritional, and metabolic system	Hyperthyroidism
Case 13	Eye and visual system	Glaucoma
Case 14	Ear, nose, throat, and oral system	Meniere disease
Case 15	Psychiatric and psychosomatic disorders	Schizophrenia
Case 16	Immune system and allergy	Systemic lupus erythematosus
Case 17	Infectious diseases	Pneumonia
Case 18	Oncology	Cervical cancer

Study Participants

GIM or GM experts were recruited to evaluate the validity of the cases created with ChatGPT-4. Since the cases covered various specialties, the evaluators were physicians with cross-specialty knowledge in GIM or GM, all of whom had experience in medical education. The participants were recruited through mailing lists from the Japanese Society of Hospital General Medicine (JSHGM) [21], the Japan Primary Care Association (JPCA) [22], and the JHospitalist Network (JHN) [23], aiming to disseminate GIM education nationwide. Consent for participation was obtained through a Google Form.

Questionnaire and Survey Distribution

The survey, created in Google Forms, included questions about the responding physicians' backgrounds and questions evaluating the AI-generated cases. Background questions covered sex, year of medical license acquisition, specialty qualifications, hospital size, and work location. The evaluation of the generated cases focused on 6 aspects, which are IQ, IA, EU, CM, TA, and DD (Table 2). IQ and IA were assessed on a binary scale (yes or no), and EU, CM, TA, and DD were rated on a 5-point Likert scale (1: strongly disagree; 2: disagree; 3: neither agree nor disagree; 4: agree; 5: strongly agree). Binary responses were analyzed by converting yes to 1 and no to 0 (Multimedia Appendix 2).

XSL•FO
RenderX

Table 2. Contents and explanations of the 6 main questions of the questionnaire.

Item	Question	Measurement method	Scale explanation
Information quality	Do you think the medical history and physical findings provide enough quality information to recall the diagnosis?	Binary (yes or no)	Enough quality information forms the basis for accurate diagnosis process, thus answering yes or no clarifies the evaluator's stance on the quality of information.
Information accuracy	Is the information presented in the case accurate and without contradictions?	Binary (yes or no)	Accurate information ensures reliability and effectiveness in medical education, and answering yes or no clarifies the evaluator's stance on the accuracy of the information.
Education usefulness	Do you consider the quality of information in this case sufficient for educational purposes?	Likert scale (1-5)[a]	The usefulness of clinical vignettes in an educational context has a strong subjective element, so a Likert scale is used to capture finer impressions.
Clinical match	Does this case information reflect the medical history and physical findings you would encounter in clinical practice?	Likert scale (1-5)	Imitating realistic clinical scenarios enables learners to better prepare for situations they might face in the field. A variety of opinions and clinical experiences is important, so the Likert scale is adopted.
Terminology accuracy	Is the case information presented using appropriate medical terminology and expressions?	Likert scale (1-5)	Even if the information is accurate, the language may not be, which is why this item was included. A Likert scale is used to grade the level of language generated.
Diagnosis difficulty	How difficult do you find the diagnosis of this case?	Likert scale (1-5)	The difficulty of diagnosis serves as an indicator of the case's complexity. A Likert scale is used to more precisely assess the level of diagnostic difficulty.

[a]1: strongly disagree; 5: strongly agree.

Respondents who rated the IQ insufficient were asked to specify reasons from among 7 options (inadequate medical history, unclear medical history, incorrect medical history, inadequate physical examination findings, unclear physical examination findings, incorrect physical examination findings, and others), allowing for multiple responses. Those who found the IA insufficient provided reasons in free-text format.

To reduce survey fatigue, the questionnaire was divided into 3 parts, each covering 6 cases, with a week allocated for each part. Responses were collected over 3 weeks, with reminders sent to nonresponders to increase response rates.

Item-Based Data Analysis

For the 6 main items in the survey, response trends were evaluated by comparing response rates. The overall mean, SD, and 95% CI values were calculated for these rates across the 18 cases to gauge general trends and identify outliers. The calculation of mean and SD values is crucial because it helps understand the central tendency and variability of data, which supports the reliability and generalizability of the findings. All statistical analyses were performed using R (version 4.4.0; R Foundation for Statistical Computing).

Case-Based Data Analysis

For each case, the mean and 95% CI values were calculated for responses to the 6 main items to understand case-specific response trends. Chi-square tests were conducted on binary data (IQ and IA) to evaluate whether the observed differences between groups were significant. The chi-square test was specifically chosen for its efficacy in analyzing categorical data, and it was used to determine if variations in responses were due to chance or if they reflected true differences in the medical applicability of AI-generated cases.

For items scored on a 5-point Likert scale (EU, CM, TA, and DD), which typically do not adhere to a normal distribution, the Shapiro-Wilk test was first used to confirm the nonparametric distribution of the data. Since all 4 assessed items did not follow a normal distribution, nonparametric Kruskal-Wallis tests were conducted.

In instances of significant findings, post hoc analyses were carried out using Mann-Whitney *U* tests with Bonferroni correction to adjust for multiple comparisons. This approach allowed the effective assessment of the significance of differences in perceptions across different cases, highlighting specific cases that elicited higher or lower evaluations from medical professionals.

The Mann-Whitney *U* test was used for the Likert scale items due to its appropriateness in handling data that do not meet the assumptions of normality, thus providing a more accurate measure of central tendencies across diverse case scenarios. The significant outcomes derived from these tests provide information about the consistency and variation of clinical judgments among the cases, offering critical insights into the quality of AI-generated case presentations.

Medical Experience and Response Trends

To evaluate the relationship between medical experience and response trends, scatter plots were created, and regression lines were drawn. Linear regression analysis was conducted to

XSL•FO
RenderX

determine if there was a significant association of response trends with years of medical licensure, treating the year of licensure as an independent variable and the average score for each assessment indicator as a dependent variable, calculating the slope (regression coefficient), intercept, coefficient of determination (*R* value), and *P* value.

Qualitative Analysis

Thematic analysis of free-text responses regarding reasons for deeming IA insufficient was performed using ChatGPT-4 [24]. ChatGPT-4 processed the free-text survey results, generating a list of codes and corresponding quotations related to the research question. Themes and subthemes were then developed from these codes. Coding and theme development were validated and, if necessary, revised by 2 authors (HT and KS) using the results obtained from ChatGPT-4 [25]. The coauthors responsible for this task were physicians knowledgeable in convergent mixed methods research [26,27].

Ethical Considerations

This study was reviewed and approved by the Juntendo University School of Medicine Research Ethics Committee, approval E23-0245, on November 10, 2023.

Results

Respondents

The participants were recruited through mailing lists from the JSHGM (2325 members), the JPCA (4607 members), and the JHN (3965 members), gathering 73 respondents. All 73 respondents were confirmed to be suitable. Of these, 97% (71/73) completed all surveys. The 71 participants included 64 (90%) male respondents and 7 (10%) female respondents, with licensure years ranging from 1976 to 2017. By specialty, there were 61 GIM experts, 5 GM experts, and 5 experts with both qualifications. Hospital sizes were diverse, including 35 (49%) experts from hospitals with more than 500 beds, 16 (23%) from those with 201-500 beds, 11 (15%) from those with 101-200 beds, 3 (4%) from those with fewer than 100 beds, and 6 (8%) from clinics. Respondents came from 28 (60%) of the 47 prefectures in Japan, with 1 participant from outside Japan.

Item-Specific Questionnaires

Across the 18 cases, 76.8% (982/1278) of respondents found IQ sufficient, and 67.9% (868/1278) found IA sufficient. For the EU, 45.9% (587/1278) of respondents rated the cases as highly educational, with scores of 4. Another 15.1% (193/1278) awarded the highest score of 5. Conversely, around 19% (246/1278) expressed skepticism, giving scores of 1 or 2. CM saw a strong consensus, with over half of the participants (671/1278, 52.5%) rating the cases as highly relevant clinically, with scores of 4. Another 16% (201/1278) awarded the highest score of 5. The minority, about 13% (163/1278), gave scores of 1 or 2. TA was highly rated, with 46.2% (590/1278) of physicians expressing confidence in the accuracy of the language used (score 4), 58.7% (750/1278) expressing overall satisfaction with the TA (scores 4 and 5), and 19.6% (251/1278) providing lower scores (1 or 2). The responses to DD were more varied, since 59.5% (760/1278) of respondents found the cases relatively straightforward (scores 1 and 2), whereas higher difficulty levels (scores 4 and 5) were less frequently selected, at 13.2% (168/1278; Table 3).

Average ratings on the binary scale were 0.77 (95% CI 0.75-0.79) for IQ and 0.68 (95% CI 0.65-0.71) for IA. On the 5-point Likert scale, the averages were 3.55 (95% CI 3.49-3.60) for EU, 3.70 (95% CI 3.65-3.75) for CM, 3.49 (95% CI 3.44-3.55) for TA, and 2.34 (95% CI 2.28-2.40) for DD (Table 4).

XSL•FO
RenderX

Table 3. Percentage of all responses to 6 items (information quality, information accuracy, education usefulness, clinical match, terminology accuracy, and diagnosis difficulty).

Category and answer	Responses (N=1278), n (%)
Information quality	
Yes	982 (76.8)
No	297 (23.2)
Information accuracy	
Yes	868 (67.9)
No	410 (32.1)
Education usefulness	
1	28 (2.2)
2	218 (17.1)
3	252 (19.7)
4	587 (45.9)
5	193 (15.1)
Clinical match	
1	15 (1.2)
2	148 (11.6)
3	243 (19)
4	671 (52.5)
5	201 (15.7)
Terminology accuracy	
1	32 (2.5)
2	219 (17.1)
3	277 (21.7)
4	590 (46.2)
5	160 (12.5)
Diagnosis difficulty	
1	281 (22)
2	479 (37.5)
3	350 (27.4)
4	139 (10.9)
5	29 (2.3)

Table 4. Summary statistics of physician evaluations for AI-generated case scenarios. Information quality and information accuracy were evaluated on a binary scale of 0 or 1. Education usefulness, clinical match, terminology accuracy, and diagnosis difficulty were assessed using a Likert scale ranging from 1 to 5.

Item	Value, mean (95% CI)	Value, SD
Information quality (0 or 1)	0.77 (0.72-0.79)	0.42
Information accuracy (0 or 1)	0.68 (0.65-0.71)	0.47
Educational usefulness, Likert scale (1-5)[a]	3.55 (3.49-3.60)	1.01
Clinical match, Likert scale (1-5)	3.7 (3.65-3.75)	0.91
Terminology accuracy, Likert scale (1-5)	3.49 (3.44-3.55)	1
Diagnosis difficulty, Likert scale (1-5)	2.34 (2.28-2.40)	1.01

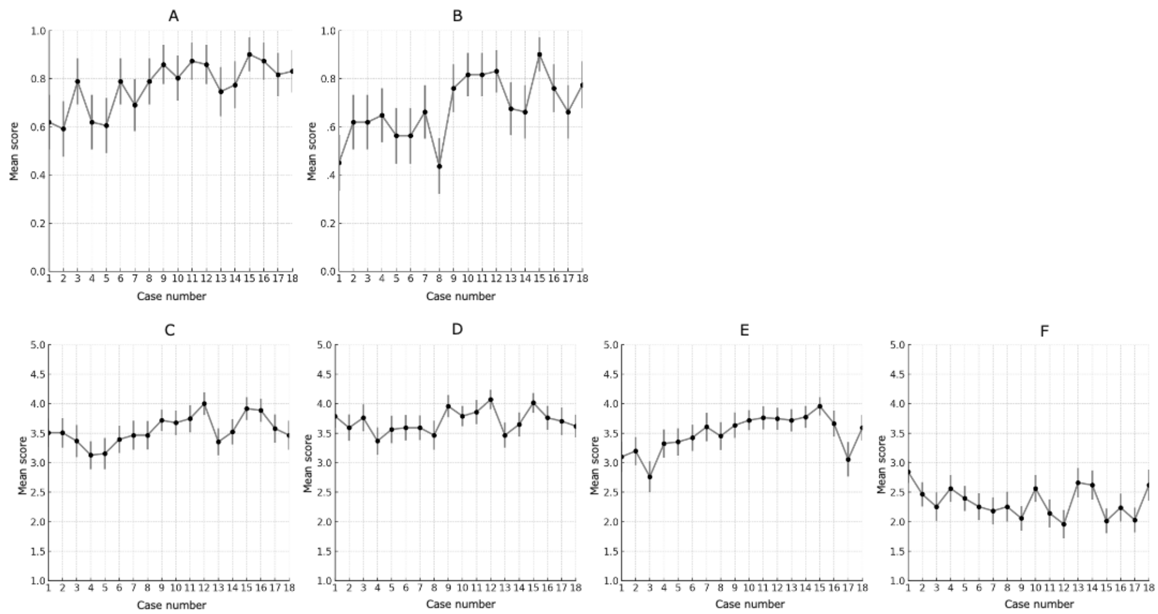
[a]1: strongly disagree; 5: strongly agree.

XSL•FO
RenderX

Case-Specific Questionnaires

Mean and 95% CI values for the 6 items (IQ, IA, EU, TA, CM, and DD) were analyzed for each of the 18 cases (Multimedia Appendix 3; Figure 1). Data analysis for each case showed significant differences in responses across all 6 items (*P*<.001 after Bonferroni correction) using chi-square tests for IQ and IA and Kruskal-Wallis tests for EU, CM, TA, and DD. Significance after correction for multiple comparisons varied, with 6 variations (6/153, 3.9% in all pairs) for IQ, 18 variations (18/153, 11.8% in all pairs) for IA, 9 variations (9/153, 5.9% in all pairs) for EU, 5 variations (5/153, 3.3% in all pairs) for CM, 22 variations (22/153, 14.4% in all pairs) for TA, and 16 variations (16/153, 10.5% in all pairs) for DD (Multimedia Appendix 4). Cases most frequently showing significant differences were case 15 (23 times), case 1 (22 times), case 12 (15 times), case 3 (13 times), and case 4 (10 times), with case 15 having the lowest *P* value combinations across all 6 main items.

Figure 1. Mean and 95% CI values per case for the 6 items: (A) information quality, (B) information accuracy, (C) education usefulness, (D) clinical match, (E) terminology accuracy, and (F) diagnosis difficulty. Information quality and information accuracy were evaluated on a binary scale of 0 or 1. Education usefulness, clinical match, terminology accuracy, and diagnosis difficulty were assessed using a Likert scale ranging from 1 to 5.



Comparing the highest and lowest average scores for each item showed significant differences: (1) IQ: 0.31 between case 15 (SD 0.90) and case 2 (SD 0.59; *P*<.001); (2) IA: 0.46 between case 15 (SD 0.90) and case 8 (SD 0.44; *P*<.001); (3) EU: 0.87 between case 12 (SD 4.00) and case 4 (SD 3.13; *P*<.001); (4) CM: 0.70 between case 12 (SD 4.07) and case 4 (SD 3.37; *P*<.001); (5) TA: 1.20 between case 15 (SD 3.96) and case 3 (SD 2.76; *P*<.001); and (6) DD: 0.89 between case 1 (SD 2.85) and case 12 (SD 1.96; *P*<.001).

Reasons for Insufficient IQ

Overall, 23.2% (297/1278) of respondents who scored the IQ insufficient cited incorrect patient history (23/1278, 1.8% of all cases) and incorrect physical findings (19/1278, 1.5% of all cases). Other responses indicated that, although AI-generated cases were useful for generating patient histories, improvements were needed in generating physical findings, the naturalness of language, and the accuracy of medical terminology (Multimedia Appendix 5).

Reasons for Insufficient IA

In exploring the reasons for inadequate IA as reported by respondents (410/1278, 32.1%), thematic analysis of open-ended responses identified 5 primary themes, which are (1) documentation clarity and precision, (2) consistency and reliability of clinical information, (3) appropriateness and contextual relevance, (4) comprehensiveness of diagnostic and treatment insights, and (5) patient-centered reporting (Table 5).

XSL•FO
RenderX

Table 5. Results of thematic analysis of the reasons for respondents' answers that information is insufficiently accurate.

Themes and subthemes	Quotes
Documentation clarity and precision	
Detail and specificity	I would like to know about gastrectomy.
Appropriate terminology	Do not use the expression "mesh-like sensation."
Consistency and reliability of clinical information	
Avoiding contradictions	Patient had a checkup at a nearby clinic, and no abnormalities were found. The details of the examination are unknown, but it is likely that a blood count was performed even if vitamins were not measured. Given that iron supplements were prescribed, it can be inferred that anemia was observed in the blood test. The statement "no abnormalities were found" is contradictory.
Ensuring accuracy in descriptions	The description "oral cavity: erythema of the tongue, normal dental health" is hard to understand.
Appropriateness and contextual relevance	
Contextual relevance to the patient's condition	It mentions obstetric history despite being about a male.
Practicality in clinical settings	I think the case itself is typical, but it seems unlikely that there would be time to conduct such a detailed physical examination on a patient experiencing severe chest pain accompanied by shortness of breath and cold sweats, and who has abnormally high blood pressure.
Comprehensiveness in diagnosis and treatment insights	
Diagnostic clarity	The name of the prescribed antibiotic is needed.
Logical treatment choices	It says an antianxiety medication was prescribed, but it is clearly a case of auditory hallucinations. It is unlikely any doctor would prescribe just an antianxiety medication in such a situation.
Patient-centered reporting	
Incorporating patient history and experience	It is strange to get a pneumococcal vaccine at 60 with only a history of high blood pressure.
Detailed symptom documentation	The main complaint is fatigue, but the details of the fatigue (changes in ADL[a], IADL[b], etc) are not described.

[a]ADL: activities of daily living.

[b]IADL: instrumental activities of daily living.

Documentation Clarity and Precision

Concerning documentation clarity and precision, issues were highlighted regarding the vagueness of specific information, such as details of surgeries and explanations of adjunct treatments. It was also noted that consistent use of medical terminology is demanded, with unclear or incorrect use of specialized terms leading to misunderstanding of information.

Consistency and Reliability of Clinical Information

For the consistency and reliability of clinical information, reported instances raised doubts about the trustworthiness of information, including contradictions in clinical findings and discrepancies in physical examinations. Medical documents should contain information relevant to the specific situations or conditions of patients, yet instances of unnecessary or irrelevant information were observed.

Appropriateness and Contextual Relevance

Concerns about appropriateness and contextual relevance were particularly noted in examples, such as the practicality of clinical tests in emergencies and the inclusion of information unrelated to patients' medical histories.

Comprehensiveness of Diagnostic and Treatment Insights

In diagnostic and treatment insights, comprehensive and detailed information is required. However, instances were observed where descriptions of specific medications were lacking or the rationale for treatment choices was questioned. It was pointed out that, in AI-generated case scenarios, detailed clinical data and clear justifications for treatment choices are crucial. Comprehensive documentation of patients' histories and experiences is essential for delivering patient-centered care, yet deficiencies were noted in the detailed reporting of specific symptoms or the consistency of patients' actions and histories, indicating insufficient patient-centered perspectives in reporting.

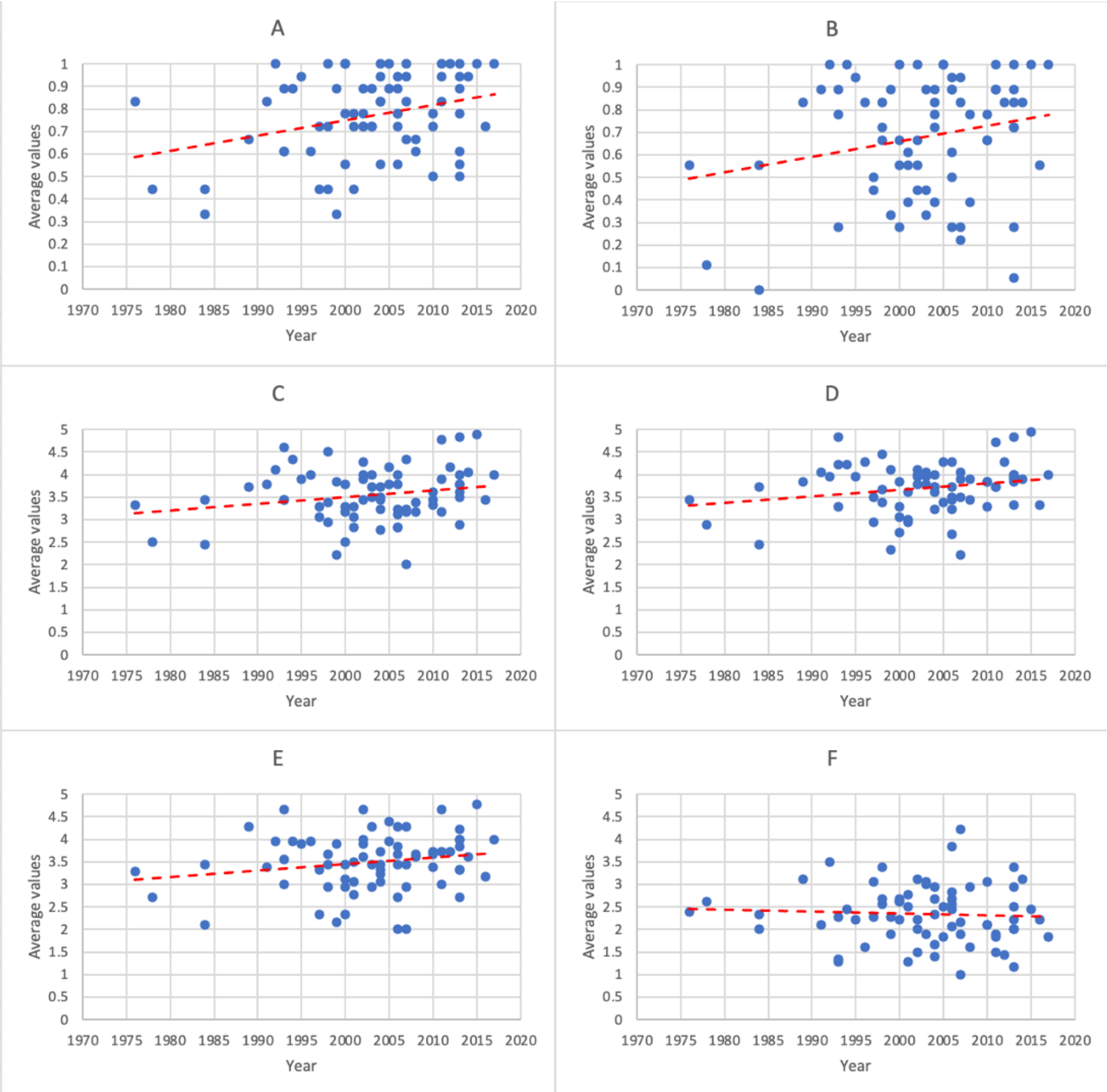
Medical Experience and Response Trends

In the scatter plots and regression lines of the years since obtaining a medical license, the 6 main items, IQ, IA, EU, CM, and TA, were all rated lower by physicians with longer careers and higher by those with shorter careers. DD, although mostly horizontal, was slightly inclined downward, indicating a trend where physicians with longer experience rated it more difficult, and those with shorter experience rated it easier (Figure 2). Linear regression analysis showed a significant association for IQ (P=.01). IA (P=.06), EU (P=.07), CM (P=.06), and TA (P=.10) did not show significant associations, although the P

XSL•FO
RenderX

values were low. The difficulty of diagnosis (*P*=.62) showed no relationship with the length of medical experience (Multimedia Appendix 6).

Figure 2. Scatter plot with regression lines showing the relationship between years since obtaining the medical license and average values for the 6 evaluated items: (A) information quality, (B) information accuracy, (C) education usefulness, (D) clinical match, (E) terminology accuracy, and (F) diagnosis difficulty.



Discussion

Principal Findings

In this study, Japanese medical cases generated by ChatGPT-4 were evaluated by GIM or GM experts. A high response rate (71/73, 97%) and a diverse participant demographic in terms of years of medical licensure, hospital size, and geographical location of affiliated organizations support the reliability of current findings. Evaluations across 6 key items (IQ, IA, EU, CM, TA, and DD) indicated that AI-generated medical cases possess a certain level of quality and accuracy suitable for use as clinical educational materials.

Overall, 76.8% (982/1278) of GIM or GM experts rated the IQ of the AI-generated cases as adequate. A very small percentage of cases were noted for having clear errors in medical history (23/1278, 1.8%) and physical examination findings (19/1278, 1.5%), with the total percentage of cases with clear errors in either item being only 3.3% (42/1278). This suggests that, despite some mistakes and lack of information, the majority of specialists found no significant issues with the quality of the generated cases, indicating that ChatGPT-4's case generation likely has fundamental reliability and accuracy. Similar to this study, research using ChatGPT-4 to generate 202 clinical vignettes in Japanese also involved evaluation by three physicians for medical and linguistic accuracy. It was found that 97% (196/202) of these clinical vignettes required some

XSL•FO
RenderX

modifications to be deemed usable, supporting these findings [28].

The 5-point Likert scale used rates “1” as the lowest and “5” as the highest evaluation, with “3” representing a neutral or undecided assessment. Each item rated as “4” suggests effectiveness. The scores for EU (3.55), CM (3.70), and TA (3.49) were between 3 and 4, indicating a level requiring modifications for practical educational use. In addition, the score for DD ranged between 2 and 3, suggesting it was easier than average. This implies that, with appropriate modifications, even relatively simple clinical vignettes could be effectively used for educational purposes.

The analysis of responses scored as insufficient IA (410/1278, 32.1%) showed that the AI-generated cases sometimes failed to provide medical information deemed necessary by GIM or GM experts, lacking specific information depending on the disease or not aligning with what the GIM or GM experts considered relevant clinical information. This suggests that, although ChatGPT-4 can generate disease information to some extent, it may not accurately represent actual clinical scenarios. Furthermore, instances of inappropriate use of Japanese language and expressions were also pointed out, highlighting the need for verification of the appropriateness and accuracy of medical information, representation of clinical scenarios, and use of Japanese language when using ChatGPT-4–generated cases for educational purposes.

The analysis of the responses to the 18 cases across the 6 key items showed significant variance through multigroup chi-square and Kruskal-Wallis tests. Specifically, case 15 was included in the combination that showed significance for all 6 items but was rated the third easiest in terms of diagnostic difficulty, and it ranked in the top 2 for the other 5 items, indicating a high evaluation. This suggests that case 15, a psychiatric case, was recognized from the medical history as a psychiatric disorder, and physical examination findings were not involved in the diagnosis. Among the reasons for inadequate IA ratings, the responses that the cases produced by ChatGPT-4 indicated that the history was accurate, but that the physical examination findings remained a challenge, supporting the reason why case 15 received a high rating.

Comparing cases with the highest and lowest mean values across all 6 items, not only were there significant differences across all items, but there were also substantial differences between the highest and lowest values. For instance, there was a 0.46 difference in the accuracy of information between case 15 and case 8, representing a significant difference, with 33 (46%) out of 71 respondents answering “yes.” Similarly, a significant difference of 1.20 was observed in TA between case 15 and case 3. These results suggest that, although AI-generated cases generally maintained a certain level of accuracy, there was significant variability in quality across cases for the 6 key items.

Analysis of scatter plots and regression lines of the relationship between years of medical licensure and response trends per case suggested a potential correlation between the length of medical practice and response tendencies. Not only IQ, which was significant on linear regression analysis, but also IA, EU, CM, and TA had low *P* values, suggesting that longer-practicing physicians developed more stringent criteria over time due to their increased knowledge and experience. It might also indicate that less experienced physicians are more receptive to new technologies and tools, valuing the utility of AI-generated cases more highly.

Limitations

This study focused on 18 cases of basic diseases and did not evaluate the maintenance of IQ and accuracy in complex cases. In addition, the evaluations were conducted by GIM or GM experts without obtaining assessments from specialists in various fields. Actual interaction and testing with learners are necessary to assess the usefulness of teaching clinical vignettes, but no interaction or testing with learners was conducted in this study. It is also important to note that this study was based on the use of ChatGPT-4 and that different outcomes might have been observed with other AI models, such as Claude by Anthropic or Llama by Meta. The evaluation structure was designed to ensure a comprehensive assessment of the AI-generated clinical vignettes. However, the absence of clear evaluative standards for respondents remains a limitation, potentially leading to variability in their interpretations and affecting the validity of the findings. The proportion of female respondents in this study was 9.5% (7/71). According to the 2020 data from the Ministry of Health, Labour, and Welfare, female physicians make up 22.8% (77,546/339,623) of all physicians in Japan, indicating a disproportionately higher number of male respondents in this study [29]. Finally, the questionnaire used was newly created and did not undergo a pilot test.

In this study, it was suggested that when creating clinical vignettes using the current ChatGPT-4, user corrections are necessary. Given the potential risks associated with the long-term use of AI, such as the homogenization of medical knowledge and the perpetuation of errors present in the training data, implementing this approach may be crucial in mitigating these risks.

Conclusions

This study showed that, although ChatGPT-4–generated medical cases contain minor mistakes, the likelihood of significant errors is low, and they possess a certain level of quality and accuracy of information. However, when evaluating individual cases, there is considerable variability in accuracy, underscoring the need for verification of the provision of appropriate medical information, representation of clinical scenarios, and accuracy of the Japanese language when using these AI-generated cases for educational purposes.

Acknowledgments

The authors would like to express their gratitude to the 71 general internists and family medicine specialists who participated in this study.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts used for generating clinical vignettes with ChatGPT-4.
[DOCX File , 21 KB - mededu_v10i1e59133_app1.docx]

Multimedia Appendix 2
Questionnaire form for case 1.
[DOCX File , 23 KB - mededu_v10i1e59133_app2.docx]

Multimedia Appendix 3
Physicians' assessments of artificial intelligence–generated cases, with mean (95% CI) values. Information quality and information accuracy were evaluated on a binary scale of 0 or 1. Education usefulness, clinical match, terminology accuracy, and diagnosis difficulty were assessed using a Likert scale ranging from 1 to 5.
[XLSX File (Microsoft Excel File), 11 KB - mededu_v10i1e59133_app3.xlsx]

Multimedia Appendix 4
Results of chi-square tests for information quality and information accuracy and of Kruskal-Wallis tests for education usefulness, clinical match, terminology accuracy, and diagnosis difficulty.
[DOCX File , 18 KB - mededu_v10i1e59133_app4.docx]

Multimedia Appendix 5
Summary of reasons for deeming the information quality insufficient.
[XLSX File (Microsoft Excel File), 9 KB - mededu_v10i1e59133_app5.xlsx]

Multimedia Appendix 6
Results of linear regression analysis.
[XLSX File (Microsoft Excel File), 9 KB - mededu_v10i1e59133_app6.xlsx]

References

1. Editorial. Will ChatGPT transform healthcare? Nat Med 2023 Mar 14;29(3):505-506. [doi: 10.1038/s41591-023-02289-5] [Medline: 36918736]
2. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023 Apr 5;6:1166014 [FREE Full text] [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]
3. Wong RS, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023 Nov 21;9:e47274 [FREE Full text] [doi: 10.2196/47274] [Medline: 37988149]
4. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ 2023 Dec 6;9:e52202 [FREE Full text] [doi: 10.2196/52202] [Medline: 38055323]
5. Shimizu I, Kasai H, Shikino K, Araki N, Takahashi Z, Onodera M, et al. Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study. JMIR Med Educ 2023 Nov 30;9:e53466 [FREE Full text] [doi: 10.2196/53466] [Medline: 38032695]
6. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
7. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
8. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 9;11:e48808 [FREE Full text] [doi: 10.2196/48808] [Medline: 37812468]
9. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam. Ann Ist Super Sanita 2023 Oct 24;59(4):267-270 [FREE Full text] [doi: 10.4415/ANN_23_04_05] [Medline: 38088393]

XSL•FO
RenderX

10. Islam MR, Urmi TJ, Mosharrafa RA, Rahman MS, Kadir MF. Role of ChatGPT in health science and research: a correspondence addressing potential application. Health Sci Rep 2023 Oct 11;6(10):e1625 [FREE Full text] [doi: 10.1002/hsr2.1625] [Medline: 37841943]
11. Nguyen J, Pepping CA. The application of ChatGPT in healthcare progress notes: a commentary from a clinical and research perspective. Clin Transl Med 2023 Jul;13(7):e1324 [FREE Full text] [doi: 10.1002/ctm2.1324] [Medline: 37394880]
12. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. Cureus 2023 Apr 14;15(4):e37589 [FREE Full text] [doi: 10.7759/cureus.37589] [Medline: 37197105]
13. Haze T, Kawano R, Takase H, Suzuki S, Hirawa N, Tamura K. Influence on the accuracy in ChatGPT: differences in the amount of information per medical field. Int J Med Inform 2023 Dec;180:105283. [doi: 10.1016/j.ijmedinf.2023.105283] [Medline: 37931432]
14. Khoiriyah U, Wijaya DP. Exploring problem-based learning (PBL) and case-based learning (CBL) in stimulating cognitive skills among medical students: analysis of verbal interaction. IIUM Medical Journal Malaysia 2022 Oct 01;21(4):45-52 [FREE Full text] [doi: 10.31436/imjm.v21i4.2066]
15. Brentnall J, Thackray D, Judd B. Evaluating the clinical reasoning of student health professionals in placement and simulation settings: a systematic review. Int J Environ Res Public Health 2022 Jan 14;19(2):936 [FREE Full text] [doi: 10.3390/ijerph19020936] [Medline: 35055758]
16. Bansal A, Singh D, Thompson J, Kumra A, Jackson B. Developing medical students' broad clinical diagnostic reasoning through GP-facilitated teaching in hospital placements. Adv Med Educ Pract 2020 May 25;11:379-388 [FREE Full text] [doi: 10.2147/AMEP.S243538] [Medline: 32547289]
17. Piot M, Attoe C, Billon G, Cross S, Rethans J, Falissard B. Simulation training in psychiatry for medical education: a review. Front Psychiatry 2021 May 21;12:658967 [FREE Full text] [doi: 10.3389/fpsyt.2021.658967] [Medline: 34093275]
18. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877 [FREE Full text] [doi: 10.2196/49877] [Medline: 37948112]
19. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research, 3rd edition. Los Angeles, CA: SAGE Publications; 2017.
20. Medical Education Model Core Curriculum Expert Research Committee. Model core curriculum for medical education in Japan. Ministry of Education, Culture, Sports, Science and Technology-Japan. 2022. URL: https://www.mext.go.jp/content/20230323-mxt_igaku-000028108_00003.pdf [accessed 2024-07-04]
21. Japanese Society of Hospital General Medicine. URL: http://hgm-japan.com [accessed 2024-04-02]
22. Japan Primary Care Association. URL: https://www.primarycare-japan.com [accessed 2024-04-02]
23. JHospitalist network. URL: http://hospitalist.jp [accessed 2024-04-02]
24. Perkins M, Roe J. Academic publisher guidelines on AI usage: a ChatGPT supported thematic analysis. F1000Res 2024 Jan 16;12:1398 [FREE Full text] [doi: 10.12688/f1000research.142411.2] [Medline: 38322309]
25. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Oct 2;16(1):1-13. [doi: 10.1177/1609406917733847]
26. Shikino K, Ide N, Kubota Y, Ishii I, Ito S, Ikusaka M, et al. Effective situation-based delirium simulation training using flipped classroom approach to improve interprofessional collaborative practice competency: a mixed-methods study. BMC Med Educ 2022 May 27;22(1):408 [FREE Full text] [doi: 10.1186/s12909-022-03484-7] [Medline: 35624492]
27. Kondo T, Miyachi J, Jönsson A, Nishigori H. A mixed-methods study comparing human-led and ChatGPT-driven qualitative analysis in medical education research. Nagoya J Med Sci (forthcoming) 2024;86(4).
28. Yanagita Y, Yokokawa D, Uchida S, Li Y, Uehara T, Ikusaka M. Can AI-generated clinical vignettes in Japanese be used medically and linguistically? medRxiv. Preprint posted on online on March 2, 2024 [FREE Full text] [doi: 10.1101/2024.02.28.24303173]
29. Table 2-42. number and average age of physicians, dentists and pharmacists by category of facility and occupation, sex and age group. Ministry of Health, Labour and Welfare. 2023. URL: https://www.mhlw.go.jp/english/database/db-hh/2-2.html [accessed 2024-07-04]

Abbreviations

AI: artificial intelligence
CM: clinical match
DD: diagnosis difficulty
EU: education usefulness
GIM: general internal medicine
GM: general medicine
IA: information accuracy
IQ: information quality
JHN: JHospitalist Network
JPCA: Japan Primary Care Association

XSL•FO
RenderX

JSHGM: Japanese Society of Hospital General Medicine
TA: terminology accuracy

Edited by B Lesselroth, G Eysenbach; submitted 04.04.24; peer-reviewed by RS Gomaa Mahmoud, Z Hou; comments to author 02.05.24; revised version received 22.05.24; accepted 27.06.24; published 13.08.24.

Please cite as:
Takahashi H, Shikino K, Kondo T, Komori A, Yamada Y, Saita M, Naito T
Educational Utility of Clinical Vignettes Generated in Japanese by ChatGPT-4: Mixed Methods Study
JMIR Med Educ 2024;10:e59133
URL: https://mededu.jmir.org/2024/1/e59133
doi:10.2196/59133
PMID:39137031

©Hiromizu Takahashi, Kiyoshi Shikino, Takeshi Kondo, Akira Komori, Yuji Yamada, Mizue Saita, Toshio Naito. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.08.2024. 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.

XSL•FO
RenderX

Original Paper

Understanding Health Care Students' Perceptions, Beliefs, and Attitudes Toward AI-Powered Language Models: Cross-Sectional Study

Ivan Cherrez-Ojeda[1,2*], MSC, MD; Juan C Gallardo-Bastidas[3*], DDS; Karla Robles-Velasco[1,2*], MD; María F Osorio[1,2*], MD; Eleonor Maria Velez Leon[4*], DDS; Manuel Leon Velastegui[5*], DDS; Patrícia Pauletto[6*], PhD; F C Aguilar-Díaz[7*], DDS; Aldo Squassi[8*], DDS; Susana Patricia González Eras[9*], DDS; Erita Cordero Carrasco[10*], DDS; Karol Leonor Chavez Gonzalez[11*], DDS; Juan C Calderon[1,2*], MD; Jean Bousquet[12,13,14*], PhD; Anna Bedbrook[14*], MD; Marco Faytong-Haro[2,15,16*], MA

[1]Universidad Espiritu Santo, Samborondon, Ecuador
[2]Respiralab Research Group, Guayaquil, Ecuador
[3]School of Dentistry, Universidad Católica de Santiago de Guayaquil, Guayaquil, Ecuador
[4]Facultad de Odontología Universidad Católica de Cuenca, Cuenca, Ecuador
[5]Universidad Nacional de Chimborazo, Riobamba, Ecuador
[6]Universidad de Las Américas (UDLA), Quito, Ecuador
[7]Departamento Salud Pública, Escuela Nacional de Estudios Superiores, Universidad Nacional Autónoma de México, Guanajuato, Mexico
[8]Universidad de Buenos Aires, Facultad de Odontologìa, Cátedra de Odontología Preventiva y Comunitaria, Buenos Aires, Argentina
[9]Universidad Nacional de Loja, Loja, Ecuador
[10]Departamento de cirugía y traumatología bucal y maxilofacial, Universidad de Chile, Santiago, Chile
[11]Universidad Politécnica Salesiana Sede Guayaquil, Guayaquil, Ecuador
[12]Institute of Allergology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
[13]Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Allergology and Immunology, Berlin, Germany
[14]MASK-air, Montpellier, France
[15]Universidad Estatal de Milagro, Cdla Universitaria "Dr. Rómulo Minchala Murillo", Milagro, Ecuador
[16]Ecuadorian Development Research Lab, Daule, Ecuador
[*]all authors contributed equally

Corresponding Author:
Ivan Cherrez-Ojeda, MSC, MD
Universidad Espiritu Santo
Km. 2.5 via Samborondon
Samborondon, 0901952
Ecuador
Phone: 593 999981769
Email: ivancherrez@gmail.com

Abstract

Background: ChatGPT was not intended for use in health care, but it has potential benefits that depend on end-user understanding and acceptability, which is where health care students become crucial. There is still a limited amount of research in this area.

Objective: The primary aim of our study was to assess the frequency of ChatGPT use, the perceived level of knowledge, the perceived risks associated with its use, and the ethical issues, as well as attitudes toward the use of ChatGPT in the context of education in the field of health. In addition, we aimed to examine whether there were differences across groups based on demographic variables. The second part of the study aimed to assess the association between the frequency of use, the level of perceived knowledge, the level of risk perception, and the level of perception of ethics as predictive factors for participants' attitudes toward the use of ChatGPT.

XSL•FO
RenderX

Methods: A cross-sectional survey was conducted from May to June 2023 encompassing students of medicine, nursing, dentistry, nutrition, and laboratory science across the Americas. The study used descriptive analysis, chi-square tests, and ANOVA to assess statistical significance across different categories. The study used several ordinal logistic regression models to analyze the impact of predictive factors (frequency of use, perception of knowledge, perception of risk, and ethics perception scores) on attitude as the dependent variable. The models were adjusted for gender, institution type, major, and country. Stata was used to conduct all the analyses.

Results: Of 2661 health care students, 42.99% (n=1144) were unaware of ChatGPT. The median score of knowledge was "minimal" (median 2.00, IQR 1.00-3.00). Most respondents (median 2.61, IQR 2.11-3.11) regarded ChatGPT as neither ethical nor unethical. Most participants (median 3.89, IQR 3.44-4.34) "somewhat agreed" that ChatGPT (1) benefits health care settings, (2) provides trustworthy data, (3) is a helpful tool for clinical and educational medical information access, and (4) makes the work easier. In total, 70% (7/10) of people used it for homework. As the perceived knowledge of ChatGPT increased, there was a stronger tendency with regard to having a favorable attitude toward ChatGPT. Higher ethical consideration perception ratings increased the likelihood of considering ChatGPT as a source of trustworthy health care information (odds ratio [OR] 1.620, 95% CI 1.498-1.752), beneficial in medical issues (OR 1.495, 95% CI 1.452-1.539), and useful for medical literature (OR 1.494, 95% CI 1.426-1.564; *P*<.001 for all results).

Conclusions: Over 40% of American health care students (1144/2661, 42.99%) were unaware of ChatGPT despite its extensive use in the health field. Our data revealed the positive attitudes toward ChatGPT and the desire to learn more about it. Medical educators must explore how chatbots may be included in undergraduate health care education programs.

(JMIR Med Educ 2024;10:e51757) doi:10.2196/51757

KEYWORDS

artificial intelligence; ChatGPT; education; health care; students

Introduction

Background

Artificial intelligence (AI) and machine learning technologies have transformed various sectors of contemporary society, including health care [1]. Among these developments, AI-powered large language models (LLMs) such as OpenAI's ChatGPT have shown significant promise in revolutionizing numerous aspects of health care services [2]. ChatGPT is a variation of OpenAI's language model that generates humanlike writing in a conversational situation [3].

As of January 2023, the population using ChatGPT exceeded 100 million [4]. While ChatGPT was not originally intended for application in health care settings, it is possible that some of these users comprise students or health care practitioners [5]. Consequently, the insights derived from their interactions with ChatGPT may offer valuable information in patient communication, information management, electronic health records, diagnostics, decision-making assistance, and, potentially, therapeutic interventions [6].

LLMs have shown to be beneficial to health care provision [7]. ChatGPT has demonstrated strong, human-level performance supporting decision-making, data management, and patient education in many specialties, such as internal medicine, surgery, and oncology [8,9]. The upcoming generations of health professionals comprise students who undergo training in conditions with plenty of easily accessible technology resources [10]. Some students may assume roles as directors of health institutes, whereas others may engage in research or work as health care professionals. Nevertheless, it is crucial to recognize that the quality of education received will directly impact the caliber of professionals in the future. Consequently, it is imperative to understand the interests that occupy their thoughts concerning the use of tools such as LLMs. This comprehension is essential in determining how these tools can either enhance or fail to enhance their academic and educational competencies as well as their professional application soon after [11].

Objectives

In light of this, the primary aim of our study was to assess the frequency of ChatGPT use, the perceived level of knowledge, the perceived risks associated with its use, and the ethical issues, as well as attitudes toward the use of ChatGPT in the context of health care education. The second part of the study aimed to assess the association between the frequency of use, the level of perceived knowledge, the level of risk perception, and the level of perception of ethics as predictive factors for participants' attitudes toward the use of ChatGPT.

Methods

Design

This study used a cross-sectional survey among students of health care–related college programs across the Americas to assess their perceptions, attitude, patterns of use, and further learning regarding ChatGPT. This study was conducted from May to June 2023 across all participating countries.

Sample Size Calculation

The sample size for this study was calculated using the following formula: $n = (\text{Estimated Design Effect Factor} \times Np[1 - p])/(d2/Z2_{1-\alpha/2} \times [N - 1] + p \times [1 - p])$. Accounting for a population size of 1 million, a hypothetical frequency of 50% with a 5% margin of error, and a confidence level of 99.99%, the calculated sample size was 1512.

XSL•FO
RenderX

Recruitment

Our study focused on individuals aged >18 years enrolled in diverse health care–related college programs such as medicine, nursing, dentistry, nutrition and dietetics, and medical laboratory science. Through a convenience sampling method, we gathered responses from 2661 participants. We adopted a multifaceted recruitment approach to ensure a varied sample of health care students. We reached out to potential participants through email, student networks, social media, on-campus events, academic institutions, and student associations.

We expanded our sample by including universities across the Americas, specifically in Argentina, Mexico, Colombia, Chile, and Ecuador. By disseminating study links to these institutions, we achieved a diverse representation of health care students from different countries and fields.

Bias

To minimize potential biases, we adopted a comprehensive recruitment strategy targeting a wide range of universities across the Americas, hence reducing selection bias. Response bias was mitigated by conducting anonymous surveys, encouraging honest responses from the participants. In addition, to limit information bias, the survey questions were designed to be straightforward and used standardized Likert-scale responses.

Questionnaire

The questionnaire was developed following the recommendations by Passmore et al [12] and Eysenbach [13]. A steering committee composed of 4 experts and heads from 4 specialized centers worldwide reviewed the literature and developed the survey items, which integrated all constructs to be assessed. The first section of the survey gathered the demographics and medical education of the participants. The second section of the survey aimed to assess the students' perceptions, attitudes, patterns of use, and further learning regarding ChatGPT.

The perception domain was further categorized into *self-perceived knowledge*, *ethics*, and *beliefs of perceived risk* subdomains. The subdomain of self-perceived knowledge was assessed on a 5-point Likert scale ranging from 1 (*no knowledge*) to 5 (*superior knowledge*). The scale of self-perception of knowledge about ChatGPT was recategorized as follows: (1) "No knowledge"—this category included participants who either answered "No" to the question "Have you heard of ChatGPT before?" or selected "No Knowledge" in response to the question "How would you rate your knowledge of ChatGPT and its applications in health care?"; (2) "Minimal knowledge"—participants falling into this category included those who answered with options such as "Minimal" or "Basic knowledge" on the Likert scale; and (3) "Adequate knowledge"—this category encompassed participants who selected options such as "Adequate" or "Superior" knowledge on the Likert scale.

The *ethical perception* subdomain featured 3 items, which respondents were asked to score on a 5-point Likert scale ranging from 1 (*totally unethical*) to 5 (*totally ethical*). The *beliefs of perceived risk* subdomain had 3 items, which respondents were asked to score on a 5-point Likert scale (1 [*strongly disagree*] to 5 [*strongly agree*]). The attitude domain included 5 statements reflecting evaluations and opinions on ChatGPT. On a 5-point Likert scale, respondents were asked to score these statements (1 [*strongly disagree*] to 5 [*strongly agree*]). The domain of further learning consisted of 4 questions inquiring as to whether respondents wanted to learn more about ChatGPT. Respondents were asked to choose the resources or educational materials that they believed would be the most beneficial in learning about ChatGPT and its potential applications in health care. Those who did not want to learn more about ChatGPT were requested to explain their reasons.

In total, 2 questions assessed the "Pattern of Use" domain: one assessing the frequency of use using a 5-point Likert scale ranging from 1 (*less than once a month*) to 5 (*more than once a day*) and one assessing the applications of ChatGPT in health care settings with a choice of 8 alternatives.

The questionnaire is shown in Multimedia Appendix 1. A pilot study was performed by the steering committee with colleagues and a sample of 20 students. After drafting the survey, it was distributed to the study population in May and June 2023. The survey was available in English and Spanish.

Ethical Considerations

Ethics approval was obtained from the Human Research Ethics Committee from Ecuador with approval HCK-CEISH-2022-006. All participants provided informed consent to take part in the study. They were informed about the purpose of the research, their rights as participants, and the voluntary nature of their participation. We ensured the privacy and confidentiality of participant data throughout the study. The survey responses were anonymized, and no personally identifiable information was collected. No compensation was provided to participants for their involvement in the study. It is important to note that the approval obtained from the Human Research Ethics Committee in Ecuador was deemed sufficient to expand recruitment to all Latin American countries included in the study. This decision was made based on the similarity of ethical standards and regulations across these countries, as well as the collaborative nature of the research conducted within the region.

Variables

Demographic Variables

The demographic variables selected for this study are pivotal for examining the diversity of health care students' attitudes toward using ChatGPT. They are used in both the descriptive (for sample composition purposes) and regression (as control variables) tables. Each variable is thoughtfully coded to capture the nuanced differences among the survey participants, facilitating a detailed analysis of their responses.

Age was recorded as a continuous variable. This allowed for precise analysis of trends across different age groups, helping identify whether younger students are more adept and receptive to AI technologies such as ChatGPT compared to their older counterparts [14].

Gender was categorized into several groups: male, female, nonbinary or third gender, prefer not to say, and other. This

XSL•FO
RenderX

categorization ensured that the study could address and respect the diversity of gender identities. It allowed for an analysis of whether perceptions of ChatGPT vary significantly across different gender groups, which could indicate targeted approaches for technology integration based on gender-specific preferences or concerns [15].

The type of university was divided into public and private. This classification helped investigate whether the institutional context influences students' familiarity with and attitudes toward ChatGPT. Differences in resources, exposure to technology, and educational priorities between public and private universities might contribute to distinct attitudes observed among the students from these institutions [16].

Region was split into Central America and South America. By distinguishing between these 2 regions, the study could explore regional differences that might affect students' acceptance and use of AI technologies. Such differences could stem from varying levels of technology integration in health care education, regional cultural attitudes toward technology, and economic factors [17].

The field of study was specified as medicine, nursing, nutrition, dentistry, therapy, psychology, pharmacology, and other. This detailed categorization allowed the study to determine whether students in certain fields are more likely to perceive ChatGPT as a beneficial tool [18]. For instance, fields requiring up-to-date information and quick data retrieval might show higher appreciation for AI assistance compared to fields that are more focused on personal patient interactions [19].

Outcome Variables

The outcomes of this study focused on the health care students' attitudes toward using ChatGPT quantified through a series of statements. These statements were designed to capture various dimensions of the perceived utility and reliability of ChatGPT in health care contexts. Each outcome variable was measured using Likert scales ranging from "strongly disagree" to "strongly agree" in order to have a granular view of respondents' attitudes and, through detailed statistical analysis, assess trends and influences on these perceptions.

Specifically, the outcomes assessed were (1) "I think that ChatGPT makes my job easier."—this statement evaluated the perceived practical utility of ChatGPT in simplifying tasks within health care settings; (2) "ChatGPT can be beneficial in health care settings."—this statement assessed broader benefits, looking at whether students believe ChatGPT can positively impact health care environments; (3) "ChatGPT provides trustworthy health care information or guidance."—this statement measured trust in the accuracy and reliability of the information provided by ChatGPT; (4) "ChatGPT is a useful tool when I need to search for information on specific medical questions."—this statement evaluated the usefulness of ChatGPT as a resource for specific, actionable medical inquiries; and (5) "ChatGPT is a useful tool when I need to search for medical literature."—this outcome explored the utility of ChatGPT in supporting academic and professional research within medical fields.

Focusing on these specific attitudes toward using ChatGPT helps us understand how health care students perceive the integration of AI into their practices. The statements target various dimensions of AI's role—from enhancing efficiency and providing reliable information to supporting academic research—highlighting areas where ChatGPT could be particularly impactful or face resistance. This nuanced approach not only sheds light on current acceptance levels but also pinpoints areas where further education or system improvements might increase trust in and the utility of AI applications within health care environments.

Exposure (Predictor) Variables

Overview

In this study, several key predictor variables were used to explore the factors influencing health care students' attitudes toward using ChatGPT. These predictors included knowledge of ChatGPT, perceptions of risk, ethical considerations, and the frequency of use of ChatGPT. A detailed overview of each predictor is presented in the following sections.

Knowledge About ChatGPT

For the regression model, this predictor measured the participants' self-reported knowledge about ChatGPT, assessing their understanding of its functionalities and potential applications in health care. It was quantified using a 5-point Likert scale ranging from 1 (*no knowledge*) to 5 (*superior knowledge*). The understanding of ChatGPT's functionalities and potential applications is crucial as it directly influences how students perceive its utility and limitations [20]. Higher levels of knowledge might correlate with more positive attitudes as students are better able to appreciate the benefits and manage the limitations of AI in health care [21].

Beliefs of Perceived Risk

This variable is a composite score derived from the median of the agreement on a 5-point scale with three specific statements assessing perceived risks associated with AI: (1) "I think my job could be replaced in the future because of AI," (2) "In the future, ChatGPT (or some similar technology) will play an even more important role in my job," and (3) "Using AI like ChatGPT in clinical practice raises ethical concerns."

Perceptions of risk are vital to consider because they shape how students weigh the advantages against the potential drawbacks of using AI technologies [22]. Concerns about job security, the increasing role of AI in health care, and ethical implications could negatively influence their attitudes toward ChatGPT, making it essential to analyze how these perceptions impact their overall acceptance [23].

Ethics

The ethical factors were assessed by calculating the median score based on the replies' level of agreement, ranging from *totally ethical* to *totally unethical* on a 5-point scale, to the following three statements that address ethical concerns about using AI in health care: (1) "Revising the language of a scientific manuscript?" (2) "Writing text in a scientific manuscript?" (3) "The sole source of information for the clinical practice?"

XSL•FO
RenderX

Ethical considerations are paramount in the adoption of any new technology, especially in sensitive fields such as health care. Evaluating how students perceive the ethical dimensions of using ChatGPT for tasks such as manuscript writing or as a clinical information source can provide insights into the ethical acceptability of AI tools in professional health care practices [24].

Frequency of Use

The frequency of use was directly measured by asking participants how often they used ChatGPT, with options on a 5-point Likert scale ranging from 1 (*less than once a month*) to 5 (*more than once a day*). The frequency of use is indicative of both familiarity and dependency on the technology. Regular use of ChatGPT might suggest greater comfort and perceived utility, possibly leading to more favorable attitudes [25]. Conversely, infrequent use might indicate skepticism or perceived inadequacies in the technology's ability to meet professional needs [26].

Statistical Analysis

Descriptive Analysis

In the descriptive analysis, we examined the demographic information and survey responses of the participants. This part of the analysis comprised 2 main components. First, the demographic characteristics of the participants were assessed and stratified according to the participants' self-rated knowledge of AI. These categories of knowledge were "No knowledge," "Minimal Knowledge," and "Adequate Knowledge." Demographic variables such as age, gender, type of university (public vs private), region, and major were analyzed across these knowledge strata. Statistical significance for differences across the knowledge categories was tested using a chi-square test for categorical variables and an ANOVA for continuous variables, with a *P* value of <.05 indicating statistical significance.

In the second part of the descriptive analysis, given the ordinal nature of the variables, we assessed the range, median, and IQR of scores for each item in the survey. The survey items were grouped into 3 primary domains: perception, ethics, and attitudes, with the perception domain further divided into 2 subdomains: knowledge and beliefs of perceived risk. In addition, the frequency of use of ChatGPT for various tasks was analyzed. Each item was assessed on a Likert scale ranging from 1 to 5 except for the use tasks, which were reported as percentages. The total median scores for each domain and subdomain were calculated and included in the report. This analysis helped provide a clear picture of the participants' perceptions, ethical considerations, attitudes, and use habits related to ChatGPT.

Regression Analysis

Our analysis of the impact of perception scores on attitude variables involved the use of multiple ordinal logistic regression models. Each model evaluated the attitudes of health care students toward the use of ChatGPT, with individual attitude statements serving as dependent variables. These statements included perceptions of ChatGPT in terms of its ease of use, its utility in health care settings, the trustworthiness of its health information, its usefulness in finding answers to specific medical questions, and its helpfulness in searching for medical literature.

For each attitude statement, three perception subdomains were considered as independent variables: knowledge, beliefs of risk, and ethical considerations. The coefficient, SE, 1-tailed *t* test, and *P* value were all calculated for each perception subdomain under each attitude statement. All models were adjusted for control variables, including gender, whether the institution attended was private or public, the field of study, and the country of the student. All analyses were carried out using Stata (version 18.0; StataCorp).

Missing Data

Although our web-based survey, which required complete responses, effectively eliminated the need to handle missing data, the self-selecting nature of web-based surveys could introduce some bias. Participants more comfortable with or having better access to technology might be overrepresented. However, the completeness of the data set ensured the accuracy of our analysis and the robustness of the findings.

Sensitivity Analyses

In the analytical procedure, we used a set of 20 ordinal logistic regression models. Importantly, SEs were clustered by country to account for potential intracountry correlations. The proportional odds assumption, pivotal for the conventional interpretation of ordinal logistic regression, was violated in half (10/20, 50%) of these models. This breach was primarily attributed to the coefficient of the main predictor in the affected models.

To address this violation and offer a more fitting statistical representation, we used the partial proportional odds model for instances in which the main predictor was unconstrained. Even after this adjustment, our results suggested that the interpretation did not differ significantly from models in which every coefficient was constrained, even when faced with assumption violations. Due to this slight difference in interpretation, and in the interest of consistency, we chose to present the outcomes of all models using ordinal logit with all constraints.

For further refinement of our analysis, and to account for potential clustering effects, we introduced random-intercept and slope models. In this setup, schools were treated as nested entities within countries. This multilevel modeling approach produced results that differed only minimally from those of our initial models, underscoring the reliability of our findings.

Results

Demographic Information

This study included 2661 health care students in total. Most were female (n=1764, 66.29%), in dentistry (n=1466, 55.09%), from South America (n=2442, 91.77%), and from private universities (n=1836, 68.99%), as indicated in Table 1. The average age was 21.65 (SD 3.42) years. Multimedia Appendix 2 provides a full overview of the sample's demographics.

XSL•FO
RenderX

Table 1. Demographic information (N=2661).

Variable	No knowledge (n=1142)	Minimal knowledge (n=578)	Adequate knowledge (n=941)	Total	*P* value
Age (y), mean (SD)	22.01 (3.41)	21.45 (3.81)	21.34 (3.12)	21.34 (3.12)	<.001
Gender, n (%)					<.001
Male	277 (31.66)	203 (23.2)	395 (45.14)	875 (32.88)	
Female	858 (48.61)	371 (21.02)	536 (30.37)	1765 (66.32)	
Nonbinary or third gender	2 (25)	2 (25)	4 (50)	8 (0.01)	
Prefer not to say	2 (20)	2 (20)	6 (60)	10 (0.01)	
Other	3 (100)	0 (0)	0 (0)	3 (0.01)	
Type of university, n (%)					<.001
Public	397 (48.18)	175 (21.24)	252 (30.58)	824 (30.96)	
Private	745 (40.56)	403 (21.94)	689 (37.51)	1837 (69.04)	
Region, n (%)					.004
Central America	116 (53.21)	43 (19.72)	59 (27.06)	218 (8.2)	
South America	1026 (42)	535 (21.9)	882 (36.1)	2443 (91.8)	
Major, n (%)					<.001
Medicine	212 (23.85)	223 (25.08)	454 (51.07)	889 (33.4)	
Nursing	36 (73.47)	4 (8.16)	9 (18.37)	49 (1.84)	
Nutrition	24 (41.38)	13 (22.41)	21 (36.21)	58 (2.17)	
Dentistry	777 (53)	286 (19.51)	403 (27.49)	1466 (55.09)	
Therapy	18 (40.91)	7 (15.91)	19 (43.18)	44 (1.65)	
Psychology	19 (42.22)	19 (42.22)	7 (15.56)	45 (1.69)	
Pharmacology	13 (86.67)	1 (6.67)	1 (6.67)	15 (0.56)	
Other	43 (45.26)	25 (26.32)	27 (28.42)	95 (3.57)	

Perception of Knowledge, Beliefs of Perceived Risks, and Ethics

Among all participants, 42.92% (1142/2661) did not know about ChatGPT. Male students knew more about ChatGPT than female students (598/875, 68.3% vs 907/1765, 51.39%, *P*<.001). Most of the group of participants who had adequate knowledge of ChatGPT were from South America. With the exception of medicine and therapy students, most health care students were unaware of ChatGPT (Table 1).

Table 2 presents findings from our survey assessing participants across multiple domains related to their perception, attitudes, and use of AI, with a particular focus on ChatGPT. In the "Perception" domain, participants were queried about their knowledge, with scores ranging from 1 to 5. They reported a median score of 2.00, which implies a minimal knowledge of ChatGPT. Delving into beliefs about the perceived risk linked to AI, respondents "somewhat agreed" that using ChatGPT raises potential ethical concerns and that AI will play a more important role in their jobs in the future.

Moving to the "Ethics" domain, participants considered the use of ChatGPT for writing text within a scientific manuscript and using ChatGPT as the sole information source for clinical practice "neither ethical nor unethical." In terms of "Attitudes" toward ChatGPT, the median score was 4.00 among all statements, showing that most participants "somewhat agreed" with the advantages and utility of ChatGPT in health care contexts.

The "Use" domain had respondents spotlight the frequency with which they engaged with ChatGPT, reporting a median score of 2.00 (once a month) on a scale of 1 to 5, with an IQR of 1.00-3.00. Regarding distinct tasks, most participants used ChatGPT for homework support (1078/1519, 70.97%), research paper writing (637/1519, 41.94%), and medical and health care education (349/1519, 22.98%); for more information, see Multimedia Appendix 3.

XSL•FO
RenderX

Table 2. Range, median, and IQR of the scores of the survey domains[a].

Items	Scores, median (IQR; range)
Perception	
Knowledge	2.00 (1.00-3.00; 1-5)
Beliefs of perceived risk	
"I think my job could be replaced in the future because of AI."	3.00 (1.50-4.50; 1-5)
"In the future, ChatGPT (or some similar technology) will play an even more important role in my job."	4.00 (3.50-4.50; 1-5)
"Using AI like ChatGPT in clinical practice raises ethical concerns."	4.00 (3.50-4.50; 1-5)
Total median score	3.28 (2.76-3.81; 1-5)
Ethics	
Revising the language of a scientific manuscript	2.00 (1.00-3.00; 1-5)
Writing text in a scientific manuscript	3.00 (2.50-3.50; 1-5)
The sole source of information for clinical practice	3.00 (2.00-4.00; 1-5)
Total median score	2.61 (2.11-3.11; 1-5)
Attitudes	
I think that ChatGPT makes my job easier.	4.00 (3.48-4.52; 1-5)
ChatGPT can be beneficial in health care settings.	4.00 (3.00-5.00; 1-5)
ChatGPT provides trustworthy health care information or guidance.	4.00 (3.50-4.50; 1-5)
ChatGPT is a useful tool when I need to search for information on specific medical questions.	4.00 (3.00-5.00; 1-5)
ChatGPT is a useful tool when I need to search for medical literature.	4.00 (3.00-5.00; 1-5)
Total median score	3.89 (3.44-4.34; 1-5)
Use	
Frequency of use	2.00 (1.00-3.00; 1-5)

[a]n=1519, which corresponds to students who were aware of ChatGPT.

Further Learning Regarding ChatGPT

Of the participants willing to learn more about ChatGPT, 67.98% (1809/2661) wanted to learn about the applications of ChatGPT in particular cases of medical practice, followed by homework support and understanding the benefits and limits of ChatGPT (Table 3). Less than 30% (745/2661, 27.99%) were interested in learning about "data privacy and security measures" and "ethical considerations." Participants found that the most interesting educational materials for learning more about this topic were research articles and case studies (426/2661, 69.16%), internet-based demonstrations or hands-on experience (1301/2661, 48.91%), workshops or conferences (1211/2661, 45.52%), and webinars or web-based courses (968/2661, 36.37%).

Table 3. Further learning domain showing aspects of ChatGPT and its applications in health care that students are more interested in learning about (N=2661).

Aspect	Students, n (%)
Specific use cases in medical practice	1832 (68.85)
Academic homework support	1241 (46.6)
Potential benefits and limitations	1158 (43.51)
Integration with existing health care systems	1078 (40.51)
Data privacy and security measures	755 (28.38)
Ethical considerations	750 (28.2)
Other	37 (1.39)

The main reasons for the 16.49% (439/2661) of participants who did not want to learn more about ChatGPT were lack of time (1234/2661, 46.37%); preference to consult with peers, mentors, and teachers (617/2661, 23.19%); not enough knowledge about these technologies (492/2661, 18.5%); and lack of relevance to their medical specialty (335/2661, 12.59%; Table 4).

XSL•FO
RenderX

Table 4. Reasons for lack of interest in learning more about ChatGPT and its potential applications in health care (N=2661).

Reason	Students, n (%)
Lack of time	1233 (46.37)
I prefer to consult with my peers, mentors, and teachers	617 (23.19)
Not enough knowledge of these technologies	492 (18.5)
Lack of relevance to my medical specialty	336 (12.65)
Skepticism about the benefits of AI[a] in health care	299 (11.24)
Already overwhelmed with existing medical knowledge and skills	249 (9.37)
Difficulty or discomfort using computer technology	155 (5.85)
Other	143 (5.39)

[a]AI: artificial intelligence.

Association Between Perception (Knowledge, Belief, and Ethics) and Frequency of Use and Attitude

The ordinal logistic regression analysis (Tables 5 and 6) illustrates the relationship between predictors such as knowledge, beliefs about risks, ethics, frequency of use, age, gender, institution type, and professional background and their impact on health care students' perceptions of ChatGPT's utility.

An enhanced understanding of ChatGPT consistently showed a positive correlation with more favorable views across all outcomes. For instance, as knowledge increased, the odds of believing that ChatGPT makes one's job easier went up, with odds ratios (ORs) ranging from 1.259 (95% CI 1.047-1.513) to 1.468 (95% CI 1.289-1.672). This trend persisted across other perceptions, such as ChatGPT's potential benefits in health care settings and its trustworthiness in providing health care information.

Beliefs about risk followed a distinctive pattern. Those with heightened risk beliefs felt that ChatGPT made their job easier and could play a beneficial role in health care settings, including obtaining information on medical questions and as a tool for searching medical literature, as evidenced by ORs of 2.040 (95% CI 1.765-2.358), 1.106 (95% CI 1.031-1.186), 1.179 (95% CI 1.110-1.255), and 1.138 (95% CI 1.076-1.203), respectively. This finding suggests that recognizing potential risks does not negate belief in the tool's utility. Ethical considerations played a significant role. Students with higher ethical concerns perceived ChatGPT's potential in health care more favorably. The ORs for these associations were notable, especially in the context of trustworthiness and specific medical queries (OR 1.620, 95% CI 1.498-1.752).

The frequency of ChatGPT use was a significant determinant. Regular users were more optimistic about its utility, which was evident across all outcomes, such as its benefits in health care (OR 1.540, 95% CI 1.420-1.670) and its efficacy in searching for medical information (OR 1.438, 95% CI 1.311-1.577).

Age influenced perceptions. Older individuals generally had a higher OR across the outcome variables, suggesting a more positive perception of ChatGPT's utility in their profession. Gender-based analysis revealed that female individuals, compared to male individuals, were generally more likely to believe that ChatGPT can help in their job. However, perceptions varied when it came to broader benefits in health care and other outcomes. Those identifying as nonbinary or third gender or those who preferred not to specify their gender showcased diverse perceptions, sometimes differing from those of both male and female individuals.

Institutional type and major played a role. Individuals from private institutions, compared to their public institution counterparts, had varied perceptions. Students from nursing and nutrition exhibited unique outlooks on ChatGPT, highlighting the influence of professional background on shaping perceptions.

XSL•FO
RenderX

Table 5. Estimates from ordinal logistic regression models for the effect of perception scores on attitude variables[a].

Variables and outcomes	Predictor: knowledge					Predictor: beliefs of risk				
	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.
Knowledge (1-5), OR[b] (SE)	1.259[c] (1.047-1.513)	1.468[d] (1.289-1.672)	1.480[d] (1.357-1.614)	1.448[d] (1.400-1.498)	1.298[d] (1.134-1.486)	—[e]	—	—	—	—
Beliefs of risk median (1-5), OR (SE)	—	—	—	—	—	2.040[d] (1.765-2.358)	1.106[c] (1.031-1.186)	1.062[f] (1.013-1.113)	1.179[d] (1.110-1.255)	1.138[d] (1.076-1.203)
Age in years, OR (SE)	1.032[c] (1.010-1.054)	1.036[d] (1.014-1.058)	1.018 (0.993-1.044)	1.028[f] (1.004-1.053)	1.043[d] (1.024-1.063)	1.033[f] (1.006-1.059)	1.049[d] (1.029-1.069)	1.015[f] (1.003-1.027)	1.042[d] (1.025-1.060)	1.041[c] (1.012-1.071)
Female (reference: male), OR (SE)	1.163[d] (1.118-1.209)	0.712[d] (0.696-0.729)	0.823[c] (0.745-0.909)	0.863[f] (0.782-0.953)	0.895 (0.783-1.024)	1.072 (0.977-1.176)	0.678[d] (0.636-0.722)	0.822[c] (0.710-0.950)	0.809 (0.651-1.005)	0.818[f] (0.695-0.961)
Nonbinary or third gender (reference: male), OR (SE)	0.471[d] (0.456-0.486)	1.662[d] (1.548-1.784)	0.916 (0.843-0.995)	0.509[d] (0.492-0.527)	0.921[f] (0.865-0.981)	0.313[d] (0.289-0.339)	2.648[d] (2.552-2.748)	1.104 (0.997-1.224)	0.526[d] (0.471-0.586)	1.900[d] (1.800-2.006)
Prefer not to say (reference: male), OR (SE)	0.384[d] (0.367-0.402)	0.438[d] (0.424-0.452)	0.953 (0.876-1.036)	0.657[d] (0.631-0.684)	1.291[d] (1.150-1.450)	0.482[d] (0.426-0.544)	0.461[d] (0.439-0.485)	1.310[d] (1.202-1.428)	0.557[d] (0.508-0.611)	0.913[f] (0.850-0.980)
Private institution (reference: public), OR (SE)	0.917 (0.791-1.063)	1.009 (0.917-1.111)	1.095 (0.911-1.316)	1.237[d] (1.096-1.396)	1.348[d] (1.104-1.646)	0.939 (0.768-1.148)	1.283[c] (1.074-1.534)	1.206 (0.984-1.478)	1.608[d] (1.328-1.946)	1.402[c] (1.105-1.779)
Nursing (reference: medicine), OR (SE)	1.519[d] (1.322-1.745)	0.956[f] (0.921-0.992)	1.705[d] (1.595-1.822)	1.867[d] (1.784-1.954)	3.552[d] (2.516-5.015)	1.973[d] (1.855-2.098)	0.485[d] (0.470-0.501)	0.559[d] (0.543-0.575)	1.486[d] (1.436-1.537)	3.909[d] (3.525-4.336)
Nutrition (reference: medicine), OR (SE)	0.879[d] (0.843-0.917)	0.437[d] (0.432-0.442)	0.340[d] (0.335-0.345)	0.503[d] (0.495-0.511)	1.287[d] (1.238-1.338)	0.676[d] (0.629-0.726)	0.271[d] (0.257-0.287)	0.175[d] (0.168-0.182)	0.336[d] (0.312-0.362)	1.401[d] (1.264-1.553)
Dentistry (reference: medicine), OR (SE)	0.884 (0.785-0.996)	0.949[f] (0.909-0.990)	1.275[d] (1.210-1.344)	0.922[d] (0.884-0.961)	1.113[d] (1.052-1.178)	0.957 (0.810-1.131)	1.031 (0.939-1.133)	1.197[d] (1.139-1.257)	0.923[f] (0.863-0.987)	1.235[d] (1.115-1.366)
Therapy (reference: medicine), OR (SE)	0.861 (0.470-1.578)	0.938 (0.878-1.002)	1.712[d] (1.412-2.075)	1.551[d] (1.054-2.282)	1.292[c] (1.029-1.622)	1.020 (0.729-1.428)	0.868 (0.710-1.060)	1.776[f] (1.032-3.056)	1.317[d] (1.212-1.432)	0.948 (0.703-1.279)
Psychology (reference: medicine), OR (SE)	0.848 (0.666-1.079)	0.223[d] (0.198-0.252)	0.428[c] (0.338-0.541)	0.413[d] (0.356-0.480)	0.642 (0.467-0.882)	0.675[f] (0.477-0.955)	0.116[d] (0.069-0.195)	0.450 (0.155-1.307)	0.325[d] (0.176-0.599)	1.064 (0.778-1.456)
Pharmacology (reference: medicine), OR (SE)	0.0912[d] (0.090-0.092)	1.946[d] (1.692-2.238)	5.509[d] (1.706-17.787)	2.603[d] (1.974-3.432)	1.368[d] (1.103-1.697)	0.236[d] (0.198-0.281)	1.703[d] (1.429-2.028)	4.976[d] (4.039-6.135)	2.740[d] (2.177-3.449)	1.506[d] (1.324-1.711)
Other, OR (SE)	1.513[d] (1.168-1.960)	1.073 (0.832-1.384)	1.667[d] (1.428-1.946)	1.196[f] (0.972-1.472)	1.261 (0.877-1.812)	1.307[f] (1.005-1.699)	1.337 (0.790-2.261)	1.751[d] (1.565-1.958)	1.166 (0.802-1.696)	1.329 (0.730-2.418)

XSL•FO
RenderX

Variables and outcomes	Predictor: knowledge					Predictor: beliefs of risk				
	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.
/cut 1	0.245[d] (0.223-0.269)	0.0945[d] (0.091-0.098)	0.181[d] (0.170-0.193)	0.142[d] (0.136-0.149)	0.203[d] (0.178-0.232)	0.966 (0.694-1.344)	0.0246[d] (0.017-0.035)	0.0537[d] (0.038-0.076)	0.0749[d] (0.027-0.209)	0.100[d] (0.040-0.254)
/cut 2	0.688 (0.527-0.898)	0.336[d] (0.294-0.384)	0.764 (0.593-0.984)	0.479[d] (0.408-0.562)	0.710 (0.469-1.076)	3.060[d] (2.266-4.133)	0.151[d] (0.116-0.198)	0.282[d] (0.197-0.404)	0.370[c] (0.194-0.705)	0.492 (0.221-1.095)
/cut 3	2.505[d] (0.964-6.507)	1.704[c] (0.879-3.305)	2.997[d] (1.086-8.272)	1.845[c] (0.936-3.635)	2.772[d] (0.613-12.538)	12.66[d] (9.459-16.945)	0.857 (0.657-1.119)	0.952 (0.727-1.245)	1.234 (0.733-2.077)	1.868 (0.873-3.995)
/cut 4	16.42[d] (0.036-7565.397)	9.010[d] (0.235-345.812)	19.50[d] (0.044-8571.641)	11.36[d] (0.221-583.882)	14.13[d] (0.006-33,909.829)	96.46[d] (72.024-129.153)	5.061[d] (3.721-6.883)	6.823[d] (5.038-9.235)	9.192[d] (4.894-17.271)	9.435[d] (4.865-18.302)

[a]Observations: predictor (knowledge): “I think that ChatGPT makes my job easier” n=863, “ChatGPT can be beneficial in health care settings” n=1513, “ChatGPT provides trustworthy health care information or guidance” n=1507, “ChatGPT is a useful tool when I need to search for information on specific medical questions” n=1501, and “ChatGPT is a useful tool when I need to search for medical literature” n=1490. Predictor (beliefs of risk): “I think that ChatGPT makes my job easier” n=861, “ChatGPT can be beneficial in health care settings” n=860, “ChatGPT provides trustworthy health care information or guidance” n=856, “ChatGPT is a useful tool when I need to search for information on specific medical questions” n=854, and “ChatGPT is a useful tool when I need to search for medical literature” n=849.

[b]OR: odds ratio.

[c]*P*<.01.

[d]*P*<.001.

[e]Not applicable.

[f]*P*<.05.

XSL•FO
RenderX

Table 6. Estimates from ordinal logistic regression models for the effect of perception scores on attitude variables (continuation)[a].

Variables and outcomes	Predictor: ethics					Predictor: frequency of use				
	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.
Ethics median (1-5), OR[b] (IQR)	1.439[c] (1.376-1.505)	1.495[c] (1.452-1.539)	1.620[c] (1.498-1.752)	1.476[c] (1.430-1.523)	1.494[a] (1.426-1.564)	—[d]	—	—	—	—
ChatGPT use frequency (1-5), OR (IQR)	—	—	—	—	—	1.320[c] (1.199-1.454)	1.540[c] (1.420-1.670)	1.365[c] (1.321-1.410)	1.438[c] (1.311-1.577)	1.396[c] (1.302-1.497)
Age in years, OR (IQR)	1.030[e] (1.011-1.049)	1.035[e] (1.010-1.060)	1.015 (0.982-1.049)	1.029 (0.999-1.060)	1.043[c] (1.021-1.065)	1.035[c] (1.014-1.057)	1.061[c] (1.036-1.087)	1.022 (0.993-1.051)	1.051[c] (1.034-1.068)	1.046[c] (1.022-1.071)
Female (reference: male), OR (IQR)	1.105[f] (1.018-1.198)	0.682[c] (0.648-0.718)	0.782[e] (0.670-0.912)	0.827[f] (0.711-0.961)	0.870 (0.742-1.019)	1.230[c] (1.114-1.358)	0.797[c] (0.768-0.827)	0.939 (0.816-1.080)	0.952 (0.752-1.204)	0.955 (0.805-1.132)
Nonbinary or third gender (reference: male)	0.599[c] (0.568-0.631)	1.984[c] (1.902-2.071)	1.223[c] (1.155-1.294)	0.685[c] (0.621-0.755)	1.110[e] (1.041-1.183)	0.380[c] (0.369-0.392)	2.174[c] (1.780-2.655)	0.892 (0.791-1.005)	0.334[c] (0.326-0.342)	1.583[c] (1.406-1.782)
Prefer not to say (reference: male), OR (IQR)	0.499[c] (0.438-0.567)	0.387[c] (0.358-0.418)	0.822[c] (0.739-0.913)	0.588[c] (0.547-0.631)	1.141[e] (1.036-1.257)	0.369[c] (0.357-0.381)	0.469[c] (0.454-0.484)	1.246[c] (1.099-1.412)	0.492[c] (0.473-0.511)	0.833[c] (0.778-0.892)
Private institution (reference: public), OR (IQR)	0.951 (0.791-1.142)	1.077 (0.956-1.213)	1.197[c] (1.091-1.313)	1.311[c] (1.247-1.379)	1.438[c] (1.261-1.639)	0.937 (0.803-1.094)	1.257[e] (1.011-1.563)	1.181[f] (0.974-1.432)	1.563[c] (1.286-1.900)	1.403[c] (1.087-1.810)
Nursing (reference: medicine), OR (IQR)	1.662[c] (1.602-1.726)	1.006 (0.986-1.027)	1.713[c] (1.589-1.846)	2.084[c] (2.042-2.125)	3.732[c] (3.504-3.971)	1.425[c] (1.351-1.503)	0.350[c] (0.340-0.360)	0.458[c] (0.448-0.468)	1.237[c] (1.192-1.284)	3.504[c] (2.496-4.918)
Nutrition (reference: medicine), OR (IQR)	0.867[c] (0.824-0.913)	0.392[c] (0.376-0.408)	0.307[c] (0.291-0.323)	0.459[c] (0.445-0.472)	1.218[c] (1.177-1.260)	0.901[c] (0.858-0.947)	0.307[c] (0.303-0.311)	0.187[c] (0.185-0.189)	0.378[c] (0.371-0.385)	1.545[c] (1.354-1.763)
Dentistry (reference: medicine), OR (IQR)	0.838[f] (0.712-0.987)	0.844[c] (0.799-0.893)	1.152[c] (1.095-1.213)	0.842[c] (0.799-0.887)	1.028 (0.980-1.078)	0.992 (0.857-1.149)	1.258[c] (1.088-1.455)	1.358[c] (1.227-1.503)	1.036 (0.972-1.104)	1.412[c] (1.261-1.581)
Therapy (reference: medicine), OR (IQR)	0.815 (0.536-1.240)	0.968 (0.757-1.239)	1.895[c] (1.539-2.335)	1.552[c] (1.467-1.642)	1.348 (0.960-1.893)	0.925 (0.582-1.469)	0.906 (0.715-1.148)	1.820[c] (1.001-3.309)	1.292 (0.876-1.905)	0.947 (0.830-1.081)
Psychology (reference: medicine), OR (IQR)	0.819 (0.596-1.126)	0.197[c] (0.112-0.345)	0.404[f] (0.173-0.941)	0.397[c] (0.287-0.548)	0.651 (0.325-1.303)	0.818 (0.628-1.066)	0.142[c] (0.134-0.151)	0.519 (0.302-0.893)	0.375[c] (0.306-0.459)	1.265 (0.924-1.731)

XSL•FO
RenderX

Variables and outcomes	Predictor: ethics					Predictor: frequency of use				
	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.	I think that ChatGPT makes my job easier.	ChatGPT can be beneficial in health care settings.	ChatGPT provides trustworthy health care information or guidance.	ChatGPT is a useful tool when I need to search for information on specific medical questions.	ChatGPT is a useful tool when I need to search for medical literature.
Pharmacology (reference: medicine), OR (IQR)	0.0920[c] (0.084-0.101)	2.064[c] (1.929-2.210)	6.336[c] (5.501-7.294)	2.787[c] (2.492-3.117)	1.545[c] (1.397-1.709)	0.0774[c] (0.077-0.078)	2.121[c] (1.631-2.758)	4.864[c] (1.748-13.531)	2.125[c] (1.456-3.102)	1.196[e] (1.018-1.405)
Other, OR (IQR)	1.394[c] (1.221-1.590)	0.886 (0.707-1.111)	1.450[c] (1.287-1.636)	1.012 (0.851-1.203)	1.085 (0.824-1.428)	1.493[c] (1.291-1.727)	1.434 (0.772-2.664)	1.866[c] (1.191-2.923)	1.276 (0.931-1.749)	1.432 (0.713-2.877)
/cut 1	0.280[c] (0.187-0.418)	0.0852[c] (0.047-0.153)	0.187[c] (0.085-0.408)	0.134[c] (0.091-0.198)	0.265[c] (0.183-0.384)	0.259[c] (0.228-0.294)	0.145[c] (0.135-0.156)	0.115[c] (0.110-0.120)	0.130[c] (0.116-0.146)	0.178[c] (0.151-0.210)
/cut 2	0.797 (0.559-1.137)	0.305[c] (0.155-0.602)	0.807 (0.382-1.709)	0.457[c] (0.288-0.725)	0.941 (0.642-1.379)	0.726 (0.528-0.997)	0.909 (0.653-1.266)	0.612[f] (0.464-0.807)	0.645 (0.444-0.938)	0.881 (0.445-1.746)
/cut 3	2.980[c] (2.090-4.250)	1.549 (0.817-2.939)	3.221[e] (1.576-6.580)	1.761[f] (1.027-3.019)	3.744[c] (2.487-5.635)	2.677[c] (0.815-8.797)	5.434[c] (0.553-53.412)	2.119[c] (0.839-5.355)	2.176[e] (0.787-6.018)	3.425[e] (0.254-46.156)
/cut 4	20.14[c] (13.437-30.175)	8.187[c] (4.554-14.717)	21.32[c] (10.848-41.888)	10.79[c] (6.527-17.832)	19.60[c] (12.642-30.387)	18.23[c] (0.002-164,263.740)	35.59[c] (0.000-431,675,496.970)	15.94[c] (0.040-6402.678)	17.24[c] (0.001-441,561.358)	18.22[c] (0.000-5,099,024.943)

[a]Observations: predictor (ethics): "I think that ChatGPT makes my job easier" n=863, "ChatGPT can be beneficial in health care settings" n=1513, "ChatGPT provides trustworthy health care information or guidance" n=1507, "ChatGPT is a useful tool when I need to search for information on specific medical questions" n=1501, and "ChatGPT is a useful tool when I need to search for medical literature" n=1490. Predictor (frequency of use): "I think that ChatGPT makes my job easier=863, ChatGPT can be beneficial in health care settings" n=861, "ChatGPT provides trustworthy health care information or guidance" n=860, "ChatGPT is a useful tool when I need to search for information on specific medical questions" n=858, and "ChatGPT is a useful tool when I need to search for medical literature" n=853.

[b]OR: odds ratio.

[c]*P*<.001.

[d]Not applicable.

[e]*P*<.01.

[f]*P*<.05.

Discussion

Principal Findings

The aim of this study was to determine the perception, attitudes, and uses of ChatGPT among health care students, as well as their willingness to learn more about it. Given that chatbots powered by AI are widely accepted by students [27], our findings provide critical insights into the possibilities of integrating them into undergraduate health care teaching programs. More than half (1419/2661, 53.32%) of the participants knew about ChatGPT according to our data, with male students being more knowledgeable than female students. In May 2023, the Pew Research Center released the findings of a web-based study that showed that, compared to our results (1142/2661, 42.92%), 33% of young people had never heard of ChatGPT. Most participants felt that they knew little to nothing about ChatGPT [28]. According to the study by Buabbas et al [29], 84% of Kuwaiti medical students did not have any training on the use of AI. It is worth noting that >80% of our participants (2160/2661, 81.17%) indicated an interest in learning more about ChatGPT's health care applications, with time restrictions being the primary barrier to learning more for 39.98% (1064/2661) of them.

Despite the widespread use of AI chatbots such as ChatGPT for self-diagnosing illnesses (up to 78%) [30] and the recognition of the value and user-friendliness of the information they provide, health care career students in the Americas maintained a neutral stance on whether ChatGPT will replace their jobs.

XSL•FO
RenderX

They neither agreed nor disagreed with the notion. This aligns with the findings of the studies by Buabbas et al [29] and Moldt et al [31], where 78.7% and 83% of participants, respectively, expressed skepticism about AI eventually replacing the roles of physicians in the future.

Only 22.98% (349/1519) of our students reported using AI for medical and health care education and training, but >70% (1101/1519, 72.48%) said that they used it for homework support. Although some colleges prohibit the use of ChatGPT and consider it plagiarism [32], teachers are investigating its utility during learning. For example, the students of Mullen [33] used ChatGPT to improve the quality of an essay in English (their nonnative language), and the participants felt that the experience left them better equipped to produce future academic output without the use of these tools.

Our study revealed that health care students displayed positive attitudes and acceptance toward ChatGPT and that most were willing to learn more about it, similar to the studies by Buabbas et al [29] and Moldt et al [31]. Although we did not inquire about the specific version of ChatGPT used by participants, and as ChatGPT's primary function is not to be used as a web search engine, it is evident that, within the context of higher education, particularly in the field of health, there has been a significant increase in the adoption of disruptive technologies [34], including ChatGPT, as both formal or informal tools for enhancing skills and achieving educational objectives [35].

Respondents perceived ChatGPT as a valuable tool in health care settings, highlighting its usefulness in providing information on specific medical questions and facilitating access to relevant literature. Interestingly, the attitudes toward ChatGPT appeared to be influenced by the participants' self-perceived knowledge about the chatbot. Those who had a better understanding of ChatGPT tended to perceive it as providing trustworthy health care information or guidance. Notably, participants' willingness to use ChatGPT in the health care setting is heavily influenced by the level of trust they have in the system [6]. Interestingly, we found a significant association between increased perceived risk scores and the following attitude statement: "ChatGPT provides trustworthy health care information or guidance." Establishing trust is crucial to ensuring the responsible and effective use of ChatGPT, thereby maximizing its benefits while mitigating any associated risks.

Indeed, this study revealed that users' attitudes toward ChatGPT are positively influenced by the frequency of use. Individuals who use ChatGPT more frequently have higher possibilities of believing that ChatGPT makes their job easier and finding it beneficial in health care settings, as well as considering it a useful tool for searching specific medical questions and medical literature. Despite students being somewhat concerned about the perceived risk of the ethical implications of using ChatGPT, they still used it once a month, especially for homework support, research paper writing support, medical or health care education and training, and mental health support. Our study differs from previous research, and Firaina and Sulisworo [36] found that most respondents preferred frequent use of ChatGPT.

Despite the many changes that have occurred in medicine over the last few decades, medical education is still largely based on traditional teaching methods [37,38]. The release of ChatGPT caused concerns and debates in health care due to ethical issues, misinformation, misuse, and challenges in practice and academic writing. Concerns include the quality and dependability of medical information, the chatbot model's transparency, the ethics of user information, and potential biases in the ChatGPT algorithms [35]. While several studies have demonstrated ChatGPT's ability to answer medical questions [39-42], many correct answers have been deemed inadequate [39,40].

Limitations

Our study has several limitations that must be considered when interpreting the results. First, our sampling strategy did not capture all health care students from the Americas. Despite our efforts to include universities across the Americas, we encountered a limited recruitment response from Central America. This low number may limit the representativeness of our findings for this specific region. As a result, the findings from Central America should be considered as preliminary and require validation through larger-scale research conducted in this region. Second, this study was cross-sectional in nature, and, therefore, we cannot establish causality among perceptions, beliefs, ethics, and attitudes. Longitudinal studies are needed to determine the temporal relationship between these variables. Third, although, during the course of this study, there were 2 available versions of ChatGPT (3.5 and 4.0), the participants were not specifically queried on which version they used. However, given their status as students, it can be reasonably deduced that they predominantly used the free version rather than the premium version. The disparities between the 2 versions lie mostly in the payment requirement associated with version 4.0. It has been said that this particular version offers enhanced safety measures, more valuable responses, and a heightened comprehension of the contextual nuances pertaining to the posed queries. On the basis of the aforementioned findings, certain worries emerge regarding the potential use of ChatGPT by students within their educational institutions but in an informal manner despite the absence of official integration of ChatGPT as an explicitly disruptive technological tool within their educational system. It is also possible that academic institutions are incorporating this technology within their instructional settings. At present, there remain unanswered inquiries pertaining to the subject matter. However, these discoveries indicate potential gaps in knowledge, warranting an assessment of whether the acquired information satisfies the minimum criteria for quality in the field of health and possesses genuine value in terms of gathering competent professionals in the near future.

Conclusions

The current debate revolves around the potential advantages and disadvantages of incorporating ChatGPT and other LLMs into the teaching and learning process. The age of AI has arrived. It is important to be aware of how it may be used and misused. Research in health care education looks bright in the future due to the essential integrity that drives the vast majority of researchers. A medical educator must remain current with the rapid advancements in technology and consider how they affect

XSL•FO
RenderX

their teaching practices, curriculum development, and evaluation techniques.

Acknowledgments

The authors would like to extend special thanks to all members of the RespiraLab Research Group for their initial input on this project. The authors would like to also express their gratitude to Universidad Espíritu Santo for their continuous support in their research endeavors. RespiraLab Research Group financed this study. The sponsor did not design the study or collect, analyze, or interpret the data.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

ICO, KRV, and JCGB contributed to conceptualization, funding acquisition, methodology, resources, project administration, supervision, validation, writing—original draft, and writing—review and editing. ICO, KRV, MFH, and MFO contributed to data curation, formal analysis, software, supervision, validation, writing—original draft, and writing—review and editing. EMVL, MLV, PP, FCAD, AS, SPGE, ECC, KLCG, JCC, JB, and AB contributed to conceptualization, investigation, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Original version of the survey administered to the participants during the study.
[DOCX File , 30 KB - mededu_v10i1e51757_app1.docx]

Multimedia Appendix 2
Detailed breakdown of the demographic characteristics of a sample population consisting of 2661 individuals. The data include the composition of the sample population in terms of age, gender, type of university, geographic region, and academic majors.
[DOCX File , 16 KB - mededu_v10i1e51757_app2.docx]

Multimedia Appendix 3
Distribution of ChatGPT use by health care students across various academic and clinical activities, indicating the range of use and the percentage of students who use ChatGPT for each activity.
[DOCX File , 15 KB - mededu_v10i1e51757_app3.docx]

References

1. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
2. Waisberg E, Ong J, Kamran SA, Masalkhi M, Zaman N, Sarker P, et al. Bridging artificial intelligence in medicine with generative pre-trained transformer (GPT) technology. J Med Artif Intell 2023 Aug;6:13 [FREE Full text] [doi: 10.21037/jmai-23-36]
3. OpenAI documentation. OpenAI. URL: https://platform.openai.com/docs/overview [accessed 2023-07-21]
4. Milmo D. ChatGPT reaches 100 million users two months after launch. The Guardian. 2023 Feb 2. URL: https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app [accessed 2023-07-21]
5. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023 Jul 25;330(4):315-316. [doi: 10.1001/jama.2023.9651] [Medline: 37410482]
6. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023 Jun 14;25:e47184 [FREE Full text] [doi: 10.2196/47184] [Medline: 37314848]
7. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus 2023 May;15(5):e39305 [FREE Full text] [doi: 10.7759/cureus.39305] [Medline: 37378099]
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]

XSL•FO
RenderX

9. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
10. Heston TF. Prompt engineering for students of medicine and their teachers. Preprint posted online on August 8, 2023 [FREE Full text]
11. Biswas S. ChatGPT and the future of medical writing. Radiology 2023 Apr;307(2):e223312 [FREE Full text] [doi: 10.1148/radiol.223312] [Medline: 36728748]
12. Passmore C, Dobbie AE, Parchman M, Tysinger J. Guidelines for constructing a survey. Fam Med 2002 Apr;34(4):281-286. [Medline: 12017142]
13. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34 [FREE Full text] [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
14. Barbul M, Bojescu I. Generations' perception towards the interaction with AI. In: Proceedings of the BASIQ International Conference. 2023 Presented at: BASIQ 2023; June 8-10, 2023; Constan a, Romania URL: https://doaj.org/article/f928a5c54e734e86babca43bf2f52f21 [doi: 10.24818/basiq/2023/09/041]
15. Menon D, Shilpa K. "Chatting with ChatGPT": analyzing the factors influencing users' intention to use the Open AI's ChatGPT using the UTAUT model. Heliyon 2023 Nov;9(11):e20962 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e20962] [Medline: 37928033]
16. Abdaljaleel M, Barakat M, Alsanafi M, Salim NA, Abazid H, Malaeb D, et al. Author correction: a multinational study on the factors influencing university students' attitudes and usage of ChatGPT. Sci Rep 2024 Apr 09;14(1):8281 [FREE Full text] [doi: 10.1038/s41598-024-59011-9] [Medline: 38594508]
17. Agostinis G, Parthenay K. Exploring the determinants of regional health governance modes in the Global South: a comparative analysis of Central and South America. Rev Int Stud 2021 May 17;47(4):399-421. [doi: 10.1017/s0260210521000206]
18. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
19. Khairatun Hisan U, Miftahul Amri M. ChatGPT and medical education: a double-edged sword. J Pedagogy Educ Sci 2023 Mar 11;2(01):71-89. [doi: 10.56741/jpes.v2i01.302]
20. Abouammoh N, Alhasan K, Raina R, Malki KA, Aljamaan F, Tamimi I, et al. Exploring perceptions and experiences of ChatGPT in medical education: a qualitative study among medical college faculty and students in Saudi Arabia. Preprint posted online on July 16, 2023 [FREE Full text] [doi: 10.1101/2023.07.13.23292624]
21. Ahmad MN, Abdallah SA, Abbasi SA, Abdallah AM. Student perspectives on the integration of artificial intelligence into healthcare services. Digit Health 2023 May 31;9:20552076231174095 [FREE Full text] [doi: 10.1177/20552076231174095] [Medline: 37312954]
22. Ghotbi N, Ho MT. Moral awareness of college students regarding artificial intelligence. Asian Bioeth Rev 2021 Sep 03;13(4):421-433 [FREE Full text] [doi: 10.1007/s41649-021-00182-2] [Medline: 34616496]
23. Bankins S, Formosa P. The ethical implications of artificial intelligence (AI) for meaningful work. J Bus Ethics 2023 Feb 11;185(4):725-740. [doi: 10.1007/s10551-023-05339-7]
24. Abdulai AF, Hung L. Will ChatGPT undermine ethical values in nursing education, research, and practice? Nurs Inq 2023 Jul 26;30(3):e12556. [doi: 10.1111/nin.12556] [Medline: 37101311]
25. Koonchanok R, Pan Y, Jang H. Public attitudes toward ChatGPT on Twitter: sentiments, topics, and occupations. Soc Netw Anal Min 2024 May 20;14(1):106 [FREE Full text] [doi: 10.1007/s13278-024-01260-7]
26. Ali JK, Shamsan MA, Hezam TA, Mohammed AA. Impact of ChatGPT on learning motivation: teachers and students' voices. J Eng Stud Arabia Felix 2023 Mar 07;2(1):41-49. [doi: 10.56540/jesaf.v2i1.51]
27. Koivisto M. Tutoring postgraduate students with an AI-based chatbot. Int J Adv Corp Learn 2023 Mar 13;16(1):41-54. [doi: 10.3991/ijac.v16i1.35437]
28. Vogels EA. A majority of Americans have heard of ChatGPT, but few have tried it themselves. Pew Research Center. 2023 May 24. URL: https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/ [accessed 2023-07-21]
29. Buabbas AJ, Miskin B, Alnaqi AA, Ayed AK, Shehab AA, Syed-Abdul S, et al. Investigating students' perceptions towards artificial intelligence in medical education. Healthcare (Basel) 2023 May 01;11(9):1298 [FREE Full text] [doi: 10.3390/healthcare11091298] [Medline: 37174840]
30. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564 [FREE Full text] [doi: 10.2196/47564] [Medline: 37195756]
31. Moldt JA, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec 28;28(1):2182659 [FREE Full text] [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
32. Yadava OP. ChatGPT-a foe or an ally? Indian J Thorac Cardiovasc Surg 2023 May 28;39(3):217-221 [FREE Full text] [doi: 10.1007/s12055-023-01507-6] [Medline: 37124601]

XSL•FO
RenderX

33. Mullen M. Structured use of an AI chatbot to support student development of English for academic purposes. University of the West of Scotland. 2023 Jul 13. URL: https://research-portal.uws.ac.uk/en/publications/structured-use-of-an-ai-chatbot-to-support-student-development-of [accessed 2024-07-30]
34. Gejendhiran S, Anicia SA, Vignesh S, Kalaimani M. Disruptive technologies - a promising key for sustainable future education. Procedia Comput Sci 2020;172:843-847. [doi: 10.1016/j.procs.2020.05.121]
35. Chow JC, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023 Apr 5;6:1166014 [FREE Full text] [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]
36. Firaina R, Sulisworo D. Exploring the usage of ChatGPT in higher education: frequency and impact on productivity. Buletin Edukasi Indonesia 2023 Mar 11;2(01):39-46. [doi: 10.56741/bei.v2i01.310]
37. Pfeifer CM. A progressive three-phase innovation to medical education in the United States. Med Educ Online 2018 Dec 20;23(1):1427988 [FREE Full text] [doi: 10.1080/10872981.2018.1427988] [Medline: 29353536]
38. Weggemans MM, van Dijk B, van Dooijeweert B, Veenendaal AG, Ten Cate OT. The postgraduate medical education pathway: an international comparison. GMS J Med Educ 2017 Nov 15;34(5):Doc63 [FREE Full text] [doi: 10.3205/zma001140] [Medline: 29226231]
39. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023 Apr;185:109729 [FREE Full text] [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
40. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
41. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad015 [FREE Full text] [doi: 10.1093/jncics/pkad015] [Medline: 36929393]
42. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]

Abbreviations

AI: artificial intelligence
LLM: large language model
OR: odds ratio

Edited by T de Azevedo Cardoso; submitted 10.08.23; peer-reviewed by R Vieira, YD Cheng; comments to author 08.09.23; revised version received 26.09.23; accepted 30.04.24; published 13.08.24.

Please cite as:
Cherrez-Ojeda I, Gallardo-Bastidas JC, Robles-Velasco K, Osorio MF, Velez Leon EM, Leon Velastegui M, Pauletto P, Aguilar-Díaz FC, Squassi A, González Eras SP, Cordero Carrasco E, Chavez Gonzalez KL, Calderon JC, Bousquet J, Bedbrook A, Faytong-Haro M
Understanding Health Care Students' Perceptions, Beliefs, and Attitudes Toward AI-Powered Language Models: Cross-Sectional Study
JMIR Med Educ 2024;10:e51757
URL: https://mededu.jmir.org/2024/1/e51757
doi:10.2196/51757
PMID:39137029

©Ivan Cherrez-Ojeda, Juan C Gallardo-Bastidas, Karla Robles-Velasco, María F Osorio, Eleonor Maria Velez Leon, Manuel Leon Velastegui, Patrícia Pauletto, F C Aguilar-Díaz, Aldo Squassi, Susana Patricia González Eras, Erita Cordero Carrasco, Karol Leonor Chavez Gonzalez, Juan C Calderon, Jean Bousquet, Anna Bedbrook, Marco Faytong-Haro. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.08.2024. 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.

XSL•FO
RenderX

Original Paper

Performance of ChatGPT in the In-Training Examination for Anesthesiology and Pain Medicine Residents in South Korea: Observational Study

Soo-Hyuk Yoon[1], MD; Seok Kyeong Oh[2], MD, PhD; Byung Gun Lim[2], MD, PhD; Ho-Jin Lee[1], MD, PhD

[1]Department of Anesthesiology and Pain Medicine, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea

[2]Department of Anesthesiology and Pain Medicine, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Ho-Jin Lee, MD, PhD
Department of Anesthesiology and Pain Medicine
Seoul National University Hospital
Seoul National University College of Medicine
Daehak-ro 101, Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 220720039
Fax: 82 27478363
Email: hjpainfree@snu.ac.kr

Abstract

Background: ChatGPT has been tested in health care, including the US Medical Licensing Examination and specialty exams, showing near-passing results. Its performance in the field of anesthesiology has been assessed using English board examination questions; however, its effectiveness in Korea remains unexplored.

Objective: This study investigated the problem-solving performance of ChatGPT in the fields of anesthesiology and pain medicine in the Korean language context, highlighted advancements in artificial intelligence (AI), and explored its potential applications in medical education.

Methods: We investigated the performance (number of correct answers/number of questions) of GPT-4, GPT-3.5, and CLOVA X in the fields of anesthesiology and pain medicine, using in-training examinations that have been administered to Korean anesthesiology residents over the past 5 years, with an annual composition of 100 questions. Questions containing images, diagrams, or photographs were excluded from the analysis. Furthermore, to assess the performance differences of the GPT across different languages, we conducted a comparative analysis of the GPT-4's problem-solving proficiency using both the original Korean texts and their English translations.

Results: A total of 398 questions were analyzed. GPT-4 (67.8%) demonstrated a significantly better overall performance than GPT-3.5 (37.2%) and CLOVA-X (36.7%). However, GPT-3.5 and CLOVA X did not show significant differences in their overall performance. Additionally, the GPT-4 showed superior performance on questions translated into English, indicating a language processing discrepancy (English: 75.4% vs Korean: 67.8%; difference 7.5%; 95% CI 3.1%-11.9%; *P*=.001).

Conclusions: This study underscores the potential of AI tools, such as ChatGPT, in medical education and practice but emphasizes the need for cautious application and further refinement, especially in non-English medical contexts. The findings suggest that although AI advancements are promising, they require careful evaluation and development to ensure acceptable performance across diverse linguistic and professional settings.

(JMIR Med Educ 2024;10:e56859) doi:10.2196/56859

KEYWORDS

AI tools; problem solving; anesthesiology; artificial intelligence; pain medicine; ChatGPT; health care; medical education; South Korea

XSL•FO
RenderX

Introduction

ChatGPT is an artificial intelligence (AI) service for conversations based on the generated pretrained transformer and a large-scale generative language model [1]. Since the release of ChatGPT, numerous attempts have been made to apply it in health care practices [2]. In this context, its medical knowledge and thinking skills have been evaluated through a range of medical examinations including the US Medical Licensing Examination and various specialty examinations. The results indicate a performance close to the passing threshold [3]. In the field of anesthesiology, ChatGPT has been evaluated using questions from several question banks designed for English-language board examination preparation. However, doubts remain regarding their ability to complete board examinations [4,5].

GPT-4 is the successor of GPT-3.5, which formed the basis of ChatGPT after its launch. OpenAI, the developer of ChatGPT, reported that GPT-4 not only outperformed GPT-3.5 but also often scored higher than most human test-takers, demonstrating a particularly strong performance in languages other than English [6]. Indeed, in previous studies using written board examinations for neurosurgery and ophthalmology, GPT-4 exhibited a significantly higher proportion of correct responses compared to GPT-3.5 [7,8]. The superiority of GPT-4 over GPT-3.5 was also noted in the field of anesthesiology, as assessed using 27 questions from the Royal College of Anaesthetists [9]. Furthermore, a comparative study evaluating the performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination revealed that GPT-4 achieved a significantly higher rate of correct responses [10], indicating its advanced performance in non-English languages.

Given the emergence and development of ChatGPT, it is crucial to examine the knowledge levels and reasoning abilities of AI language models in the fields of anesthesiology and pain medicine in Korea to estimate their potential to aid medical professionals. However, to date, no study has explored the performance of ChatGPT in the fields of anesthesiology and pain medicine in a Korean language context. Therefore, this study aimed to investigate the performance of ChatGPT, including both GPT-3.5 and GPT-4, using the in-training examination administered by the Korean Society of Anesthesiologists (KSA). This study also aimed to compare the performance of ChatGPT with that of CLOVA X, a new generative AI service in South Korea.

Methods

Data Source and Contents

This study evaluated the performance of AI services by using the metric "number of correct answers/number of questions" [3], using the KSA in-training examinations from 2018 to 2022, each comprising 100 annual questions. The KSA conducts annual in-training tests for residents to assess their readiness and prepare them for specialist certification examinations. Beginning in 2019, a cutoff point system was introduced: if an examinee scored below a certain threshold for each year's grade, they were considered to have failed and were required to retake the examination. The cutoff points are set at 50 in the first year of training, 55 in the second year, and 60 in the third and fourth years. The full texts of the questionnaires, correct answers, and commentaries provided by the society are accessible only to members via official websites [11]. Each question consisted of one query and five choices, each with one correct answer. Because ChatGPT only accepts text as input, we excluded questions with images, diagrams, or photographs within the question content.

To compare the performance of ChatGPT with that of the actual examinees, we requested anonymized data from the Training and Education Committee of the KSA on the scores achieved by residents over the past 5 years, both overall and for each training year.

Initially, we focused solely on evaluating the performance of ChatGPT. However, a new generative AI service, CLOVA X, was launched in South Korea by Naver Corporation in August 2023 during our study period; therefore, we decided to expand our investigation to include an examination of its performance. CLOVA X was developed based on the Korean large language model HyperCLOVA X. HyperCLOVA X was trained on a vast corpus of high-quality data primarily sourced from Korean text content. This makes the training data particularly rich in terms of Korean culture and lifestyle, unlike the more diverse multilingual data sets used for ChatGPT. In addition, HyperCLOVA X uses specific alignment techniques, such as supervised fine-tuning and reinforcement learning from human feedback, to enhance its ability to follow instructions and align with human values [12].

Ethical Considerations

The ethical review of the study was exempted by the institutional review board of Seoul National University Hospital (E-2308-102-1459). This study used only in-training examinations that are already available on the KSA website and did not involve human participants or use any personal information.

Testing Process

We compared the performances of GPT-3.5, GPT-4, and CLOVA X in solving problems in the fields of anesthesiology and pain medicine using the following process: to ensure that both models were tested under identical conditions, the following command in Korean was entered before posing the questions: "(Translated) Below are the in-training assessment questions for the specialty of anesthesiology and pain medicine. Please complete the questions and describe your solution in detail. There is only one answer for each question" (Figure S1A in Multimedia Appendix 1). Previously, there were instances in which multiple answers were provided by ChatGPT when the prompt did not explicitly state a single correct answer. In addition, answers were sometimes provided without explanation, when the prompt did not request detailed steps. Therefore, we implemented these commands to address these issues. The included questions were then individually entered into the prompt in the order of their question numbers, exactly as they were written in Korean (Multimedia Appendix 1). This is because we determined that within the same window, previously

XSL•FO
RenderX

entered questions could influence the answers to subsequent questions. If a question or choice included a table, we transcribed the content and maintained the same arrangement, using spaces and hyphens. After completing the set of questions, a new window was opened, the same command was entered, and questions from another year were entered. This process was identical for both ChatGPT and CLOVA X.

For each question input, we recorded the answers chosen by GPT-3.5, GPT-4, and CLOVA X and the explanation for the selection. After completing the problem-solving process for all the questions in both models, we compared their responses to the answer keys provided by the KSA. An answer was recorded as correct if the first response matched the actual response. It was recorded as incorrect if no answer was selected, if the answer was incorrect, or if multiple answers were selected, even if the correct answer was among them. After scoring, we calculated the overall and yearly scores for GPT-3.5, GPT-4, and CLOVA X, as well as the percentage of questions answered correctly relative to the total number of questions.

To compare the performance of GPT-4 in Korean and English, we translated the questions into English and conducted additional problem-solving. This process was conducted in 2 stages. First, the English translation was initially performed by inputting the original Korean questions one by one, along with the command, "Please translate the following into English." All questions included in this study were translated, and the translated texts were recorded separately (Figure S1A in Multimedia Appendix 2). Two authors (SHY and HJL) reviewed the accuracy of English translations. In the second stage, we entered the initial instruction commands used for the Korean questions in English into a new window (Figure S1B in Multimedia Appendix 2), followed by the translated English questions individually (Figure S1C in Multimedia Appendix 2). The process of answering the questions, recording the answers and explanations, and scoring was identical to that used for the Korean questions.

Two authors (SHY and HJL) conducted the task of having AI services solve problems, and all authors reviewed the results. Two authors (SHY and HJL) were using the paid version of ChatGPT-4 at their own expense, independently of this study. ChatGPT-3.5 and CLOVA X were used free of charge. Therefore, no additional costs were incurred when using the three programs.

Outcome Measure and Analysis

The primary outcome of this study was to assess the performance difference between GPT-3.5, GPT-4, and CLOVA-X, as measured by the overall score on the 5 years of in-training examination for residents of anesthesiology and pain medicine in Korea. Secondary outcomes included performance on the ChatGPT and CLOVA-X according to the examination year, subfields within anesthesiology and pain medicine, inclusion of clinical cases, and level of logical thinking required by the questions. Additionally, the performance of GPT-4 on the English-translated questions was compared to its performance on the original Korean version.

The analytical methods used in this study were first used to compare the overall performances of GPT-3.5, GPT-4, and CLOVA X for each year. As a performance reference, we calculated the mean and SD of the examinees' scores both overall and for each training year. However, a direct comparison of the scores was not possible because the study excluded questions involving images, diagrams, and photographs. Second, the questions were categorized into the subfields of anesthesiology and pain medicine following the taxonomy outlined by the KSA. Third, we classified the questions based on the inclusion of clinical cases or the level of logical thinking required (Figure S1B-D in Multimedia Appendix 1). A question was classified as containing a clinical case if it described a specific situation involving patient information, such as demographics, medical history, surgery, and anesthesia, requiring the use of this information to answer the question. If the question dealt only with theoretical knowledge or if there was some mention of a patient but it was not necessary to apply this information to answer the question, we classified the question as not containing a clinical case. The level of logical thinking was categorized as either first-order or higher-order problem-solving based on a previous study design that evaluated the performance of GPT-3.5 and GPT-4 on the self-assessment examination of neurosurgery [7]. A question was classified as first order if it required direct use of the conditions or circumstances of the question, simple recall of facts, selection of an answer from a set of choices, or determination of the truth or falsity of each option. When a question required additional logical steps to select the correct answer, such as estimating a diagnosis, applying guidelines, or calculating with formulas, it was classified in the higher order. Fourth, we compared the differences in GPT-4 performance between the original Korean questions and their English-translated versions. Additionally, we measured self-agreement, which refers to the number and percentage of questions for which ChatGPT chose the same answer in Korean and English, irrespective of the accuracy of the response.

During the revision process, we further analyzed the explanations for the incorrect answers of each model. A classification system from a previous study was used to categorize the reasons for each incorrect choice as logical, informational, or statistical errors [13]. In cases where two errors occurred simultaneously, both errors were identified. This process was conducted independently by two authors (SHY and HJL), and discrepancies in labeling were resolved through a post hoc discussion involving all authors.

Statistics

When comparing the performances of GPT-3.5, GPT-4, and CLOVA X in Korean, we used the Cochran *Q* test; in cases where there was a significant difference among the 3 tools, the comparison between the two groups was investigated by calculating the minimum required difference for a significant difference between the 2 groups [14]. The significance level of Cochran *Q* test for the three language models was 0.05, while a Bonferroni correction was applied to set the significance level to 0.017 when comparing the two groups, considering that there were three combinations for comparisons. Additionally, although we used the KSA classification to compare the performances

XSL•FO
RenderX

of GPT-3.5, GPT-4, and CLOVA-X across various subfields, we recognized that the number of questions per field was too limited for a statistical comparison. Descriptive statistics were used to analyze these factors. A chi-square test was conducted to compare the inclusion of clinical cases and the level of logical thinking in the questions. Finally, to compare the performance of GPT-4 in Korean and English, we used McNemar's test and calculated Poisson 95% CIs for the two performances. All statistical analyses were performed using MedCalc Statistical Software (version 18.6; MedCalc Software bvba).

Results

A total of 398 questions were included in the analysis, selected from a set of 500 questions used over the past 5 years, excluding those containing images, diagrams, or photographs. The performances of GPT-3.5, GPT-4, and CLOVA X are presented in Table 1. The overall performance of GPT-4 (67.8%) was significantly higher than that of GPT-3.5 (37.7%) and CLOVA X (37.2%), surpassing the minimum required difference of 9.1% in Cochran *Q* test. However, GPT-3.5 and CLOVA X did not show significant differences in their overall performance. In the year-by-year analysis, GPT-4 consistently showed a significantly higher performance than GPT-3.5 and CLOVA-X, except in 2022 when only the difference between GPT-4 and GPT-3.5 was significant. Multimedia Appendix 3 shows the actual scores of Korean anesthesiology residents in 2022, 2021, and 2019. However, due to the unavailability of data for 2018 and the inapplicability of the 2020 data for estimating the mean and SD of the residents' scores, these years were excluded from the analysis.

Table 1. Performances of the models in overall and yearly examinations.

Year (questions)[a]	GPT-3.5[b], n (%)	GPT-4[b], n (%)	CLOVA X[b], n (%)	*P* value[c]	GPT-4 versus GPT-3.5[c]	GPT-4 versus CLOVA X[c]	GPT-3.5 versus CLOVA X[c]
2022 (n=72)	20 (28)	49 (68)	34 (47)	<.001	S[d]	N/S[e]	N/S
2021 (n=74)	29 (39)	51 (69)	23 (31)	<.001	S	S	N/S
2020 (n=79)	28 (35)	54 (68)	22 (28)	<.001	S	S	N/S
2019 (n=85)	36 (42)	53 (62)	33 (39)	.001	S	S	N/S
2018 (n=88)	37 (42)	63 (72)	36 (41)	<.001	S	S	N/S
Total (n=398)	150 (37.7)	270 (67.8)	148 (37.2)	<.001	S	S	N/S

[a]Number of questions included in the overall and yearly examinations is presented in parentheses.

[b]Performances of ChatGPT and CLOVA X are presented as the number of correct answers for each examination, along with the percentage of correct answers out of the total number of questions in parentheses.

[c]Cochran *Q* test was conducted to compare the performance of GPT-3.5, GPT-4, and CLOVA X, and the *P* values are presented. In multiple comparisons of the two models, significance determined at a *P* value of .017 using Bonferroni correction was denoted as S or N/S.

[d]S: significant.

[e]N/S: not significant.

Table 2 presents a comparison of the performances of GPT-3.5, GPT-4, and CLOVA-X in each specific subfield of anesthesiology. A total of 21 subfields were examined based on the taxonomy of the KSA. The highest-scoring subfield was geriatric anesthesia in GPT-3.5 (58.8%), GPT-4 (88.2%), and CLOVA X (64.7%). The lowest scoring subfield was "neuromuscular blocking agents" for GPT-3.5 and CLOVA X (17.6%), and "anesthesia equipment and monitoring" for GPT-4 (37.5%).

XSL•FO
RenderX

Table 2. Performance for each subfield in anesthesiology and pain medicine.

Subfields (questions)[a]	GPT-3.5[b], n (%)	GPT-4[b], n (%)	CLOVA X[b], n (%)
Medical ethics (n=5)	1 (20)	3 (60)	2 (40)
Preanesthetic care (n=11)	5 (46)	5 (46)	5 (46)
Anesthesia equipment and monitoring (n=16)	4 (25)	6 (38)	8 (50)
Transplant anesthesia (n=19)	5 (26)	13 (68)	4 (21)
Inhalation anesthesia (n=21)	5 (24)	14 (67)	9 (43)
Obstetric anesthesia (n=25)	11 (44)	15 (60)	9 (36)
Pediatric anesthesia (n=24)	8 (33)	17 (71)	9 (38)
Ambulatory anesthesia (n=11)	6 (55)	8 (73)	3 (27)
Neuromuscular blocking agents (n=17)	3 (18)	12 (71)	3 (18)
Geriatric anesthesia (n=17)	10 (59)	15 (88)	11 (65)
Regional anesthesia (n=22)	11 (50)	15 (68)	7 (32)
Neuro-anesthesia (n=20)	9 (45)	16 (80)	11 (55)
Anesthetic pharmacology (n=11)	3 (27)	7 (64)	3 (27)
Intravenous anesthesia (n=13)	5 (39)	5 (39)	6 (46)
Cardiac anesthesia (n=14)	3 (21)	10 (71)	3 (21)
Thoracic anesthesia (n=17)	5 (29)	9 (53)	4 (24)
Fluids and transfusion (n=19)	6 (32)	14 (74)	6 (32)
Cardio-pulmonary resuscitation (n=17)	7 (41)	13 (77)	5 (29)
Pain clinic (n=57)	20 (35)	42 (74)	24 (42)
Intensive care unit (n=31)	17 (55)	23 (74)	12 (39)
Sedation or anesthesia outside the operating theater (n=11)	6 (55)	8 (73)	4 (36)
Total (n=398)	150 (37.7)	270 (67.8)	148 (37.2)

[a]Number of questions in each subfield is presented in parentheses.

[b]Performances of ChatGPT and CLOVA X are presented as the number of correct answers for each subfield along with the percentage of correct answers out of the total number of questions in parentheses.

Table 3 presents a comparison of the performances of GPT-3.5, GPT-4, and CLOVA X based on the question type. The models exhibited no significant performance differences when clinical cases were included. However, in terms of the level of logical thinking, GPT-3.5 and CLOVA X showed no significant difference, whereas GPT-4 showed a significantly higher performance for higher-order questions than for first-order questions (77% vs 64.2%; *P*=.02).

Table 3. Performance based on the inclusion of a clinical case and the level of logical thinking in the question.

Category and number of questions	GPT-3.5[a], n (%)	*P* value[b]	GPT-4[a], n (%)	*P* value[b]	CLOVA X[a], n (%)	*P* value[b]
Case		.57		.20		.11
Included (n=185)	73 (39.5)		132 (71.4)		77 (41.6)	
Not included (n=213)	77 (36.2)		138 (64.8)		71 (33.3)	
Level		.35		.02		.09
First-order (n=285)	112 (39.3)		183 (64.2)		98 (34.4)	
Higher-order (n=113)	38 (33.6)		87 (77)		50 (44.2)	

[a]Performances of ChatGPT and CLOVA X are presented as the number of correct answers for each category, along with the percentage of correct answers out of the total number of questions in parentheses.

[b]A chi-square test was conducted to compare each performance of GPT-3.5, GPT-4, and CLOVA X according to the inclusion of cases and the level of logical thinking, and the *P* values are presented.

Table 4 presents the differences in GPT-4 performance between the original Korean questions and their English versions. All examination questions translated from Korean to English using ChatGPT-4 were accurate and appropriate. Overall, GPT-4

XSL•FO
RenderX

performed significantly better on English-translated questions than on Korean originals (75.4% vs 67.8%; difference 7.5%; 95% CI 3.1%-11.9%; *P*=.001). When analyzed by year, the performance was consistently higher in English than in Korean, with the difference reaching statistical significance only in 2019 (75.3% vs 62.3%; *P*=.01). Furthermore, the overall self-agreement rate between the Korean and English-translated versions was 72.6%. In 14.1% of cases, correct answers were derived only from the English-translated version, and in 6.5% of cases, correct answers were derived solely from the original Korean questions.

Table 4. Performance of GPT-4 on Korean and English versions.

Year (questions)[a]	Korean[b], n (%)	English[b], n (%)	Difference (95% CI)[c]	*P* value[c]	Correct answers in each language[d]			Self-agreement[d], n (%)
					Both language, n (%)	Korean only, n (%)	English only, n (%)	
2022 (72)	49 (68)	54 (75)	6.9% (–4.2 to 18.1)	.33	43 (60)	6 (8)	11 (15)	49 (68)
2021 (74)	51 (69)	57 (77)	8.1% (–2.3 to 18.5)	.21	46 (62)	5 (7)	11 (15)	54 (73)
2020 (79)	54 (68)	59 (75)	6.3% (–4.4 to 17.1)	.36	47 (60)	7 (9)	12 (15)	55 (70)
2019 (85)	53 (62)	64 (75)	12.9% (3.8 to 22.0)	.01	50 (59)	3 (4)	14 (17)	62 (73)
2018 (88)	63 (72)	66 (75)	3.4% (–4.6 to 11.4)	.58	58 (66)	5 (6)	8 (9)	69 (78)
Total (398)	270 (67.8)	300 (75.4)	7.5% (3.1 to 11.9)	.001	244 (61.3)	26 (6.5)	56 (14.1)	289 (72.6)

[a]Number of questions included in the overall and yearly examinations is presented in parentheses.

[b]Performance in GPT-4 is presented as the number of correct answers for each language along with the percentage of correct answers out of the total number of questions in parentheses.

[c]McNemar's test was conducted to compare the performance of GPT-4 in Korean and English, and the differences of proportion (95% CI) with the *P* values are presented.

[d]Other variables, such as the number of correct answers in both languages, Korean only, and English only, and the self-agreement rate of ChatGPT answers when tested in Korean and English, are presented as numbers and percentages.

Table 5 presents the categorized reasons for the incorrect answers for each model. In all models, over 70% of the incorrect answers were due to informational errors, whereas less than 10% were caused by simple logical errors.

Table 5. Reasons for incorrect answers.

Category[a]	GPT-3.5	GPT-4 (Korean)	GPT-4 (English)	CLOVA X
Logical error, n (%)	24 (9.7)	11 (8.6)	7 (7.1)	4 (1.6)
Information error, n (%)	183 (73.8)	107 (83.6)	86 (87.8)	185 (74.0)
Statistical error, n (%)	3 (1.2)	1 (0.8)	1 (1.0)	3 (1.2)
Logical and information errors, n (%)	38 (15.3)	9 (7.0)	4 (4.1)	58 (23.2)
Overall, n	248	128	98	250

[a]Reasons for incorrect answers by ChatGPT and CLOVA X are presented as numbers with percentages of the total number of incorrect answers in parentheses.

Discussion

Principal Findings

This study assessed the proficiency of ChatGPT in the fields of anesthesiology and pain medicine by analyzing its performance on in-training examinations administered to Korean anesthesiology residents over the past 5 years. Our findings revealed that GPT-4 performed better in solving Korean-language problems in this field than its predecessors, GPT-3.5 and CLOVA X, which were trained using a Korean-language database. An interesting observation emerged when examination questions originally written in Korean were translated into English. In this scenario, GPT-4 exhibits higher performance levels. This suggests an enhanced capability of GPT-4 to process and respond to questions in English compared to Korean. However, it is important to note that despite this improved performance in the English-translated examinations, GPT-4 did not meet the recommended performance level for educational tools (>95%) [15].

Comparison to the Literature

In the fields of anesthesiology and pain medicine, the ChatGPT knowledge base has been rigorously evaluated using various practical questions. A previous report involving 1321 questions from the American Board of Anesthesiology (ABA) examination preparation book revealed that GPT-3.5 attained a correct answer rate of 56.2% [4]. A recent follow-up report with the same set of questions in GPT-4 discovered a remarkable improvement, with a correct answer rate of 72.1% [16]. In a separate evaluation using 3705 questions from the Fellowship of the Royal College of Anaesthetists Primary examination, GPT-3.5 achieved a

XSL•FO
RenderX

higher correct answer rate of 69.7% [5]. Furthermore, in a study that used a mock ABA examination comprising questions from the ABA website and examination preparation book, GPT-4 was the only tool among its peers, including GPT-3.5 and Google Bard, to pass all three stages of the examination [17]. However, these studies focused on English-language questions. This study differs by examining ChatGPT's performance on non-English questions, encompassing both translated versions and the original Korean questions. Additionally, this study provides a unique perspective by presenting the scoring results of Korean anesthesiology residents, facilitating a direct comparison between human performance and ChatGPT.

Additionally, our results reaffirmed the performance disparities of ChatGPT on English and Korean questions, as observed in recently reported studies in medicine. A notable study in the field of dermatology that used the Korean dermatology specialty certificate examination found that the English-translated version of the questions yielded significantly higher performance than the original Korean version (69.0% vs 57.0%) [18]. Another study assessed the performance differentials between GPT-3.5 and GPT-4 by translating cirrhosis-related questions into multiple languages including English, Korean, Mandarin, and Spanish [19]. This study revealed that GPT-4 consistently outperformed GPT-3.5 across all languages, with the performance gap being more pronounced in the Korean and Mandarin versions than in English. Notably, even GPT-4 demonstrated lower performance in Korean than in English, which is consistent with the trends observed in this study. The GPT-4 technical report by OpenAI provides further insight, indicating that while GPT-4's performance in Korean surpassed that of GPT-3.5 in English (77% vs 70.1%), it fell short compared to GPT-4's performance in English (85.5% vs 77.0%) [6]. This disparity in the language-specific performance of ChatGPT can be attributed to the predominance of English-based text in the GPT training data. This is particularly significant in the medical field, where there is more English literature than Korean literature. Consequently, the process of translating Korean questions into English for answer generation, followed by retranslation into Korean, likely affected performance. This is due to potential losses or alterations in meaning inherent in the translation process [20].

Implications of Findings

To the best of our knowledge, this is the first study to investigate the problem-solving performance of CLOVA X using medical knowledge. Although CLOVA X is a generative AI tool developed based on the Korean large-scale language model AI HyperCLOVA X, its performance in solving anesthesiology and pain medicine problems posed in Korea was inferior to that of GPT-4 and similar to that of GPT-3.5. This likely resulted from the HyperCLOVA X being trained exclusively on Korean data. The size of medical knowledge data sets likely varies by language [21], and English is presumed to contain more extensive medical knowledge data than other languages. Therefore, while CLOVA X might have advantages in processing Korean compared with ChatGPT, its limitations in specialized medical knowledge areas could be attributed to the limitations of its training data set.

In the results of the subfields of anesthesiology and pain medicine, the highest performances were observed in "geriatric anesthesia" in all three tools, whereas the subfield with the lowest performance was "neuromuscular blocking agents" in GPT-3.5 and CLOVA X, and "anesthesia equipment and monitoring" in GPT-4. This may be because the contents on neuromuscular blocking agents, anesthesia equipment, and monitoring are generally included in specialized textbooks that are less publicly accessible. However, neuromuscular blocking agents, which showed lower performance in GPT-3.5 and CLOVA X, showed higher-than-average performances in GPT-4.0. This improvement in GPT-4.0 suggests the potential for more sophisticated language-understanding models in these specialized fields. On the other hand, for questions about anesthesia equipment and monitoring or intravenous anesthesia, CLOVA X scored higher than both GPT-3.5 and GPT-4. It can be assumed that this prominent deviation from the overall score pattern reflects the differences in the data on which each language model was trained. Although this study did not have enough questions in each subfield to investigate the differences between them, further research on the performance differences of ChatGPT or CLOVA X across specific subfields is necessary for the future use of AI in anesthesiology and pain medicine.

The results of this study indicate that GPT-4 has the potential to surpass the correct answer rate of Korean anesthesiology residents for both Korean and English examination questions, thus meeting the passing criteria. Despite this achievement, GPT-4 failed to show acceptable performance as a reliable educational tool (>95%) [3,15] and also had the following limitations stemming from the fundamental operational mechanisms of large language models such as GPT [22]. Unlike human reasoning processes, these models generate responses based on probability distributions and likely word combinations rather than a genuine understanding of the learned content. Moreover, the possibility of incorrect GPT learning could not be ignored. Consequently, despite training with large data sets, the generated responses may be erroneous. Moreover, the model's tendency to provide plausible yet erroneous explanations for incorrect answers poses a significant risk of disseminating misinformation to anesthesiology residents. Therefore, we conclude that these models are inadequate for medical education applications, owing to their inherent limitations.

Opportunities for Future Work

Although GPT-4 demonstrated a higher level of knowledge in solving anesthesiology problems than Korean anesthesiology residents, its performance was not sufficiently reliable to be taken at face value. This performance shortfall is primarily attributed to the lack of training data in specialized fields such as medicine. Our analysis of incorrect answers also revealed that misinformation was the most common cause of error. If accurate information from professional medical texts is included in the training data of generative AI and is continuously updated, its performance in the medical field can be improved. However, the potential legal implications of using copyrighted textbooks, such as copyright infringement [23], further complicate the prospects of incorporating specialized medical texts into generative AI training in the future. Addressing these issues is

XSL•FO
RenderX

essential to enhance the medical knowledge of generative AI models.

Limitations

This study, while pioneering in its exploration of ChatGPT performance in Korean anesthesiology questions, had several limitations. First, the representativeness of the in-training examinations for Korean anesthesiology residents as a comprehensive measure of anesthesiology knowledge remains controversial. However, this examination was selected because it is the only test that is readily accessible to Korean anesthesiologists. Second, due to the inherent limitations of ChatGPT, our analysis excluded questions that incorporated images, diagrams, or photographs. Therefore, we were unable to directly compare the actual examination results of the residents with the performance of the AI services. Additionally, this limitation hinders our ability to fully evaluate the performance of AI services during anesthesiology examinations. This exclusion also potentially limits the scope of our findings as these elements are integral to many medical questions. Third, we used a selective data set that may not have fully captured the performance of AI across a broad range of medical scenarios in the field of anesthesiology. Future research should incorporate nonselective data sets to ensure a more comprehensive and generalizable evaluation of AI performance. Ultimately, owing to these limitations, we could only investigate a partial aspect of AI performance in understanding anesthesiology knowledge. Despite these limitations, this study is the first to assess the capabilities of ChatGPT in handling anesthesiology questions in Korean. We expect our findings to stimulate discussion and consideration among Korean anesthesiologists regarding the potential roles and limitations of AI tools, such as ChatGPT, in the field of anesthesiology. In addition, by demonstrating performance differences in GPT in English and Korean, this study raises the issue of narrowing the performance gap across different languages.

Conclusions

In summary, this study demonstrated that, although GPT-4 is advanced compared to its predecessors in processing Korean anesthesiology examination questions, it has yet to reach a level of reliability that would justify its use as a standalone educational tool in the medical domain. Specifically, our research highlights the significant performance disparity between English and Korean ChatGPT outputs, drawing attention to the challenges inherent in evaluating proficiency in non-English medical content. This investigation of the capabilities of ChatGPT in Korean anesthesiology is a pioneering effort, and the potential of this tool to assist medical professionals is promising. However, our findings necessitate a cautious approach to their application in clinical and educational settings. This study serves as a call for continued research and development in this area to enhance the performance of AI tools, such as ChatGPT, in diverse linguistic and professional contexts.

Acknowledgments

This study used data on the examination scores of residents provided by the Training and Education Committee of the Korean Society of Anesthesia.

Data Availability

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

Authors' Contributions

HJL and SHY conceptualized and designed the study. HJL and SHY contributed to data acquisition. All authors contributed to the data analysis and interpretation. HJL and SHY drafted the initial manuscript and all authors substantially revised it. All the authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of the prompts in ChatGPT (GPT-4) in Korean.
[PDF File (Adobe PDF File), 195 KB - mededu_v10i1e56859_app1.pdf]

Multimedia Appendix 2
Screenshots of the prompts in ChatGPT (GPT-4) in English. A) An example from the translation process. Each question in Korean was entered in succession, along with the translation command. B) The initial command of the testing process. The command was entered in English which was translated from the original Korean version. C) An example of testing ChatGPT with English questions.
[PDF File (Adobe PDF File), 79 KB - mededu_v10i1e56859_app2.pdf]

XSL•FO
RenderX

Multimedia Appendix 3
The actual scores of Korean anesthesiology residents.
[DOCX File , 12 KB - mededu_v10i1e56859_app3.docx]

References

1. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-11-01]
2. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
3. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG 2024;131(3):378-380. [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
4. Shay D, Kumar B, Bellamy D, Palepu A, Dershwitz M, Walz JM, et al. Assessment of ChatGPT success with specialty medical knowledge using anaesthesiology board examination practice questions. Br J Anaesth 2023;131(2):e31-e34 [FREE Full text] [doi: 10.1016/j.bja.2023.04.017] [Medline: 37210278]
5. Birkett L, Fowler T, Pullen S. Performance of ChatGPT on a primary FRCA multiple choice question bank. Br J Anaesth 2023;131(2):e34-e35 [FREE Full text] [doi: 10.1016/j.bja.2023.04.025] [Medline: 37210281]
6. OpenAI. GPT-4 technical report. ArXiv. Preprint posted online on March 15, 2023 2023:1-100. [doi: 10.48550/arXiv.2303.08774]
7. Ali R, Tang OY, Connolly ID, Sullivan PLZ, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93(6):1353-1365. [doi: 10.1227/neu.0000000000002632] [Medline: 37581444]
8. Lin JC, Younessi DN, Kurapati SS, Tang OY, Scott IU. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. Eye (Lond) 2023;37(17):3694-3695. [doi: 10.1038/s41433-023-02564-2] [Medline: 37156862]
9. Aldridge MJ, Penders R. Artificial intelligence and anaesthesia examinations: exploring ChatGPT as a prelude to the future. Br J Anaesth 2023;131(2):e36-e37 [FREE Full text] [doi: 10.1016/j.bja.2023.04.033] [Medline: 37244834]
10. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023;9:e48002 [FREE Full text] [doi: 10.2196/48002] [Medline: 37384388]
11. Residents pages. Official Website of The Korean Society of Anesthesiologists. URL: https://anesthesia.or.kr/new_record/ [accessed 2023-08-16]
12. HyperCLOVA X Team. HyperCLOVA X technical report. ArXiv. Preprint posted online on April 2, 2024 2024:1-44. [doi: 10.48550/arXiv.2404.01954]
13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
14. Sheskin DJ. The Cochran Q test. In: Handbook of Parametric and Nonparametric Statistical Procedures. 5th Edition. UK: Chapman and Hall/CRC; 2011:1119-1136.
15. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology self-assessment test. Am J Gastroenterol 2023;118(12):2280-2282. [doi: 10.14309/ajg.0000000000002320] [Medline: 37212584]
16. Shay D, Kumar B, Redaelli S, von Wedel D, Liu M, Dershwitz M, et al. Could ChatGPT-4 pass an anaesthesiology board examination? Follow-up assessment of a comprehensive set of board examination practice questions. Br J Anaesth 2024;132(1):172-174. [doi: 10.1016/j.bja.2023.10.025] [Medline: 37996275]
17. Angel MC, Rinehart JB, Cannesson MP, Baldi P. Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Anesthesiology: A Comparative Study on the American Board of Anesthesiology Examination. Anesth Analg 2024 Aug 01;139(2):349-356 [FREE Full text] [doi: 10.1213/ANE.0000000000006892] [Medline: 38640076]
18. Joh H, Kim M, Ko J, Kim J, Jue M. Evaluating the performance of ChatGPT in a dermatology specialty certificate examination: a comparative analysis between English and Korean language settings. Research Square 2023. [doi: 10.21203/rs.3.rs-3241164/v1]
19. Yeo YH, Samaan JS, Ng WH, Ma X, Ting PS, Kwak MS, et al. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. medRxiv. Preprint posted online on May 5, 2023 2023:1-20. [doi: 10.1101/2023.05.04.23289482]
20. Zhang X, Li S, Hauer B, Shi N, Kondrak G. Don't trust ChatGPT when your question is not in English: a study of multilingual abilities and types of LLMs. ArXiv. Preprint posted online on May 24, 2023 2023:1-13. [doi: 10.48550/arXiv.2305.16339]
21. Baethge C. The languages of medicine. Dtsch Arztebl Int 2008;105(3):37-40 [FREE Full text] [doi: 10.3238/arztebl.2008.0037] [Medline: 19633751]
22. Orrù G, Piarulli A, Conversano C, Gemignani A. Human-like problem-solving abilities in large language models using ChatGPT. Front Artif Intell 2023;6:1199350. [doi: 10.3389/frai.2023.1199350] [Medline: 37293238]

XSL•FO
RenderX

23. Lucchi N. ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. Eur J Risk Regul 2023:1-23. [doi: 10.1017/err.2023.59]

Abbreviations

ABA: American Board of Anesthesiology
AI: artificial intelligence
KSA: Korean Society of Anesthesiologists

Edited by B Lesselroth; submitted 28.01.24; peer-reviewed by J Bruthans, A Hidki; comments to author 06.05.24; revised version received 10.06.24; accepted 15.08.24; published 16.09.24.

Please cite as:
Yoon SH, Oh SK, Lim BG, Lee HJ
Performance of ChatGPT in the In-Training Examination for Anesthesiology and Pain Medicine Residents in South Korea: Observational Study
JMIR Med Educ 2024;10:e56859
URL: https://mededu.jmir.org/2024/1/e56859
doi:10.2196/56859
PMID:

©Soo-Hyuk Yoon, Seok Kyeong Oh, Byung Gun Lim, Ho-Jin Lee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 16.09.2024. 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.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Psychological Safety Competency Training During the Clinical Internship From the Perspective of Health Care Trainee Mentors in 11 Pan-European Countries: Mixed Methods Observational Study

Irene Carrillo[1], MSc, PhD; Ivana Skoumalová[2], PhD; Ireen Bruus[3], MSc; Victoria Klemm[4]; Sofia Guerra-Paiva[5,6], MsD; Bojana Knežević[7], PhD; Augustina Jankauskiene[8], PhD; Dragana Jocic[9], PhD; Susanna Tella[10], PhD; Sandra C Buttigieg[11], PhD; Einav Srulovici[12], PhD; Andrea Madarasová Gecková[2,13], PhD; Kaja Põlluste[14], PhD; Reinhard Strametz[4], MD; Paulo Sousa[5,6], PhD; Marina Odalovic[15], PhD; José Joaquín Mira[1,16], MPH, PhD

[1]Department of Health Psychology, Miguel Hernández University of Elche, Elche, Spain

[2]Department of Health Psychology and Research Methodology, Faculty of Medicine, Pavol Jozef Šafárik University, Kosice, Slovakia

[3]Tartu Health Care College, Tartu, Estonia

[4]Wiesbaden Institute for Healthcare Economics and Patient Safety (WiHelP), Wiesbaden Business School, RheinMain University of Applied Sciences, Wiesbaden, Germany

[5]Public Health Research Centre, National School of Public Health, NOVA University Lisbon, Lisbon, Portugal

[6]Comprehensive Health Research Center, National School of Public Health, NOVA University Lisbon, Lisbon, Portugal

[7]University Hospital Centre Zagreb, University of Zagreb, Zagreb, Croatia

[8]Pediatric Center, Institute of Clinical Medicine, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

[9]BENU Pharmacy, PHOENIX Group Serbia, Belgrade, Serbia

[10]Faculty of Social and Health Care, LAB University of Applied Sciences, Lappeenranta, Finland

[11]Department of Health Systems Management and Leadership, Faculty of Health Sciences, University of Malta, Malta, Malta

[12]Cheryl Spencer Department of Nursing, University of Haifa, Haifa, Israel

[13]Institute of Applied Psychology, Faculty of Social and Economic Sciences, Comenius University Bratislava, Bratislava, Slovakia

[14]Institute of Clinical Medicine, University of Tartu, Tartu, Estonia

[15]Faculty of Pharmacy, University of Belgrade, Belgrade, Serbia

[16]Foundation for the Promotion of Health and Biomedical Research of the Valencia Region (FISABIO), Sant Joan d'Alacant, Spain

Corresponding Author:
Irene Carrillo, MSc, PhD
Department of Health Psychology
Miguel Hernández University of Elche
Avenida de la Universidad s/n
Elche, 03202
Spain
Phone: 34 966658350
Email: icarrillo@umh.es

Related Article:

Correction of: https://mededu.jmir.org/2024/1/e64125/

(JMIR Med Educ 2024;10:e68503) doi:10.2196/68503

In "Psychological Safety Competency Training During the Clinical Internship From the Perspective of Health Care Trainee Mentors in 11 Pan-European Countries: Mixed Methods Observational Study" (JMIR Medical Education 2024;1(10): *e64125*) the authors made one revision.

The following section in the Acknowledgments section:

> *This paper was based on work from the European Cooperation in Science and Technology Action 19113, The European Researchers' Network Working on Second Victims, supported by the European Cooperation in Science and Technology [67]*

Has been corrected to:

XSL•FO
RenderX

This article is based upon work from COST Action, TheERNSTGroup, CA19113, supported by COST (European Cooperation in Science and Technology).

Additionally, the associated reference [67] (below) will be removed from the Reference List, as it will no longer be cited in the paper.

67. Home page. COST Association. URL: https://www.cost.eu/ [accessed 2024-04-29]

The correction will appear in the online version of the paper on the JMIR Publications website on November 15, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 07.11.24; this is a non–peer-reviewed article;accepted 07.11.24; published 15.11.24.

Please cite as:
Carrillo I, Skoumalová I, Bruus I, Klemm V, Guerra-Paiva S, Knežević B, Jankauskiene A, Jocic D, Tella S, Buttigieg SC, Srulovici E, Madarasová Gecková A, Põlluste K, Strametz R, Sousa P, Odalovic M, Mira JJ
Correction: Psychological Safety Competency Training During the Clinical Internship From the Perspective of Health Care Trainee Mentors in 11 Pan-European Countries: Mixed Methods Observational Study
JMIR Med Educ 2024;10:e68503
URL: https://mededu.jmir.org/2024/1/e68503
doi:10.2196/68503
PMID:

©Irene Carrillo, Ivana Skoumalová, Ireen Bruus, Victoria Klemm, Sofia Guerra-Paiva, Bojana Knežević, Augustina Jankauskiene, Dragana Jocic, Susanna Tella, Sandra C Buttigieg, Einav Srulovici, Andrea Madarasová Gecková, Kaja Põlluste, Reinhard Strametz, Paulo Sousa, Marina Odalovic, José Joaquín Mira. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.11.2024. 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.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment

Aidan Gilson[1,2], BS; Conrad W Safranek[1], BS; Thomas Huang[2], BS; Vimig Socrates[1,3], MS; Ling Chi[1], BSE; Richard Andrew Taylor[1,2*], MD, MHS; David Chartash[1,4*], PhD

[1]Section for Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, CT, United States
[2]Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT, United States
[3]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States
[4]School of Medicine, University College Dublin, National University of Ireland, Dublin, Dublin, Ireland
[*]these authors contributed equally

Corresponding Author:
David Chartash, PhD
Section for Biomedical Informatics and Data Science
Yale University School of Medicine
300 George Street
Suite 501
New Haven, CT, 06511
United States
Phone: 1 203 737 5379
Email: david.chartash@yale.edu

Related Article:

Correction of: https://mededu.jmir.org/2023/1/e45312

(JMIR Med Educ 2024;10:e57594) doi:10.2196/57594

In “How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge” (MIR Med Educ 2023;9:e45312) three additions were made to enhance discoverability.

The title originally appeared as:

How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment

And has been changed to:

How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge

In the “Objective” section of the Abstract, the following sentence:

This study aimed to evaluate the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination Step 1 and Step 2 exams, as well as to analyze responses for user interpretability.

Has been changed to read as:

This study aimed to evaluate the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams, as well as to analyze responses for user interpretability.

Finally, the abbreviation “USMLE” has been added to the Keywords section.

The correction will appear in the online version of the paper on the JMIR Publications website on February 27, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 20.02.24; this is a non–peer-reviewed article;accepted 20.02.24; published 27.02.24.

Please cite as:
Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D
Correction: How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment
JMIR Med Educ 2024;10:e57594
URL: https://mededu.jmir.org/2024/1/e57594
doi:10.2196/57594
PMID:38412478

©Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash. Originally published in JMIR Medical Education (https://mededu.jmir.org), 27.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Students' Perceived Knowledge, Skills, Attitudes, and Experience

Lena Rettinger[1,2], BSc, MSc; Peter Putz[3], Mag, Dr Rer Nat; Lea Aichinger[1], BSc, MSc; Susanne Maria Javorszky[4], BSc, MSc; Klaus Widhalm[5], MSc; Veronika Ertelt-Bach[6], Mag, MAS; Andreas Huber[7], MSc; Sevan Sargis[8], BSc, MSc; Lukas Maul[1], BSc, MSc; Oliver Radinger[9], BA, Dr; Franz Werner[1], Mag, Dr Tech; Sebastian Kuhn[2], MME, Prof Dr

[1]Health Assisting Engineering, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[2]Institute of Digital Medicine, Philipps-University & University Hospital of Giessen and Marburg, Marburg, Germany
[3]Competence Center INDICATION, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[4]Logopedics – Phoniatrics - Audiology, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[5]Physiotherapy, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[6]Occupational Therapy, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[7]Orthoptics, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[8]Midwifery, FH Campus Wien, University of Applied Sciences, Vienna, Austria
[9]Competence Center Nursing Sciences, FH Campus Wien, University of Applied Sciences, Vienna, Austria

Corresponding Author:
Lena Rettinger, BSc, MSc
Health Assisting Engineering
FH Campus Wien
University of Applied Sciences
Favoritenstrasse 226
Vienna, 1100
Austria
Phone: 43 1 606 68 77 ext 4382
Email: lena.rettinger@fh-campuswien.ac.at

Related Article:

Correction of: https://mededu.jmir.org/2024/1/e51112/

(JMIR Med Educ 2024;10:e59919) doi:10.2196/59919

In "Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Students' Perceived Knowledge, Skills, Attitudes, and Experience" (JMIR Med Educ 2024;10:e51112) an error was noted.

In the title, the word "student's" has been revised to "students'".

Therefore, the original title:

> *Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Student's Perceived Knowledge, Skills, Attitudes, and Experience*

Has been revised to:

> *Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Students' Perceived Knowledge, Skills, Attitudes, and Experience*

The correction will appear in the online version of the paper on the JMIR Publications website on April 26, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

XSL•FO
RenderX

Submitted 25.04.24; this is a non–peer-reviewed article;accepted 26.04.24; published 26.04.24.

Please cite as:
Rettinger L, Putz P, Aichinger L, Javorszky SM, Widhalm K, Ertelt-Bach V, Huber A, Sargis S, Maul L, Radinger O, Werner F, Kuhn S
Correction: Telehealth Education in Allied Health Care and Nursing: Web-Based Cross-Sectional Survey of Students' Perceived Knowledge, Skills, Attitudes, and Experience
JMIR Med Educ 2024;10:e59919
URL: https://mededu.jmir.org/2024/1/e59919
doi:10.2196/59919
PMID:38669670

©Lena Rettinger, Peter Putz, Lea Aichinger, Susanne Maria Javorszky, Klaus Widhalm, Veronika Ertelt-Bach, Andreas Huber, Sevan Sargis, Lukas Maul, Oliver Radinger, Franz Werner, Sebastian Kuhn. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.04.2024. 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.

XSL•FO
RenderX

Editorial

ChatGPT in Medical Education: A Precursor for Automation Bias?

Tina Nguyen[1], 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, as is necessary, the use of ChatGPT in medical education should be limited to certain tasks. Otherwise, having no constraints on what ChatGPT should be used for could lead to automation bias.

(JMIR Med Educ 2024;10:e50174) doi:10.2196/50174

KEYWORDS

ChatGPT; artificial intelligence; AI; medical students; residents; medical school curriculum; medical education; automation bias; large language models; LLMs; bias

Introduction

With the introduction of artificial intelligence (AI), automated processes for nearly most tasks have become the norm. In the clinical environment, AI has been used for diagnosis, prognosis, and administrative tasks. Given the popularity of other forms of AI—as seen most recently with ChatGPT, a large language model developed by the company OpenAI—there are suggestions for its potential role in medical education. Users of ChatGPT boast its efficiency and relative accuracy, such as in the generation of a patient's discharge summary or the conduction of literature reviews [1]. As advancements in medicine continue to arise, medical students are burdened with the impossible task of balancing the need to continuously learn and retain competencies and the need to provide compassionate patient care. As a result, some medical students might feel an incentive to use ChatGPT to save them time in their busy schedules. However, despite the novel acclaim, the technical and ethical issues seen with AI, such as biased results or nonsensical outputs, also plague ChatGPT. These problems become exacerbated when medical students inadvertently develop automation bias, where they overrely on AI, and continue to have this mentality when they become residents, at which point they have the potential to harm patients if the AI provides an erroneous outcome. In this editorial, I argue the justification for AI education in the medical school curriculum and how the lack of it leads to the problem of automation bias, as well as the other harms from automation bias. Subsequently, I connect the implications of students using ChatGPT with automation bias. Finally, I provide recommendations for when ChatGPT use is appropriate.

XSL•FO
RenderX

The Need for AI Education in the Medical School Curriculum

As the health care landscape has drastically changed through the years, physicians have had to quickly adapt to the digital age. Given the amount of information physicians are required to retain and the new information they must continue to learn, such as information on emerging diseases and the health data of the patients they track, physicians are expected to interact with computer systems in some capacity, whether it is for charting their patients' information or consulting clinical decision support systems. However, the lack of content on the technological systems in the health care setting inhibits prospective physicians from understanding the benefits of using these technologies, the ethical issues that can arise with their use, and future innovations, along with the wider implications of AI. In Civaner et al's [2] survey of medical students' opinions on AI education, they found that 75.6% of students had either limited or no education on the topic of AI. These participants also noted not feeling well equipped to work with AI in the clinical setting. Additionally, in Yun et al's [3] proposal for future internal medicine physicians, they suggested that these prospective physicians should be able to appreciate the roles of big data and AI in health care. Clearly, there is a desire from students, as well as residency and fellowship programs, to incorporate AI education into the medical school curriculum and training. AI education and training cannot continue to be delayed, as some forms of AI have already been deployed in the clinical setting.

Although several studies have provided proposals for implementing AI education into the medical school curriculum, they have also noted the difficulties of developing AI education, such as schedule constraints and the challenges of deciding the material that should be covered [4,5]. Additionally, this task should not solely be deferred to the attending physicians, as they themselves might not have the adequate training with AI to teach others [5]. Although these challenges serve as barriers to implementing quality AI education into the curriculum, an attempt to include at least some type of education on or educational resources about AI is needed to prepare students and potentially prevent problems in the clinical setting, as further explored in the following section. Therefore, future physicians, medical students, and residents should be trained on the use of AI in health care and other related topics, such as big data or machine learning, to understand the tools they will be working with. Even though medical students should not be expected to be experts in AI and know every technical aspect of these technologies, they should at least feel comfortable with navigating how and when to use AI.

The Problem of Automation Bias

Although AI is supposed to aid physicians in various processes to decrease their workload and give them more time with their patients, AI can also cause unintended ethical issues. One of the common ethical concerns with AI is that it can essentially be a black box, where the results from the AI are illogical, and the AI developer cannot track how it produced those erroneous results. This problem becomes exacerbated when automation bias arises. Automation bias occurs when a user overrelies on AI systems. Therefore, if a physician exhibits automation bias, then they will not question the results from the AI, potentially leading to bad medical care. In Lyell et al's [6] study, the error rate associated with a clinical decision support system when it was inaccurate was higher (86.6%) in comparison to the rate it had when it was accurate (58.8%). Although automated processes aid in decision-making and can provide accurate results, there is also the possibility of these systems providing incorrect results and causing irreversible harm on a much larger scale. An example includes the Prescription Drug Monitoring Program (PDMP), a machine learning system that provides risk scores for patients' likelihood to misuse prescription drugs, which can cause both testimonial injustice and physical harm [7,8]. Testimonial injustice, a form of epistemic injustice, develops when a patient's account of their health is unfairly dismissed by their physician [8]. Testimonial injustice invalidates the credibility of patients and further implies that their care is dependent on how physicians deem their trustworthiness [8]. A patient's risk scores can be negatively affected if their chart becomes commingled, which is also known as *overlay*, where a specific person's electronic health record erroneously pulls in the data of other patients with similar demographic characteristics and compiles these data into 1 chart [7,9]. As such, a patient with chronic pain may not receive the medication they need due to the PDMP providing an incorrect risk score. If a physician uses the risk scores of the PDMP without validating the results or considering their patients' testimonies, then physical harm, as well as patients' mistrust toward the physician and the potential deterrence of seeking health care, will ensue. Although AI can aid in the decision-making process, ultimately it is the duty of the physician to ensure that their decisions are based on sound clinical judgment. As such, if a physician with automation bias applies an erroneous outcome to a patient's 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

XSL•FO
RenderX

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

XSL•FO
RenderX

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

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
2. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022 Nov 9;22(1):772 [FREE Full text] [doi: 10.1186/s12909-022-03852-3] [Medline: 36352431]
3. Yun HC, Cable CT, Pizzimenti D, Desai SS, Muchmore EA, Vasilias J, et al. Internal medicine 2035: preparing the future generation of internists. J Grad Med Educ 2020 Dec;12(6):797-800 [FREE Full text] [doi: 10.4300/JGME-D-20-00794.1] [Medline: 33391612]
4. Ngo B, Nguyen D, vanSonnenberg E. The cases for and against artificial intelligence in the medical school curriculum. Radiol Artif Intell 2022 Aug 17;4(5):e220074 [FREE Full text] [doi: 10.1148/ryai.220074] [Medline: 36204540]
5. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048 [FREE Full text] [doi: 10.2196/16048] [Medline: 31793895]

XSL•FO
RenderX

6. Lyell D, Magrabi F, Raban MZ, Pont LG, Baysari MT, Day RO, et al. Automation bias in electronic prescribing. BMC Med Inform Decis Mak 2017 Mar 16;17(1):28 [FREE Full text] [doi: 10.1186/s12911-017-0425-5] [Medline: 28302112]
7. Nguyen T. PDMP causes more than just testimonial injustice. J Med Ethics 2023 Aug;49(8):549-550. [doi: 10.1136/jme-2023-109112] [Medline: 37217278]
8. Pozzi G. Testimonial injustice in medical machine learning. J Med Ethics 2023 Aug;49(8):536-540. [doi: 10.1136/jme-2022-108630] [Medline: 36635066]
9. Landsbach GD. Study analyzes causes and consequences of patient overlay errors. J AHIMA 2016 Sep;87(9):40-43. [Medline: 29400427]
10. Ho A. Live Like Nobody is Watching: Relational Autonomy in the Age of Artificial Intelligence Health Monitoring. New York, NY: Oxford University Press; Mar 2023.
11. Goddard K, Roudsari A, Wyatt JC. Automation bias: empirical results assessing influencing factors. Int J Med Inform 2014 May;83(5):368-375. [doi: 10.1016/j.ijmedinf.2014.01.001] [Medline: 24581700]
12. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021 Feb 19;4(1):31 [FREE Full text] [doi: 10.1038/s41746-021-00385-9] [Medline: 33608629]
13. Logg JM, Minson JA, Moore DA. Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 2019 Mar;151:90-103 [FREE Full text] [doi: 10.1016/j.obhdp.2018.12.005]
14. Lubell J. ChatGPT passed the USMLE. What does it mean for med ed? American Medical Association. 2023 Mar 3. URL: https://www.ama-assn.org/practice-management/digital/chatgpt-passed-usmle-what-does-it-mean-med-ed [accessed 2024-01-09]
15. Tiwari CK, Bhat MA, Khan ST, Subramaniam R, Khan MAI. What drives students toward ChatGPT? An investigation of the factors influencing adoption and usage of ChatGPT. Interactive Technology and Smart Education 2023 Aug 29 Online ahead of print. [doi: 10.1108/itse-04-2023-0061]
16. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb 19;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
17. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J 2023 Feb 28;64(1):1-3 [FREE Full text] [doi: 10.3325/cmj.2023.64.1] [Medline: 36864812]
18. 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 6;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
19. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
20. Wang DY, Ding J, Sun AL, Liu SG, Jiang D, Li N, et al. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. J Am Med Inform Assoc 2023 Sep 25;30(10):1684-1692. [doi: 10.1093/jamia/ocad118] [Medline: 37561535]
21. Johnston SC. Anticipating and training the physician of the future: the importance of caring in an age of artificial intelligence. Acad Med 2018 Aug;93(8):1105-1106. [doi: 10.1097/ACM.0000000000002175] [Medline: 29443717]
22. Pfeifer CM. A progressive three-phase innovation to medical education in the United States. Med Educ Online 2018 Dec;23(1):1427988 [FREE Full text] [doi: 10.1080/10872981.2018.1427988] [Medline: 29353536]
23. What to expect in medical school. Association of American Medical Colleges. URL: https://students-residents.aamc.org/choosing-medical-career/what-expect-medical-school [accessed 2024-01-09]
24. Feng S, Shen Y. ChatGPT and the future of medical education. Acad Med 2023 Aug 1;98(8):867-868. [doi: 10.1097/ACM.0000000000005242] [Medline: 37162219]
25. Yusof YAM, Taridi NM, Mustapa M, Shaharuddin S, Hamid MWA, Shakrin NNSM, et al. Student-centred approach in medical education: a review of the teaching-learning activities and the perceptions of educators on the students engagement and performance at the Faculty of Medicine and Defence Health, National Defence University of Malaysia. Advances in Human Biology 2022;12(2):101-107. [doi: 10.4103/aihb.aihb_150_21]
26. van de Ridder JMM, Shoja MM, Rajput V. Finding the place of ChatGPT in medical education. Acad Med 2023 Aug 1;98(8):867. [doi: 10.1097/ACM.0000000000005254] [Medline: 37162206]
27. Klugman CM. How health humanities will save the life of the humanities. J Med Humanit 2017 Dec;38(4):419-430. [doi: 10.1007/s10912-017-9453-5] [Medline: 28642990]
28. Blanco-Gonzalez A, Cabezon A, Seco-Gonzalez A, Conde-Torres D, Antelo-Riveiro P, Pineiro A, et al. The role of AI in drug discovery: challenges, opportunities, and strategies. arXiv Preprint posted online on Dec 8, 2022. [FREE Full text] [doi: 10.48550/arXiv.2212.08104]
29. Nabi W. Utilizing technology to address gaps in medical education. Harvard Macy Institute. 2021 Nov 1. URL: https://harvardmacy.org/blog/utilizing-technology-gaps-meded [accessed 2024-01-09]

Abbreviations

AI: artificial intelligence

XSL•FO
RenderX

PDMP: Prescription Drug Monitoring Program

Edited by K Venkatesh; submitted 21.06.23; peer-reviewed by J Kim, S Arya, M Arab-Zozani; comments to author 28.09.23; accepted 11.12.23; published 17.01.24.

Please cite as:
Nguyen T
ChatGPT in Medical Education: A Precursor for Automation Bias?
JMIR Med Educ 2024;10:e50174
URL: https://mededu.jmir.org/2024/1/e50174
doi:10.2196/50174
PMID:38231545

©Tina Nguyen. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.01.2024. 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.

XSL•FO
RenderX

Opportunities to Improve Communication With Residency Applicants: Cross-Sectional Study of Obstetrics and Gynecology Residency Program Websites

Paulina M Devlin[1], MS, MD; Oluwabukola Akingbola[2], MS, DO; Jody Stonehocker[3], MD; James T Fitzgerald[4], PhD; Abigail Ford Winkel[5], MHPE, MD; Maya M Hammoud[4,6], MBA, MD; Helen K Morgan[4,6], MD

1
2
3
4
5
6

Corresponding Author:
Helen K Morgan, MD

Abstract

Background: As part of the residency application process in the United States, many medical specialties now offer applicants the opportunity to send program signals that indicate high interest to a limited number of residency programs. To determine which residency programs to apply to, and which programs to send signals to, applicants need accurate information to determine which programs align with their future training goals. Most applicants use a program's website to review program characteristics and criteria, so describing the current state of residency program websites can inform programs of best practices.

Objective: This study aims to characterize information available on obstetrics and gynecology residency program websites and to determine whether there are differences in information available between different types of residency programs.

Methods: This was a cross-sectional observational study of all US obstetrics and gynecology residency program website content. The authorship group identified factors that would be useful for residency applicants around program demographics and learner trajectories; application criteria including standardized testing metrics, residency statistics, and benefits; and diversity, equity, and inclusion mission statements and values. Two authors examined all available websites from November 2011 through March 2022. Data analysis consisted of descriptive statistics and one-way ANOVA, with P<.05 considered significant.

Results: Among 290 programs, 283 (97.6%) had websites; 238 (82.1%) listed medical schools of current residents; 158 (54.5%) described residency alumni trajectories; 107 (36.9%) included guidance related to the preferred United States Medical Licensing Examination Step 1 scores; 53 (18.3%) included guidance related to the Comprehensive Osteopathic Medical Licensing Examination Level 1 scores; 185 (63.8%) included international applicant guidance; 132 (45.5%) included a program-specific mission statement; 84 (29%) included a diversity, equity, and inclusion statement; and 167 (57.6%) included program-specific media or links to program social media on their websites. University-based programs were more likely to include a variety of information compared to community-based university-affiliated and community-based programs, including medical schools of current residents (113/123, 91.9%, university-based; 85/111, 76.6%, community-based university-affiliated; 40/56, 71.4%, community-based; P<.001); alumni trajectories (90/123, 73.2%, university-based; 51/111, 45.9%, community-based university-affiliated; 17/56, 30.4%, community-based; P<.001); the United States Medical Licensing Examination Step 1 score guidance (58/123, 47.2%, university-based; 36/111, 32.4%, community-based university-affiliated; 13/56, 23.2%, community-based; P=.004); and diversity, equity, and inclusion statements (57/123, 46.3%, university-based; 19/111, 17.1%, community-based university-affiliated; 8/56, 14.3%, community-based; P<.001).

Conclusions: There are opportunities to improve the quantity and quality of data on residency websites. From this work, we propose best practices for what information should be included on residency websites that will enable applicants to make informed decisions.

(JMIR Med Educ 2024;10:e48518) doi:10.2196/48518

KEYWORDS

obstetrics and gynecology; residency program; residency application; website; program signals; communication best practices

XSL•FO
RenderX

Introduction

In the United States, becoming an accredited physician is a rigorous and competitive process where candidates complete undergraduate training, medical school education, and residency training in a chosen specialty. Typically, individuals first obtain an undergraduate degree to gain admittance to a medical school. Next, they must earn a medical doctorate (MD) or doctor of osteopathic medicine (DO) from an accredited medical school or an equivalent international medical degree. Finally, they must complete postgraduate residency training; to fulfill this requirement, individuals apply to a residency program in their intended specialty. In the United States, many residency applicants are medical students in their final year of training, but individuals may also apply if they previously completed an MD or DO degree or completed medical school outside the United States and obtained certification from the Educational Commission for Foreign Medical Graduates [1]. All residency programs fulfill requirements set by the Accreditation Council for Graduate Medical Education, but programs have different strengths. Residency programs may be based in large university academic centers, community medical centers, or medical centers in a community setting that are affiliated with universities and often consequently emphasize clinical service to communities versus academic pursuits in training. Applying for residency is a competitive step in the physician training process; qualified applicants often apply to programs in a matching system that algorithmically matches applicants into programs that rank the applicant. In 2022, a total of 42,549 applicants were matched into 36,943 residency positions in the National Resident Matching Program Main Residency Match, making the overall match rate for all active applicants 86.8% [2]. This match rate, however, does not illustrate the full story; there is a wide range of match rates for different types of applicants and specialties, and the number of applicants who do not match into their top programs of interest is increasing [3].

Due to this competitiveness, now more than ever, residency applicants need transparent data to make informed decisions during the residency application process. Applicants determine where to apply, and among an increasing number of specialties, they must also decide where to send program signals—electronic tokens indicating high interest in a program—at the time of application submission. In the 2022 - 2023 application cycle, 17 specialties opted to include program signaling [4-7]. Ideally, applicants should apply and send signals to programs that align with their values and priorities and to programs where they have a reasonable chance of matching [4]. Determining which programs meet these criteria is a challenge for applicants; they rely on a variety of nationally available data sources [8,9] and have particularly valued information from program websites for their application decision-making [10-12]. Therefore, our study sought to characterize content available on obstetrics and gynecology (OBGYN) residency program websites and to determine whether there were differences in website content according to program type and geographic location. Our goal was to use this information to inform best practices for residency program websites.

Methods

Study Design

This was a cross-sectional observational study of US OBGYN residency program websites. We examined programs listed on the Electronic Residency Application Service (ERAS) 2022 Participating Specialties and Programs website. All programs listed on March 22, 2022, were included. Data for whether the type or program was university-based, community-based university-affiliated, or community-based were obtained by searching for the program in the American Medical Association's Fellowship and Residency Electronic Interactive Database Access System. Data for the census region and division of programs were determined based on the US Census Bureau Regions and Divisions with State FIPS Codes document.

Two authors (PMD and OA) collected data between November 2021 and March 2022. After obtaining the list of residency programs, we searched for a website associated with the program through a direct link from the ERAS list. In cases where a link was unavailable or incorrect, a Google search was conducted to attempt to find a website. Individual programs were not contacted directly by the study team.

The authorship group identified factors that would be useful for residency applicants. This group consisted of OBGYN faculty with education leadership roles, an OBGYN resident, and an OBGYN medical student applicant. The group used experiences from these roles to iteratively create a list of factors to consider, including program demographics and learner trajectories, application criteria including standardized testing metrics, residency salary and benefits, and diversity, equity, and inclusion mission statements and values. Variables described whether particular information was available on websites and were classified as yes or no. Variable information needed to be available on the program website and its website pages, or via a direct link from the program website and pages. Each website page linked from the main page of the residency website was reviewed for content, and direct links that were judged likely to be relevant were also opened. Data were entered in a Google spreadsheet for collection. In cases of ambiguity, PMD and OA discussed the content and agreed on the determination. To confirm accuracy and interrater reliability, after completing data collection, 10% of records as determined by random number generation were checked, with no systematic errors identified. Interrater reliability was not formally calculated; however, a few data entries were incongruent. All collected variables are described in Table 1.

XSL•FO
RenderX

Table . Content of obstetrics and gynecology residency program websites and comparison by type of residency program (N=290).

Characteristic	Total programs, n (%)	U[a] programs, n (%)	CU[b] programs, n (%)	C[c] programs, n (%)	ANOVA, *P* value	Post hoc comparisons, global[d]
Website	283 (97.6)	123 (100)	108 (97.3)	52 (92.9)	.02	U>C
Medical schools of residents	238 (82.1)	113 (91.9)	85 (76.6)	40 (71.4)	<.001	U>CU and U>C
Alumni trajectories	158 (54.5)	90 (73.2)	51 (45.9)	17 (30.4)	<.001	U>CU and U>C
USMLE[e] requirements	225 (77.6)	108 (87.8)	77 (69.4)	40 (71.4)	.001	U>CU and U>C
Step 1 attempts considered	77 (26.6)	29 (23.6)	26 (23.4)	22 (39.3)	.06	N/A[f]
Step 1 program notes no minimum noted	48 (16.6)	36 (29.3)	10 (9.0)	2 (3.6)	<.001	U>CU and U>C
Step 1 range, averages, or suggestions other than passing or no minimum	64 (22.1)	26 (21.1)	27 (24.3)	11 (19.6)	.75	N/A
Step 1 any score guidance other than passing	107 (36.9)	58 (47.2)	36 (32.4)	13 (23.2)	.004	U>CU and U>C
COMLEX[g] requirements	143 (49.3)	52 (42.3)	57 (51.4)	34 (60.7)	.06	N/A
Level 1 attempts considered	39 (13.4)	9 (7.3)	16 (14.4)	14 (25.0)	.005	C>U
Level 1 program notes no minimum noted	16 (5.5)	8 (6.5)	7 (6.3)	1 (1.8)	.40	N/A
Level 1 range, averages, or suggestions other than passing or no minimum	36 (12.4)	9 (7.3)	17 (15.3)	10 (17.9)	.07	N/A
Level 1 any score guidance other than passing	53 (18.3)	17 (13.8)	24 (21.6)	12 (21.4)	.24	N/A
Discusses DACA[h] applicants	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	N/A	N/A
Indication of whether international applicants are considered[i]	185 (63.8)	93 (75.6)	57 (51.4)	35 (62.5)	<.001	U>CU
Residency mission statement	132 (45.5)	61 (49.6)	51 (45.9)	20 (35.7)	.23	N/A
Residency diversity, equity, and inclusion statement or link to departmental statement	84 (29.0)	57 (46.3)	19 (17.1)	8 (14.3)	<.001	U>CU and U>C
Fellowship availability noted or directly accessible from residency website	128 (44.1)	96 (78.0)	27 (24.3)	5 (8.9)	<.001	U>CU and U>C
Average or estimated number of applications disclosed	23 (7.9)	11 (8.9)	10 (9.0)	2 (3.6)	.41	N/A
Average or estimated interview invitations disclosed	23 (7.9)	17 (13.8)	6 (5.4)	0 (0.0)	.003	U>CU and U>C
Salary noted or direct link to salary	185 (63.8)	80 (65.0)	65 (58.6)	40 (71.4)	.25	N/A
Benefits noted or direct link to benefits	200 (69.0)	89 (72.4)	66 (59.5)	45 (80.4)	.01	C>CU
Rotations according to residency year noted	248 (85.5)	111 (90.2)	90 (81.1)	47 (83.9)	.13	N/A

XSL•FO
RenderX

Characteristic	Total programs, n (%)	U[a] programs, n (%)	CU[b] programs, n (%)	C[c] programs, n (%)	ANOVA, *P* value	Post hoc comparisons, global[d]
Indication of average or most recent ACGME[j] case numbers per resident	39 (13.4)	18 (14.6)	18 (16.2)	3 (5.4)	.13	N/A
Program-specific videos or links to social media	167 (57.6)	80 (65.0)	62 (55.9)	25 (44.6)	.03	U>C

[a]U: university-based.
[b]CU: community-based university-affiliated.
[c]C: community-based.
[d]*P*=.05.
[e]USMLE: United States Medical Licensing Examination.
[f]N/A: not applicable.
[g]COMLEX: Comprehensive Osteopathic Medical Licensing Examination of the United States.
[h]DACA: Deferred Action for Childhood Arrivals.
[i]Including discussion on visa sponsorship.
[j]ACGME: Accreditation Council for Graduate Medical Education.

Data were exported from the Google spreadsheet as an .xlsx file and uploaded into JMP Pro 17.0.0 (SAS Institute, Inc), which was used to conduct statistical analysis. Descriptive statistics and one-way ANOVA were performed to determine differences among the three types of programs using a significance level of .05. Post hoc comparisons used the Tukey-Kramer honest significant difference (global *P*=.05).

Ethical Considerations

This study was considered by the University of Michigan's IRBMED institutional review board (study identification HUM00218409). The board determined that, in accordance with the board and federal regulations, the study did not require institutional review board approval because it considered publicly available data that could not be identified with a human subject.

Results

Of 290 OBGYN residency programs, 123 (42.4%) were university-based programs, 111 (38.3%) were community-based university-affiliated, and 56 (19%) were community-based. Most programs (283/290, 97.6%) had websites. Many programs did not include information about whether standardized testing filtering metrics are applied to applications (details are in Table 1). Notably, less than half (143/290, 49.3%) included any information about the Comprehensive Osteopathic Medical Licensing Examination (COMLEX). A majority of programs (238/290, 82.1%) listed the medical school of current residents, but fewer (158/290, 54.5%) described alumni trajectories. No programs discussed whether applicants with Deferred Action for Childhood Arrivals status would be considered.

When comparing types of programs, university-based programs were more likely to include a variety of information on their websites compared to community-based university-affiliated programs and community-based programs, including medical schools of current residents (113/123, 91.9%, university-based; 85/111, 76.6%, community-based university-affiliated; 40/56, 71.4%, community-based; *P*<.001); alumni trajectories (90/123, 73.2%, university-based; 51/111, 45.9%, community-based university-affiliated; 17/56, 30.4%, community-based; *P*<.001); statements about whether the United States Medical Licensing Examination (USMLE) Step 1 is required (108/123, 87.8%, university-based; 77/111, 69.4%, community-based university-affiliated; 40/56, 71.4%, community-based; *P*=.001); statements about no minimum USMLE score (36/123, 29.3%, university-based; 10/111, 9%, community-based university-affiliated; 2/56, 3.6%, community-based; *P*<.001); any USMLE score guidance other than a passing grade (58/123, 47.2%, university-based; 36/111, 32.4%, community-based university-affiliated; 13/56, 23.2%, community-based; *P*=.004); diversity, equity, and inclusion statements (57/123, 46.3%, university-based; 19/111, 17.1%, community-based university-affiliated; 8/56, 14.3%, community-based; *P*<.001); discussion of availability of fellowships at the same institution (96/123, 78%, university-based; 27/111, 24.3%, community-based university-affiliated; 5/56, 8.9%, community-based; *P*<.001); and whether the average or estimated number of interview invitations were disclosed (17/123, 13.8%, university-based; 6/111, 5.4%, community-based university-affiliated; 0/56, 0%, community-based; *P*=.003).

On post hoc analysis, there were several characteristics with overall significantly different representation on the websites of different types of programs but not between all types of programs. On post hoc comparison, university-based programs had websites significantly more often than community-based programs, but not significantly more often than community-based university-affiliated programs (123/123, 100%, university-based; 108/111, 97.3%, community-based university-affiliated; 52/56, 92.9%, community-based; *P*=.02). University-based program websites indicated whether international applicants were considered significantly more often than community-based university-affiliated programs, but not significantly more often than community-based programs (93/123, 75.6%, university-based; 57/111, 51.4%, community-based university-affiliated; 35/56, 62.5%, community-based; *P*<.001). University-based program websites had significantly more program-specific videos or links to social media than community-based programs, but not

XSL•FO
RenderX

community-based university-affiliated programs (80/123, 65%, university-based; 62/111, 55.9%, community-based university-affiliated; 25/56, 44.6%, community-based; *P*=.03).

Additionally, on post hoc comparison of significant findings, two of the 25 characteristics studied had a different pattern of presence on program websites. Community-based program websites noted whether COMLEX Level 1 attempts were considered significantly more often than university-based program websites, but not more often than community-based-university affiliated programs (9/123, 7.3%, university-based; 16/111, 14.4%, community-based university-affiliated; 14/56, 25%, community-based; *P*=.005), and community-based program websites noted benefits or directly linked to benefits significantly more often than community-based university-affiliated programs, but not more often than university-based programs (89/123, 72.4%, university-based; 66/111, 59.5%, community-based university-affiliated; 45/56, 80.4%, community-based; *P*=.01). Further description is listed in Table 1. There were minimal differences based on geographic location.

Discussion

Principal Results

Many OBGYN residency program websites lack information that is necessary for applicants to make informed decisions about where to apply and send program signals. When comparing types of programs, we found significant differences in website content, with many factors more often included by university-based programs than by community-based university-affiliated and community-based programs. Although this study was limited to OBGYN, these findings are relevant to all specialties, especially given the need for multiple intervention points for widespread residency application reform [3].

At this important educational transition point, applicants should ideally select residency programs that will enable them to thrive, both personally and professionally, during and after residency training. Many factors should be considered in learners' self-reflection processes, including whether they want to practice in an academic or community setting, their goals for research and fellowship training, and their individual learning styles. For residency programs to facilitate this decision-making process, this information should be available on program websites, particularly given applicants' reliance on this source [10-12]. Our work suggests that community-based university-affiliated programs and community-based programs currently lag behind university-based programs in several factors on their websites; consequently, applicants may miss an opportunity to learn about whether these programs align with their needs.

Our work is particularly salient given the widespread adoption of program signaling by many specialties. Transparency around application criteria is necessary if this meaningful residency application reform is to be successful. Notably, detailed standardized testing score guidance was not included on many program websites. These criteria are especially important for applicants who have historically applied to more residency programs and had lower match rates, such as osteopathic medical school and international medical graduate applicants [3,13]. About half of the programs did not include information about alumni trajectories, which can be valuable for applicants trying to determine whether their professional goals around practice setting or fellowship align with those of prior residents. Program signaling presents an exciting opportunity for equity, but it is important for applicants to have the opportunity to send signals to programs that will consider their applications and align with their goals.

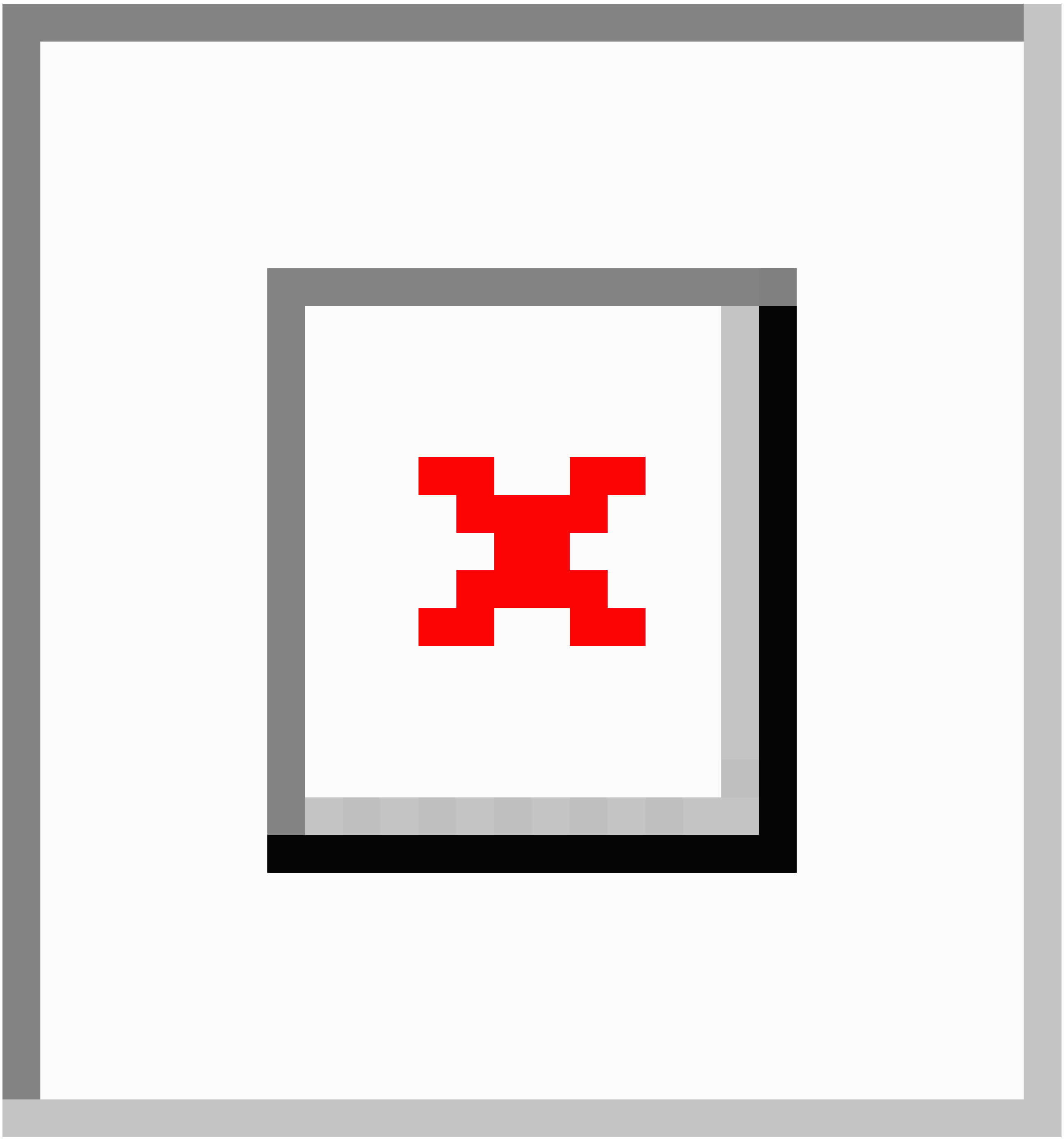
Improving transparency could also reduce residency programs' burden of reviewing large volumes of applications. By describing more criteria on websites, programs could communicate which applications will be considered—before applicants have spent resources on applying or signaling. In the National Resident Matching Program's Program Director Survey results, OBGYN residency program directors reported that an average of over 45% of applications are rejected based on standardized screening tools, before holistic review [14]. Failing to transparently describe criteria for standard screening tools can perpetuate rising application numbers and costs if applicants unknowingly apply to programs where their applications are automatically screened out of consideration.

From this work, we propose best practices for residency program websites in Figure 1. The practices are informed by the authors' perspectives as applicant, resident, and OBGYN faculty stakeholders in the residency application process. These practices include describing transparent application criteria to help applicants understand if they qualify for consideration, statements about values and outcomes that illuminate program priorities, and logistic considerations that can influence whether a program is a feasible option for an applicant. If applicants have access to this information, they may identify a more targeted list of programs to which they can apply and send signals, which will ultimately aid in improving the residency application process for applicants and programs alike.

The US residency application process needs multiple reforms to improve match rates and increase favorable outcomes for applicants [3]. Signaling may prove to be an important component of this reform, but signaling can only be successful if applicants can send informed signals to programs that align with their goals and values. One opportunity for residency programs to contribute to the success of this reform is sharing information, such as our residency website best practices, that help applicants determine whether the program aligns with their qualifications, desires, and goals.

XSL•FO
RenderX

Figure 1. Best practices of what should be included in obstetrics and gynecology residency program websites. DO: doctor of osteopathic medicine.



Limitations

Some programs may not control their website content; instead, they may follow graduate medical education or organization-specified templates. Nevertheless, our work provides important information for these groups to make choices about website content and we propose best practices to consider in Figure 1.

In this study, we collected data regarding USMLE Step 1 and COMLEX Level 1 examination scores. However, both exams have transitioned to a pass-or-fail grading system—USMLE Step 1 in January 2022 and COMLEX Level 1 in May 2022. Therefore, our data regarding USMLE Step 1 and COMLEX Level 1 scores may not apply to future applicants. Effects of a pass or fail grading system in the application process are yet to be determined, but other criteria, such as USMLE Step 2 and COMLEX Level 2 scores, may take on increasing importance. Websites must be updated to accurately reflect program requirements, so we suggest this is an excellent opportunity to provide increased information to applicants, such as clearly stating testing requirements, whether multiple attempts at exams are accepted, and if there are USMLE Step 2 or COMLEX Level 2 score thresholds or guidelines for applicants.

Comparison With Prior Work

This work aligns with findings in other specialties and illustrates key findings that will be of value given the evolving state of residency application processes. OBGYN programs' websites

XSL•FO
RenderX

had rates of listing residents' medical schools, salary, benefits, and rotation schedules that are similar to those of other specialties [15-20]. Application selection criteria were more difficult to compare because definitions varied across studies. However, like several other specialties, less than half of OBGYN residency programs included specific USMLE Step 1 score guidance [15-18,21]. Additionally, OBGYN programs, like several other specialties, do not universally indicate whether programs consider international medical graduate students and can sponsor visas [20,21]. However, some OBGYN program websites do stand out for including diversity, equity, and inclusion information and case numbers more often than some other specialties [16,22].

Our comparison of different types of programs is less common. Studies in two other specialties compared academic and non-academic programs and found academic programs included more of the characteristics they studied, which aligns with our findings in OBGYN [20,22]. Given the inherent value and differences in all programs, we believe that comparing types of residency programs presents an opportunity to understand which programs can improve in communicating with applicants.

Conclusions

In this competitive application landscape, it is crucial that applicants are provided equitable access to information that allows them to determine where to apply and send signals to optimize their success in matching at a program aligned with their values. Applicants use websites to determine residency program qualities, but the onus of deciphering the best fit should not rest entirely on them. A robust presentation of residency program personnel, curriculum, values, benefits, and application criteria can help applicants understand where their applications will be considered, and possibly where their signals are most strategic. Increased information sharing on program websites could contribute to an improved application process.

Conflicts of Interest

None declared.

References

1. About physician licensure. Federation of State Medical Boards. URL: https://www.fsmb.org/u.s.-medical-regulatory-trends-and-actions/guide-to-medical-regulation-in-the-united-states/about-physician-licensure [accessed 2024-06-06]
2. 2022 main residency match. National Resident Matching Program. 2022. URL: https://www.nrmp.org/wp-content/uploads/2022/05/2022-Main-Match-Results-and-Data_Final.pdf [accessed 2024-06-06]
3. Mott NM, Carmody JB, Marzano DA, Hammoud MM. What's in a number? Breaking down the residency match rate. N Engl J Med 2022 Apr 28;386(17):1583-1586. [doi: 10.1056/NEJMp2119716]
4. Right resident, right program, ready day one: program resources. Association of Professors of Gynecology and Obstetrics. URL: https://apgo.org/page/rrrprogram [accessed 2022-05-10]
5. Specialties participating in the supplemental ERAS® application. Association of American Medical Colleges. URL: https://students-residents.aamc.org/applying-residencies-eras/specialties-and-programs-participating-supplemental-eras-application [accessed 2022-07-06]
6. Cole JA, Ludomirsky AB. The costliness of US residency applications: moving toward preference signaling and caps. J Grad Med Educ 2022 Dec;14(6):647-649. [doi: 10.4300/JGME-D-22-00067.1] [Medline: 36591434]
7. Supplemental ERAS® application guide. Association of American Medical Colleges. 2022. URL: https://students-residents.aamc.org/media/12326/download?attachment [accessed 2022-12-31]
8. Residency directory. Association of Professors of Gynecology and Obstetrics. URL: https://tools.apgo.org/residency-directory/search-residency-directory/ [accessed 2022-10-14]
9. FREIDA residency and fellowship database. American Medical Association. URL: https://www.ama-assn.org/amaone/freida-membership [accessed 2022-10-14]
10. Mahler SA, Wagner MJ, Church A, Sokolosky M, Cline DM. Importance of residency program web sites to emergency medicine applicants. J Emerg Med 2009 Jan;36(1):83-88. [doi: 10.1016/j.jemermed.2007.10.055] [Medline: 18439790]
11. Gaeta TJ, Birkhahn RH, Lamont D, Banga N, Bove JJ. Aspects of residency programs' web sites important to student applicants. Acad Emerg Med 2005 Jan;12(1):89-92. [doi: 10.1197/j.aem.2004.08.047] [Medline: 15635145]
12. Chu LF, Young CA, Zamora AK, et al. Self-reported information needs of anesthesia residency applicants and analysis of applicant-related web sites resources at 131 United States training programs. Anesth Analg 2011 Feb;112(2):430-439. [doi: 10.1213/ANE.0b013e3182027a94] [Medline: 21081766]
13. ERAS statistics. Association of American Medical Colleges. 2022. URL: https://www.aamc.org/data-reports/interactive-data/eras-statistics-data [accessed 2022-06-20]
14. Results of the 2021 NRMP program director survey. National Resident Matching Program. 2021. URL: https://www.nrmp.org/wp-content/uploads/2021/11/2021-PD-Survey-Report-for-WWW.pdf [accessed 2022-06-20]
15. Novin S, Yi PH, Vanderplas T, Yim D, Hong K. Integrated interventional radiology residency program websites: a development in progress. AJR Am J Roentgenol 2018 Jul;211(1):211-216. [doi: 10.2214/AJR.17.19008] [Medline: 29792738]

XSL•FO
RenderX

16. Patel BG, Gallo K, Cherullo EE, Chow AK. Content analysis of ACGME accredited urology residency program webpages. Urology 2020 Apr;138:11-15. [doi: 10.1016/j.urology.2019.11.053] [Medline: 31954168]
17. Patel SJ, Abdullah MS, Yeh PC, Abdullah Z, Jayaram P. Content evaluation of physical medicine and rehabilitation residency websites. PMR 2020 Oct;12(10):1003-1008. [doi: 10.1002/pmrj.12303] [Medline: 31840922]
18. Hansberry DR, Bornstein J, Agarwal N, McClure KE, Deshmukh SP, Long S. An assessment of radiology residency program websites. J Am Coll Radiol 2018 Apr;15(4):663-666. [doi: 10.1016/j.jacr.2017.11.010] [Medline: 29273474]
19. Pollock J, Weyand J, Reyes A, et al. Descriptive analysis of components of emergency medicine residency program websites. West J Emerg Med 2021 Jul 15;22(4):937-942. [doi: 10.5811/westjem.2021.4.50135]
20. Daniel D, Vila C, Leon Guerrero CR, Karroum EG. Evaluation of adult neurology residency program websites. Ann Neurol 2021 Apr;89(4):637-642. [doi: 10.1002/ana.26016] [Medline: 33421179]
21. Markle JC, Ahmed H, Pandya K, et al. Transparency in the ophthalmology residency match: background, study, and implications. Cureus 2021 Nov;13(11):e19826. [doi: 10.7759/cureus.19826] [Medline: 34963843]
22. Chinedozi I, Martin O, Hays N, Kubicki NS, Kidd-Romero S, Kavic SM. Love at first click: surgery residency websites in the virtual era. J Surg Educ 2021;78(6):2088-2093. [doi: 10.1016/j.jsurg.2021.04.016] [Medline: 34011477]

Abbreviations

COMLEX: Comprehensive Osteopathic Medical Licensing Examination
DO: doctor of osteopathic medicine
ERAS: Electronic Residency Application Service
MD: medical doctorate
OBGYN: obstetrics and gynecology
USMLE: United States Medical Licensing Examination

Edited by B Lesselroth; submitted 26.04.23; peer-reviewed by A Santiago, CI Sartorão Filho, K George, KH Mori, S Cox; revised version received 13.06.24; accepted 19.08.24; published 21.10.24.

Please cite as:
Devlin PM, Akingbola O, Stonehocker J, Fitzgerald JT, Winkel AF, Hammoud MM, Morgan HK
Opportunities to Improve Communication With Residency Applicants: Cross-Sectional Study of Obstetrics and Gynecology Residency Program Websites
JMIR Med Educ 2024;10:e48518
URL: https://mededu.jmir.org/2024/1/e48518
doi:10.2196/48518

© Paulina M Devlin, Oluwabukola Akingbola, Jody Stonehocker, James T Fitzgerald, Abigail Ford Winkel, Maya M Hammoud, Helen K Morgan. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.10.2024. 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.

XSL•FO
RenderX

Short Paper

Impact of the COVID-19 Pandemic on Medical Grand Rounds Attendance: Comparison of In-Person and Remote Conferences

Ken Monahan[1], MD; Edward Gould[1], MD; Todd Rice[1], MD; Patty Wright[1], MD; Eduard Vasilevskis[1], MD; Frank Harrell[1], PhD; Monique Drago[1], EdD; Sarah Mitchell[1], MS

Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Ken Monahan, MD
Vanderbilt University Medical Center
1215 21st Avenue
Medical Center East - 5th Floor
Nashville, TN, 37232
United States
Phone: 1 6153222318
Email: ken.monahan@vumc.org

Abstract

Background: Many academic medical centers transitioned from in-person to remote conferences due to the COVID-19 pandemic, but the impact on faculty attendance is unknown.

Objective: This study aims to evaluate changes in attendance at medical grand rounds (MGR) following the transition from an in-person to remote format and as a function of the COVID-19 census at Vanderbilt Medical Center.

Methods: We obtained the faculty attendee characteristics from Department of Medicine records. Attendance was recorded using a SMS text message–based system. The daily COVID-19 census was recorded independently by hospital administration. The main attendance metric was the proportion of eligible faculty that attended each MGR. Comparisons were made for the entire cohort and for individual faculty.

Results: The observation period was from March 2019 to June 2021 and included 101 MGR conferences with more than 600 eligible faculty. Overall attendance was unchanged during the in-person and remote formats (12,536/25,808, 48.6% vs 16,727/32,680, 51.2%; P=.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

XSL•FO
RenderX

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.

XSL•FO
RenderX

Table 1. Conference and attendee characteristics.

Characteristics	Value	Value at the end of the observation (range during observation period)
Conferences, n		
Total during observation period	101	N/A[a]
In person (prepandemic)	47	N/A
Remote (intrapandemic)	54	N/A
Topic, n		
Cardiology	19	N/A
Endocrine	10	N/A
Gastroenterology	12	N/A
General internal medicine	15	N/A
Geriatric medicine	3	N/A
Hematology or oncology	10	N/A
Infectious disease	10	N/A
Nephrology	7	N/A
Pulmonary or critical care	7	N/A
Rheumatology	5	N/A
Speaker, n		
Internal	41	N/A
External	60	N/A
Faculty attendance[b], mean (SD)		
Total eligible to attend	579 (22)	611 (544-612)
Cardiology	100 (2)	103 (95-103)
Endocrine	25 (2)	28 (23-28)
Gastroenterology	41 (2)	43 (38-43)
General internal medicine	175 (8)	187 (161-187)
Hematology or oncology	65 (2)	69 (60-69)
Infectious disease	43 (1)	45 (40-45)
Nephrology	33 (2)	36 (31-36)
Pulmonary or critical care	46 (2)	46 (42-49)
Rheumatology	22 (1)	23 (21-23)
Assistant professor	328 (16)	349 (279-350)
Associate professor	107 (1)	109 (105-109)
Full professor	143 (11)	149 (107-151)

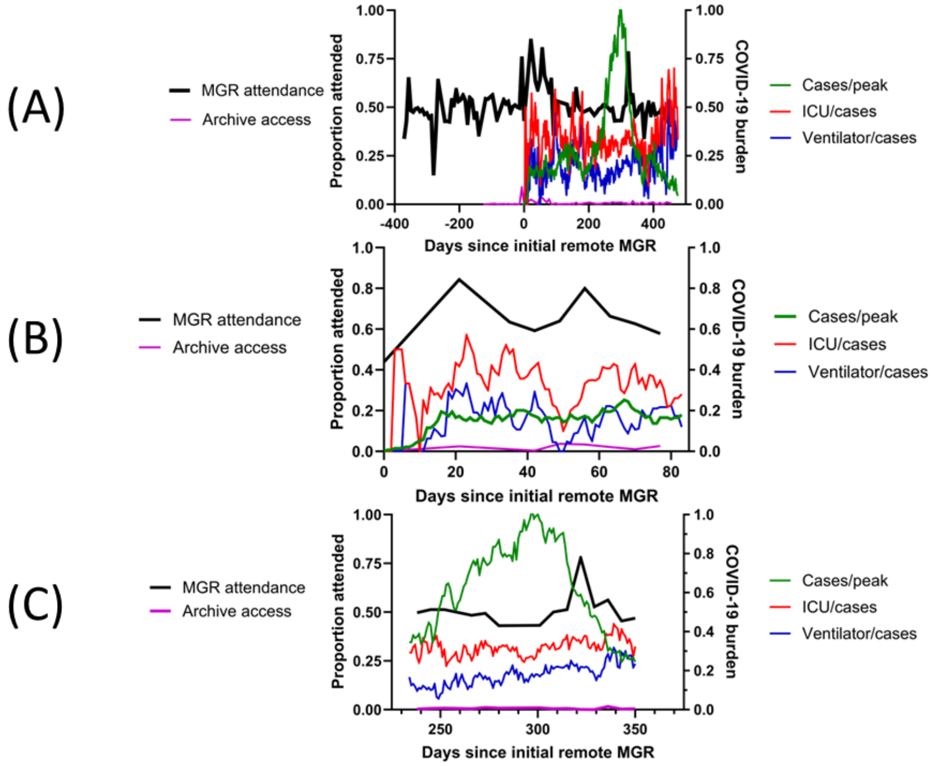
[a]N/A: not applicable.

[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%).

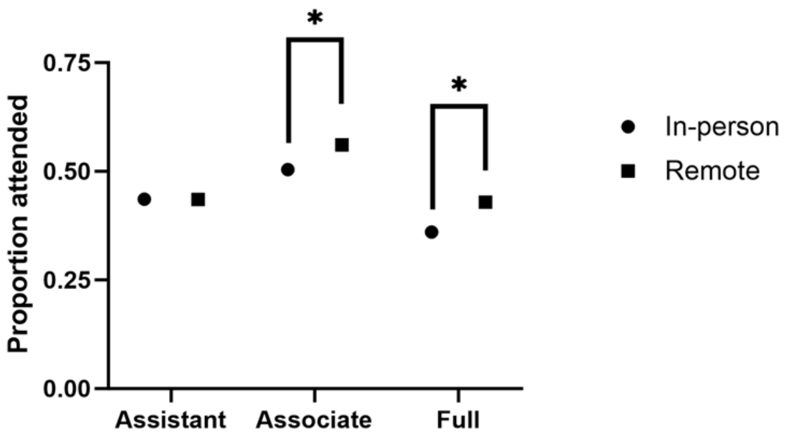
XSL•FO
RenderX

Figure 1. Time series of medical grand rounds (MGR) attendance and concurrent COVID-19 burden. (A) The entire observation period, (B) focus on the beginning of the remote format, and (C) focus on the surge. At the onset of the remote format, there is a nonsustained increase in attendance. As the COVID-19 census increased rapidly leading up to the peak census, there was no change in attendance. During the peak of the surge, there is a very small transient reduction in attendance followed by an extremely brief increase in attendance during a period of rapid decline in the COVID-19 census. Access to archived MGR lectures remained low during the entire observation period. ICU: intensive care unit.



MGR attendance stratified by academic rank across the in-person and remote formats is shown in Figure 2. Associate (3249/5788, 56.1% vs 2515/4989, 50.4%; *P*<.001) and full professor (3309/7718, 42.9% vs 2433/6757, 36%; *P*<.001) attendance was higher at MGR during the remote format relative to the in-person format.

Figure 2. Attendance at medical grand rounds stratified by academic rank. Assistant professor attendance was the same regardless of conference format, whereas associate and full professor attendance increased during the remote format relative to in person. *P<.001.



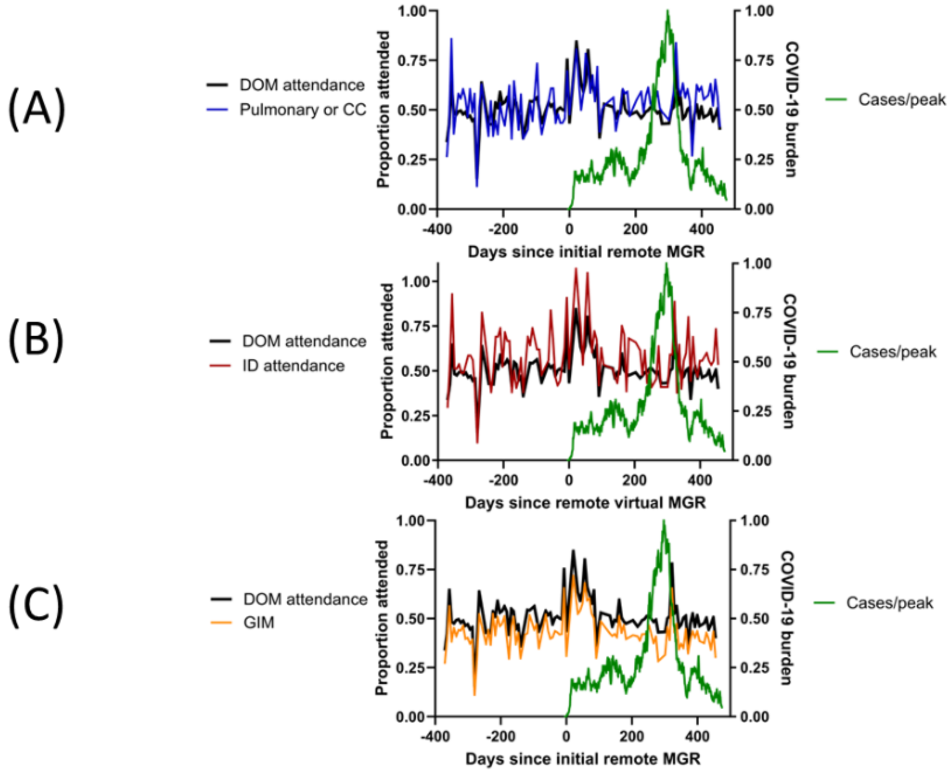
Subinterval and Subgroup Analyses

There was no difference in attendance during the surge compared to the 2 months before (October to November 2020; 2071/4218, 49.1% vs 2194/4229, 51.9%; *P*=.38) or 1 year before (December 2019 to January 2020; 2028/3990, 50.8% vs 2194/4229, 51.9%; *P*=.34).

The attendance trends of DOM subspecialties that were particularly impacted by the pandemic are superimposed on the overall DOM trend in Figure 3 for pulmonary or critical care (CC), infectious diseases (ID), and general internal medicine (GIM).

XSL•FO
RenderX

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).



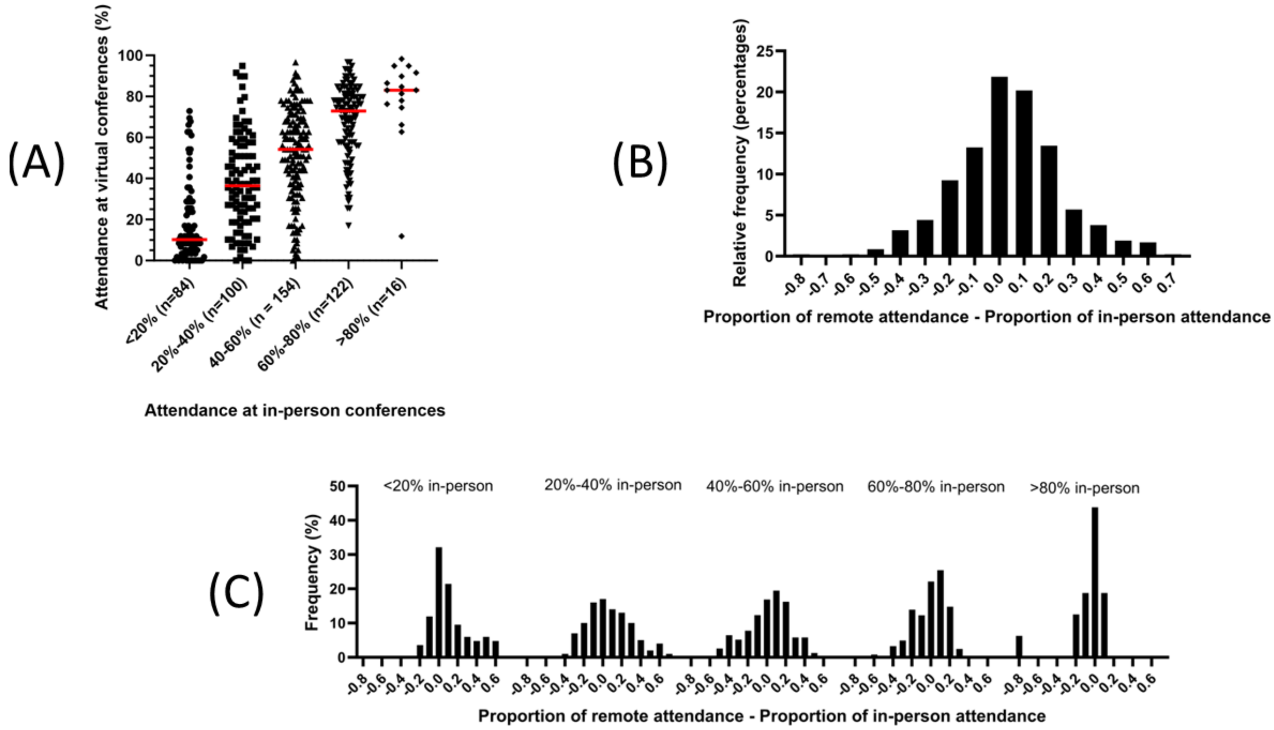
Pulmonary or CC attendance during the remote format was higher than during the in-person format (1450/2616, 55.4% vs 1004/2045, 49.1%; *P*<.001). This attendance pattern persisted while cases were rising and peaking during the surge, when demands on this portion of the faculty were likely greater than 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.

XSL•FO
RenderX

Figure 4. Individual-level attendance at in-person and remote medical grand rounds (MGR). (A) For a given level of attendance at in-person MGR, individual faculty member attendance at remote MGR fluctuated widely. (B) The distribution of the difference in attendance rates between conference formats for the entire cohort is relatively symmetric around the null, as expected given the overall lack of change. Nonetheless, extreme values of changes in attendance at the tails are noted. (C) Faculty that attended in-person MGR less frequently generally increased their attendance at remote MGR, while the opposite response was observed for those that frequently attended in-person MGR. Red bars indicate the mean.



Discussion

Principal Findings

Overall faculty attendance at MGR remained constant regardless of conference format, suggesting no disadvantage to the remote format. In addition, there may be substantial cost savings [6] and beneficial environmental impacts [7] associated with the remote format as it pertains to external speakers, who comprised the majority (60/101, 59.4%) of this cohort.

The increase in attendance of associate and full professors during the remote format may indicate fewer concurrent clinical obligations for these groups compared to their more junior colleagues. COVID-19–related MGR lectures at the beginning of the remote period may have led to the concurrent initial increase in attendance [8], but attendance quickly regressed to the mean, which was maintained even during a subsequent period of rapid rise and peak in COVID-19 burden.

Paradoxically, pulmonary or CC faculty attendance increased during the pandemic. It is possible that the attendance of the subgroup of non-ICU providers within pulmonary or CC may have increased during the pandemic while the attendance of their ICU-based colleagues declined. We speculate that the decreased attendance of the division of GIM was contributed to by lower attendance within the section of hospital medicine, perhaps because of burnout [9].

Individual faculty attendance habits did not remain static in response to the change in conference format. The pandemic or the remote format may have motivated faculty to attend MGR who did not regularly do so, thus taking the place of faculty that were unable to attend due to increased clinical or administrative responsibilities. The presence of outliers at both extremes of attendance shifts may enrich further investigations of specific drivers of conference attendance, which could inform decisions regarding conference format moving forward.

Archived conferences were infrequently accessed throughout the observation period. Encouraging asynchronous viewing may increase consumption of MGR among faculty who are unable to do so in real time. Offering attendance credit for viewing MGR asynchronously could incentivize otherwise nonattending faculty.

Limitations

This study did not use surveys or other methods of obtaining feedback from faculty regarding their attendance patterns relative to the mode of MGR presentation, as collecting these data was not feasible given the study's retrospective design.

Attendance does not guarantee the observer has learned from MGR, although mandatory evaluations may not assess this objective either [10].

Conclusions

Overall faculty attendance at MGR was neither durably affected by a pandemic-related transition from in-person to a remote format nor by concurrent COVID-19 burden, although individual attendance behaviors varied considerably. If coupled with archived conference recordings, the remote format may be an equally attended and more cost-effective option for presenting MGR in a postpandemic era.

XSL•FO
RenderX

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

1. Osler W. The natural method of teaching the subject of medicine. JAMA 1901;XXXVI(24):1673-1679. [doi: 10.1001/jama.1901.52470240001001]
2. Jattan A, Francois J. Twelve tips for adapting grand rounds for contemporary demands. Med Teach 2022;44(2):144-148. [doi: 10.1080/0142159X.2021.1898573] [Medline: 33725468]
3. Mueller PS, Litin SC, Sowden ML, Habermann TM, LaRusso NF. Strategies for improving attendance at medical grand rounds at an Academic Medical Center. Mayo Clin Proc 2003;78(5):549-553. [doi: 10.4065/78.5.549] [Medline: 12744540]
4. Yang AZ, Hyland CJ, Xiang DH, Helliwell LA, Broyles JM. Improving the quality of grand rounds in plastic surgery: in-person, hybrid, or virtual. Plast Reconstr Surg Glob Open 2023;11(Suppl 5):42 [FREE Full text] [doi: 10.1097/01.gox.0000937860.63119.52]
5. Reddy GB, Ortega M, Dodds SD, Brown MD. Virtual versus in-person grand rounds in orthopaedics: a framework for implementation and participant-reported outcomes. J Am Acad Orthop Surg Glob Res Rev 2022;6(1):e21.00308 [FREE Full text] [doi: 10.5435/JAAOSGlobal-D-21-00308] [Medline: 35044329]
6. Crossman M, Papanagnou D, Sullivan T, Zhang XC. Virtual grand rounds in COVID-19: a financial analysis. Acad Emerg Med 2021;28(4):480-482 [FREE Full text] [doi: 10.1111/acem.14224] [Medline: 33527635]
7. Monahan S, Monahan K. The potential environmental impact of external speakers' airplane travel to grand rounds conferences. Environ Health 2023;22(1):34 [FREE Full text] [doi: 10.1186/s12940-023-00989-6] [Medline: 37060082]
8. Sparkes D, Leong C, Sharrocks K, Wilson M, Moore E, Matheson NJ. Rebooting medical education with virtual grand rounds during the COVID-19 pandemic. Future Healthc J 2021;8(1):e11-e14 [FREE Full text] [doi: 10.7861/fhj.2020-0180] [Medline: 33791467]
9. Dugani SB, Geyer HL, Maniaci MJ, Fischer KM, Croghan IT, Burton C. Psychological wellness of internal medicine hospitalists during the COVID-19 pandemic. Hosp Pract (1995) 2021;49(1):47-55. [doi: 10.1080/21548331.2020.1832792] [Medline: 33012183]
10. Wecksell M, Salik I. Mandatory grand rounds evaluations: more data, less information. Cureus 2022;14(4):e24567 [FREE Full text] [doi: 10.7759/cureus.24567] [Medline: 35651415]

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:
Monahan K, Gould E, Rice T, Wright P, Vasilevskis E, Harrell F, Drago M, Mitchell S
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

XSL•FO
RenderX

©Ken Monahan, Edward Gould, Todd Rice, Patty Wright, Eduard Vasilevskis, Frank Harrell, Monique Drago, Sarah Mitchell. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.01.2024. 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.

XSL•FO
RenderX

Collaborative Development of an Electronic Portfolio to Support the Assessment and Development of Medical Undergraduates

Luiz Ricardo Albano dos Santos[1,*], PhD; Alan Maicon de Oliveira[2,*], PhD; Luana Michelly Aparecida Costa dos Santos[1,*], PhD; Guilherme José Aguilar[3,*], PhD; Wilbert Dener Lemos Costa[1,*], MSc; Dantony de Castro Barros Donato[1,*], MSc; Valdes Roberto Bollela[1,4,*], MD, PhD

1

2

3

4

*all authors contributed equally

Corresponding Author:
Luiz Ricardo Albano dos Santos, PhD

Abstract

This study outlines the development of an electronic portfolio (e-portfolio) designed to capture and record the overall academic performance of medical undergraduate students throughout their educational journey. Additionally, it facilitates the capture of narratives on lived experiences and sharing of reflections, fostering collaboration between students and their mentors.

(JMIR Med Educ 2024;10:e56568) doi:10.2196/56568

KEYWORDS

e-portfolio; education; health education; learning; medical students; medical school curriculum; medical education; student support; software

Introduction

The Brazilian curriculum guidelines for medical schools incorporate competencies in information technology, emphasizing students' co-responsibility in acquiring soft skills such as leadership, teamwork, and continuous professional development [1]. The curriculum experience must foster critical and reflective skills [2].

Ribeirão Preto Medical School at University of São Paulo, Brazil (FMRP-USP), is a 72-year-old traditional institution that initiated a curriculum change in January 2023. In this new proposal, we introduced a longitudinal axis and curricular unit called personal and professional development (PPD). The primary objective of PPD is to foster self-reflection on lived experiences, regular self-assessment, and monitoring of the students' progress in curricular and extracurricular activities, with a mentor's support.

To support the implementation of the PPD curricular unit, we collaboratively developed a software to serve as the electronic portfolio (e-portfolio) and record the overall academic performance of undergraduate medical students throughout their educational journey. An additional expectation is to encourage and guide teachers to provide and register formative assessments in their disciplines and rotations, and to document their experiences and reflections.

Methods

The collaborative development of the system involved developers, health educators, and students, which was crucial to ensure that the e-portfolio meets the needs and expectations of all stakeholders. Developers contributed technical expertise for functionality and accessibility, while educators shaped content based on educational principles. Students, as primary users, provided valuable feedback.

The main challenge in developing the e-portfolio was to create an initial set of requirements. With various participants bringing different ideas, there was a multitude of perspectives in the initial phase, which brought fundamental enrichment during development but also increased the difficulty of integrating all perspectives.

These challenges were overcome with Scrum [3] integrated with socio-technical research methodology to facilitate the collaborative environment. We implement Scrum practices, such as daily 5-minute meetings and biweekly 30-minute sprint reviews, ensuring incremental and continuous deliveries and communication between the development team and stakeholders, mainly regarding system development. Additionally, we integrated the socio-technical research methodology [4] into SCRUM, aiming to understand the software requirements as well as the various social and technological factors involved.

XSL•FO
RenderX

Regarding software development technologies, we used HTML, CSS, PHP, and the MySQL database management system.

Ethics Approval

The study received approval from the research ethics committee of the Clinical Hospital of FMRP-USP (CAAE: 67577523.1.0000.5440).

Results

The e-portfolio utilizes a web application architecture (Figure 1). Initially, we developed a structure to manage the registration of all the programs within the medical school, different curricular units, and offerings. We created a registration module for students and faculty members, allowing those to act as mentors, teachers, and discipline coordinators. Additionally, e-portfolio enables the recording of direct observed assessments in clinical settings, using preregistered forms based on methods such as mini-clinical evaluation exercise (Mini-CEx) [5], 360-degree assessment [6], One-Minute Preceptor, direct observation of procedural skills (DOPS), and case-based discussion/chart-stimulated recall (CBD/CSR) [7].

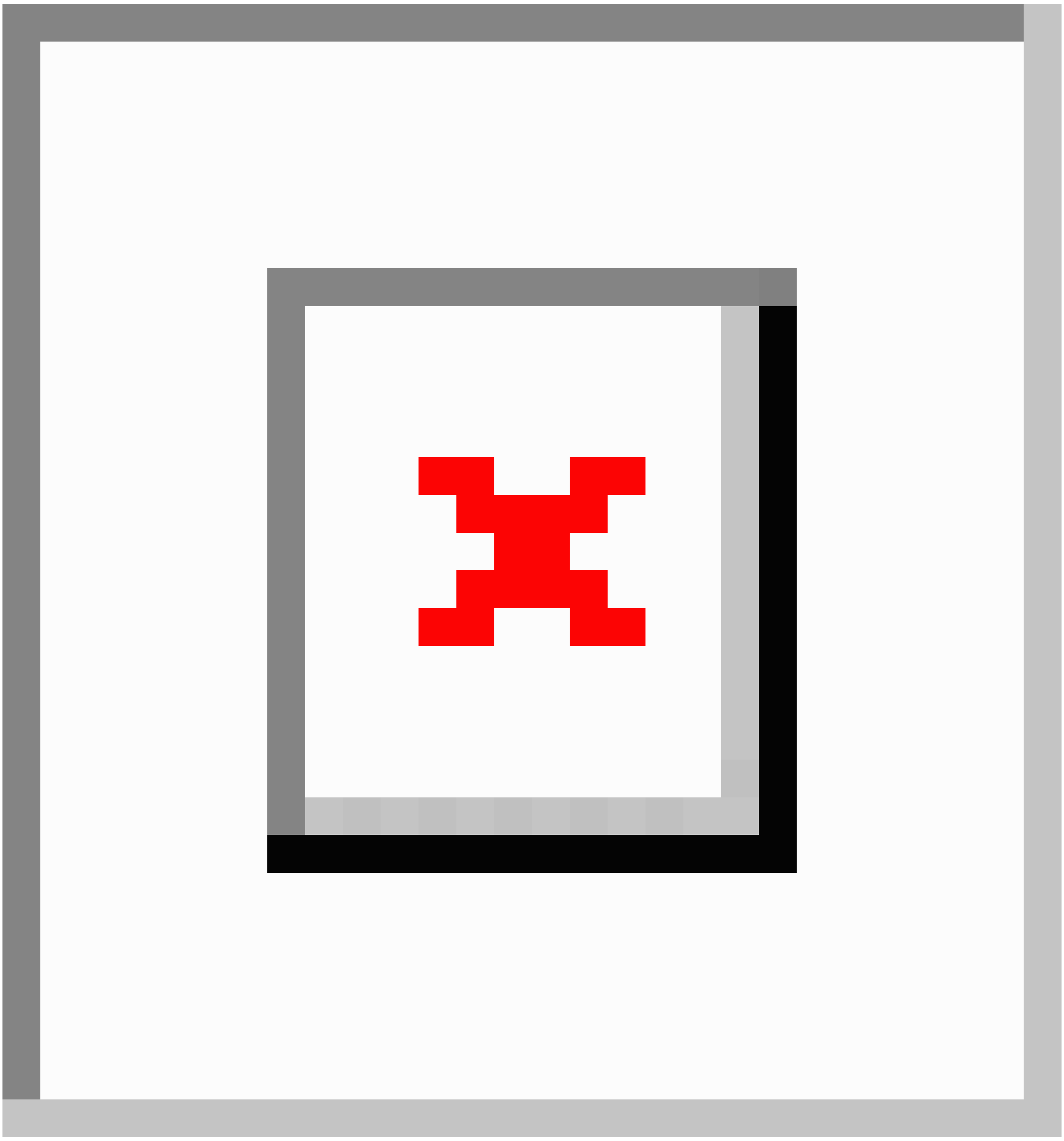
For narratives in medicine [8], there is a specific form to guide students on how to report a lived experience followed by a meaningful reflection, based on the REFLECT rubric for assessing reflective writing [9] (Figure 2).

Students are allowed to fill in data in their private profile (Figure 1), access their disciplines and received assessments, respond to formative assessments, record significant events for their education, check and compare their performance with their cohort, register extracurricular activities, and consult critical incidents recorded.

e-Portfolio enables students, discipline coordinators, and members of the student assessment committee to track assessments and feedback received, providing a longitudinal and progressive view of the student’s cognitive, psychomotor (skills), and attitudinal development (Figures 1 and 2).

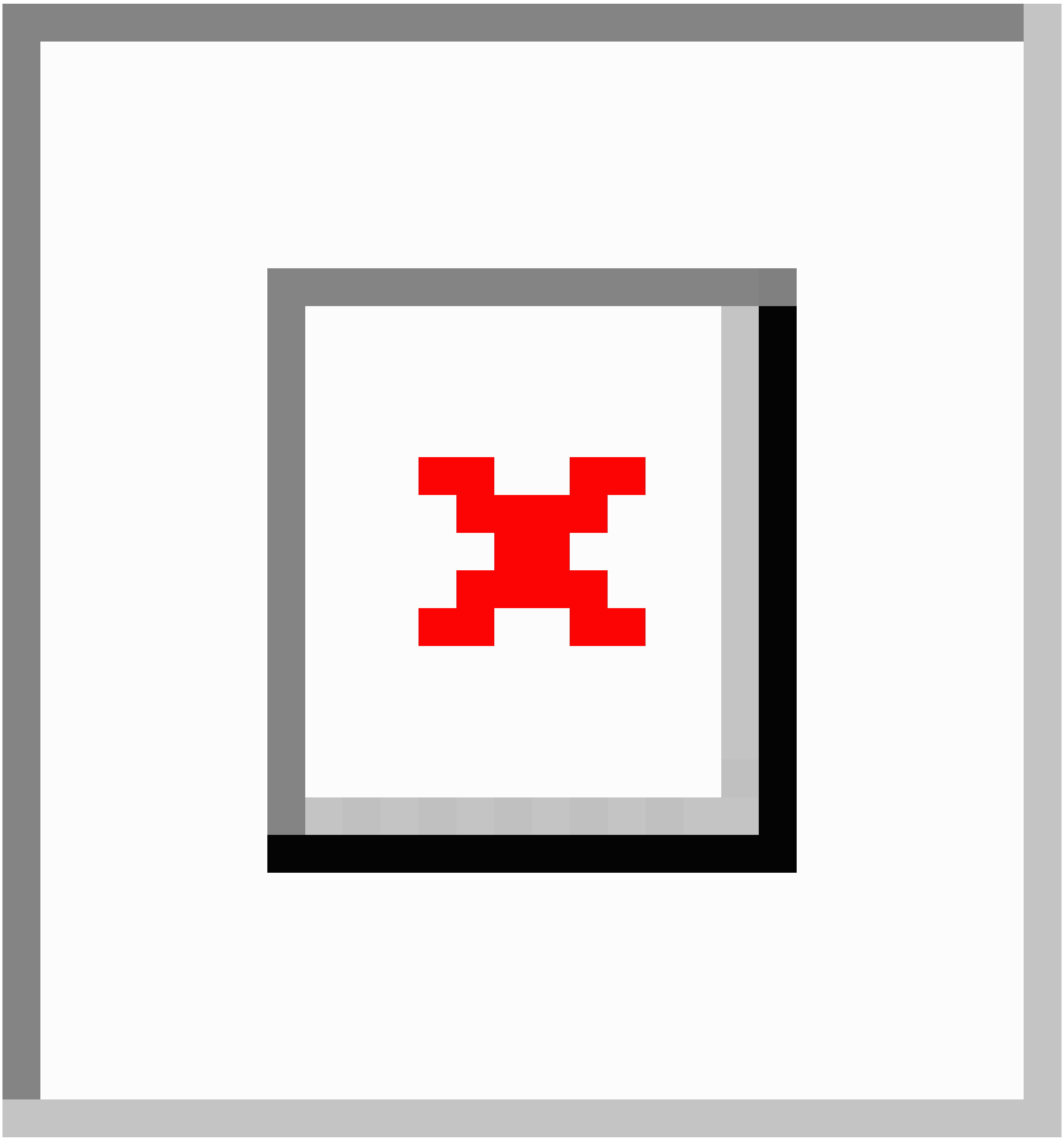
XSL•FO
RenderX

Figure 1. Profile and performance report of the medical student in the electronic portfolio (e-portfolio). (A) Profile created by the student in the e-portfolio. (B) Student’s performance in various subjects is presented in relation to the radar chart: the blue line represents a comparison with the cohort mean (depicted by the gray area).



XSL•FO
RenderX

Figure 2. Evaluation form in the electronic portfolio (e-portfolio) with narratives and the adapted REFLECT rubric to guide the medical student and the mentor.



Discussion

This work presents the successful development of an e-portfolio at FMRP-USP. The e-portfolio is continuously enhanced and updated, and it is currently in a state suitable for use in a pilot study. The use of similar tools has been recognized for stimulating personal reflection, fostering collaboration, and strengthening digital literacy among students, encouraging active participation in the learning process [10].

The application of Scrum offered an adaptable framework, promoting efficient collaboration among stakeholders. Additionally, socio-technical research methods, such as qualitative interviews involving in-depth conversations with individuals or groups to explore their experiences related to technology, provided valuable insights into the needs and dynamics of end users in the educational context. The use of Scrum with socio-technical research methods enables a more integrated, collaborative, and reflective approach during development.

Future Steps

We intend to evaluate e-portfolio usability, effectiveness, acceptance, and satisfaction in practical contexts with the objective of consistently enhancing the system and its outcomes.

XSL•FO
RenderX

Acknowledgments

We would like to express our gratitude to the Ribeirão Preto Medical School, University of São Paulo, and the startup Intersection (Ribeirão Preto, Brazil) for their partnership in software development. We would like to thank Prof Francisco S Guimarães for all the support and follow-up with the medical students and mentors. Additionally, we extend our appreciation to the Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq). This study was financed in part by CNPq (process no.: 001). Furthermore, this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Authors' Contributions

LRAS and VRB contributed to the study concept and design, data acquisition, analysis, interpretation, and manuscript writing. AMO, LMACS, GJA, WDLC, and DCBD contributed to the interpretation, manuscript writing, and critical review of the manuscript for important intellectual content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Weinberger SE, Smith LG, Collier VU, Education Committee of the American College of Physicians. Redesigning training for internal medicine. Ann Intern Med 2006 Jun 20;144(12):927-932. [doi: 10.7326/0003-4819-144-12-200606200-00124] [Medline: 16601254]
2. Elshami WE, Abuzaid MM, Guraya SS, David LR. Acceptability and potential impacts of innovative E-Portfolios implemented in E-Learning systems for clinical training. J Taibah Univ Med Sci 2018 Dec;13(6):521-527. [doi: 10.1016/j.jtumed.2018.09.002] [Medline: 31435372]
3. Sutherland J. Scrum: The Art of Doing Twice the Work in Half the Time, 1st edition: Crown Business; 2014.
4. Fuggetta A. Software process: a roadmap. In: ICSE '00: Proceedings of the Conference on The Future of Software Engineering: Association for Computing Machinery; 2000:25-34. [doi: 10.1145/336512.336521]
5. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med 1995 Nov 15;123(10):795-799. [doi: 10.7326/0003-4819-123-10-199511150-00008] [Medline: 7574198]
6. Jani H, Narmawala W, Ganjawale J. Evaluation of competencies related to personal attributes of resident doctors by 360 degree. J Clin Diagn Res 2017 Jun;11(6):JC09-JC11. [doi: 10.7860/JCDR/2017/25907.10027] [Medline: 28764199]
7. Furney SL, Orsini AN, Orsetti KE, Stern DT, Gruppen LD, Irby DM. Teaching the one-minute preceptor. a randomized controlled trial. J Gen Intern Med 2001 Sep;16(9):620-624. [doi: 10.1046/j.1525-1497.2001.016009620.x] [Medline: 11556943]
8. Charon R. Narrative and medicine. N Engl J Med 2004 Feb 26;350(9):862-864. [doi: 10.1056/NEJMp038249] [Medline: 14985483]
9. Wald HS, Borkan JM, Taylor JS, Anthony D, Reis SP. Fostering and evaluating reflective capacity in medical education: developing the REFLECT rubric for assessing reflective writing. Acad Med 2012 Jan;87(1):41-50. [doi: 10.1097/ACM.0b013e31823b55fa] [Medline: 22104060]
10. Mudau PK, Modise MMP. Using e-portfolios for active student engagement in the ODeL environment. JITE:Res 2022;21:425-438. [doi: 10.28945/5012]

Abbreviations

CAPES: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for the Improvement of Higher Education Personnel)
CBD/CSR: case-based discussion/chart-stimulated recall
CNPq: Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development)
DOPS: direct observation of procedural skills
e-portfolio: electronic portfolio
FMRP-USP: Ribeirão Preto Medical School, University of São Paulo
Mini-CEx: mini-clinical evaluation exercise
PPD: personal and professional development

XSL•FO
RenderX

Edited by SR Mogali; submitted 19.01.24; peer-reviewed by A Arbabisarjou, IS Lima, K Lacerda; revised version received 27.02.24; accepted 04.03.24; published 04.04.24.

Please cite as:
dos Santos LRA, de Oliveira AM, dos Santos LMAC, Aguilar GJ, Costa WDL, Donato DDCB, Bollela VR
Collaborative Development of an Electronic Portfolio to Support the Assessment and Development of Medical Undergraduates
JMIR Med Educ 2024;10:e56568
URL: https://mededu.jmir.org/2024/1/e56568
doi:10.2196/56568

© Luiz Ricardo Albano dos Santos, Alan Maicon de Oliveira, Luana Michelly Aparecida Costa dos Santos, Guilherme José Aguilar, Wilbert Dener Lemos Costa, Dantony de Castro Barros Donato, Valdes Roberto Bollela. Originally published in JMIR Medical Education (https://mededu.jmir.org), 4.4.2024. 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.

XSL•FO
RenderX

The Performance of ChatGPT-4V in Interpreting Images and Tables in the Japanese Medical Licensing Exam

Soshi Takagi[1,*], BA; Masahide Koda[2,*], MD, PhD; Takashi Watari[3,4,*], MHQS, MD, PhD

1

2

3

4

*all authors contributed equally

Corresponding Author:
Takashi Watari, MHQS, MD, PhD

(JMIR Med Educ 2024;10:e54283) doi:10.2196/54283

KEYWORDS

ChatGPT; medical licensing examination; generative artificial intelligence; medical education; large language model; images; tables; artificial intelligence; AI; Japanese; reliability; medical application; medical applications; diagnostic; diagnostics; online data; web-based data

Introduction

OpenAI's ChatGPT, a leading large language model (LLM), has shown promise for medical purposes. The program can pass the United States Medical Licensing Examination (USMLE) and the Japanese Medical Licensing Exam (JMLE) [1-3]. However, previous studies regarding this software have focused on its text-based capabilities. ChatGPT-4 Vision (ChatGPT-4V), announced on September 25, 2023, includes image input features, potentially expanding the medical applications of the program [4]. To assess the multimodal performance of ChatGPT-4V in medicine, its performance on JMLE questions involving clinical images and tables was tested.

Methods

Overview

ChatGPT-4V was used to complete the 117th JMLE in the Japanese language (Figure S1 in Multimedia Appendix 1). Its responses were compared to the passing criteria and mean human examinee score of the JMLE. This study, conducted from October 12 to 14, 2023, used the September 25, 2023, version of the LLM (ChatGPT-4V) with a knowledge cutoff date of January 2022 (Multimedia Appendix 2 [5]). Human examinees' correct response rates were obtained from statistics based on reports from actual JMLE examinees, calculated by medu4, a preparatory school for the JMLE [5,6].

Statistical Analysis

The mean and 95% CIs of the test scores are provided. A one-sample proportion test was used to compare the correct response rate of the human examinees with that of ChatGPT-4V. Statistical significance was set at *P*<.05 for all 2-tailed tests. All statistical analyses were conducted using Stata statistical software (version 17; StataCor).

Ethical Considerations

This study used previously available web-based data and did not include human participants. Therefore, Shimane University's Institutional Review Board did not mandate ethics approval.

Results

Evaluation Outcomes

The responses to 386 questions from the 117th JMLE were used in this study. Using the Ministry of Health, Labor, and Welfare criteria, GPT-4V scored 85.1% on the essential knowledge section and 76.5% on the other sections of the JMLE, meeting the passing criteria [6]. For text-only questions, ChatGPT-4V achieved a correct response rate of 84.5%, similar to the mean human examinee score (Table 1). The correct response rate for questions with images was 71.9% for ChatGPT-4V, 13.1 points below the mean human examinee score (*P*<.001). The correct response rate for questions with tables (including figures) was 35.0% for ChatGPT-4V, which was significantly lower than the mean human examinee score (83.9%; *P*<.001).

XSL•FO
RenderX

Table . Correct response rates of ChatGPT-4 Vision (ChatGPT-4V) and human examinees on the Japanese Medical Licensing Examination (JMLE).

Characteristics		Total, n (%)	Examinees[a], mean	GPT-4V, mean	95% CI	Difference	*P* value
All questions		386 (100)	84.9	78.2	74.1-82.4	−6.7	.003
Question category							
	Essential knowledge	96 (24.9)	89.6	83.3	75.9-90.8	−6.3	.04
	General clinical knowledge	144 (37.3)	83.1	70.8	63.4-78.3	−12.3	<.001
	Specific diseases	146 (37.8)	83.5	82.2	76.0-88.4	−1.3	.67
Type							
	General	190 (49.2)	84.6	78.9	73.2-84.7	−5.7	.03
	Clinical	149 (38.6)	84.1	77.2	70.4-83.0	−6.9	.02
	Clinical sentence	47 (12.2)	88.5	78.7	67.0-90.4	−9.8	.04
Imaging and table questions							
	Text only	252 (65.3)	84.9	84.5	80.1-89.0	−0.4	.87
	With images	114 (29.5)	85.0	71.9	63.7-80.2	−13.1	<.001
	With tables	20 (5.2)	83.9	35.0	14.1-55.9	−48.9	<.001

[a]The correct response rates of human examinees are based on a survey of actual human examinees, reported by medu4, a preparatory school for the JMLE [5].

Discussion

Principal Results

Although ChatGPT-4V demonstrated proficiency in text-centric questions, the correct response rates were significantly lower for image and table-oriented questions. ChatGPT-4V may have poorer text comprehension skills compared to ChatGPT-4, even when image processing is not required [7]. Additionally, a language bias may obscure the image context when interpreting images and texts simultaneously, potentially leading to an overreliance on prior text information, even when it contradicts the image context, a phenomenon called “hallucination” [8]. These factors may have led to ChatGPT-4V’s lower rate of correct responses to questions involving images.

Furthermore, responding to questions with tables requires interpreting the Japanese characters within the tables. OpenAI has verified that its GPT-4V model misrecognizes symbols, including image characters [4]. Previous studies have noted that GPT-4V relies on text-based information rather than an analysis of tables when answering questions [8]. In addition, the program’s performance diminishes when interpreting characters in non-Latin languages [9]. These factors may explain the observed decline in performance when interpreting tables containing Japanese characters.

The multimodal LLM GPT-4V is unreliable in interpreting information presented in image or tables, especially for medical purposes [4]. Further development of the program is required for diagnostic applications.

Limitations

This study has several limitations. First, different results may be obtained even when using the same methods owing to the inherent randomness of ChatGPT or version changes in ChatGPT. A report indicates that test results can vary with repeated responses from ChatGPT [10]. Furthermore, when providing images to ChatGPT, we did not remove blank spaces, indicating that the quality of images sent to ChatGPT could also affect the outcomes. Second, the JMLE includes options that, if selected twice or more, will result in failure. However, these options are not publicly disclosed, making them unaccounted for in this study [5]. Finally, although this study focused on ChatGPT, ongoing advancements in other multimodal LLMsshould also be considered.

Conclusions

ChatGPT-4V successfully passed the 117th JMLE, demonstrating proficiency in handling including image- and table-based questions. However, more developments are needed to improve its ability to interpret tables. Further research should assess the safety and efficacy of ChatGPT-4V as a multimodal LLM in supporting medical practice, facilitating learning in clinical environments and advancing medical education.

XSL•FO
RenderX

Acknowledgments

We would like to thank Dr Kota Sakaguchi, Shimane University Hospital, for his careful support throughout this study. We would also like to thank Dr Sanjay Saint, a professor at the University of Michigan, for his numerous contributions and support in this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional statistics.
[DOCX File, 2285 KB - mededu_v10i1e54283_app1.docx]

Multimedia Appendix 2
Detailed methodology.
[DOCX File, 17 KB - mededu_v10i1e54283_app2.docx]

References

1. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/blog/chatgpt/ [accessed 2023-11-30]
2. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
3. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ 2023 Jun 29;9:e48002. [doi: 10.2196/48002] [Medline: 37384388]
4. GPT-4V(Ision) system card. OpenAI. 2023. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf [accessed 2023-10-26]
5. Announcement of successful passage of the 117th National Medical Examination (Japanese) [Article in Japanese]. Ministry of Health. 2023. URL: https://www.mhlw.go.jp/general/sikaku/successlist/2023/siken01/about.html [accessed 2023-10-26]
6. Searching questions [Article in Japanese]. Medu4. 2023. URL: https://medu4.com/quizzes/search [accessed 2023-10-26]
7. Wu Y, Wang S, Yang H, et al. An early evaluation of GPT-4V(ision). arXiv. Preprint posted online on Oct 25, 2023 URL: https://arxiv.org/abs/2310.16534 [accessed 2024-05-14]
8. Liu F, Lin K, Li L, Wang J, Yacoob Y, Wang L. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. arXiv. Preprint posted online on Jun 26, 2023 URL: https://arxiv.org/abs/2306.14565 [accessed 2024-05-14]
9. Shi Y, Peng D, Liao W, Lin Z, Chen X, Liu C, et al. Exploring OCR capabilities of GPT-4V(ision): a quantitative and in-depth evaluation. arXiv. Preprint posted online on Oct 25, 2023 URL: https://arxiv.org/abs/2310.16809 [accessed 2024-05-14]
10. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]

Abbreviations

ChatGPT-4V: ChatGPT 4 Vision
JMLE: Japanese Medical Licensing Examination
LLM: large language model
USMLE: United States Medical Licensing Examination

Edited by G Eysenbach, TDA Cardoso; submitted 06.11.23; peer-reviewed by F Liu, L Zhu, T Ma; revised version received 09.04.24; accepted 22.04.24; published 23.05.24.

Please cite as:
Takagi S, Koda M, Watari T
The Performance of ChatGPT-4V in Interpreting Images and Tables in the Japanese Medical Licensing Exam
JMIR Med Educ 2024;10:e54283
URL: https://mededu.jmir.org/2024/1/e54283
doi:10.2196/54283

XSL•FO
RenderX

© Soshi Takagi, Masahide Koda, Takashi Watari. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.5.2024. 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.

XSL•FO
RenderX

The Utility of Wearable Cameras in Developing Examination Questions and Answers on Physical Examinations: Preliminary Study

Sho Fukui[1,2,3], MPH, MD; Taro Shimizu[4], MPH, MBA, MD, PhD; Yuji Nishizaki[5], MPH, MD, PhD; Kiyoshi Shikino[6,7], MD, MHPE, PhD; Yu Yamamoto[8], MD; Hiroyuki Kobayashi[9], MD, PhD; Yasuharu Tokuda[10,11], MPH, MD

1
2
3
4
5
6
7
8
9
10
11

Corresponding Author:
Yuji Nishizaki, MPH, MD, PhD

Abstract

To assess the utility of wearable cameras in medical examinations, we created a physician-view video-based examination question and explanation, and the survey results indicated that these cameras can enhance the evaluation and educational capabilities of medical examinations.

(JMIR Med Educ 2024;10:e53193) doi:10.2196/53193

KEYWORDS

medical education; medical technology; wearable device; wearable camera; medical examination; exam; examination; exams; examinations; physical; resident physicians; wearable; wearables; camera; cameras; video; videos; innovation; innovations; innovative; recording; recordings; survey; surveys

Introduction

Wearable devices have been increasingly used in medicine [1]. Wearable video cameras differ from conventional cameras in that they simulate the perspectives of health care professionals rather than the view of observers. In medical education, wearable video cameras have shown their usefulness in patient interviews [2], virtual physical examination training [3], educational live-streaming ward rounds [4], basic clinical procedures (eg, vascular access) [2], and endoscopic and surgical procedures [5,6]. Wearable cameras, capable of capturing highly realistic situations, can be effective in assessing practical knowledge and providing educational feedback. However, they have not been used in medical examinations. This study aimed to examine the utility of wearable cameras in creating examination questions and answers.

Methods

Development of an Examination Question and Its Explanation

We developed a single examination question focusing on physical examination skills for resident physicians. In October 2021, authors YN and TS created a simulated outpatient case of appendicitis: a middle-aged man with abdominal pain and localized peritoneal irritation in the right lower quadrant. A volunteer physician played the role of the simulated patient. A physician examined the patient with a wearable camera on his head, recording physician-patient interactions. A compact wide-angle wearable camera (Insta360 ONE R) was used to reproduce a high-resolution physician view, including peripheral view fields (Figure 1).

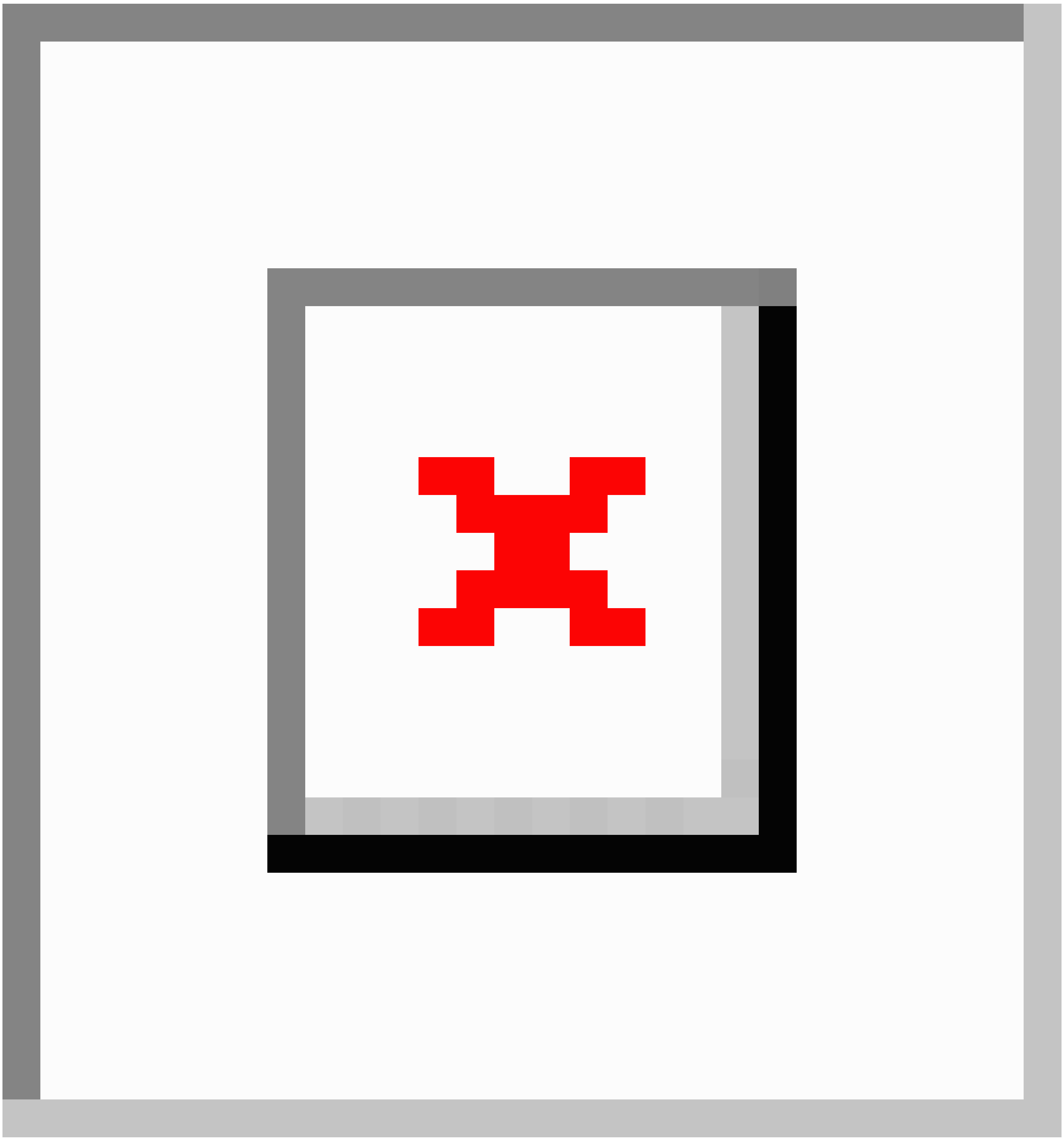
Using the recorded footage, we created 5 concise videos (about 10 seconds each) depicting various physical examination scenes, including (1) indirect abdominal percussion, (2) checking

XSL•FO
RenderX

peritoneal irritation by coughing in the supine position, (3) direct abdominal percussion, (4) the heel drop test, and (5) abdominal palpation; Multimedia Appendices 1, 2, 3, 4, and 5. The examination question asked for the correct sequence of the physical examination. Based on the patient's position (standing to supine position) and the invasiveness of the examination procedure, 4-2-1-3-5 was considered a correct answer upon the authors' agreement.

Additionally, we produced an explanatory physician-view answer video in which an experienced physician (TS) explained the proper sequence and key points in abdominal examinations (Multimedia Appendix 6).

Figure 1. Schematic description of (A) recording physical examination with a wearable camera and (B) its recoded image from the physician's view.



Study Participants, Examination, and Subsequent Survey

The General Medicine In-Training Examination (GM-ITE) for the academic year 2021, a validated nationwide computer-based examination in Japan, was conducted in January and February 2022 [7]. After completing GM-ITE, participants were given the voluntary option to participate in this study on their computer monitors. If agreed, they were requested to answer the question and view the explanatory video. To evaluate the utility of wearable cameras from the examinees' perspective, participants were asked to complete a subsequent questionnaire survey using the same computers.

XSL•FO
RenderX

Data Analysis

We described participant characteristics, examination results, and survey results using descriptive statistics.

Ethical Consideration

We obtained informed consent from the person who played the simulated patient's role and all other participants before the examination. All data were anonymized, and no honorarium was provided to participants. This study was approved by the Ethics Review Board of the Japan Organization of Advancing Medical Education (approval number: 21 - 10).

Results

A total of 43 resident physicians from multiple Japanese institutions who completed the examination and survey were included. Of these, 28 (65.1%) participants were postgraduate year-1 and 15 (34.9%) were postgraduate year-2 residents; 19 (44%) participants correctly answered the question.

In the postexamination survey, 32 (74%) participants agreed that they could envision real patients better compared to text-based questions (question 1); 26 (61%) were satisfied with the question (question 2); 29 (67%) stated that physician-view videos were more suitable for evaluating clinical competency than observer-view videos (question 3); and 34 (79%) answered that physician-view explanatory video was a more effective educational approach than text-based explanations (question 4; Table 1).

Table . Results of the survey about the examination and explanatory videos.

Survey questions		Total (N=43), n (%)
Question 1: Are you able to envision real patients better with this examination compared to a text-based question?		
	Strongly agree	11 (26)
	Agree	21 (49)
	Neutral	5 (12)
	Disagree	5 (12)
	Strongly disagree	1 (2)
Question 2: Are you satisfied with this question?		
	Strongly agree	8 (19)
	Agree	18 (42)
	Neutral	13 (30)
	Disagree	4 (9)
	Strongly disagree	0 (0)
Question 3: Are physician-view videos more suitable for evaluating clinical competency than observer-view videos?		
	Strongly agree	13 (30)
	Agree	16 (37)
	Neutral	11 (26)
	Disagree	3 (7)
	Strongly disagree	0 (0)
Question 4: Is an explanatory video from a physician's viewpoint more effective for learning the content than traditional text-based explanations?		
	Strongly agree	13 (30)
	Agree	21 (49)
	Neutral	5 (12)
	Disagree	4 (9)
	Strongly disagree	0 (0)

Discussion

This study used wearable cameras to create examination questions and subsequent answer explanations; the survey suggested the potential utility of physician-view videos in medical examinations.

This study's results align with previous research, which showed the effectiveness of chest-mounted point-of-view footage over observer-view videos in physical examination training [3]. For

XSL•FO
RenderX

teaching physical examination, video-based e-learning was superior to illustrated text-based e-learning [8]. Moreover, a clinical simulation video successfully assessed clinical competencies across multiple domains in resident physicians [9]. Wearable cameras can provide learners with “immersion,” a sense that one is participating in realistic experiences, which enhances situated learning [10]. Physician-view videos of real clinical situations may emphasize diverse (eg, nonverbal) information. Furthermore, physician-view videos can motivate examinees to learn more actively by regarding themselves as practitioners rather than observers.

This pilot study had limitations. We used a simple subjective survey in a small cohort of volunteer participants. Additionionally, the participants’ detailed characteristics were not collected. More quantitative research with objective outcomes will be required to verify the educational value of incorporating wearable cameras into medical examinations.

Acknowledgments

The authors would like to express their deep appreciation to Dr Soshi Mano for his cooperation in assisting with this study. Generative artificial intelligence was not used in our manuscript. This study was supported by the Health, Labor, and Welfare Policy Grants of Research on Region Medical (21IA2004) from the Ministry of Health, Labor, and Welfare (MHLW). The MHLW did not participate in any part of the study process, including the designing, data analysis, data interpretation, review, and approval of the manuscript.

Data Availability

The data used in this study are not available because participants of this study did not consent to public sharing of their data.

Conflicts of Interest

FS, TS, KS, and YY received honoraria from Japan Institute for Advancement of Medical Education Program (JAMEP) as General Medicine In-Training Examination (GM-ITE) exam preparers. YN received an honorarium from JAMEP as GM-ITE project manager. YT is the JAMEP director and received an honorarium from JAMEP as a speaker at a JAMEP lecture. KS and HK also received honoraria from JAMEP as speakers at JAMEP lectures.

Multimedia Appendix 1
Abdominal examination video 1.
[MP4 File, 7552 KB - mededu_v10i1e53193_app1.mp4]

Multimedia Appendix 2
Abdominal examination video 2.
[MP4 File, 7647 KB - mededu_v10i1e53193_app2.mp4]

Multimedia Appendix 3
Abdominal examination video 3.
[MP4 File, 5176 KB - mededu_v10i1e53193_app3.mp4]

Multimedia Appendix 4
Abdominal examination video 4.
[MP4 File, 5349 KB - mededu_v10i1e53193_app4.mp4]

Multimedia Appendix 5
Abdominal examination video 5.
[MP4 File, 7173 KB - mededu_v10i1e53193_app5.mp4]

Multimedia Appendix 6
Answer explanation video.
[MP4 File, 138428 KB - mededu_v10i1e53193_app6.mp4]

References

1. Iqbal MH, Aydin A, Brunckhorst O, Dasgupta P, Ahmed K. A review of wearable technology in medicine. J R Soc Med 2016 Oct;109(10):372-380. [doi: 10.1177/0141076816663560] [Medline: 27729595]
2. Kwon OY. Online clinical skills education using a wearable action camera for medical students. J Exerc Rehabil 2022 Dec;18(6):356-360. [doi: 10.12965/jer.2244460.230] [Medline: 36684536]

XSL•FO
RenderX

3. Teitelbaum D, Xie M, Issa M, et al. Use of wearable point-of-view live streaming technology for virtual physical exam skills training. Can Med Educ J 2022 Jul;13(3):64-66. [doi: 10.36834/cmej.73076] [Medline: 35875435]
4. Mill T, Parikh S, Allen A, et al. Live streaming ward rounds using wearable technology to teach medical students: a pilot study. BMJ Simul Technol Enhanc Learn 2021;7(6):494-500. [doi: 10.1136/bmjstel-2021-000864] [Medline: 35520979]
5. Hinchcliff M, Kao M, Johnson K. The importance of technical skills assessment during an airway foreign body removal course. Int J Pediatr Otorhinolaryngol 2019 Feb;117:1-5. [doi: 10.1016/j.ijporl.2018.11.007] [Medline: 30579061]
6. Lee B, Chen BR, Chen BB, Lu JY, Giannotta SL. Recording stereoscopic 3D neurosurgery with a head-mounted 3D camera system. Br J Neurosurg 2015 Jun;29(3):371-373. [doi: 10.3109/02688697.2014.997664] [Medline: 25620087]
7. Nagasaki K, Nishizaki Y, Nojima M, et al. Validation of the general medicine in-training examination using the professional and linguistic assessments board examination among postgraduate residents in Japan. Int J Gen Med 2021;14:6487-6495. [doi: 10.2147/IJGM.S331173] [Medline: 34675616]
8. Buch SV, Treschow FP, Svendsen JB, Worm BS. Video- or text-based e-learning when teaching clinical procedures? A randomized controlled trial. Adv Med Educ Pract 2014;5:257-262. [doi: 10.2147/AMEP.S62473] [Medline: 25152638]
9. Shikino K, Nishizaki Y, Fukui S, et al. Development of a clinical simulation video to evaluate multiple domains of clinical competence: cross-sectional study. JMIR Med Educ 2024 Feb 29;10:e54401. [doi: 10.2196/54401] [Medline: 38421691]
10. Dede C. Immersive interfaces for engagement and learning. Science 2009 Jan 2;323(5910):66-69. [doi: 10.1126/science.1167311] [Medline: 19119219]

Abbreviations

GM-ITE: General Medicine In-Training Examination

Edited by TDA Cardoso; submitted 29.09.23; peer-reviewed by A Arbabisarjou, P Vemavarapu, S Pesälä; revised version received 19.06.24; accepted 24.06.24; published 19.07.24.

Please cite as:
Fukui S, Shimizu T, Nishizaki Y, Shikino K, Yamamoto Y, Kobayashi H, Tokuda Y
The Utility of Wearable Cameras in Developing Examination Questions and Answers on Physical Examinations: Preliminary Study
JMIR Med Educ 2024;10:e53193
URL: https://mededu.jmir.org/2024/1/e53193
doi:10.2196/53193

© Sho Fukui, Taro Shimizu, Yuji Nishizaki, Kiyoshi Shikino, Yu Yamamoto, Hiroyuki Kobayashi, Yasuharu Tokuda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.7.2024. 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.

XSL•FO
RenderX

Research Letter

Using AI Text-to-Image Generation to Create Novel Illustrations for Medical Education: Current Limitations as Illustrated by Hypothyroidism and Horner Syndrome

Ajay Kumar[1], MBBS, MSc; Pierce Burr[1], BSc, MBChB, MRCSEd; Tim Michael Young[1], BSc, MBBS, PGCME, PhD

Queen Square Institute of Neurology, University College London, London, United Kingdom

Corresponding Author:
Tim Michael Young, BSc, MBBS, PGCME, PhD
Queen Square Institute of Neurology
University College London
Number 7 Queen Square
London, WC1N 3BG
United Kingdom
Phone: 44 2031082781
Email: t.young@ucl.ac.uk

Abstract

Our research letter investigates the potential, as well as the current limitations, of widely available text-to-image tools in generating images for medical education. We focused on illustrations of important physical signs in the face (for which confidentiality issues in conventional patient photograph use may be a particular concern) that medics should know about, and we used facial images of hypothyroidism and Horner syndrome as examples.

(JMIR Med Educ 2024;10:e52155) doi:10.2196/52155

KEYWORDS

artificial intelligence; AI; medical illustration; medical images; medical education; image; images; illustration; illustrations; photo; photos; photographs; face; facial; paralysis; photograph; photography; Horner's syndrome; Horner syndrome; Bernard syndrome; Bernard's syndrome; miosis; oculosympathetic; ptosis; ophthalmoplegia; nervous system; autonomic; eye; eyes; pupil; pupils; neurologic; neurological

Introduction

Artificial intelligence (AI) has become integral in medicine, outperforming skilled radiologists in certain domains [1]. However, there is limited exploration of AI's potential in producing illustrations for medical education [2,3]. Confidentiality concerns can limit traditional patient photo use, especially when facial features are essential [4]. Using widely available AI text-to-image tools, we aimed to create images portraying distinct facial signs important for medical trainees—hypothyroidism (myxedema) and Horner syndrome [5,6]. These tools generate unique, high-quality images based on text prompts, utilizing learned probability distributions rather than pre-existing images [7].

Methods

ChatGPT was used to generate prompts for the two AI text-to-image tools used in this study—DALL·E 2 and Midjourney (Multimedia Appendix 1) [8-10], with which the prompts were used to generate images for hypothyroidism and Horner syndrome. The images were assessed and selected, using the following suitability criteria:

1. Images were excluded if any of the following features were present: insufficient coverage of the face, blurred images, a lack of realistic or humanoid features, a lack of continuity of edges, background noise, cloning errors, and geometrical and shadow inconsistencies.
2. Remaining images were accepted if they adequately represented the facial features of hypothyroidism or Horner syndrome, as judged by the coauthors (all were experienced physicians).

If adequate images could not be generated via the above methods, additional prompts, which were not generated with ChatGPT, were used. If adequate images were still not generated, then secondary editing via Microsoft Paint and GNU

XSL•FO
RenderX

Image Manipulation Program (GIMP) was performed on the best image to try and meet the criteria listed above.

Results

Facial Features of Hypothyroidism

Using ChatGPT, the following text prompt was generated (restricted to the DALL·E 2 prompt word limit):

> *Generate an image depicting a middle-aged Caucasian woman with hypothyroidism presenting with facial myxedema. The woman should be shown in a frontal view, focusing on her face, scalp, and neck, without any makeup. The face must be very rounded and extreme scalp balding with coarse hair. Skin looks dry and pale. Outer eyebrows have a paucity of hairs, eyelids look very puffy. She looks tired.*

The prompt was used to generate 120 images. Of these, 53 were removed, using our preset exclusion criteria. Of the remaining 67, only 17 met some of the criteria for adequately representing facial features of hypothyroidism. The best image was selected as Figure 1 [9], with no additional editing needed.

Figure 1. Artificial intelligence text-to-image production of facial features typical of hypothyroidism (myxedema) showing classical clinical features, including a rounded face with dry, pale skin; puffy eyelids; a general appearance of tiredness; and partial balding with coarse hair and loss of hair in the eyebrows (especially in the outer third). This image was produced by using DALL·E 2 [9] alone and without additional editing.

Horner Syndrome

The following prompt was obtained from ChatGPT:

> *Create an illustrative depiction of a patient displaying Horner's syndrome, emphasizing the key clinical features, such as ptosis (drooping of the upper eyelid), miosis (constricted pupil), and anhidrosis (lack of sweating) on one side of the face. Ensure the image is clear and medically accurate, aiding in the understanding of this neurological condition.*

Of the 120 images, 85 met our exclusion criteria, but none met our inclusion criteria, even after alternative prompts and DALL·E 2 were used. We therefore selected the best image (produced by Midjourney) and then performed secondary editing with Microsoft Paint and GIMP (Figure 2 [10]). This produced an image of Horner syndrome that was judged as adequate.

Figure 2. Generated illustration of Horner syndrome. Image 1 was produced by using Midjourney [10]. Image 2 shows the result after minor image editing (as described in our *Methods* section) to attenuate the key teaching features, which are labeled in image 3 (A: ptosis; B: miosis; C: apparent enophthalmos; D: upside-down ptosis).

XSL•FO
RenderX

Discussion

We aimed to explore the potential, as well as the current limits, of AI text-to-image generation in producing illustrations of medical conditions affecting the face. Without the use of high-quality medical images, it can be more challenging to teach others about these important conditions [11]. We showed that AI text-to-image generation is readily possible for hypothyroidism—a condition with symmetrical features. However, for Horner syndrome—a condition with asymmetrical features—adequate images could only be produced after some additional slight editing, reflecting a possible limiting factor of these tools. Ours are the first AI-generated images of classical facial features of hypothyroidism and Horner syndrome that we are aware of.

Confidentiality has become an increasing concern in the use of medical images over the last few decades. Text-to-image tools have ethical issues, including issues of consent for the original photos used to train these tools. Additionally, issues of accuracy are key. Nonmedics might be misled on medical signs by using such tools. Targets for future research are the potential for biases with these tools and the danger of stereotypes being perpetuated. Despite these limitations, AI-generated images may enhance case-based learning, allowing students to study and analyze a diverse range of medical cases. Text-to-image tools show exciting potential and may allow easier access to high-quality images in medical education [12,13].

Conflicts of Interest

None declared.

Multimedia Appendix 1

Tools used in this article (all prompts entered in English).

[DOCX File , 13 KB - mededu_v10i1e52155_app1.docx]

References

1. Plesner LL, Müller FC, Nybing JD, Laustrup LC, Rasmussen F, Nielsen OW, et al. Autonomous chest radiograph reporting using AI: estimation of clinical impact. Radiology 2023 May;307(3):e222268. [doi: 10.1148/radiol.222268] [Medline: 36880947]
2. Williams MC, Williams SE, Newby DE. Artificial intelligence-based text-to-image generation of cardiac CT. Radiol Cardiothorac Imaging 2023 Apr 6;5(2):e220297 [FREE Full text] [doi: 10.1148/ryct.220297] [Medline: 37274418]
3. Adams LC, Busch F, Truhn D, Makowski MR, Aerts HJWL, Bressem KK. What does DALL-E 2 know about radiology? J Med Internet Res 2023 Mar 16;25:e43110 [FREE Full text] [doi: 10.2196/43110] [Medline: 36927634]
4. Hill K. Consent, confidentiality and record keeping for the recording and usage of medical images. J Vis Commun Med 2006 Jun;29(2):76-79. [doi: 10.1080/01405110600863365] [Medline: 16928590]
5. Siskind SM, Lee SY, Pearce EN. Investigating hypothyroidism. BMJ 2021 Apr 27;373:n993. [doi: 10.1136/bmj.n993] [Medline: 33906834]
6. Amonoo-Kuofi HS. Horner's syndrome revisited: with an update of the central pathway. Clin Anat 1999;12(5):345-361. [doi: 10.1002/(SICI)1098-2353(1999)12:5<345::AID-CA5>3.0.CO;2-L] [Medline: 10462732]
7. Zhang C, Zhang C, Zhang M, Kweon IS. Text-to-image diffusion models in generative AI: a survey. arXiv Preprint posted online on Apr 2, 2023. [FREE Full text] [doi: 10.48550/arXiv.2303.07909]
8. ChatGPT. OpenAI. URL: https://chat.openai.com [accessed 2024-01-10]
9. DALL·E 2. OpenAI. URL: https://openai.com/dall-e-2 [accessed 2024-01-10]
10. Midjourney. Midjourney Inc. URL: https://www.midjourney.com/home/ [accessed 2023-07-03]
11. Sagoo MG, Vorstenbosch MATM, Bazira PJ, Ellis H, Kambouri M, Owen C. Online assessment of applied anatomy knowledge: the effect of images on medical students' performance. Anat Sci Educ 2021 May;14(3):342-351. [doi: 10.1002/ase.1965] [Medline: 32289198]
12. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
13. Amri MM, Hisan UK. Incorporating AI tools into medical education: harnessing the benefits of ChatGPT and Dall-E. Journal of Novel Engineering Science and Technology 2023 Apr 24;2(02):34-39 [FREE Full text] [doi: 10.56741/jnest.v2i02.315]

Abbreviations

AI: artificial intelligence
GIMP: GNU Image Manipulation Program

XSL•FO
RenderX

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 24.08.23; peer-reviewed by U Kanike, Anonymous; comments to author 22.09.23; revised version received 12.01.24; accepted 29.01.24; published 22.02.24.

Please cite as:
Kumar A, Burr P, Young TM
Using AI Text-to-Image Generation to Create Novel Illustrations for Medical Education: Current Limitations as Illustrated by Hypothyroidism and Horner Syndrome
JMIR Med Educ 2024;10:e52155
URL: https://mededu.jmir.org/2024/1/e52155
doi:10.2196/52155
PMID:38386400

©Ajay Kumar, Pierce Burr, Tim Michael Young. Originally published in JMIR Medical Education (https://mededu.jmir.org), 22.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Medical Education and Artificial Intelligence: Web of Science–Based Bibliometric Analysis (2013-2022)

Shuang Wang[*], MMed; Liuying Yang[*], MMed; Min Li, BMed; Xinghe Zhang, PhD; Xiantao Tai, MMed

Second Clinical Medical College, Yunnan University of Chinese Medicine, Kunming, China

[*]these authors contributed equally

Corresponding Author:
Xiantao Tai, MMed

Abstract

Background: Incremental advancements in artificial intelligence (AI) technology have facilitated its integration into various disciplines. In particular, the infusion of AI into medical education has emerged as a significant trend, with noteworthy research findings. Consequently, a comprehensive review and analysis of the current research landscape of AI in medical education is warranted.

Objective: This study aims to conduct a bibliometric analysis of pertinent papers, spanning the years 2013 - 2022, using CiteSpace and VOSviewer. The study visually represents the existing research status and trends of AI in medical education.

Methods: Articles related to AI and medical education, published between 2013 and 2022, were systematically searched in the Web of Science core database. Two reviewers scrutinized the initially retrieved papers, based on their titles and abstracts, to eliminate papers unrelated to the topic. The selected papers were then analyzed and visualized for country, institution, author, reference, and keywords using CiteSpace and VOSviewer.

Results: A total of 195 papers pertaining to AI in medical education were identified from 2013 to 2022. The annual publications demonstrated an increasing trend over time. The United States emerged as the most active country in this research arena, and Harvard Medical School and the University of Toronto were the most active institutions. Prolific authors in this field included Vincent Bissonnette, Charlotte Blacketer, Rolando F Del Maestro, Nicole Ledows, Nykan Mirchi, Alexander Winkler-Schwartz, and Recai Yilamaz. The paper with the highest citation was "Medical Students' Attitude Towards Artificial Intelligence: A Multicentre Survey." Keyword analysis revealed that "radiology," "medical physics," "ehealth," "surgery," and "specialty" were the primary focus, whereas "big data" and "management" emerged as research frontiers.

Conclusions: The study underscores the promising potential of AI in medical education research. Current research directions encompass radiology, medical information management, and other aspects. Technological progress is expected to broaden these directions further. There is an urgent need to bolster interregional collaboration and enhance research quality. These findings offer valuable insights for researchers to identify perspectives and guide future research directions.

(JMIR Med Educ 2024;10:e51411) doi:10.2196/51411

KEYWORDS

artificial intelligence; medical education; bibliometric analysis; CiteSpace; VOSviewer

Introduction

The concept of artificial intelligence (AI), referring to machines and systems capable of emulating human intelligence, was first introduced at an academic conference in 1956. Its extensive research fields encompass numerous domains, including intelligent expert systems, language processing, intelligent data retrieval, and intelligent control. AI stands as one of the three groundbreaking technologies of the 21st century, sharing the pedestal with genetic engineering and nanoscience technologies [1-3]. The ultimate aim of AI is to facilitate the use of machines in replicating and expanding human intelligence. In doing so, machines are empowered to listen, see, speak, think, and make decisions in a manner akin to humans, thus elevating the quality of human life [4,5].

The sustained evolution of AI has resulted in a paradigm shift in medical practice, transitioning from traditional methods to digital health care, with AI finding applications in diverse realms of medical and health care. AI can generate pathological diagnostic reports through integrated data analysis, aid psychologists in diagnosing mental disorders by simulating human thinking patterns, and perform imaging evaluations via deep learning. Moreover, AI can be used to manage clinical patients, and deliver doctor-prescribed treatment plans through records of patient history and treatment processes [6]. Research in AI has demonstrated that the output-input ratio in the medical field holds more promise than other disciplines [7]. As such, the advancement of medical education is imperative, and, over the past several decades, research and development in the application of AI in medical education has escalated [8].

XSL•FO
RenderX

Bibliometrics serves as a tool for the quantitative analysis of published literature, determining the relationship between research statements and emerging research frontiers, based on co-occurrence, citation, and cocitation [9]. Numerous global bibliometric analyses have been conducted using CiteSpace and VOSviewer in recent years. These analyses have focused on the comprehensive rehabilitation statuses and research trends of diseases such as cancer, ankylosing spondylitis, motor and neuropathic pain, and osteoarthritis [10-13]. However, to the best of our knowledge, a bibliometric analysis of AI's application in medical education has yet to be implemented.

Consequently, this study leverages CiteSpace and VOSviewer to assess the current research status and emergent trends of AI in medical education over the past decade.

Methods

All data for this research were procured from the Web of Science. The search parameters for data retrieval encompassed the topics "artificial intelligence" and "medical education" (refer to Table 1), with a publication date range from 2013 to 2022. The search results were subsequently analyzed using CiteSpace and VOSviewer. CiteSpace, a visual analysis software developed by Chaomei Chen, was used to analyze the total number of papers related to the topic, the trend of changes over the years, the frequency of keywords, and centrality. This software allowed for a more convenient and intuitive analysis of the structure, rules, and distribution of subject knowledge. A scientific knowledge map facilitated the identification of research hotspots, progress, and the current situation within a specific field. VOSviewer, a software tool primarily oriented toward document data processing, enabled the analysis of the country, institution, author, journal, keywords, and co-occurrence knowledge graph of country, institution, journal, and document in the literature. Each node on the knowledge graph represented a unique element, with the connection width between nodes indicating collaboration strength, node size reflecting the number of publications, and larger nodes indicating more frequent releases.

Table . Search queries.

Set	Results, n	Search query
#1	140,447	(((TS[a]=(generative AI))[b] OR TS=(AI)) OR TS=(Artificial Intelligence)) OR TS=(generative Artificial Intelligence) Indexes=Web of Science, timespan=2013-2022
#2	93,678	(TS=(medical education) Indexes=Web of Science, timespan=2013-2022
#3	580	#1 and #2

[a]TS: topic.

[b]AI: artificial intelligence.

The papers for this study were downloaded in .txt format from the Web of Science database. Two expert researchers examined the title, keywords, and abstract, and screened the papers based on inclusion and exclusion criteria. In cases of disagreement or difficulty in paper inclusion, a third reviewer made the final decision via discussion. Initially, a total of 580 papers were searched, of which 385 papers that did not meet the study's topic were excluded, resulting in the retention of 195 papers.

Ethical Considerations

According to the Regulations of the People's Republic of China on Ethical Review of Science and Technology (Trial), Number 167 of the State Science and Technology Development Supervision (2023), scientific research activities involving humans or other animals need to undergo ethical review. This thesis does not involve humans or other animals, nor does it pose risks to life and health, the ecological environment, public order, or sustainable development. Therefore, ethical approval is not required.

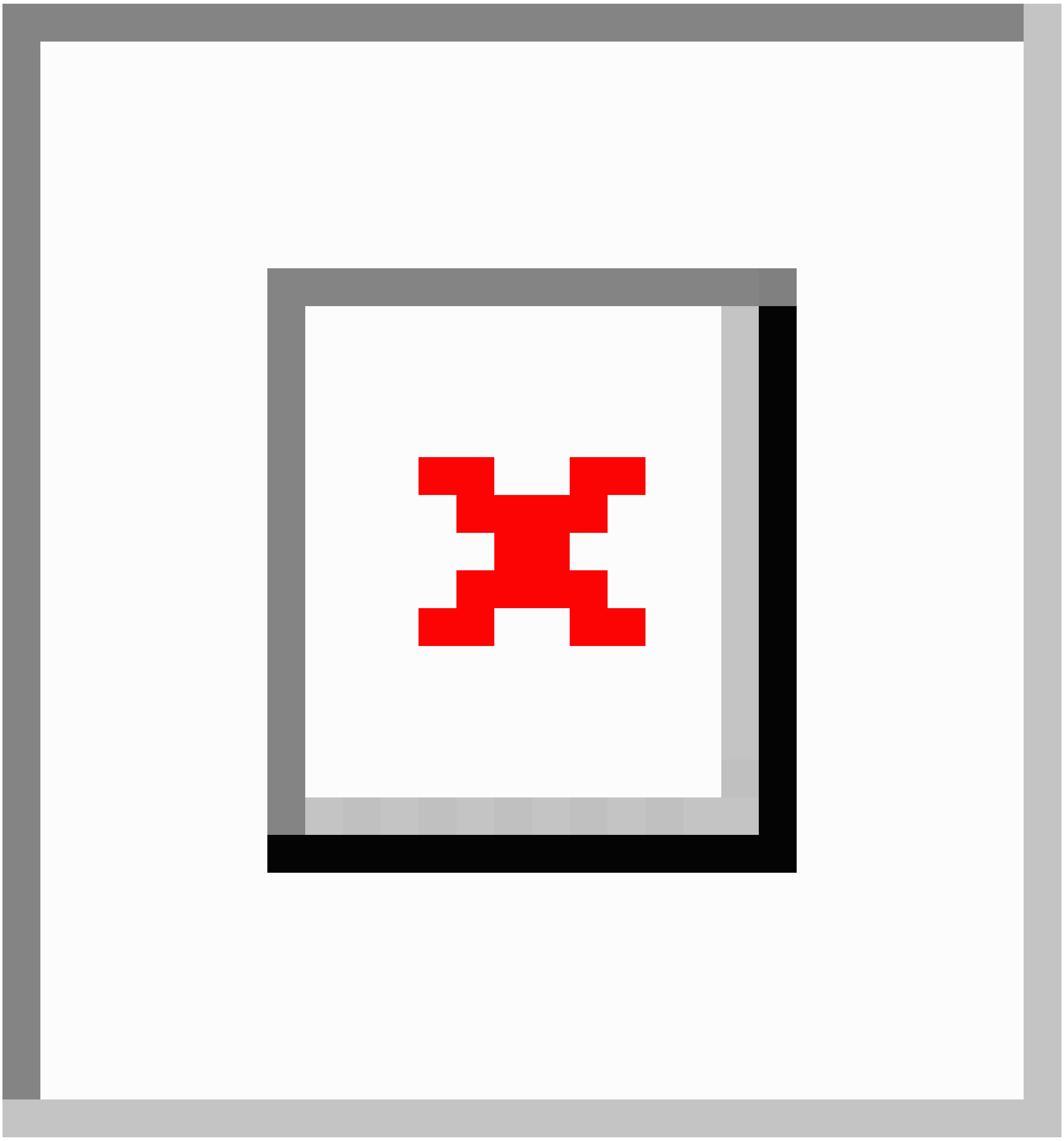
Results

Annual Publications

Figure 1 shows that a total of 195 papers on AI and medical education have been published in the past decade, showing an overall upward trend. The publications saw a significant surge from 2020 to 2021, reaching a peak in 2021, although the number of related papers published in 2022 decreased. The development of AI presented unprecedented opportunities and challenges to the medical and health industry. Medical education, being the cornerstone of medical industry development, can benefit from the application of AI, driving continual innovation.

XSL•FO
RenderX

Figure 1. Chart of the number of years issued.



National Analysis

Based on a comprehensive national analysis, 57 countries globally contributed to the exploration of AI within the field of medical education from 2013 to 2022. The United States took the lead by publishing 66 papers, thereby establishing itself as the most actively engaged country in this domain. The subsequent countries, albeit with lesser contributions, were Canada (24 papers), China (17 papers), England (13 papers), Singapore (12 papers), Australia (12 papers), India (9 papers), Germany (8 papers), the Netherlands (8 papers), and Spain (7 papers). The most cited countries were the United States (845 citations), Singapore (489 citations), and China (435 citations). When evaluated in terms of total link strength, the United States (44), the Netherlands (29), and Belgium (26) emerged as the top 3 countries (Table 2). Figure 2 shows that a clear inclination of North American and European countries toward the application of AI in medical education is evident, possibly due to their technological advancement. The United States has been a front-runner in this arena, publishing a multitude of relevant papers. Concurrently, it has fostered collaborative relationships with various countries for related research.

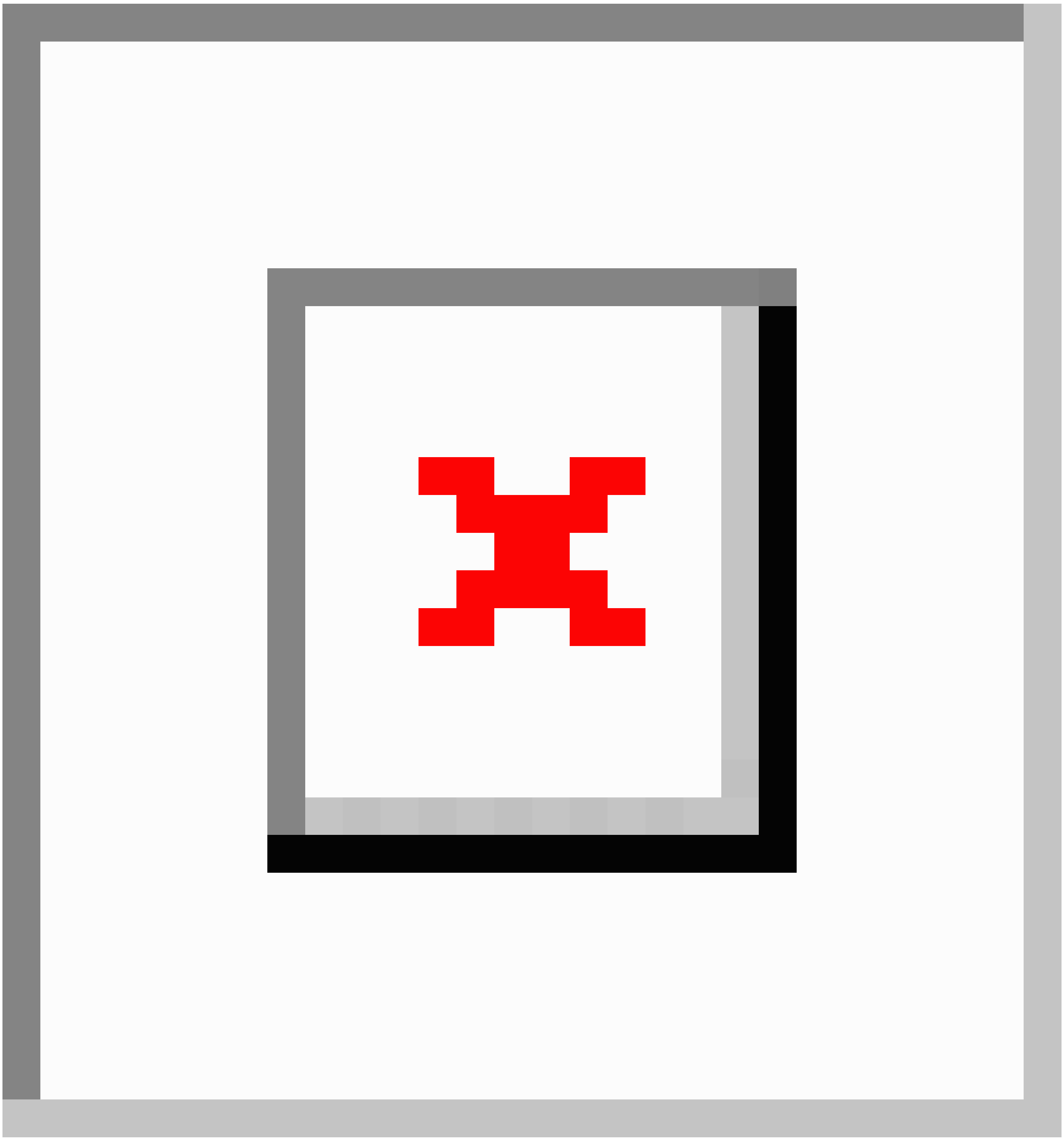
XSL•FO
RenderX

Table . Top 10 publications, centrality, and citations of countries.

Rank	Documents	Countries	Citations	Countries	Total link strength	Countries
1	66	United States	845	United States	44	United States
2	24	Canada	489	Singapore	29	The Netherlands
3	17	People’s Republic of China	435	People’s Republic of China	26	Belgium
4	13	England	371	Canada	23	Germany
5	12	Australia	155	England	22	England
6	12	Singapore	108	Spain	20	France
7	9	India	101	Germany	19	Italy
8	8	Germany	94	The Netherlands	19	Switzerland
9	8	The Netherlands	94	Belgium	18	Spain
10	7	Spain	85	Iran	16	Greece

XSL•FO
RenderX

Figure 2. National and regional co-occurrence map.



Institutional Analysis

Shifting the focus to an institutional analysis reveals that from 2013 to 2022, 77 institutions were engaged in research on AI in medical education. The two institutions that topped the list in terms of the number of publications were Harvard Medical School and the University of Toronto, each with 7 contributions, followed by McGill University and the National University of California, San Francisco (5 contributions each) (Table 3). The institutions receiving the most citations were Nanyang Technological University (396 citations), McGill University (149 citations), and the University of Chicago (127 citations). Figure 3 shows that Leiden University and Harvard Medical School demonstrated more collaboration with other institutions, both exhibiting a link strength of 15.

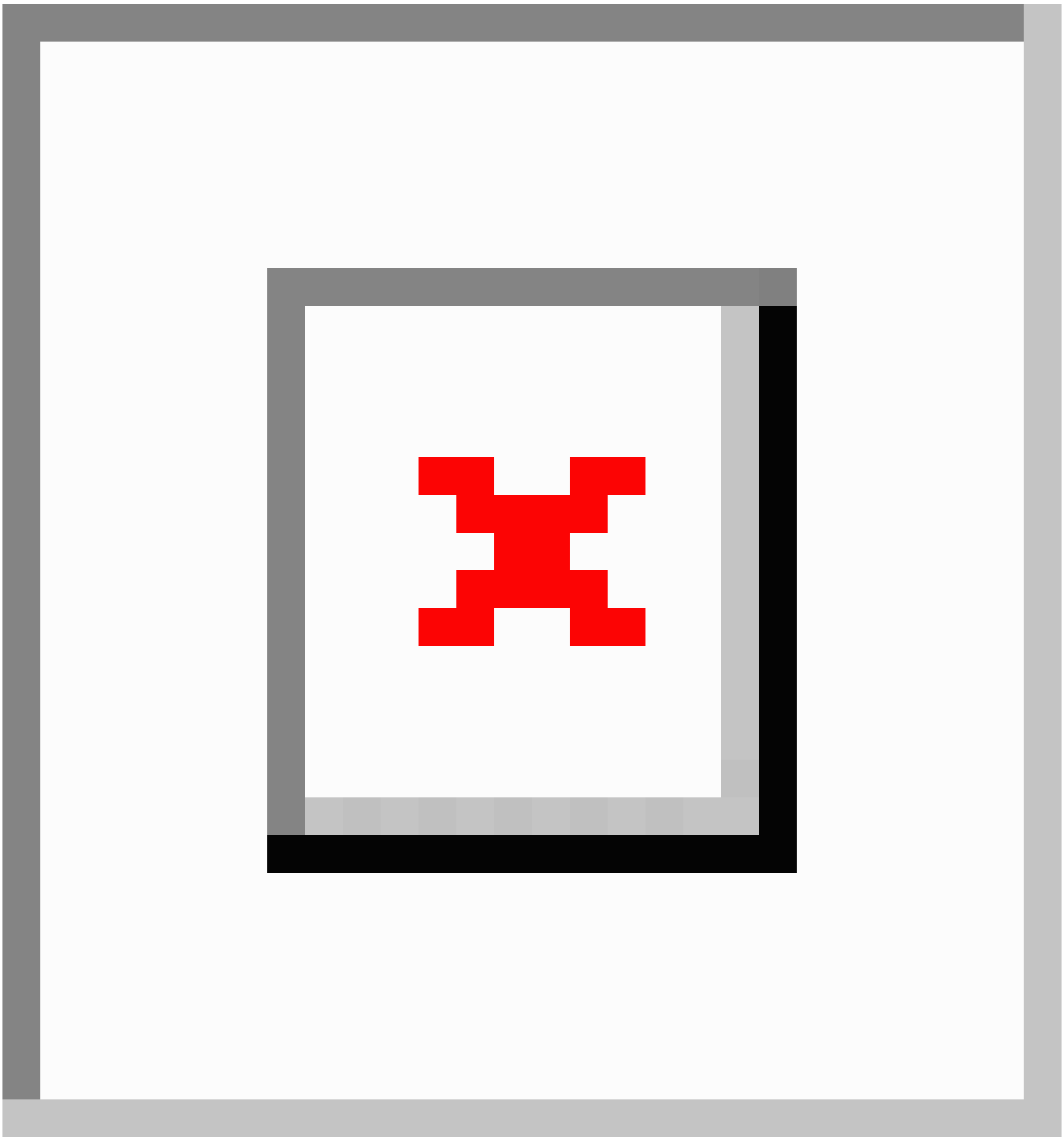
XSL•FO
RenderX

Table . Top 10 publications, centrality, and citations of organizations.

Rank	Documents	Organization	Citations	Organization	Total link strength	Organization
1	7	Harvard Medical School	396	Nanyang Technological University	15	Leiden University
2	7	University of Toronto	149	McGill University	15	Harvard Medical School
3	5	McGill University	127	University of Chicago	11	Oregon Health and Science University
4	5	National University Singapore	104	University of British Columbia	10	University of Toronto
5	5	Oregon Health and Science University	86	Guy's and St Thomas' NHS Foundation Trust	9	University of British Columbia
6	5	Queens University	83	Kings College London	9	Stanford University
7	5	Stanford University	68	University California San Francisco	9	Queens University
8	5	University of California San Francisco	67	National University Singapore	8	Imperial College London
9	4	Emory University	66	Sultan Qaboos University	8	Johns Hopkins University
10	4	Leiden University	60	University of Maryland	7	Ludwig Maximilians University Munchen

XSL•FO
RenderX

Figure 3. Organizations co-occurrence map.



Author Analysis

In the span of the last decade, research on AI and medical education has seen the involvement of a total of 53 authors. The authors most frequently contributing to the documents included Vincent Bissonnette, Charlotte Blacketer, Rolando F Del Maestro, Nicole Ledwos, Nykan Mirchi, Alexander Winkler-Schwartz, and Recai Yilmaz, each writing 3 papers. The authors garnering the highest citations encompassed the same group, with each achieving 143 citations (Table 4). As discerned from the VOSviewer image, there are no researchers with a significantly high number of publications, indicating that the volume of published papers remains relatively minimal. Figure 4 shows that research in this field is still nascent, with no particular research team outperforming others.

XSL•FO
RenderX

Table . Top 10 publications, centrality, and citations of authors.

Rank	Documents	Author	Citations	Author	Total link strength	Author
1	3	Bissonnette, Vincent	143	Bissonnette, Vincent	22	Bacchi, Stephen
2	3	Blacketer, Charlotte	143	Del Maestro, Rolando F	22	Duggan, Paul
3	3	Del Maestro, Rolando F	143	Ledwos, Nicole	22	Gallagher, Steve
4	3	Ledwos, Nicole	143	Mirchi, Nykan	22	Licinio, Julio
5	3	Mirchi, Nykan	143	Winkler-Schwartz, Alexander	22	Parnis, Roger
6	3	Winkler-Schwartz, Alexander	143	Yilmaz, Recai	22	Perry, Seth W
7	3	Yilmaz, Recai	56	Culp, Melissa P	22	Symonds, Ian
8	2	Bacchi, Stephen	56	Mollura, Daniel J	22	Tan, Yiran
9	2	Bulatov, Sergey	47	Sapci, A Hasan	22	Thomas, Josephine
10	2	Caliskan, S Ayhan	47	Sapci, H Aylin	22	Wagner, Morganne

XSL•FO
RenderX

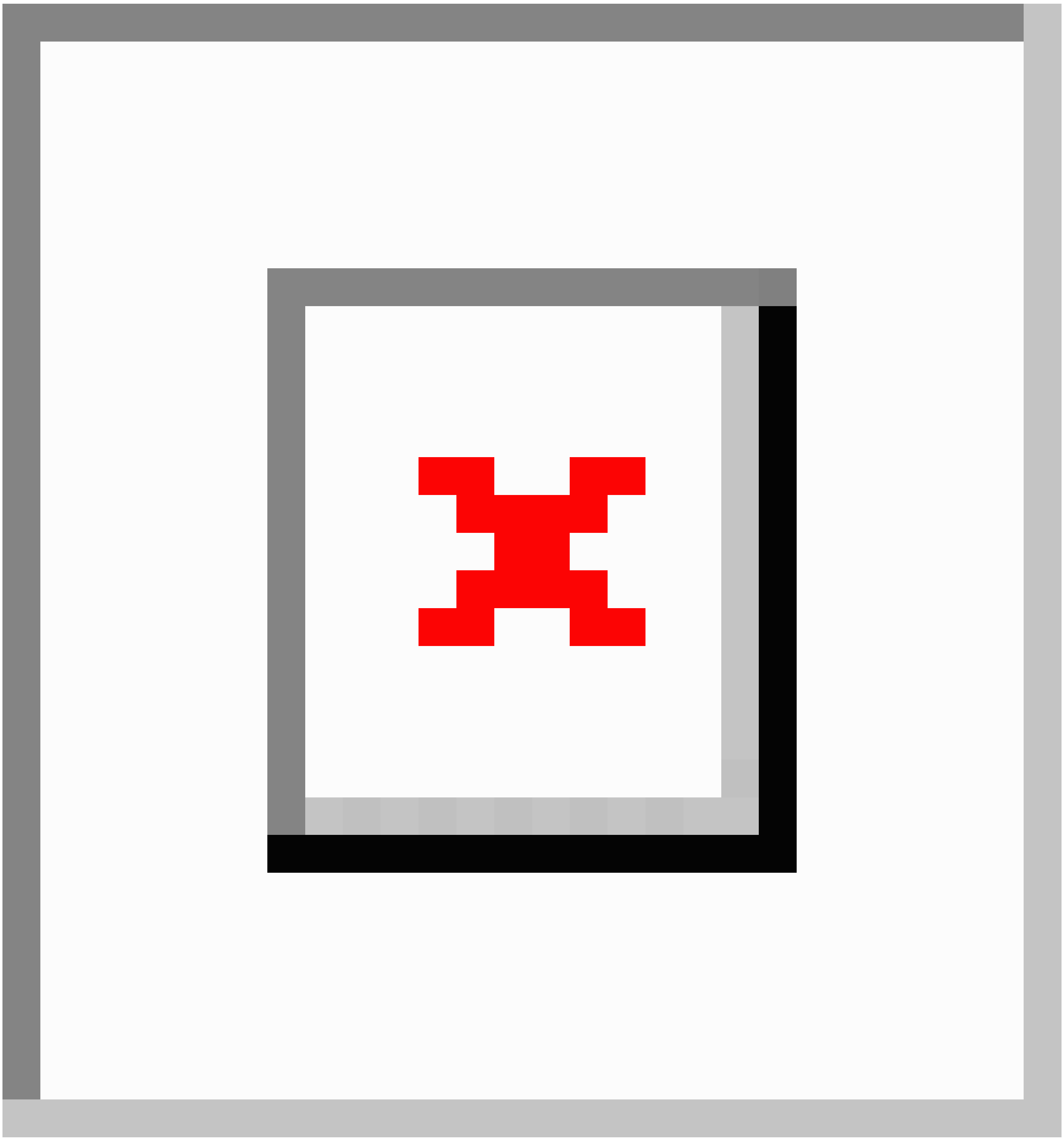


Figure 4. Authors' co-occurrence map.

References Analysis

In accordance with Table 5, there are 15 papers that serve as primary references in the research of AI and medical education. The paper titled "Medical Students' Attitude Towards Artificial Intelligence: A Multicenter Survey" emerged as the most frequently cited and most pertinent literature, garnering 36 and 109 citations, respectively. It primarily evaluates the attitudes of undergraduate medical students toward radiology and medical AI.

XSL•FO
RenderX

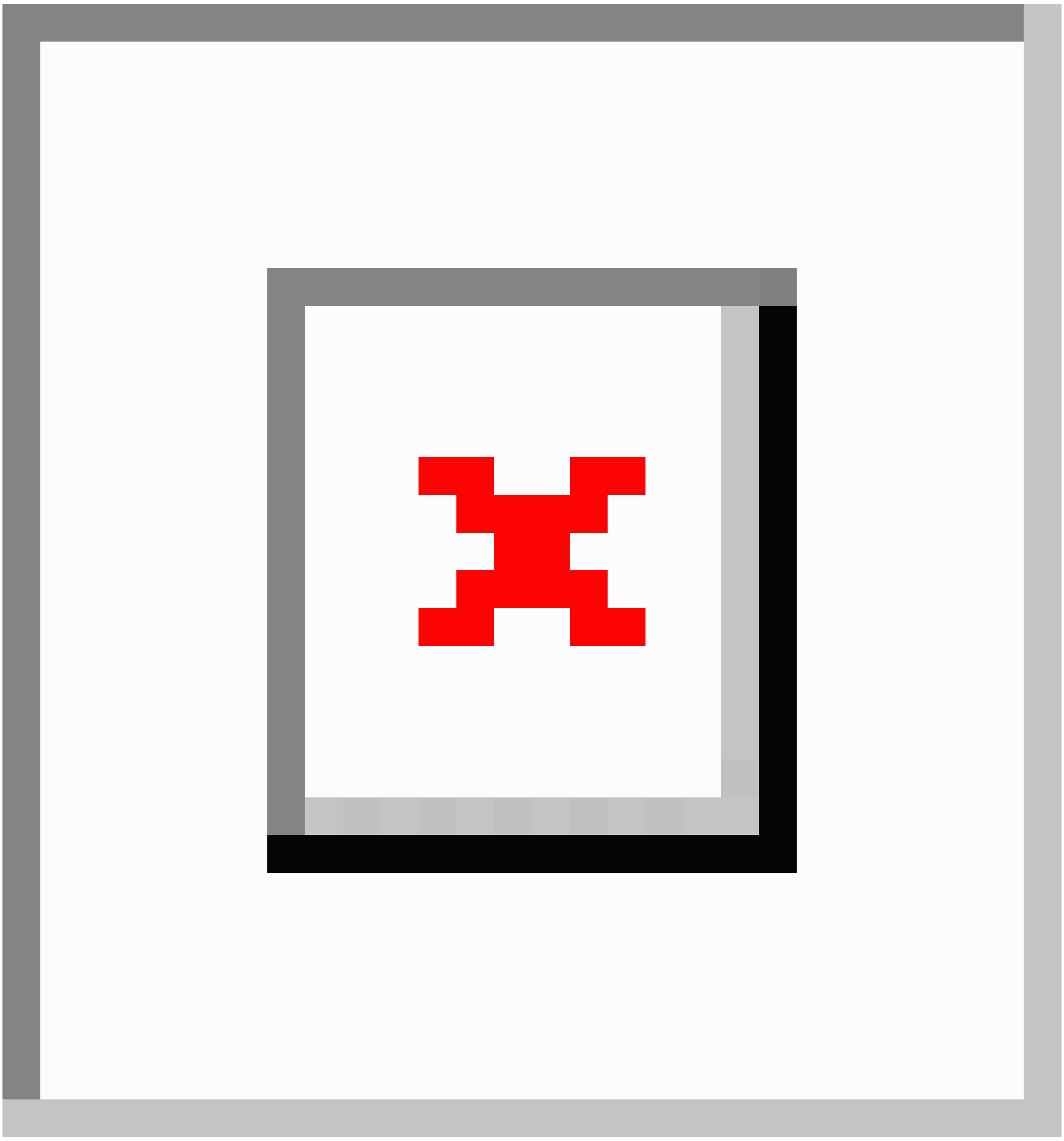
Table . Top 10 publications, centrality, and citations of cited reference.

Rank	Citations	Cited reference, year	Total link strength	Cited reference, year
1	36	Dos Santos et al [14], 2019	109	Dos Santos et al [14], 2019
2	23	Kolachalama and Garg [15], 2018	103	Wartman and Combs [16], 2018
3	23	Sit et al [17], 2020	98	Kolachalama and Garg, 2018 [15]
4	21	Gong et al [18], 2019	96	Sit et al [17], 2019
5	21	Wartman and Combs [16], 2018	85	Masters [19], 2019
6	19	Paranjape K et al [20], 2019	81	Paranjape K et al [20], 2019
7	19	Topol [21], 2019	78	Topol [21], 2019
8	16	Chan and Zary [8], 2019	78	Wartman and Combs [22], 2019
9	16	Masters [19], 2019	78	McCoy et al [23], 2020
10	15	Wartman and Combs [22], 2019	75	Park et al [24], 2019

The papers “Machine Learning and Medical Education” and “Attitudes and Perceptions of UK Medical Students Towards Artificial Intelligence and Radiology: A Multicenter Survey” are the second most frequently cited. The papers “Medical Education Must Move From the Information Age to the Age of Artificial Intelligence” and “Machine Learning and Medical Education” occupy the second position in terms of total link strength. Figure 5 illustrates this information.

XSL•FO
RenderX

Figure 5. Cited reference co-occurrence map.



Keywords Analysis

The study examining AI and medical education from 2013 to 2022 concentrated on 39 primary keywords (Table 6). Figure 6 shows that AI (100), education (47), and medical education (45) have the highest frequency and connection intensity.

XSL•FO
RenderX

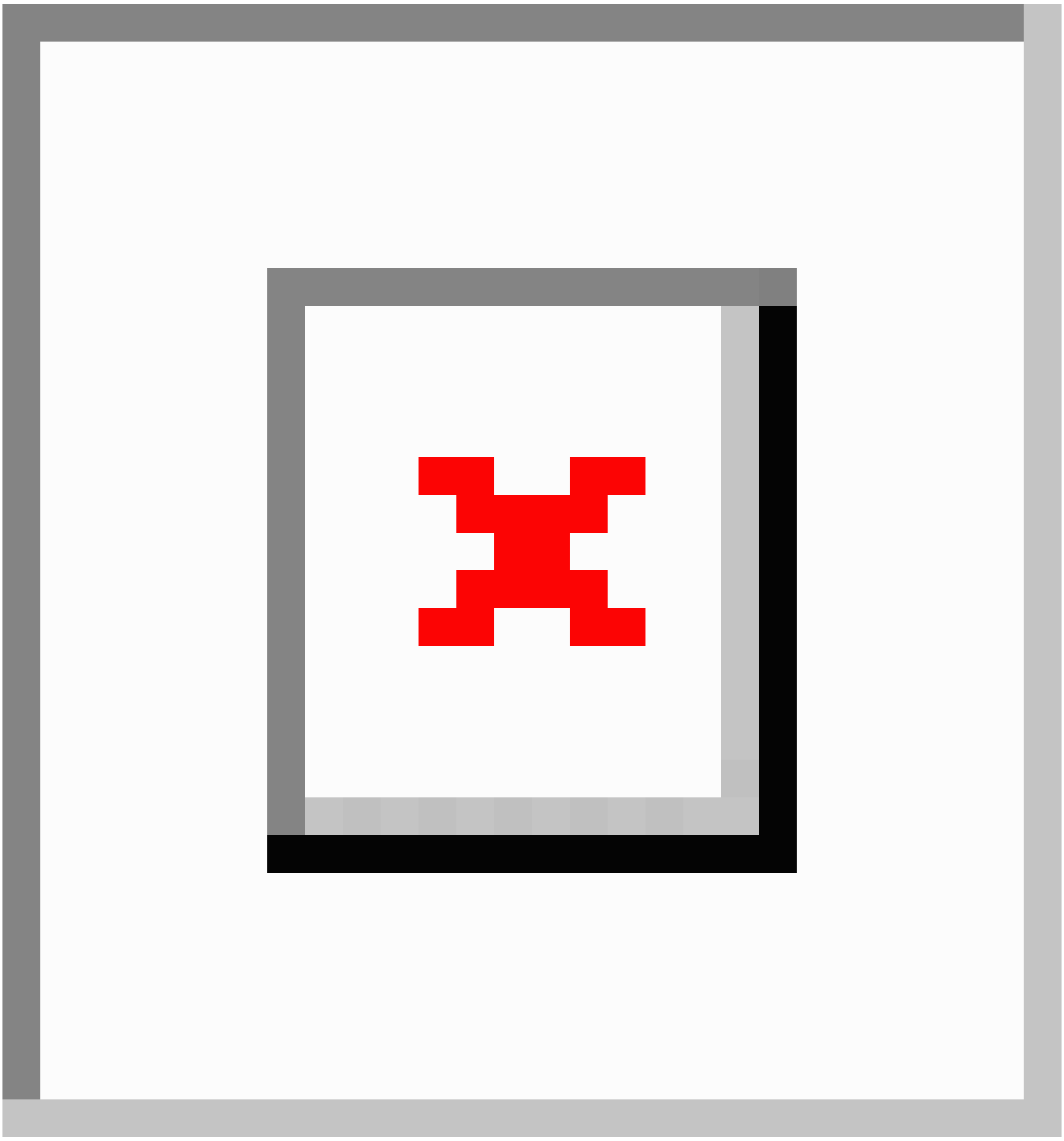
Table . Top 10 keywords related to AI in medical education.

Rank	Occurrence (%)	Keywords	Total link strength	Keywords
1	100	AI[a]	259	AI[a]
2	47	Education	131	Education
3	45	Medical education	114	Medical education
4	33	Machine learning	107	Machine learning
5	23	Technology	94	Technology
6	15	Radiology	56	Curriculum
7	14	Artificial intelligence	43	Radiology
8	13	Curriculum	43	Artificial-intelligence
9	12	Health	41	Performance
10	12	Medical students	38	Health

[a]AI: artificial intelligence.

XSL•FO
RenderX

Figure 6. Keywords co-occurrence map.

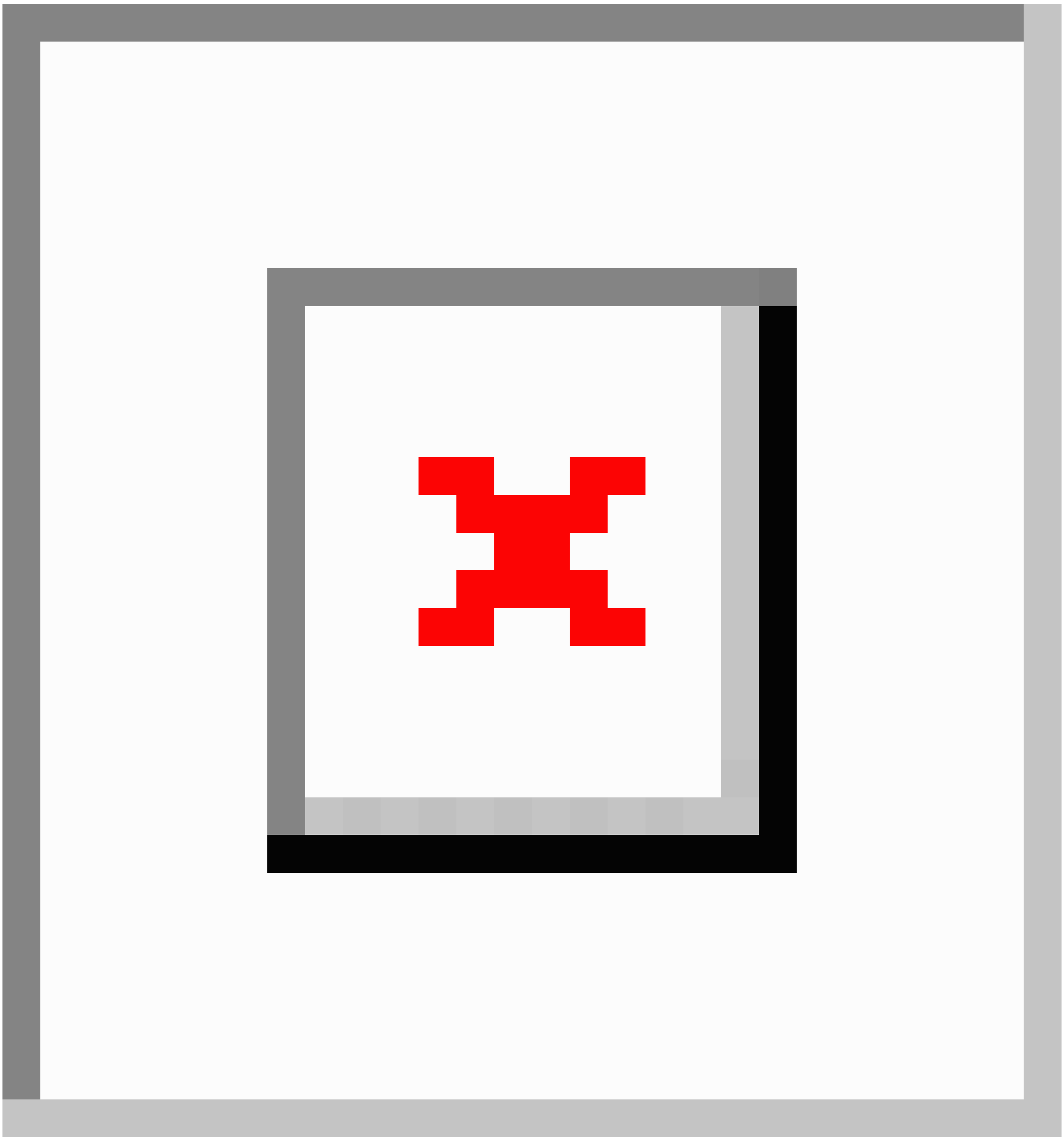


Research Status

Figure 7 shows that the analysis of references with high citation frequency and centrality enables us to understand highly respected research results in the application of AI in medical education.

XSL•FO
RenderX

Figure 7. Research status map.



In clusters 0 and 1, the swift advancement of AI has led to its application across all medical sectors, notably radiology [25-27]. Despite radiologists, residents, and medical students increasingly recognizing the importance of understanding AI, medical education that targets future radiologists is only just commencing [14,20,28]. Current investigations fall into 3 categories, that are (1) methods to facilitate medical students in learning AI knowledge, (2) using AI technology to augment radiology teaching efficiency and assist medical students in identifying clinical images, and (3) medical students' attitudes toward AI application in radiology. An AI curriculum (Artificial Intelligence in Radiology [AI-RADS]) has been devised to equip residents devoid of computing backgrounds with basic AI knowledge and its radiology application. The curriculum was highly rated (9.8 out of 10) by residents for overall satisfaction and significantly increased students' confidence in interpreting AI-related journal papers. There was a marked improvement in residents' comprehension of AI's fundamental concepts [29]. Some institutions emphasize integrating AI frameworks to strengthen radiology education. For example, after scanning, the patient's condition will be interpreted by artificial intelligence to give a preliminary diagnosis. AI assigns cases to interns whose personal profiles indicate that they will benefit the most. Interns cooperate with artificial intelligence and use equivalent tools for diagnosis. Interns and attending radiologists elaborate on the final report. AI uses natural language processing to anonymize new cases, add them to the teaching archive, and update the personal profiles of trainees after new cases are

XSL•FO
RenderX

completed. When trainees review cases similar to new cases, AI will provide them with corresponding cases from the teaching archive.[30]. As this framework continues to evolve, it may be possible to achieve "precise medical education" tailored to the individual learning styles and needs of the students [30]. A multicenter survey assessing UK medical students' attitudes and perceptions of AI and radiology revealed that students recognize the significance of AI and are eager to engage [17]. This prompts the need to integrate relevant AI courses into medical education to acquaint students with practical AI applications and constraints, thereby maintaining their learning enthusiasm and preventing AI-related panic.

Natural language processing is an important direction in the fields of computer science and AI. It studies various theories and methods that enable effective communication between humans and computers using natural language. Its main function here is to distinguish rare cases

In cluster 2, eHealth refers to the use of information and communication technologies to fulfill health care needs in various domains, including AI, telemedicine, Internet of Things, connected devices, and mobile health (mHealth) [31]. eHealth technologies provide access to health care in rural areas and support the management of numerous health conditions [32-36]. Following the release of the World Health Organization's national eHealth strategy tool in 2012, it is imperative for future medical students to receive eHealth education and training. Current medical education primarily includes conceptual courses while neglecting practical training [37]. While emphasizing the inclusion of eHealth in medical education, it is also important to recognize the potential adverse outcomes of over-reliance on AI technology [38]. Hence, identifying the optimal eHealth application areas in health care is necessary [39].

In cluster 3, the integration of medical education and AI holds significant value and potential beyond radiology, extending into surgical education and surgery. AI's earliest medical applications were in image-based specialties, such as radiology, pathology, ophthalmology, and dermatology. However, its application in procedural professions such as surgery may require more time [40,41]. The benefits of AI application in surgery mainly include integrating preoperative, intraoperative, and postoperative data to improve the accuracy of the clinical decision-making system and predict postoperative complications more efficiently and applying surgical knowledge and education to interact with surgeons and patients through virtual or augmented reality. For instance, virtual reality simulators were initially used in laparoscopic surgery training [42]. A study involving 176 medical students was conducted to assess the accuracy of robot-assisted virtual surgical simulations after integrated deep learning, showing improved accuracy [43]. In 2022 and 2023, AI application breakthroughs were achieved in oral and maxillofacial surgery education [44] and orthopedic surgery [45]. While AI proves beneficial in surgery and surgical education, especially in surgical ability assessment, it raises questions regarding whether AI can ever match the intelligence and audacity of the human educators. Although advanced AI teaching tools can be incorporated into surgical education, current technology cannot fully replace multifaceted surgeons or surgical educators. Addressing the transparency and responsibility of AI application in medical education and resolving ethical issues may require more time and effort.

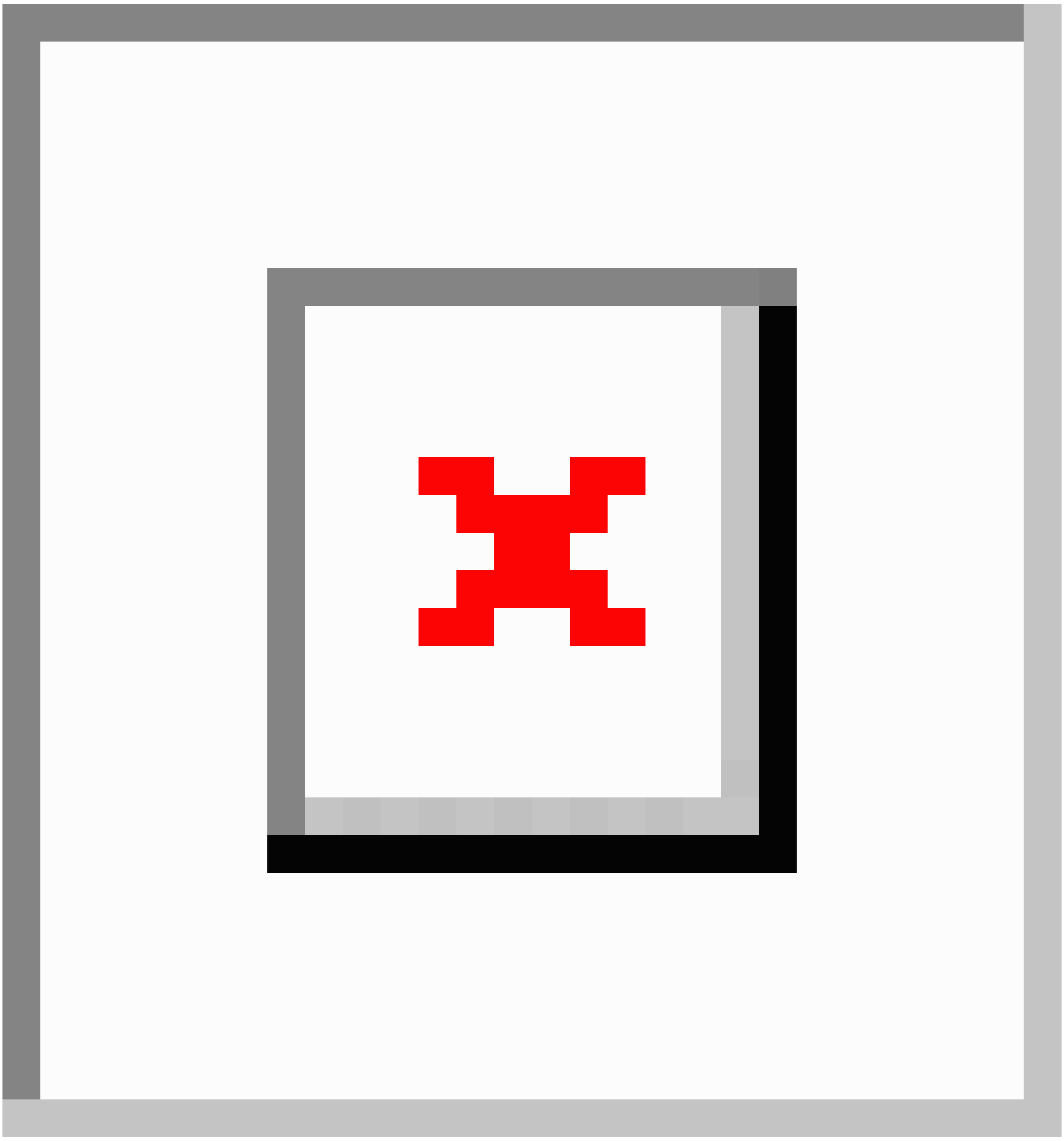
In cluster 5, the rapid AI development profoundly impacts medical education. Modern medical education must accommodate various health care systems, including digital health systems and big data generation in a highly connected world [46]. A Canadian survey of medical students' perceptions of AI's impact on radiology in 2018 showed that anxiety induced by the prospect of AI replacing radiologists deterred many students from considering radiology [18]. The radiology community should appreciate AI's potential impact on the profession, educate students appropriately about AI's role, and ensure radiology's viability as a long-term career option. While AI's benefits in medicine include eliminating human bias and enhancing pattern recognition and decision-making, its drawbacks, such as the inability to provide warmth and empathy to patients and absorb the wisdom of human educators, should not be underestimated. The confusion about whether AI's role in medical education is supplementary or replacement-based is another concern [47]. In summary, while AI promises great advances and changes in medicine, it also poses numerous challenges and problems. The medical community needs to proactively address these challenges, leverage AI technology benefits, and promote continuous innovation and improvement in medical services.

Research Frontier

Figure 8 shows that big data has a significant intensity of 2.01, firmly at the top of the list, and has become the focus of medical education in the past 3 years. The emergence and proliferation of COVID-19 in 2019 ushered in the big data epoch in medicine, with telemedicine systems, clinical intelligent decision-making, and management systems taking on pivotal roles.

XSL•FO
RenderX

Figure 8. Top 20 keywords with strongest citation bursts.



First, the advent of big data has catalyzed the innovation of medical teaching paradigms: what does the future hold for medical education in the digital age? A study conducted by Han et al zeroes in on a future medical education model that leans heavily on big data, cutting-edge technology, and AI, with the aim to cultivate a new breed of medical students who display enhanced humanistic attributes, co-operation capacity, patient-needs sensitivity, and societal and global orientation [46].

Second, big data has stimulated innovation in clinical medicine models: the integration of advanced technologies like machine learning, clinical intelligent decision and management systems, and electronic medical records has propelled shifts, innovation, and advancement within clinical medicine paradigms. The study by Kolachalama and Garg posits that AI, fueled by machine learning algorithms, is an emerging computer science branch that is swiftly gaining traction in health care. AI is anticipated to play an instrumental role in precision medicine and health [15]. In 2022, Watson and Wilkinson released a paper entitled “Digital Healthcare in COPD Management: A Narrative Review on the Advantages, Pitfalls, and Need for Further Research,” illustrating the vast potential of digital health care innovation [48]. During the COVID-19 pandemic, it was expected that big data would mitigate the workload for doctors interpreting digital data, enhance their diagnostic and prognostic abilities, equip clinicians with intelligent decision-making and management systems, and offer patients optimal clinical care and self-management strategies.

XSL•FO
RenderX

Undeniably, big data, akin to many emergent tools, is a double-edged sword. Ensuring its tailored use and dialectical treatment constitutes a crucial aspect of digital health, striving to exploit its merits while circumventing its demerits. The pursuit of enduring, comprehensive, and precise population health data management emerges as a long-term strategy.

The recent surge in terms indicates that "management" is intimately linked to "big data." Confronting the colossal medical data of today, the incorporation of AI technology can enhance management efficiency in spheres, such as hospital medical management, disease surgery management, and chronic disease management, among others. AI algorithms are used to scrutinize data pertaining to patients' hospitalization duration, hospitalization route, and climatic and temporal factors, which effectively curtail the hospitalization duration and significantly rectify issues, such as the misallocation of medical resources [49]. Leveraging a diabetic retinopathy automatic grading and training system furnished with an AI-driven diagnosis algorithm to groom budding doctors can augment diagnostic accuracy, thereby strengthening DR management [50]. Surgical video, a crucial data source for medical education, should be systematically stored and managed. A system intended to assist doctors in managing surgical videos can heighten the efficiency of continuing education by dissecting surgical videos and marking critical segments or frames to generate AI reports [51].

Discussion

In this investigation, a bibliometric evaluation of 195 pertinent papers over the preceding decade was meticulously executed using CiteSpace and VOSviewer. This research illustrates the findings related to countries, institutions, authors, citations, and keywords using tables and diagrams, offering an analytical perspective on the current research status and emerging frontiers in this domain. The outcomes were exhaustively analyzed.

Initially, examining the annual publication count, authors, institutions, and countries, it was identified that from 2019 onwards, global interest and recognition of AI's applicability in medical education experienced an upswing. Second, superficially, collaboration in this arena might appear limited, an aspect that can be attributed to this field's unique nature and the diverse modalities of medical education across different regions. For future progress, it is recommended that countries focus on harmonizing their approaches while acknowledging their differences, fostering collective advancement, and advocating for a mutual elevation of medical education standards.

Furthermore, an evaluation of the current research status and prevalent research themes highlighted that the extent of AI technology integration in medical education is significantly inadequate, with a rather limited focus area. Consequently, it is advocated that future efforts should aim at active exploration to unearth novel advancements.

Finally, AI, being inherently enigmatic, evokes uncertainty among both educators and learners about its future potentialities. Therefore, the immediate concern should be to strategically leverage its potential while mitigating its drawbacks, which, indeed, becomes the highest priority for future advancement.

Some limitations should be considered. The search strategies used can potentially yield divergent results, and the strategy opted for in this study might not encompass all pertinent literature. With the swift advancement of AI, several papers in this domain were brought to light in 2023. However, the temporal span of this study extends from 2013 to 2022, thus excluding the contributions from 2023.

The study highlights the promising potential of AI in medical education research, emphasizing the need for enhanced interregional collaboration and improved research quality. These insights provide valuable guidance for future research directions.

Acknowledgments

The authors would like to give their heartfelt thanks to all the people who helped them with this paper. All authors are grateful for the support of all present and future participants or participants as well as institutions. This study is funded by Major Science and Technology Special Plan of Science and Technology Department of Yunnan Province (project number 202102AA100016), Yunnan Provincial Department of Science and Technology—Yunnan University of Chinese Medicine Joint Special Project of Applied Basic Research (project number 201901AI070004), Yunnan Provincial Department of Science and Technology—Yunnan University of Chinese Medicine Joint Special Project of Applied Basic Research (project number 202101AZ070001-059), and Key Laboratory of Acupuncture and Massage for Prevention and Treatment of Encephalopathy in Universities of Yunnan Province (project number 2019YGZ04). The funding agencies do not play any role in the design, collection, analysis, or writing manuscript.

Data Availability

The data sets generated or analyzed in this study will not be publicly available. Consent and ethical approval do not include a provision for the sharing of data from this study.

Authors' Contributions

XT and XZ were the main investigators, mainly responsible for the overall framework and design of the paper. SW contributed to data processing and mapping. LY and ML supervised article writing and table design. All authors participated in the revision and approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Dzobo K, Adotey S, Thomford NE, Dzobo W. Integrating artificial and human intelligence: a partnership for responsible innovation in biomedical engineering and medicine. OMICS 2020 May;24(5):247-263. [doi: 10.1089/omi.2019.0038] [Medline: 31313972]
2. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc 2020 Oct;92(4):807-812. [doi: 10.1016/j.gie.2020.06.040] [Medline: 32565184]
3. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol 2019 Mar 4;28(2):73-81. [doi: 10.1080/13645706.2019.1575882]
4. Patnaik PR. Synthesizing cellular intelligence and artificial intelligence for bioprocesses. Biotechnol Adv 2006 Mar;24(2):129-133. [doi: 10.1016/j.biotechadv.2005.08.002] [Medline: 16171965]
5. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 2021 Aug;25(3):1315-1360. [doi: 10.1007/s11030-021-10217-3] [Medline: 33844136]
6. Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract 2018 Mar;68(668):143-144. [doi: 10.3399/bjgp18X695213] [Medline: 29472224]
7. Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. Artif Intell Med 2009 May;46(1):5-17. [doi: 10.1016/j.artmed.2008.07.017] [Medline: 18790621]
8. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019 Jun 15;5(1):e13930. [doi: 10.2196/13930] [Medline: 31199295]
9. Qin Y, Zhang Q, Liu Y. Analysis of knowledge bases and research focuses of cerebral ischemia-reperfusion from the perspective of mapping knowledge domain. Brain Res Bull 2020 Mar;156:15-24. [doi: 10.1016/j.brainresbull.2019.12.004] [Medline: 31843561]
10. Stout NL, Alfano CM, Belter CW, et al. A bibliometric analysis of the landscape of cancer rehabilitation research (1992-2016). J Natl Cancer Inst 2018 Aug 1;110(8):815-824. [doi: 10.1093/jnci/djy108] [Medline: 29982543]
11. Akyol A, Kocyigit BF. Ankylosing spondylitis rehabilitation publications and the global productivity: a web of science-based bibliometric analysis (2000-2019). Rheumatol Int 2021 Nov;41(11):2007-2014. [doi: 10.1007/s00296-021-04836-0] [Medline: 33797569]
12. Chen YM, Wang XQ. Bibliometric analysis of exercise and neuropathic pain research. J Pain Res 2020 Jun;13:1533-1545. [doi: 10.2147/JPR.S258696] [Medline: 32612381]
13. Wang SQ, Wang JX, Zhang C, et al. What you should know about osteoarthritis rehabilitation: a bibliometric analysis of the 50 most-cited articles. Geriatr Orthop Surg Rehabil 2020 Nov;11:2151459320973196. [doi: 10.1177/2151459320973196] [Medline: 33240559]
14. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019 Apr;29(4):1640-1646. [doi: 10.1007/s00330-018-5601-1] [Medline: 29980928]
15. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digit Med 2018 Sep;1:54. [doi: 10.1038/s41746-018-0061-1] [Medline: 31304333]
16. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018 Aug;93(8):1107-1109. [doi: 10.1097/ACM.0000000000002044] [Medline: 29095704]
17. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights Imaging 2020 Feb 5;11(1):14. [doi: 10.1186/s13244-019-0830-7] [Medline: 32025951]
18. Gong B, Nugent JP, Guest W, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: a national survey study. Acad Radiol 2019 Apr;26(4):566-577. [doi: 10.1016/j.acra.2018.10.007] [Medline: 30424998]
19. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
20. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
21. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. N Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
22. Reimagining medical education in the age of AI. AMA J Ethics ;21(2):E146-E152. [doi: 10.1001/amajethics.2019.146]
23. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med 2020;3:86. [doi: 10.1038/s41746-020-0294-7] [Medline: 32577533]
24. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? J Educ Eval Health Prof 2019;16:18. [doi: 10.3352/jeehp.2019.16.18] [Medline: 31319450]

XSL•FO
RenderX

25. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510. [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
26. Goddard P, Leslie A, Jones A, Wakeley C, Kabala J. Error in radiology. Br J Radiol 2001 Oct;74(886):949-951. [doi: 10.1259/bjr.74.886.740949] [Medline: 11675313]
27. Boland GWL, Guimaraes AS, Mueller PR. The radiologist's conundrum: benefits and costs of increasing CT capacity and utilization. Eur Radiol 2009 Jan;19(1):9-11. [doi: 10.1007/s00330-008-1159-7] [Medline: 18766347]
28. Ooi SKG, Makmur A, Soon AYQ, et al. Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey. Singapore Med J 2021 Mar;62(3):126-134. [doi: 10.11622/smedj.2019141] [Medline: 31680181]
29. Lindqwister AL, Hassanpour S, Lewis PJ, Sin JM. AI-RADS: an artificial intelligence curriculum for residents. Acad Radiol 2021 Dec;28(12):1810-1816. [doi: 10.1016/j.acra.2020.09.017] [Medline: 33071185]
30. Duong MT, Rauschecker AM, Rudie JD, et al. Artificial intelligence for precision education in radiology. Br J Radiol 2019 Nov;92(1103):20190389. [doi: 10.1259/bjr.20190389] [Medline: 31322909]
31. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. Mhealth 2017;3:38. [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
32. Speyer R, Denman D, Wilkes-Gillan S, et al. Effects of telehealth by allied health professionals and nurses in rural and remote areas: a systematic review and meta-analysis. J Rehabil Med 2018 Feb 28;50(3):225-235. [doi: 10.2340/16501977-2297] [Medline: 29257195]
33. So CF, Chung JW. Telehealth for diabetes self-management in primary healthcare: a systematic review and meta-analysis. J Telemed Telecare 2018 Jun;24(5):356-364. [doi: 10.1177/1357633X17700552] [Medline: 28463033]
34. Xiao Q, Wang J, Chiang V, et al. Effectiveness of mHealth interventions for asthma self-management: a systematic review and meta-analysis. Stud Health Technol Inform 2018;250:144-145. [Medline: 29857410]
35. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis. Asian Pac J Cancer Prev 2018 Jul 27;19(7):1747-1752. [doi: 10.22034/APJCP.2018.19.7.1747] [Medline: 30049182]
36. Lee Y, Ragguett RM, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 2018 Dec 1;241:519-532. [doi: 10.1016/j.jad.2018.08.073] [Medline: 30153635]
37. Echelard JF, Méthot F, Nguyen HA, Pomey MP. Medical student training in eHealth: scoping review. JMIR Med Educ 2020 Sep 11;6(2):e20027. [doi: 10.2196/20027] [Medline: 32915154]
38. McDonald L, Ramagopalan SV, Cox AP, Oguz M. Unintended consequences of machine learning in medicine? F1000Res 2017 Sep;6:1707. [doi: 10.12688/f1000research.12693.1] [Medline: 29250316]
39. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. JAMA 2019 Jan 1;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
40. Kirubarajan A, Young D, Khan S, Crasto N, Sobel M, Sussman D. Artificial intelligence and surgical education: a systematic scoping review of interventions. J Surg Educ 2022 Mar;79(2):500-515. [doi: 10.1016/j.jsurg.2021.09.012] [Medline: 34756807]
41. Sheikh AY, Fann JI. Artificial intelligence: can information be transformed into intelligence in surgical education? Thorac Surg Clin 2019 Aug;29(3):339-350. [doi: 10.1016/j.thorsurg.2019.03.011] [Medline: 31235303]
42. Ritter EM, Park YS, Durning SJ, Tekian AS. The impact of simulation based training on the fundamentals of endoscopic surgery performance examination. Ann Surg 2023 Mar 1;277(3):e699-e706. [doi: 10.1097/SLA.0000000000005088] [Medline: 34310356]
43. Moglia A, Morelli L, D'Ischia R, et al. Ensemble deep learning for the prediction of proficiency at a virtual simulator for robot-assisted surgery. Surg Endosc 2022 Sep;36(9):6473-6479. [doi: 10.1007/s00464-021-08999-6] [Medline: 35020053]
44. Krishnan DG. Artificial intelligence in oral and maxillofacial surgery education. Oral Maxillofac Surg Clin North Am 2022 Nov;34(4):585-591. [doi: 10.1016/j.coms.2022.03.006] [Medline: 36224076]
45. St Mart JP, Goh EL, Liew I, Shah Z, Sinha J. Artificial intelligence in orthopaedics surgery: transforming technological innovation in patient care and surgical training. Postgrad Med J 2023 Jun 30;99(1173):687-694. [doi: 10.1136/postgradmedj-2022-141596] [Medline: 37389584]
46. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec 11;19(1):460. [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
47. Grunhut J, Marques O, Wyatt ATM. Needs, challenges, and applications of artificial intelligence in medical education curriculum. JMIR Med Educ 2022 Jun 7;8(2):e35587. [doi: 10.2196/35587] [Medline: 35671077]
48. Watson A, Wilkinson TMA. Digital healthcare in COPD management: a narrative review on the advantages, pitfalls, and need for further research. Ther Adv Respir Dis 2022 Jan;16:17534666221075493. [doi: 10.1177/17534666221075493] [Medline: 35234090]
49. Nas S, Koyuncu M. Emergency department capacity planning: a recurrent neural network and simulation approach. Comput Math Methods Med 2019 Nov;2019:4359719. [doi: 10.1155/2019/4359719] [Medline: 31827585]

XSL•FO
RenderX

50. Qian X, Jingying H, Xian S, et al. The effectiveness of artificial intelligence-based automated grading and training system in education of manual detection of diabetic retinopathy. Front Public Health 2022 Nov;10:1025271. [doi: 10.3389/fpubh.2022.1025271] [Medline: 36419999]
51. Kim D, Hwang W, Bae J, Park H, Kim KG. Video archiving and communication system (VACS): a progressive approach, design, implementation, and benefits for surgical videos. Healthc Inform Res 2021 Apr;27(2):162-167. [doi: 10.4258/hir.2021.27.2.162] [Medline: 34015882]

Abbreviations

AI: artificial intelligence
AI-RADS: Artificial Intelligence in Radiology
mHealth: mobile health

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna; submitted 31.07.23; peer-reviewed by G Diedenhofen, S Pesälä; revised version received 21.02.24; accepted 30.04.24; published 10.10.24.

Please cite as:
Wang S, Yang L, Li M, Zhang X, Tai X
Medical Education and Artificial Intelligence: Web of Science–Based Bibliometric Analysis (2013-2022)
JMIR Med Educ 2024;10:e51411
URL: https://mededu.jmir.org/2024/1/e51411
doi:10.2196/51411

© Shuang Wang, Liuying Yang, Min Li, Xinghe Zhang, Xiantao Tai. Originally published in JMIR Medical Education (https://mededu.jmir.org), 10.10.2024. 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.

XSL•FO
RenderX

Challenges and Needs in Digital Health Practice and Nursing Education Curricula: Gap Analysis Study

Karen Livesay[1,*], PhD; Ruby Walter[1,*], PhD; Sacha Petersen[1,*], PhD; Robab Abdolkhani[1,*], PhD; Lin Zhao[1,*], PhD; Kerryn Butler-Henderson[1,2,*], PhD

1

2

*all authors contributed equally

Corresponding Author:
Karen Livesay, PhD

Abstract

Background: Australian nursing programs aim to introduce students to digital health requirements for practice. However, innovation in digital health is more dynamic than education providers' ability to respond. It is uncertain whether what is taught and demonstrated in nursing programs meets the needs and expectations of clinicians with regard to the capability of the nurse graduates.

Objective: This study aims to identify gaps in the National Nursing and Midwifery Digital Health Capability Framework , based on the perspectives of clinical nurses, and in nurse educators' confidence and knowledge to teach. The findings will direct a future co-design process.

Methods: This study triangulated the findings from 2 studies of the Digital Awareness in Simulated Health project and the National Nursing and Midwifery Digital Capability Framework. The first was a qualitative study that considered the experiences of nurses with digital health technologies during the COVID-19 pandemic, and the second was a survey of nurse educators who identified their confidence and knowledge to teach and demonstrate digital health concepts.

Results: The results were categorized by and presented from the perspectives of nurse clinicians, nurse graduates, and nurse educators. Findings were listed against each of the framework capabilities, and omissions from the framework were identified. A series of statements and questions were formulated from the gap analysis to direct a future co-design process with nursing stakeholders to develop a digital health capability curriculum for nurse educators.

Conclusions: Further work to evaluate nursing digital health opportunities for nurse educators is indicated by the gaps identified in this study.

(JMIR Med Educ 2024;10:e54105) doi:10.2196/54105

KEYWORDS

nursing; digital health; capability; workforce; framework; nursing education; education; digital health practice; clinicians; nurse; nurse graduates; clinical nurses; nurses; nurse educators; teach; teaching; learning; nursing students; student; students

Introduction

It is widely recognized that digital health technologies have advanced at a rate greater than education about digital health [1]. Indeed, digital health has barely been established in the nursing curriculum, let alone evaluated to match what is needed in the clinical setting [2]. This is a somewhat unusual phenomenon wherein the application of a practice has happened in advance of the evidence that supports it. Digital technologies are constantly evolving to improve access efficiency, safety, and communication, and in turn, the scope of nursing informatics is rapidly evolving at the intersection of health care and information technology [3]. The adoption and optimization of electronic health records (EHRs) continue to be a major focus in nursing informatics, with a growing knowledge base on the importance of user-centered design to improve the usability and functionality of EHRs, ultimately leading to better care and safety [4]. The COVID-19 pandemic accelerated the use of technologies such as EHRs, which led to reported burnout in the nursing profession [5]. Further during the pandemic, there was an increase in the adoption of telehealth technologies, with nurse clinicians playing a crucial role in facilitating telemedicine care, including remote monitoring and consultation [6]. The integration of telehealth into nursing practice has raised the importance of developing telehealth capabilities in nursing graduates and clinicians [7]. However, within health care, digital technologies have been adopted at a pace faster than they can be taught. As a result, digital health technologies are firmly embedded in clinical practice (eg, the electronic medical record, telehealth, and remote monitoring) while they are rarely used

XSL•FO
RenderX

or taught in the nursing curriculum [8]. There is consensus [9] that digital health technologies will improve the safety, efficiency, and quality of health care when implemented appropriately. It is essential that nursing graduates are not only skilled in the safe use of these technologies but also aware of the professional, ethical, and potential benefits and risks of these technologies.

The Australian Nursing and Midwifery Accreditation Council Registered Nurse Accreditation standards [10] state that digital health in the curriculum should be informed by the domains of the National Nursing and Midwifery Digital Health Capability Framework [11], at the level of implementation the extent to which these domains are applied is difficult to assess. In order to determine what education is needed at an undergraduate level to provide adequate digital awareness, knowledge, and skills, first it must be understood what is known, what is taught, and what health care expects of graduate nurses.

The Digital Awareness in Simulated Health (DASH) project has, through several successive and iterative phases, examined the current digital health education needs in nursing training in order to address knowledge and skill needs for nurses in the clinical arena. The project aimed to uncover and quantify what is needed in nursing entry to practice degrees regarding digital health through the different lenses of nursing, nurse educators, nurse clinicians, and nurse graduates. The project has been designed based on the Learning Health System model and includes 3 main cyclical phases: practice to data, data to knowledge, and knowledge to practice. The three phases of the project were (1) a systematic review [12]; (2) an interview study of nurses in clinical practice about their digital health use, digital health application, and related challenges during COVID-19, as well as the needs of the graduate nurse [13]; and (3) a survey of nurse educators in Australian universities regarding their knowledge and confidence in teaching aspects of digital health to entry to practice nursing students [14]. The aim of this paper is to present an analysis of the gaps identified through the triangulation of this past research. The process of identifying the gaps was important to assist in understanding ways to develop, plan, and implement educational strategies to bridge the divide and meet the digital health needs within clinical nursing practice [15].

Methods

Overview

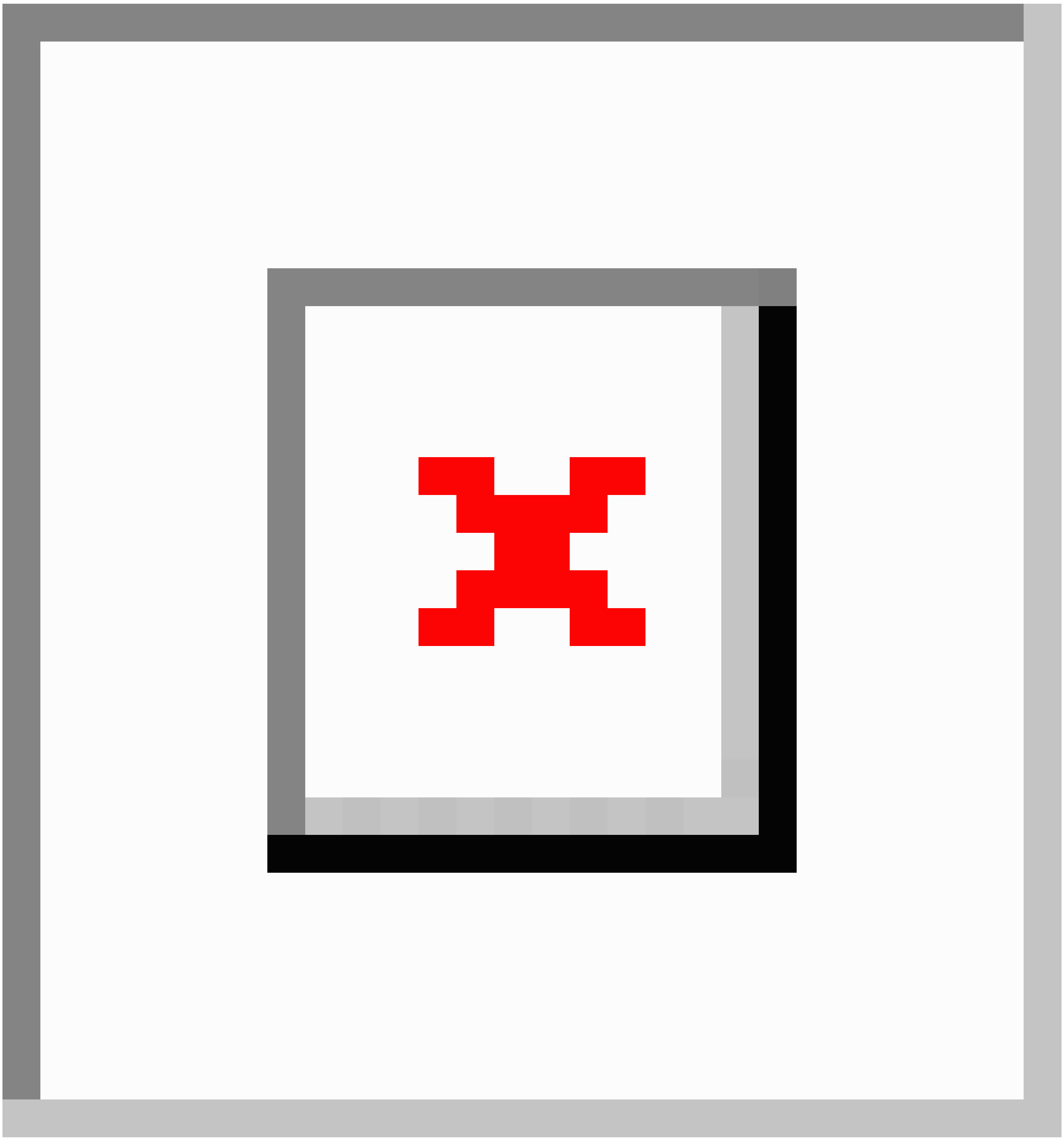
This gap analysis triangulated the findings from 2 studies of the DASH project. The source of the data is described in each of the papers previously cited in this paper. Although processes for gap analysis can be found in the literature, they represent a wide context of settings and study types that did not directly translate to the anticipated needs of this project. In fact, the learning needs analysis literature provided more direction in developing a process to identify the gaps [16]. In particular, the learning needs process paid attention to the range of contexts of study participants, focused on knowledge skills and attitudes, considered capabilities, and provided an opportunity to reflect on resources [15]. All these elements were important to understand in determining the requirements for nurses entering practice and the capabilities of nurse educators to teach and provide practice opportunities to learners. In addition to the findings from the interviews and survey, the results were triangulated against the National Nursing and Midwifery Digital Capability Framework [11]. (This framework represents the work of key stakeholders in digital health within Australia and aims to define the digital health knowledge skills and attitudes required for nursing and midwifery practice while providing a basis against which to tailor learning.

The gap analysis was undertaken by the DASH project team. The research team consisted of 4 nurse educators, a digital health academic, and a postdoctoral research fellow. The team collectively reviewed the data from both the interview and survey studies to discern what gaps existed between the interview data, the survey data, and the framework. The review resulted in several questions that formed the basis of the methods, as follows: (1) How knowledgeable and confident are nurse educators to provide the education expected by the framework? (2) What are the barriers encountered by nurse clinicians reflected in the expectations of the framework? (3) What are the nurse clinicians' expectations of nurse graduates' digital health capability reflected in the expectations of the framework? (4) What do nurse clinicians identify as challenges that are identified in the framework, but nurse educators lack confidence in teaching? (5) What do nurse clinicians identify as challenges that are missing from the framework? (6) What do nurse educators identify they do not have confidence or knowledge in teaching, but nurse clinicians do not identify it as an adoption challenge and therefore does it need to be in the framework? (7) What do nurse clinicians identify as graduate competencies that are not in the framework? (8) What do nurse clinicians identify as graduate competencies that are in the framework, but educators lack the confidence and knowledge in teaching?

These questions were mapped in Figure 1 to show how the gaps were identified through the triangulation process, whereby the number represents each of the above questions, and those in green examined items present in the framework as opposed to those in orange that identified items that were missing from the framework.

XSL•FO
RenderX

Figure 1. Model of gap analysis.



The procedure to answer each question involved several steps. The first step was to identify the contexts and tools of the nurses who participated in the interview study. There were a range of contexts including nurses in clinical settings, chief nursing information officers, clinical educators, nurse representatives in digital health vendor companies, and nurse representatives in government. The types of digital tools participants had identified using during the COVID-19 pandemic were also identified. The second step involved mapping the barriers to digital health use encountered by those stakeholders during the COVID-19 pandemic to the framework (questions 1, 2, and 3). The survey was devised using concepts from both the framework and the TIGER (Technology Informatics Guiding Education Reform) core nursing informatics competency framework [17], with the method described in a paper by Zhao et al [14]. Survey items that respondents (n=119) had reported "no or minimal knowledge and confidence" were categorized as high where more than 70% of responses chose no or minimum knowledge, medium where 31% - 69% of responses chose no or minimum knowledge, and low where less than 30% of responses indicated no or minimum knowledge. To identify only those items that relate to the Australian framework and remove those that only relate to the TIGER framework, these high and medium items were mapped against the items identified by the nurse educator that appeared in the framework (question 1). The items identified by nurse clinicians as challenges to adoption (question 2) and requirements of the nurse graduate (question 3) in the interviews were mapped to items in the framework. Last, the items

XSL•FO
RenderX

identified as challenges to adoption identified by nurse clinicians were mapped against those items that nurse educators identified as lacking the knowledge or confidence to teach (question 4), which resulted in a list of items that need to be taught, but nurse educators lack the knowledge or confidence to teach.

Once these items were mapped to each other and the framework (questions 1-4), the gaps could be identified. This included the items the nurse clinicians identified as challenges that were missing from the framework (question 5); the items nurse educators identified they do not have confidence or knowledge in teaching, but nurse clinicians did not identify as an adoption challenge and raise whether they need to be in the framework (question 6); and the items nurse clinicians identified as nurse graduate competencies that are not in the framework (question 7).

The last part of the triangulation was to identify the items that nurse graduates require to work in health, as identified by nurse clinicians; map them to the framework; and then map them against items that the nurse educators lacked the confidence and knowledge to teach, as identified by the nurse educators themselves (question 8). This resulted in a list of items that nurse graduates need to know, but nurse educators lack the knowledge or confidence to teach. The steps of the procedure linked to the question numbers are summarized in Table 1.

Table . Procedure for each of the method questions.

Question number	Procedures
Questions 1, 2, and 3	Mapped the qualitative nurse clinician data to the capability statements.
Question 4	Mapped items from the capability statements linked to the nurse clinician's qualitative data to the nurse educator survey items that scored M[a] or H[b].
Question 5	Identified the qualitative nurse clinician data that did not have a competency statement assigned.
Question 6	Compared the H and M items from the nurse educators to the qualitative nurse clinician data where there is no corresponding item in the framework.
Question 7	Mapped the qualitative nurse clinician data about expectations for nurse graduates and identified omissions in the framework.
Question 8	Identified the qualitative nurse clinician data about nurse graduates that mapped to the framework and the nurse educator data that scored an H or an M.

[a]M: medium.

[b]H: high.

Ethical Considerations

Institutional ethics approval (via the Low Risk Committee; reference number: 2022-25054-16817) was provided for each phase of the studies cited. The studies undertaken used informed consent and the ability for participants to opt out in accordance with the ethical standards of the Low Risk Committee and with the Helsinki Declaration.

Results

The results of the gap analysis were analyzed and presented through the 3 lenses: nurse educator, nurse clinician, and nurse graduate.

Nurse Educator

Several items were identified by nurse educators that they lacked confidence and knowledge to teach and were related to only the Australian framework (question 1), as summarized in Table 1. No items relating to the domain of "Technology" were identified. The numbers in brackets show the capability statements in the framework the items were mapped against, and the number of participants that reported the lack of knowledge or confidence are reported as medium (M) or high (H), as defined in the *Methods* section.

Nurse Clinician

There was a strong alignment between the challenges identified by the nurse clinician and the framework across all the domains (question 2; Table 2). Several areas were identified as challenges by nurse clinicians that were not apparent in the framework (question 5). These were classified as either resource or nursing informatics specialist items, as summarized in the last row of Table 2.

There were several areas that nurse clinicians had identified as challenges that were aligned with the framework and nurse educators identified they have a lack of confidence to teach (question 4; summarized in Table 3). When examined at a domain level, every domain except "Technology" contained competency statements that were identified by the nurse clinicians and nurse educators. However, not all statements in these other 4 domains were identified by the nurse clinician and the nurse educators, indicating not all statements are an identified challenge or they are a challenge, but the nurse educators are confident to teach (Table 4). Areas that nurse educators identified they do not have confidence in but were not identified as a challenge by nurse clinicians (question 6; Table 3) were limited and did not include the domains of "Leadership and advocacy" or "Data and information quality." The "Technology" domain featured the most, however, items related to information systems (eg, radiology, pharmacy, and

XSL•FO
RenderX

laboratory) were identified by nurse educators but not by nurse clinicians.

Table . Alignment of items nurse educators report they lack confidence and knowledge to teach and the framework.

National Nursing and Midwifery Digital Capability Framework domain	Items nurse educators report they lack confidence and knowledge (numbers in brackets represent the capability statement number in the framework; M[a]=31% - 69%, H[b]≥70%)
Digital professionalism	• Prescribing and referral rights as it relates to electronic identity (M) (1.2) • Cybersecurity and risk management (M) (1.3) • Cultural and socioeconomic factors in digital health (M) (1.2) • Dynamic consent (H) (1.2) • Digital identity (M) (1.2)
Leadership and advocacy	• Digital health governance (M) (2.2, 2.3) • Patient digital health advocacy (M) (2.1)
Data and information quality	• Data, information, and knowledge management (M) (3.2) • Processes for reporting quality and safety issues (M) (3.2) • Data capture (M) (3.1) • Errors in data entry (M) (3.1) • Smart phrases (H) (3.1) • Smart links (H) (3.1) • Quality management (M) (3.2)
Information-enabled care	• Information management in clinical research (M) (4.2)

[a]M: medium.
[b]H: high.

XSL•FO
RenderX

Table . Alignment of challenges identified by nurse clinicians and the framework.

National Nursing and Midwifery Digital Capability Framework domain	Challenges identified by nurse clinicians
Digital professionalism	• Fear and demotivation in interacting and using a new technology due to lack of preparedness (1.2) • Lack of digital health literacy in the senior nursing workforce (1.1, 1.2) • Lack of consistent and continuous formal training (1.1, 1.2, 1.3) • Lack of time for appropriate training (1.1, 1.2) • New technologies led to the emergence of new roles for nurses that required new skill set (1.1, 1.2, 1.3)
Leadership and advocacy	• Lack of nurses' involvement in critical decision making in digital health implementation (2.2) • Lack of effective communication among nurses and other stakeholders in using digital health (2.2, 2.3) • Lack of communication between managers and ward nurses to understand nurse-specific needs in using digital health (2.2) • Current legislations are not applicable nationwide (2.1, 2.3) • Lack of legislation to support data transfer between primary and acute care settings (2.1, 2.3) • Lack of involvement of external experts in using digital health technologies (2.3)
Data and information quality	• Heavy load of digital documentation and nurse shortage to do that (3.2) • Lack of access to and use of PROMS[a] to improve remote management (3.2) • The user interface was challenging for immediate clinical actions (3.1) • Difficulties in data collection from siloed technologies that are not integrated into the EMRs[b] (3.1, 3.2) • Lack of nurse evaluation of the implemented digital health services (3.3) • Lack of feedback and measurements of nurse performance in the digital health systems (3.2, 4.2)
Information-enabled care	• Lack of feedback and measurements of nurse performance in the digital health systems (3.2, 4.2)
Technology	• Difficulty in communication between nurses and patients in using mobile apps (5.1) • Interaction with various screens in telehealth consultations is overwhelming (5.1) • Challenges in using interpreters in telemedicine appointments (5.1) • Lack of strategies on how to improve access to telemedicine care by culturally and linguistically diverse background communities (5.2) • Lack of organizational approach to identify the practice problems that can be solved by a particular technology (5.1, 5.3) • Inability of digital health systems to store and analyze a large volume of collected data (5.1) • Inability to troubleshoot devices (5.3) • Difficulties in reporting errors (5.3)
Items not aligned with framework	• Lack of chief nursing informatics officer roles • Lack of use of informatics workforce in technology implementations • Lack of economists' perspectives on digital health business models • Lack of time to manage the digital content for quality assurance. • Lack of funding for continuous evaluation • Lack of workforce to know and conduct the evaluation. • Interruptions in nurses' workflows due to lack of computers at bedsides • Lack of internet connectivity in distant areas • Interoperability challenges among various devices • Difficulties in the infrastructure network • More cumbersome training in settings that were new to digital health

XSL•FO
RenderX

[a]PROM: Patient reported outcome measure

[b]EMR: Electronic medical record

Table . Comparison of challenges identified by nurse clinicians that align with the framework and areas that nurse educators lack knowledge or confidence to teach.

National Nursing and Midwifery Digital Capability Framework domain	Challenges that nurse educators lack the confidence to teach (numbers in brackets represent the capability statement number in the framework; M[a]=31% - 69%, H[b]≥70%)	Areas nurse educators lack the confidence to teach but were not identified as a challenge by nurse clinicians (numbers in brackets represent the capability statement number in the framework; M=31% - 69%, H≥70%)
Digital professionalism	• Prescribing and referral rights as it relates to electronic identity (M) (1.2) • Cultural and socioeconomic factors in digital health (M) (1.2) • Dynamic consent (H) (1.2) • Digital identity (M) (1.2)	• Cybersecurity and risk management (M) (1.3)
Leadership and advocacy	• Digital health governance (M) (2.2, 2.3) • Patient digital health advocacy (M) (2.1)	—[c]
Data and information quality	• Data, information, and knowledge management (M) (3.2) • Processes for reporting quality and safety issues (M) (3.2) • Data capture (M) (3.1) • Errors in data entry (M) (3.1) • Smart phrases (H) (3.1) • Smart links (H) (3.1) • Quality management (M) (3.2)	—
Information-enabled care	• Information management in clinical research (M) (4.2)	• Big data analytics (H) (4.2)
Technology	—	• Interoperability (H) (5.1) • Troubleshooting (M) (5.3) • Clinical decision support systems (M) (5.1) • Robotic surgeries (H) (5.1) • Blockchain networks (H) (5.1)

[a]M: medium.

[b]H: high.

[c]Not applicable.

Nurse Graduate

There were a significant number of items the nurse clinicians identified the nurse graduate needs to know that align with the framework (question 3), summarized in . No items were identified in the domains of “Data and information quality” and “Technology.” More importantly, when all the results were triangulated, it identified the items the nurse clinician identified the nurse graduate needs to know, yet the nurse educators lack the knowledge or confidence to teach (question 8; Table 5). Only 1 item was identified—“Students should learn about rules and regulations of data security, privacy, and social media in using digital health (1.2, 1.3).” Clinicians wanted students and graduates to learn about rules and regulations of data security, privacy, and social media in using digital health. This corresponded to the survey item cybersecurity and risk management, which scored 41% for knowledge and 35% for confidence to teach by nurse educators.

XSL•FO
RenderX

Table . Comparison of items that were identified by nurse clinicians that nurse graduates need to know, that aligned with the framework, and that nurse educators lack knowledge or confidence to teach.

National Nursing and Midwifery Digital Capability Framework domain	Nurse clinician expectations of nurse graduates (numbers in brackets represent the capability statement number in the framework; M[a]=31% - 69%, H[b]≥70%)
Digital professionalism	• Students should learn about rules and regulations of data security, privacy, and social media in using digital health (1.2, 1.3)[c] • Students should be taught about nursing's digital health capabilities before coming into practice (1.2, 4.3) • The use of academic EMR[d] should be a requirement in nursing programs (1.1) • Nursing students should learn about digital health systems in more detail than only data entry. For example, about data exchange, security, and analytics (1.1) • Students should be taught about real-world digital health challenges in nursing in addition to the theoretical concepts (1.1) • Universities should foster digital health training to be responsive to the new generation of technologies (1.1)
Leadership and advocacy	• There is a need for investment in the digitally enabled nursing workforce, as they are the only providers in the remote areas of Australia (2.1, 2.2)
Information-enabled care	• The concept of a multidisciplinary approach should be embedded in digital health training for nursing students. They need to learn how to interact with internal and external stakeholders (4.1, 4.2) • Universities can embed training content about analytics to foster critical thinking and curiosity among nurses about digital health technologies (4.1, 4.2) • Students should be taught about nursing's digital health capabilities before coming into practice. (1.2, 4.3)

[a]M: medium.

[b]H: high.

[c]Identified as a need by nurse clinicians but nurse educators do not have the confidence or knowledge to teach.

[d]EMR: electronic health record.

Discussion

Principal Findings

The triangulation of results from previous studies examined the items identified by nurse clinicians as challenges in digital health adoption in practice, capability needs of nurse graduates, and items nurse educators lack confidence or knowledge to teach (Tables 2-5). These were then mapped to the Australian National Nursing and Midwifery Digital Health Capability Framework [11], and several gaps were identified. These gaps were the main findings of this paper.

A significant difference between the framework and the other reference point, the TIGER study [17] for the development of survey items for nurse educators, is the extent of detail outlining elements within the competency domains in the latter. The international recommendations of the TIGER study identify capabilities associated with recognized nursing informatics roles, which the Australian framework does not identify. The framework with 5 domains and limited detail is open to interpretation as to which domain an area of digital knowledge would best be mapped to. The nurse researchers in the study required advice from digital health academics on the team to clarify the best fit when undertaking this mapping, indicating nurse educators may not be able to implement the framework within their own curriculum without digital health expertise guidance.

The alignment between the challenges identified by the nurse clinician and the framework (Table 4) confirmed the necessity of those items in the framework. Further to the criticism listed above about the framework being open to interpretation is that users of the framework are unable to gauge the depth of knowledge required to achieve capability, as these will vary depending on the statement and the functions of the Clinical Nurse. These items, identified by nurse clinicians (Table 3), can guide users of the framework on the depth of knowledge required for nurse graduates and nurse clinicians for these particular items. The small number of items identified by both the nurse educator and the nurse clinicians, and in particular the lack of items related to the domain "Technology" (Table 3, last row), should not be assumed to mean the other statements are not relevant to practice and should be removed from the framework. Instead, this finding can guide the depth of knowledge required for the nurse graduate. Some areas related to technology are subject to their own programs of graduate study, for example, cybersecurity, and nurses entering practice in a field like this would require a high level of knowledge. The identification of different information systems by nurse

XSL•FO
RenderX

educators (Table 4), which were not identified as a challenge by nurse clinicians, may be related to the vendor nature of these information systems. Training is generally provided by the vendor for tendered information system adoption, whereas noninformation system capabilities that a graduate nurse needs to have to be able to work safely and responsibly with technology or data needs to be developed either through graduate training or on the job.

Nurse clinician expectations of nurse graduates did not map to the framework, with the exception of 1 item (Table 5). The significance of this is speculative but likely relates to the questions used in the interview study [13]. As no reference was made to the framework, the responses were not provided that specifically addressed it. Additionally, the questions were broad, and the answers provided also lacked detail. In general, nurse clinician responses generally focussed on global capabilities, such as using an electronic medical record, rather than detailed responses. Further, the framework did not identify several areas identified by nurse clinicians related to resources (Table 3). These resources could be categorized as either physical or human. For example, clinicians spoke about specialist digital or informatics roles that provide support for digital health initiatives, which is not addressed in the framework. Additionally, nurse clinicians spoke about being able to access the correct tools appropriate for the task, including having access to secure internet services. While Standard 5.1 covers recommending appropriate digital technologies and staff and consumers being able to use these where available, the framework does not address a requirement for availability. The growth of digital capacity among nurses in practice will continue to be hampered by under-resourcing. Nurse clinicians called out expertise and human support by other nurses as a barrier in practice. These items cannot be addressed educationally and therefore were not included in the survey of nurse educators (Table 2). Nonetheless, the silence in the framework may result in a missed opportunity to recognize and support specialist nurse practice in digital health. Booth et al [8] suggest support from all levels of nurse leadership to invest in resources and champion and support nurses in their practice and research. The complexity of health environments and the rapid rate of change and development in health care, inevitably result in nurses with a wide range of knowledge and confidence in digital technology use and complex differences in demands and access to digital tools [8,18].

There are several implications for both education and practice. While this gap analysis has identified there was only 1 item that the nurse clinicians identified is required for the nurse graduate but the nurse educator lacked the knowledge or confidence to teach ("Students should learn about rules and regulations of data security, privacy, and social media in using digital health"), there are several items across the domains that nurse educators reported they will be unable to teach but are important skills or knowledge for practice. Given the nurse researchers in this project needed to consult with the digital health experts on the team to interpret elements of the framework highlighted this challenge. It implies that digital health expertise may be required for graduate training providers to meet the Australian Nursing and Midwifery Accreditation Council Registered Nurse Accreditation standards requirements related to digital health. The practicality is most graduate training providers do not have this resource available and it was not the intention of the framework. Nurse educators need to be upskilled in digital health, so they have the confidence and knowledge to design and deliver the necessary digital health curriculum.

The nurse clinicians identified practice barriers relevant to their own context during the COVID-19 pandemic. Nazeha et al [18] recommend that frameworks for digital health be updated regularly in line with innovation. However, the applicability of digital tools and technologies is never going to be universal or static and will always be context-specific. Therefore, considering those items that nurse clinicians did not identify as barriers and nurse educators lacked confidence or knowledge to teach, may be as simple as being out of context of experience for those practitioners and educators. For example, robotic surgery is a very defined digital technology, whereas EHRs are a generic concept. Educators may have an oversight or awareness of a specific digital tool but have answered negatively in the survey as their knowledge is global rather than specific and would not be sufficient to teach or demonstrate to learners.

The items that educators lacked confidence or knowledge to teach should be examined for their value in a crowded, busy curriculum. Nurses, as lifelong learners, continue to develop, and many enter and exit the education system more than once across their careers [19]. Postgraduate study requirements may include specialist knowledge not apparent for entry to practice minimum standards. Risling [20] predicted exponential increases in technology use in the coming decade and warned that nurse educators need to lead the evolution in practice and education. Changes to curriculum, while challenging, will need to be carefully considered for their worth at the same time as recognizing the rapidity of change likely to be required.

The findings from the interview and survey studies informing this analysis may have been influenced by the mixed progress of digital health roll out nationally. A co-design process to develop an educational intervention will be undertaken as the next stage of this DASH research project. The following questions and statements will form the basis for discussion with a broader panel to validate the findings of this analysis and develop strategies to overcome the challenges and weaknesses in clinical practice and educational delivery settings:

- Digital health expertise and guidance are required for nurse educators to develop curricula in digital health.
- The framework requires augmentation to describe the depth of knowledge and experience.
- Should specific technologies be inherent within the framework?
- How should differences in experience, exposure, and resources related to digital health be addressed in nursing education?
- Which of the items that educators lacked knowledge or confidence to teach should be addressed in the curriculum for nurse educators teaching entry-to-practice nursing courses?

XSL•FO
RenderX

Limitations

In the interview study, we did not ask what skills students were noted to have in digital health, rather, the approach was aspirational for what clinicians desired. The extent to which those aspirations for students are achieved is unknown. It is suggested that a further study investigating the actual capability of nurse graduates in digital health be undertaken.

Conclusion

This analysis took the outputs of 2 studies investigating the digital health perspectives of nurses in practice environments and their expectations of graduates' digital health capabilities, with the second paper investigating the capabilities of nurse educators to teach and practice digital health, and mapped these findings with the National Nursing and Midwifery Digital Capability Framework. The outcomes of this analysis will inform a co-design process to create a curriculum for nurse educators to uplift capability in teaching and simulation for entry-to-practice nursing programs in Australia. A series of 8 questions directed the process to triangulate the findings and identify which factors were and were not included in the framework. A series of statements and questions were then formed from the analysis as recommendations to direct the co-design phase of this national research project.

Acknowledgments

This study is part of a larger project that was funded by the Victorian Higher Education State Investment Fund in 2021.

Conflicts of Interest

None declared.

References

1. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967. [doi: 10.2196/12967] [Medline: 31456579]
2. Troncoso EL, Breads J. Best of both worlds: digital health and nursing together for healthier communities. Int Nurs Rev 2021 Dec;68(4):504-511. [doi: 10.1111/inr.12685] [Medline: 34133028]
3. Kaihlanen AM, Elovainio M, Virtanen L, et al. Nursing informatics competence profiles and perceptions of health information system usefulness among registered nurses: a latent profile analysis. J Adv Nurs 2023 Oct;79(10):4022-4033. [doi: 10.1111/jan.15718] [Medline: 37243421]
4. Desai AV, Michael CL, Kuperman GJ, et al. A novel patient values tab for the electronic health record: a user-centered design approach. J Med Internet Res 2021 Feb 17;23(2):e21615. [doi: 10.2196/21615] [Medline: 33595448]
5. Melnick ER, West CP, Nath B, et al. The association between perceived electronic health record usability and professional burnout among US nurses. J Am Med Inform Assoc 2021 Jul 30;28(8):1632-1641. [doi: 10.1093/jamia/ocab059] [Medline: 33871018]
6. Cipriano PF, Boston-Leary K, Mcmillan K, Peterson C. The US COVID-19 crises: facts, science and solidarity. Int Nurs Rev 2020 Dec;67(4):437-444. [doi: 10.1111/inr.12646] [Medline: 33428227]
7. Rutledge CM, Gustin T. Preparing nurses for roles in telehealth: now is the time!. Online J Issues Nurs 2021 Jan;26(1). [doi: 10.3912/OJIN.Vol26No01Man03]
8. Booth RG, Strudwick G, McBride S, O'Connor S, Solano López AL. How the nursing profession should adapt for a digital future. BMJ 2021 Jun;373:1190. [doi: 10.1136/bmj.n1190]
9. National digital health workforce and education roadmap V1.0. Australian Digital Health Agency. 2020. URL: https://www.digitalhealth.gov.au/sites/default/files/2020-11/Workforce_and_Education-Roadmap.pdf [accessed 2024-08-22]
10. Registered nurse accreditation standards 2019. Australian Nursing & Midwifery Accreditation Council. 2019. URL: https://anmac.org.au/sites/default/files/2024-06/registerednurseaccreditationstandards2019_0.pdf [accessed 2024-08-22]
11. Williamson L, Dobroff N, Jones A, et al. National Nursing and Midwifery Digital Health Capability Framework. Australian Digital Health Agency. 2020. URL: https://www.digitalhealth.gov.au/sites/default/files/2020-11/National_Nursing_and_Midwifery_Digital_Health_Capability_Framework_publication.pdf [accessed 2024-08-22]
12. Abdolkhani R, Petersen S, Walter R, Zhao L, Butler-Henderson K, Livesay K. The impact of digital health transformation driven by COVID-19 on nursing practice: systematic literature review. JMIR Nurs 2022 Aug 30;5(1):e40348. [doi: 10.2196/40348] [Medline: 35867838]
13. Livesay K, Petersen S, Walter R, Zhao L, Butler-Henderson K, Abdolkhani R. Sociotechnical challenges of digital health in nursing practice during the COVID-19 pandemic: national study. JMIR Nurs 2023 Aug 16;6:e46819. [doi: 10.2196/46819] [Medline: 37585256]
14. Zhao L, Abdolkhani R, Walter R, Petersen S, Butler-Henderson K, Livesay K. National survey on understanding nursing academics' perspectives on digital health education. J Adv Nurs 2024 Apr 1. [doi: 10.1111/jan.16163] [Medline: 38558473]
15. Hudson A, Ellis-Cohen E, Davies S, et al. The value of a learning needs analysis to establish educational priorities in a new clinical workforce. Nurse Educ Pract 2018 Mar;29:82-88. [doi: 10.1016/j.nepr.2017.11.016] [Medline: 29220645]

XSL•FO
RenderX

16. Conducting a learning needs analysis (LNA) in high-risk industries. Instructional Design Australia. 2018. URL: https://instructionaldesign.com.au/learning-needs-analysis/ [accessed 2024-08-22]
17. Hübner U, Shaw T, Thye J, et al. An international recommendation framework of core competencies in health informatics for nurses. Methods Inf Med 2018 Jun;57(S 01):e30-e42. [doi: 10.3414/ME17-01-0155] [Medline: 29956297]
18. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 5;22(11):e22706. [doi: 10.2196/22706] [Medline: 33151152]
19. Qalehsari MQ, Khaghanizadeh M, Ebadi A. Lifelong learning strategies in nursing: a systematic review. Electron Physician 2017 Oct;9(10):5541-5550. [doi: 10.19082/5541] [Medline: 29238496]
20. Risling T. Educating the nurses of 2025: technology trends of the next decade. Nurse Educ Pract 2017 Jan;22:89-92. [doi: 10.1016/j.nepr.2016.12.007] [Medline: 28049072]

Abbreviations

DASH: Digital Awareness in Simulated Health
EHR: electronic health record
TIGER: Technology Informatics Guiding Education Reform

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna; submitted 30.10.23; peer-reviewed by FA Dhabbari, M Kleib, S Fenton; revised version received 15.02.24; accepted 13.05.24; published 13.09.24.

Please cite as:
Livesay K, Walter R, Petersen S, Abdolkhani R, Zhao L, Butler-Henderson K
Challenges and Needs in Digital Health Practice and Nursing Education Curricula: Gap Analysis Study
JMIR Med Educ 2024;10:e54105
URL: https://mededu.jmir.org/2024/1/e54105
doi:10.2196/54105

© Karen Livesay, Ruby Walter, Sacha Petersen, Robab Abdolkhani, Lin Zhao, Kerryn Butler-Henderson. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.9.2024. 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.

XSL•FO
RenderX

An Approach to the Design and Development of an Accredited Continuing Professional Development e-Learning Module on Virtual Care

Vernon Curran, PhD; Robert Glynn, MEd; Cindy Whitton, MEd; Ann Hollett, MA

Faculty of Medicine, Memorial University of Newfoundland, St John's, NL, Canada

Corresponding Author:
Vernon Curran, PhD

Abstract

Virtual care appointments expanded rapidly during COVID-19 out of necessity and to enable access and continuity of care for many patients. While previous work has explored health care providers' experiences with telehealth usage on small-scale projects, the broad-level adoption of virtual care during the pandemic has expounded opportunities for a better understanding of how to enhance the integration of telehealth as a regular mode of health care services delivery. Training and education for health care providers on the effective use of virtual care technologies are factors that can help facilitate improved adoption and use. We describe our approach to designing and developing an accredited continuing professional development (CPD) program using e-learning technologies to foster better knowledge and comfort among health care providers with the use of virtual care technologies. First, we discuss our approach to undertaking a systematic needs assessment study using a survey questionnaire of providers, key informant interviews, and a patient focus group. Next, we describe our steps in consulting with key stakeholder groups in the health system and arranging committees to inform the design of the program and address accreditation requirements. The instructional design features and aspects of the e-learning module are then described in depth, and our plan for evaluating the program is shared as well. As a CPD modality, e-learning offers the opportunity to enhance access to timely continuing professional education for health care providers who may be geographically dispersed across rural and remote communities.

(JMIR Med Educ 2024;10:e52906) doi:10.2196/52906

KEYWORDS

virtual care; continuing professional development; needs assessment; remote care; medical education; continuing medical education; CME; CPD; PD; professional development; integration; implementation; training; eHealth; e-health; telehealth; telemedicine; ICT; information and communication technology; provider; providers; healthcare professional; healthcare professionals; accreditation; instructional; teaching; module; modules; e-learning; eLearning; online learning; distance learning

Introduction

Most provincial health care systems across Canada responded to the COVID-19 pandemic with a rapid adoption of digital tools and technologies, including virtual care appointments. The Canadian Institute for Health Information (CIHI) reported that between March to September 2020, the percentage of patients availing virtual care services increased from 6% to 56% [1]. Virtual care refers to the delivery of health care services digitally or at a distance using information and communications technology [2-4]. During COVID-19, a variety of virtual care types were used, with synchronous and asynchronous appointments being the most common [3,4]. Synchronous virtual care refers to communication between the health care provider and patient that occurs in real time and can include the use of telephone or videoconferencing. Asynchronous communication does not occur live and may include the use of e-mail messaging, messages left for patients in a portal site, and e-consultations [3-5]. Considering the goal of reducing COVID-19 exposure during the recent global pandemic, virtual care proved to be most effective in that it minimized exposure and risk to health care providers by reducing the movement of people [2-7]. In addition, virtual care helped patients stay home who may have otherwise traveled to a health care site and incurred the risk of unnecessary exposure [3,4,6,7]. Virtual care was also used for control and triage during COVID-19, remote monitoring of patients, treatment and management, and provision of online health services [3,4].

In Canada, considerable work during and after the COVID-19 pandemic was undertaken to develop guidelines to inform physicians, health care providers, and patients on how they could best use virtual care. The Canadian Medical Association (CMA) and Royal College of Physicians and Surgeons of Canada (RCPSC) developed resources for both physicians and health care providers, as well as patients. The "Virtual Care Playbook" provided virtual care guidance for providers and connected patients to the "Virtual Care Guidelines for Patients" [8,9]. Canada Health Infoway has also undertaken significant work regarding virtual care support for physicians and health care providers. In particular, Infoway's "Clinician Change

XSL•FO
RenderX

Management" project provided support in the form of virtual care tools and training [10]. The Canadian Medical Protective Association (CMPA) has also supported providers by disseminating virtual care informational resources for physicians and health care providers through their website [11].

With the rapid introduction of virtual care across many jurisdictions during the COVID-19 pandemic, both health care providers and patients alike were not always adequately trained on how to use virtual care appropriately. Previous research has suggested that a lack of training around virtual care tools and software was a challenge for providers. A lack of understanding and training may have contributed to lower confidence levels among providers and a reluctance to use virtual care, thereby negatively influencing virtual care adoption [3,4,12-15]. Adjusting clinical approaches to caring for patients remotely can also be challenging, including how to virtually examine patients by videoconferencing or telephony systems [3,4,16,17]. The use of new digital health systems like virtual care also requires knowledge and competence in how to incorporate the technology within a provider's practice workflow. This includes understanding how to use the technology effectively, as well as the privacy and security issues surrounding the use of the technology. Providers also need to be able to adapt their techniques and clinical acumen to build rapport with their patients while using virtual care technologies. Given this, consideration of the potential continuing professional development (CPD) needs of health care providers is critical to ensuring that proper support systems and training are available to enable and empower providers to adopt and use virtual care most effectively and efficiently.

e-Learning has been defined as any educational intervention mediated electronically via the internet [18] and has become a popular modality for providing CPD in the health professions, with offerings across a diverse array of topics and subject areas [18,19]. The advantages and benefits of e-learning have been described as including lower costs, widespread distribution, increased accessibility to information, frequent content updates, and personalized instruction in terms of content and pace of learning [18]. Several systematic reviews of e-learning effectiveness in health professions' education, including CPD, have been published. Key findings of these reviews suggest that e-learning is associated with large positive effects when compared with no interventions [20]; e-learning can be as effective as traditional methods of teaching and instruction [20-22]; e-learning and traditional educational interventions take similar time to participate in or complete [23]; and interactivity, practice exercises, repetition, and feedback are important design features of effective e-learning approaches and appear to be associated with improved learning outcomes [23].

CPD encompasses the multiple educational and developmental activities pursued by health care providers to maintain and enhance their knowledge, skills, performance, and relationships in the provision of health care [4,24]. For many regulated health care providers around the world, CPD participation is often mandated and required throughout the extensive postlicensure phase of the provider's career. It is viewed as a key means for providers to stay current and up to date with evidence-based practices in their professional field. The evidence for CPD participation suggests that health care providers who participate in formal CPD activities are more likely to provide better care than their peers who do not participate [4,25]. CPD, which is designed to be interactive, practice-based, and longitudinal in nature, is also believed to yield better outcomes [4,25]. A needs assessment–driven approach to the development of CPD is more likely to lead to a change in practice, largely as a result of the learning being directly linked to personal and practice needs [4,26].

Edirippulige and Armfield [4,27] reviewed a number of studies describing the delivery and evaluation of education and training in telehealth. They identified 9 peer-reviewed studies describing education and training in telehealth that included several CPD-level courses on telehealth. Online learning was the most common delivery format described across the studies, with course duration ranging from 1 week to 6 months [4]. More recently, several studies conducted during the COVID-19 pandemic have reported on CPD in virtual care also using online delivery formats [28-31]. Both synchronous and asynchronous modalities were used in providing CPD on virtual care or telehealth; however, the most common delivery format was the use of web conferencing (eg, Zoom and Skype). Topics covered across these programs included introduction to virtual care, advantages and disadvantages of virtual care, types of virtual care, ensuring privacy during appointments, and legal and technological requirements for virtual care. One interesting method described by Hayden et al [30] was the use of web conferencing to facilitate simulated telehealth appointments with standardized patients. Participants found the use of standardized patients to simulate a virtual care appointment enhanced their confidence in focused telehealth skills. The use of online learning formats was perceived favorably by participants across the studies and was found to be particularly useful in accommodating the busy schedules of providers [28].

The purpose of this paper is to describe our efforts to design and develop an accredited e-learning CPD module on virtual care for physicians and health care providers. First, we discuss our approach to systematically exploring the needs of health care providers in learning to use virtual care effectively and efficiently in their practices. We describe results from a survey questionnaire we administered to a sample of health care providers in Newfoundland and Labrador, Canada, findings from key informant interviews with several experts in virtual care, and key themes emerging from a focus group with patient representatives. We then describe our approach to designing and developing this e-learning module, including key interactivity and design features to foster effective learning. Finally, we describe our approach and plan to evaluate the effectiveness and impact of this e-learning module on health care providers' adoption and use of virtual care. The work described in the paper was undertaken by our team with the Office of Professional & Educational Development (OPED), Faculty of Medicine, Memorial University of Newfoundland. The Faculty of Medicine at Memorial University has long been a pioneer in research and development in the fields of telemedicine, tele-education, and digital learning for physicians and rural health care providers. Our Professional Development

XSL•FO
RenderX

office was one of the first CPD units in North America to introduce accredited e-learning programming for physicians through our MDcme [32] learning management platform [33].

Methods

Needs Assessment Study

We undertook a needs assessment study initially as a first phase of our project to design and develop an e-learning CPD module on virtual care. The needs assessment encompassed a web-based survey, key informant interviews with experts in virtual care, and a focus group with patient representatives [3,4,34]. The goal of the web-based survey was to explore the experiences, perceptions, and satisfaction of health care providers with the adoption and use of virtual care during COVID-19. We developed and distributed this survey to physicians, nurses, and allied health professionals across the province of Newfoundland and Labrador, Canada, to explore their CPD needs and preferences as well.

In total, 51% of respondents (n=432) in our survey indicated they were currently offering virtual care and a majority (68.9%) reported it had improved their work experience [3]. The telephone was the most used method and respondents reported the most comfort and satisfaction with telephone appointments [3]. The most challenging aspects of telephone appointments were the inability to conduct physical exams to the degree required, the inability to assess physical health status, and the patient's or client's cell phone service being unreliable [3]. Respondents rated the importance of a variety of CPD topics on effective use of virtual care, and the highest rated topics included compliance with regulatory standards or rules for virtual care, understanding boundaries (eg, personal telephone numbers used to call patients or clients), and developing and maintaining competency and professionalism while engaging in virtual care [3]. Other important topics for virtual care CPD included CPD on how to use the technology, the best or easiest platforms for providing virtual care and how to use them effectively, and assessment skills and aids for doing assessments virtually. Ethical issues and legalities of virtual care were also identified by respondents as valuable as well [3].

The second component of our needs assessment was a qualitative study to explore experts' ascribed opinions on health care providers' CPD needs in virtual care [4]. We conducted semistructured interviews with a purposive sample of key informants representing Canadian provincial and national organizations with expertise in virtual care delivery. According to the key informant respondents, lack of training specific to virtual care tools and software was a challenge for health care providers, particularly videoconferencing appointments. All key informants identified technology as a main barrier or challenge, not only for health care providers but also for administrative staff. The main areas of knowledge, skills, and abilities deemed most important for health care providers in adopting and using virtual care identified by the key informants included effective use of technology, knowledge of how to integrate technology and virtual care in the practice workflow, privacy and security aspects of the technology, and adaptation of examination skills to virtual care and how to build effective rapport with patients [4].

A focus group study was also conducted with a purposive sample of patient representatives to explore patients' experiences and perspectives on the adoption and use of virtual care during COVID-19, and identify the education and informational needs of patients [34]. The findings from the patient focus group were useful in informing the types of topics to include in CPD on virtual care. Patient respondents felt that virtual care was beneficial and enabled greater convenience, flexibility, and access to health care services. Key barriers and challenges in adopting and using virtual care appeared to primarily arise from patients' lack of knowledge, understanding, and familiarity with it. Cost, technological access, connectivity, and low digital literacy were challenges for some patients, particularly in rural communities and among older patients. Patient education and support were critical and needed to be inclusive, easy to understand, and include information regarding privacy, security, consent, and the technology itself.

Approach to Mainpro+ and MainCert Accreditations for CPD Credit

The OPED, Faculty of Medicine at Memorial University is an accredited provider of CPD that targets the needs and competency development of health care providers within Newfoundland and Labrador and beyond. OPED is an accredited provider of university CPD by the Committee on Accreditation of Continuing Medical Education (CACME) and the Association of Faculties of Medicine of Canada (AFMC). As an accredited CPD provider, OPED is permitted to accredit CPD activities that meet the administrative, educational, and ethical standards of the College of Family Physicians of Canada (CFPC) Mainpro+ Certification program [35] and the RCPSC Maintenance of Certification program [36]. Key requirements for accrediting CPD activities include a needs assessment and the formation of a scientific planning committee (SPC) to oversee and advise on the development of the accredited CPD activity. An SPC is a group of target audience representatives responsible for identifying the educational needs of the intended target audience; developing educational objectives; selecting educational methods; selecting speakers, moderators, facilitators, and/or authors; developing and delivering content; and evaluating the outcomes of an accredited CPD activity. Requirements for accredited e-learning activities also include a means for participants to interact with the material, with each other, and with faculty members or a facilitator and the ability for participants to track their progress, provide evaluation feedback, register, and receive a record of registration. Such programs must also be offered within a definitive period of time communicated before the start of the program.

e-Learning Module Design

We ensured our e-learning module met the requirements of Newfoundland and Labrador's primary health care providers by establishing 2 guiding committees during its design and development. The first, an advisory committee, ensured alignment with policy and practices within the provincial health care system. This committee included representation from the Newfoundland and Labrador Centre for Health Information,

XSL•FO
RenderX

the provincial government's department of health, Memorial University's Faculty of Medicine, the College of Physicians and Surgeons of Newfoundland and Labrador, the College of Registered Nurses of Newfoundland and Labrador, and the Newfoundland and Labrador Medical Association. The second committee structure, an SPC, oversaw the design and development of the module and was responsible for ensuring that the learning experience reflected the needs of the primary health care providers [35]. This committee included a family physician, registered nurse, nurse practitioner, specialist physician, and emergency medicine physician.

We provided the advisory committee with the information collected through our needs assessment process and asked the members to offer feedback in terms of system-level needs. We then engaged with the SPC to review the needs assessment findings and advisory committee feedback and to develop a set of learning objectives we would use to guide the development of the module. Next, we engaged with several subject matter experts in virtual care to draft instructional materials and activities that would enable us to meet our stated objectives. This instructional material was used to develop a prototype of the e-learning module that we shared with the SPC and advisory committee for review and feedback. We compiled the feedback received and adjusted the prototype accordingly. We then proceeded to launch the prototype on the MDcme learning platform, a proprietary learning management system developed by OPED to house our accredited CPD activities. The MDcme environment provides user registration, asynchronous communications, technical support, and transcript or certificate issuance. The module is developed as a series of web pages using PHP (Hypertext Preprocessor) scripting and leverages responsive design to adapt its presentation based on the device used to access. The module will undergo an annual review process during which the assessment and evaluation data are reviewed, and any requisite modifications will be made, including updates and modifications to content and approach.

Van Hecke et al [37] developed the Criteria for Reporting on Development and Evaluation of Professional Training interventions in Healthcare (CRe-DEPTH) as a way to systematically report on the development and evaluation of training interventions for health care professionals. These criteria consist of 12 items representing 4 categories, which are the development of the training, characteristics of the training, characteristics of the providers, and assessment of the training outcomes. The following description of the e-learning module on virtual care outlines aspects of the development and evaluation of this educational program according to these criteria.

In developing the e-learning module, we followed the phases of the ADDIE model of instructional design. The ADDIE model is a systematic instructional design framework widely used in the creation and development of educational and training programs. The acronym "ADDIE" stands for 5 sequential key stages in the instructional design process, which are Analysis, Design, Development, Implementation, and Evaluation [38]. While sometimes criticized as being too linear in its approach, we have found that this framework delivers a consistent approach to educational development and aligns well with the requirements of the Mainpro+ and Maintenance of Certification accreditation programs. Gagne's "Nine Events of Instruction" model was also followed as an overarching approach in the development of content for the module [39].

We adopted an asynchronous e-learning design for this module. The asynchronous model assumes that learners taking the program will access the content at different times and from different locations. This approach allows primary care providers across the province to access the instructional material at their convenience, thereby providing the flexibility needed to balance professional learning with varied work hours, family, and other personal or professional commitments [40-42]. Results from our needs assessment survey of potential participants indicated that a large proportion of survey respondents preferred "E-Modules (self-paced/online learning)" as the delivery format [3].

Learners access the module by creating an account on the MDcme platform. The e-learning module provides a 90-minute introduction to the delivery of virtual care in a primary care setting and addresses the learning objectives, which are: (1) describe the benefits and key considerations of conducting virtual care appointments; (2) identify the technological requirements and setup required to conduct optimal virtual care; (3) recognize how to integrate virtual care delivery into your existing practice workflows; (4) discuss the clinical implications for delivering optimal virtual care encounters; (5) explain how to prepare patients for virtual care sessions; and (6) summarize the key regulatory and legal considerations in providing virtual care in Newfoundland and Labrador.

Our experience has been that a 60- to 90-minute duration for online CPD modules is an appropriate length to increase completion rates and reduce participant attrition.

The module is organized into three primary sections: (1) virtual care technologies, (2) the incorporation of virtual care into one's practice, and (3) the regulatory landscape. While the content is structured in a sequential fashion for learners to progress through, a comprehensive course menu tree is available, enabling learners to access any section of the module whenever they wish (Figure 1).

The module design uses several strategies to enhance learner engagement and support multimodal learning [43-45]. First, a variety of media are used in the presentation of module content, including text, images or graphics, and short video clips (Figure 2). Second, user interface design elements such as clickable tabs and dropdowns or flyouts are used where appropriate to encourage the learner to physically interact with the module. Finally, several interactive instructional design components are included, such as pre- and posttest assessments and interactive case scenarios. Several accessibility standards are also included in module design, including the use of descriptive alt tags for all graphical elements and the inclusion of closed captions for all audio or video elements for people with hearing impediments.

The pre- and posttest assessments are interactive quizzes that present a number of multiple choice questions designed to evaluate learner knowledge of the subject matter and enable self-assessment and reflection, as well as several Likert scale

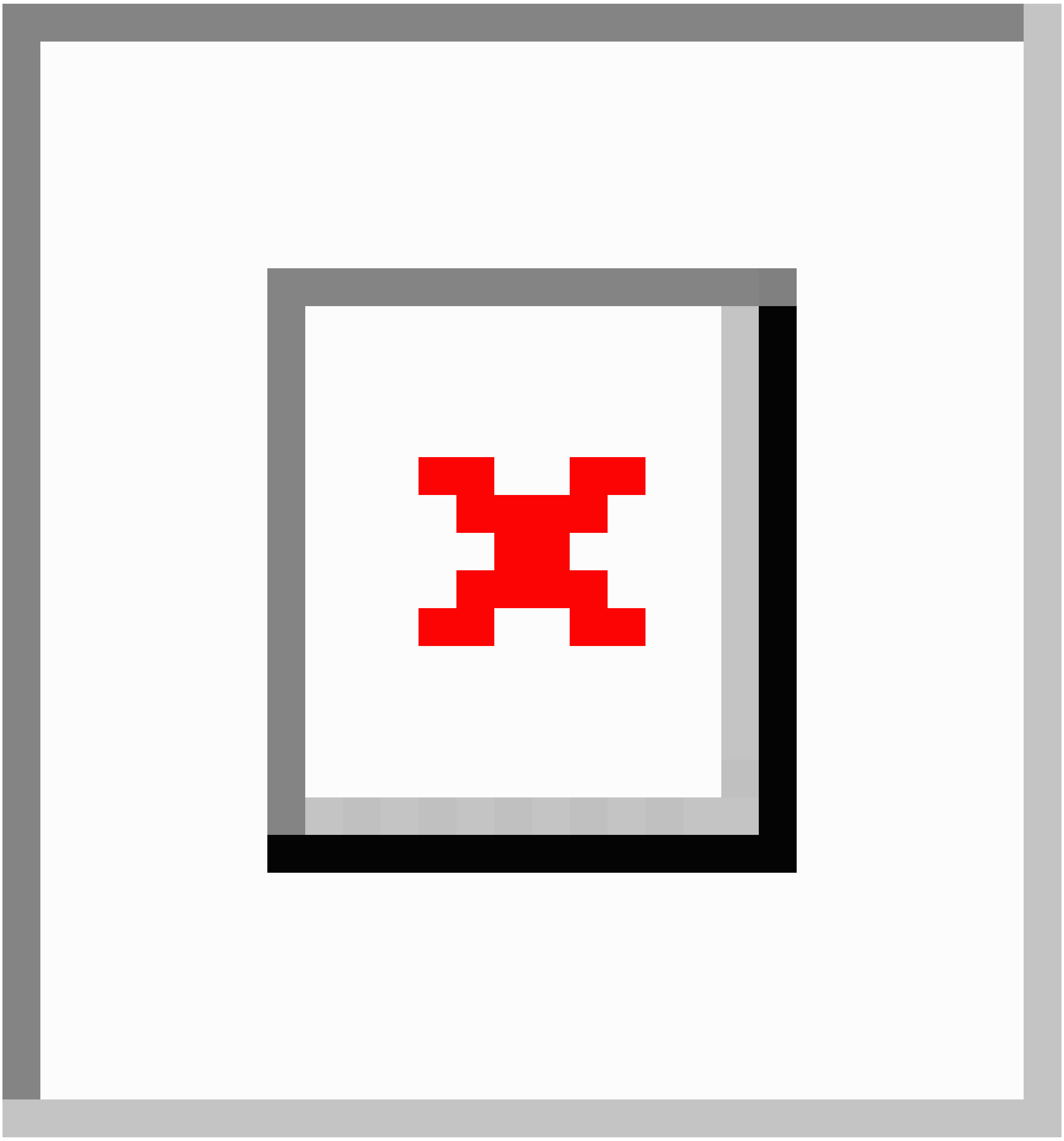
XSL•FO
RenderX

measures of learner confidence in performing the learning objectives (Figure 3). The assessment is presented once at the beginning of the module and then again at the end of the learning experience. The learner receives immediate feedback after submitting each assessment; correct or incorrect data are presented as feedback to the pretest, and correct or incorrect data along with a brief rationale for the correct response are presented as feedback to the posttest.

An interactive case scenario is presented as a final learning activity in the module (Figure 4 and Figure 5). The case scenario models the application of module content to the primary care practice context. The learner is first presented with an overall scenario and then asked a series of "What would you do?" questions designed to prompt reflection. The learner enters their response and is presented with immediate feedback including the response of peer learners in the system as well as a model answer summarizing how the concepts covered in the module could be applied to the given situation.

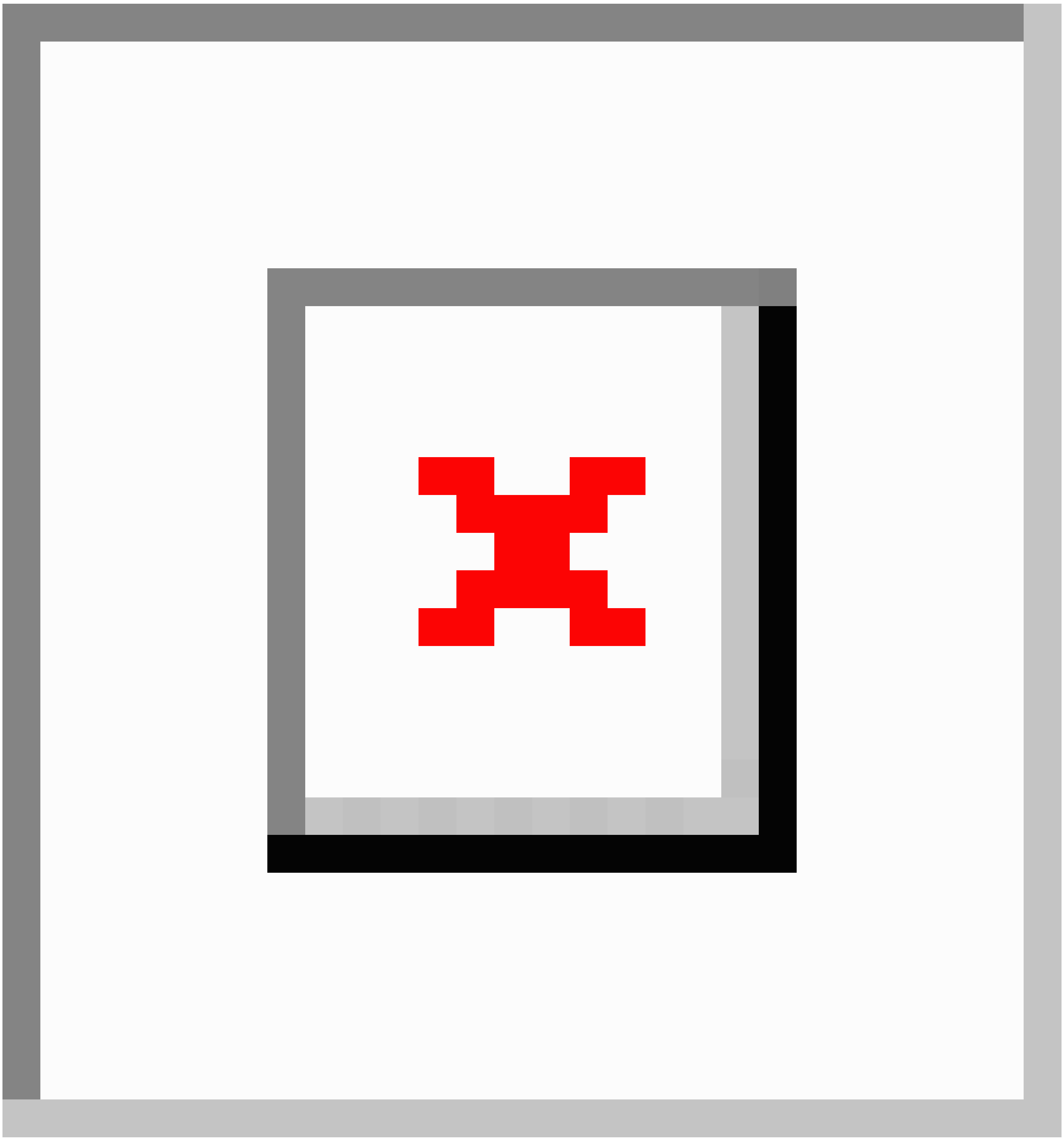
Given the asynchronous approach used in the module design, learners are able to view peer responses to the interactive case scenario but are not able to engage in a dialog with other learners taking the course. Learners can interact with subject matter experts if they have questions related to the content presented in the module. In that case, a learner can enter a question or comment through the "Ask the Expert" feature in the module and will receive a response via email within 48 hours.

Figure 1. Menu interface.



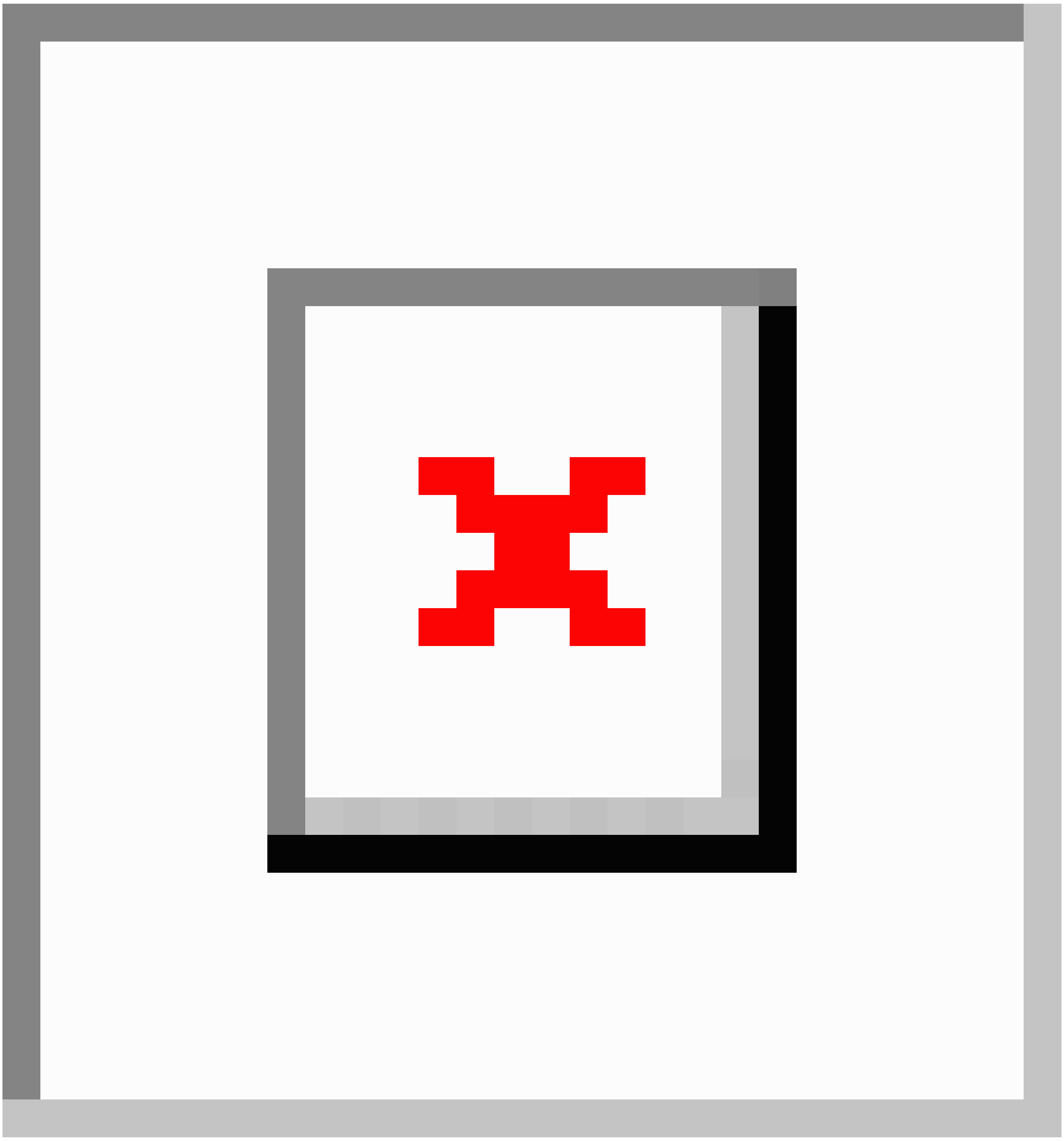
XSL•FO
RenderX

Figure 2. Example of video tutorial.



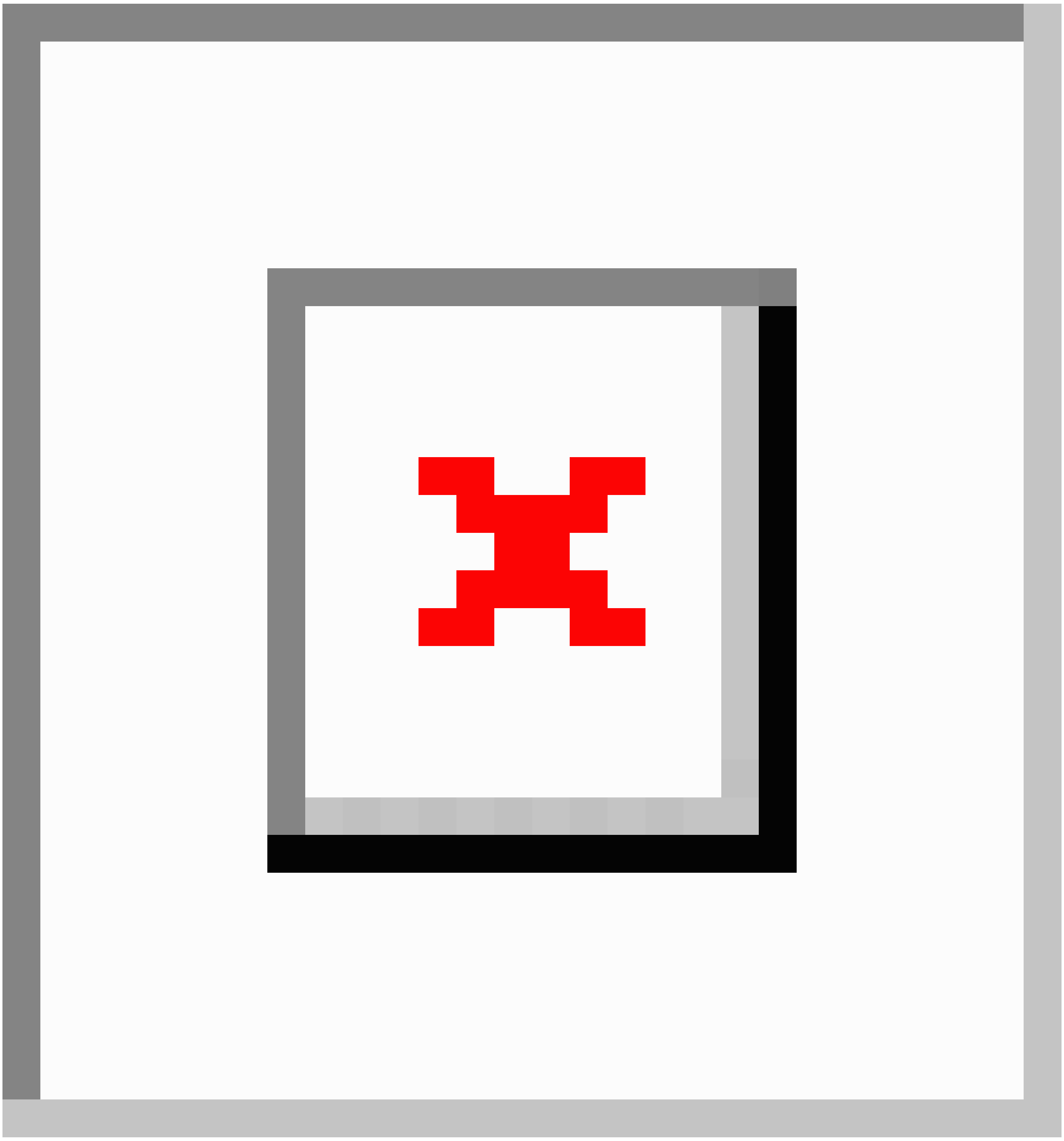
XSL•FO
RenderX

Figure 3. Learning assessment.



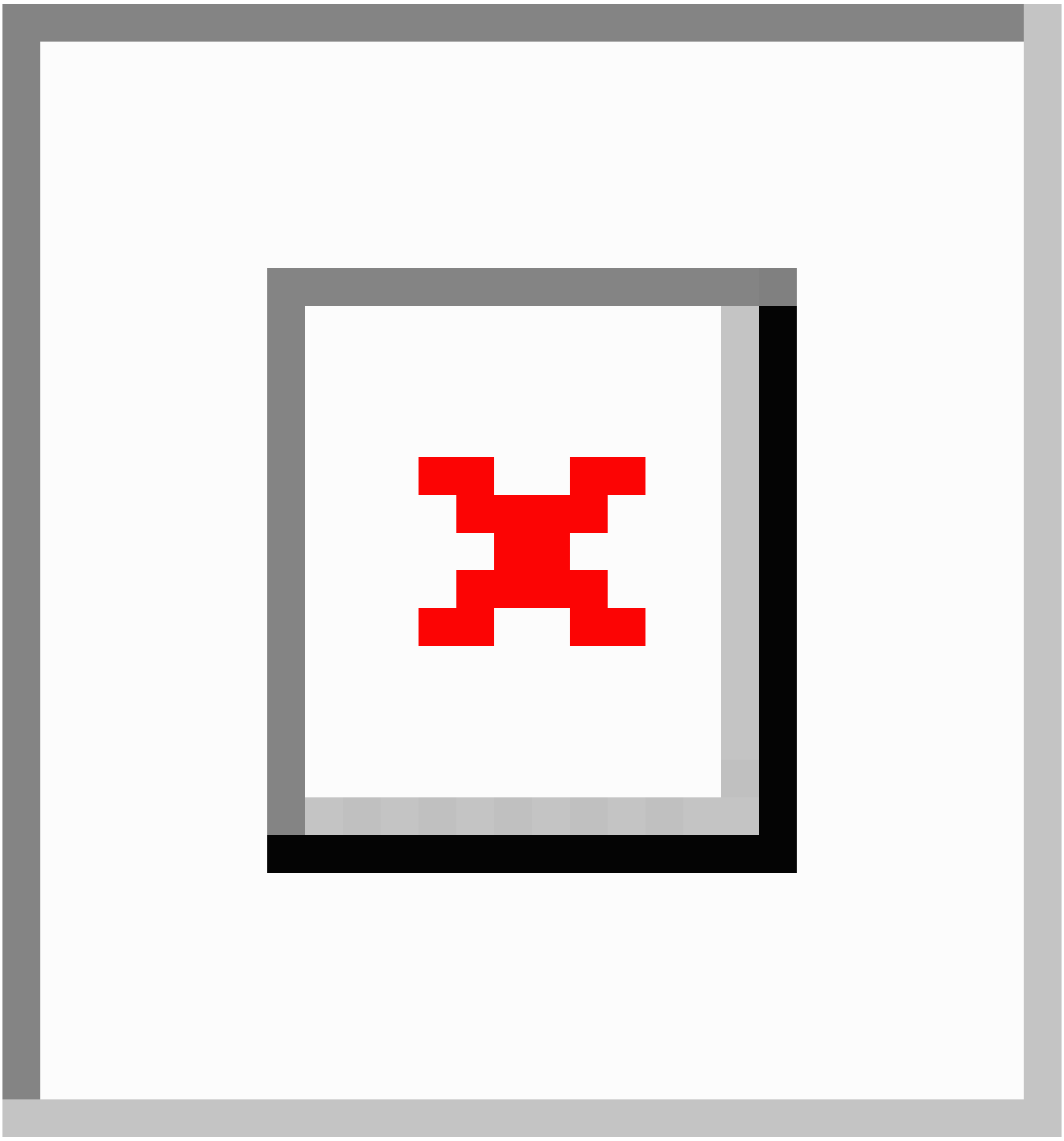
XSL•FO
RenderX

Figure 4. Interactive case scenario.



XSL•FO
RenderX

Figure 5. Interactive case scenario (continued).



Evaluation Approach

An evaluation approach has been designed around Curran and Fleet's [46] use of Kirkpatrick's levels of evaluation. The levels comprising the evaluation approach of the e-learning module include pre- and postknowledge and confidence assessments, satisfaction surveys, and a postmodule outcomes survey.

Pre- and Postknowledge and Confidence Assessments

A pre- and posttest assessment is embedded directly into the e-learning module. The assessment includes measures of both learner knowledge of the content covered and learner confidence in the ability to achieve the stated learning objectives. The knowledge items were developed by content experts and consist of a bank of one-best-answer multiple-choice questions. The confidence items consist of several statements related to the learning objectives for the online module.

Satisfaction Surveys

An online satisfaction survey is provided to the learner at the end of the module. A combination of close- and open-ended questions related to the module content and overall impressions are used to gauge satisfaction and to allow for continuous improvement of subsequent deliveries of the program. The survey enables participants to provide feedback on the module related to relevancy, appropriateness, practicality of the content, and whether they would recommend the module to others.

XSL•FO
RenderX

Postmodule Outcomes Survey

An online survey will be distributed to participants 6-8 months after the completion of the module. The purpose of this survey will be to further explore the impact of participation in the module on participants' adoption and use of virtual care in their practice.

Discussion

Through our systematic needs assessment study, we were able to specify several areas of knowledge, skills, and/or abilities that would be most helpful for physicians and health care providers as they sought to adopt and use virtual care in their practices and patient care. Respondents highlighted 3 main areas. First, the use of technology necessitates knowledge of how to integrate technology and virtual care into the practice workflow. This includes knowing how to use technology and knowledge relating to the privacy and security of the systems being used. There is an increased emphasis for providers to ensure they are meeting the standard of care, adequately obtaining consent, and embracing values of equity and fairness. Next, respondents identified the importance of being able to adapt clinical skills to virtual care and building rapport through good communication with patients. Finally, providers need to be able to adapt their examination skills for virtual care environments.

According to Edirippulige and Armfield [4,27], because using telehealth implies a change in practice, it should be supported by an appropriate level of evidence-based education for health care providers. An appropriate way to do this should start with educating and training future health care providers by incorporating telehealth education as a standard component in the prelicensure curriculum. At a CPD level, online education may be particularly attractive for busy practitioners who choose to participate in short CPD courses to develop knowledge and skills. However, it also seems that the practice of virtual care requires certain hands-on skills. Practical sessions can be helpful in developing such skills, as well as the observation of real-life or simulated virtual care appointments to gain exposure to the modality [22]. Our approach involves the development and provision of an accredited CPD e-learning module, designed to enhance the confidence and competencies of primary health care providers in virtual care adoption and use in their practices. An ongoing evaluation will be conducted with the findings used to improve e-learning approaches to teaching this important area for health care providers and health care delivery systems around the world.

The current evidence surrounding the most effective e-learning modalities is limited by the fact that the reported program designs differ with variation in the types of modalities used to deliver virtual care CPD. There are also limited studies on the effectiveness of asynchronous approaches like those described in this paper. This variation makes it difficult to draw conclusions around the most effective approach, although future comparative type studies could contribute to our understanding of the most effective approaches or combinations of modalities. Another notable observation of the existing literature is the general lack of evaluation at a "knowledge level." Most evaluation studies have not reported assessment of knowledge as a key evaluative outcome from virtual care CPD, whether online or in person. Calleja et al [47] suggest this lack of standardized knowledge evaluation is common among virtual care training programs. The field would benefit from more consistent application of systematic evaluation frameworks, such as Kirkpatrick's [48] or Moore et al's [49] models of evaluation.

The need for virtual care is greater than ever, and health care providers must receive appropriate and meaningful education and training to understand the best ways to conduct virtual care appointments. The current evidence suggests online CPD approaches have been a more common approach, particularly during and after the COVID-19 pandemic. Online CPD on virtual care appears to have been well received by participants; however, there is a lack of evidence surrounding the effectiveness of interactive asynchronous online learning designs like that described in this paper. Asynchronous designs afford greater convenience and flexibility for providers in accessing CPD at times that are best for them. An adaptation of Kirkpatrick's [46] levels of evaluation model is being applied to understand the effectiveness of our asynchronous design, and this will offer further evidence around this online learning modality for CPD on virtual care.

Acknowledgments

We would like to acknowledge the many different health care provider representatives and stakeholders from key governmental and professional association organizations advising on the needs assessment study, design, and development of the e-learning module. Ms Megan Clemens also contributed to the literature review cited in the paper.

This work was supported by the Newfoundland and Labrador Centre for Health Information and Department of Health and Community Services, Government of Newfoundland and Labrador.

Data Availability

The data generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

XSL•FO
RenderX

Authors' Contributions

VC and AH were responsible for the conception and design of the needs assessment study and reporting of results. VC, RG, and CW were responsible for project conceptualization and overseeing design and development of the e-learning module. VC and RG were responsible for manuscript writing as well as final approval of the manuscript. AH and CW reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Health workforce in Canada: highlights of the impact of COVID-19. Canadian Institute for Health Information. 2021. URL: https://www.cihi.ca/en/health-workforce-in-canada-highlights-of-the-impact-of-covid-19/increase-in-virtual-care-services [accessed 2022-04-15]
2. Bokolo AJ. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. J Med Syst 2020 Jun 15;44(7):132. [doi: 10.1007/s10916-020-01596-5] [Medline: 32542571]
3. Curran VR, Hollett A, Peddle E. Virtual care and COVID-19: a survey study of adoption, satisfaction and continuing education preferences of healthcare providers in Newfoundland and Labrador, Canada. Front Digit Health 2023 Jan 25;4:970112. [doi: 10.3389/fdgth.2022.970112] [Medline: 36761449]
4. Curran V, Hollett A, Peddle E. Training for virtual care: what do the experts think? Digit Health 2023 May 30;9:20552076231179028. [doi: 10.1177/20552076231179028] [Medline: 37274369]
5. Bokolo AJ. Implications of telehealth and digital care solutions during COVID-19 pandemic: a qualitative literature review. Inform Health Soc Care 2021 Mar 2;46(1):68-83. [doi: 10.1080/17538157.2020.1839467] [Medline: 33251894]
6. Bokolo AJ. Exploring the adoption of telemedicine and virtual software for care of outpatients during and after COVID-19 pandemic. Ir J Med Sci 2021 Feb;190(1):1-10. [doi: 10.1007/s11845-020-02299-z] [Medline: 32642981]
7. Shachak A, Alkureishi MA. Virtual care: a 'Zoombie' apocalypse? J Am Med Inform Assoc 2020 Nov 1;27(11):1813-1815. [doi: 10.1093/jamia/ocaa185] [Medline: 32940711]
8. Canadian Medical Association, College of Family Physicians of Canada, Royal College of Physicians and Surgeons of Canada. Virtual care guidelines for patients. CMA Digital Library - Canadian Medical Association. 2020. URL: https://www.cma.ca/sites/default/files/pdf/Patient-Virtual-Care-Guide-E.pdf [accessed 2022-04-15]
9. Canadian Medical Association, College of Family Physicians of Canada, Royal College of Physicians and Surgeons of Canada. Virtual care playbook. CMA Digital Library - Canadian Medical Association. 2021. URL: https://www.cma.ca/sites/default/files/pdf/Virtual-Care-Playbook_mar2020_E.pdf [accessed 2022-04-15]
10. Clinician change management: supporting clinicians with virtual care tools and training. Canada Health Infoway. 2022. URL: https://www.infoway-inforoute.ca/en/clinicians-health-workforce/clinician-change-management [accessed 2022-04-15]
11. Telehealth and virtual care. Canadian Medical Protective Association. 2021. URL: https://www.cmpa-acpm.ca/en/covid19/telehealth-and-virtual-care [accessed 2022-04-15]
12. Wong A, Bhyat R, Srivastava S, Boissé Lomax L, Appireddy R. Patient care during the COVID-19 pandemic: use of virtual care. J Med Internet Res 2021 Jan 21;23(1):e20621. [doi: 10.2196/20621] [Medline: 33326410]
13. Appleton R, Williams J, Vera San Juan N, et al. Implementation, adoption, and perceptions of telemental health during the COVID-19 pandemic: systematic review. J Med Internet Res 2021 Dec 9;23(12):e31746. [doi: 10.2196/31746] [Medline: 34709179]
14. James HM, Papoutsi C, Wherton J, Greenhalgh T, Shaw SE. Spread, scale-up, and sustainability of video consulting in health care: systematic review and synthesis guided by the NASSS framework. J Med Internet Res 2021 Jan 26;23(1):e23775. [doi: 10.2196/23775] [Medline: 33434141]
15. Tully L, Case L, Arthurs N, Sorensen J, Marcin JP, O'Malley G. Barriers and facilitators for implementing paediatric telemedicine: rapid review of user perspectives. Front Pediatr 2021 Mar;9:630365. [doi: 10.3389/fped.2021.630365] [Medline: 33816401]
16. Khoshrounejad F, Hamednia M, Mehrjerd A, et al. Telehealth-based services during the COVID-19 pandemic: a systematic review of features and challenges. Front Public Health 2021 Jul;9:711762. [doi: 10.3389/fpubh.2021.711762] [Medline: 34350154]
17. Canadian Medical Association, Canada Health Infoway. 2021 National Survey of Canadian Physicians: quantitative market research report. Canada Health Infoway. 2021. URL: https://www.infoway-inforoute.ca/en/component/edocman/3935-2021-national-survey-of-canadian-physicians/view-document [accessed 2022-04-15]
18. Vaona A, Banzi R, Kwag KH, et al. E-learning for health professionals. Cochrane Database Syst Rev 2018 Jan 21;1(1):CD011736. [doi: 10.1002/14651858.CD011736.pub2] [Medline: 29355907]
19. Ruiz JG, Teasdale TA, Hajjar I, Shaughnessy M, Mintzer MJ. The Consortium of E-Learning in Geriatrics Instruction. J Am Geriatr Soc 2007 Mar;55(3):458-463. [doi: 10.1111/j.1532-5415.2007.01095.x] [Medline: 17341252]

XSL•FO
RenderX

20. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
21. Lahti M, Hätönen H, Välimäki M. Impact of e‐learning on nurses' and student nurses knowledge, skills, and satisfaction: a systematic review and meta‐analysis. Int J Nurs Stud 2014 Jan;51(1):136-149. [doi: 10.1016/j.ijnurstu.2012.12.017] [Medline: 23384695]
22. Sinclair PM, Kable A, Levett-Jones T, Booth D. The effectiveness of internet-based e-learning on clinician behaviour and patient outcomes: a systematic review. Int J Nurs Stud 2016 May;57:70-81. [doi: 10.1016/j.ijnurstu.2016.01.011] [Medline: 27045566]
23. Cook DA, Levinson AJ, Garside S. Time and learning efficiency in internet-based learning: a systematic review and meta-analysis. Adv Health Sci Educ Theory Pract 2010 Dec;15(5):755-770. [doi: 10.1007/s10459-010-9231-x] [Medline: 20467807]
24. Kitto SC, Bell M, Goldman J, et al. (Mis)perceptions of continuing education: insights from knowledge translation, quality improvement, and patient safety leaders. J Contin Educ Health Prof 2013;33(2):81-88. [doi: 10.1002/chp.21169] [Medline: 23775908]
25. Al-Ismail MS, Naseralallah LM, Hussain TA, et al. Learning needs assessments in continuing professional development: a scoping review. Med Teach 2023 Feb;45(2):203-211. [doi: 10.1080/0142159X.2022.2126756] [Medline: 36179760]
26. Shannon S. Needs assessment for CME. Lancet 2003 Mar 15;361(9361):974. [doi: 10.1016/S0140-6736(03)12765-1] [Medline: 12649005]
27. Edirippulige S, Armfield NR. Education and training to support the use of clinical telehealth: a review of the literature. J Telemed Telecare 2017 Feb;23(2):273-282. [doi: 10.1177/1357633X16632968] [Medline: 26892005]
28. Felker BL, Towle CB, Wick IK, McKee M. Designing and implementing TeleBehavioral health training to support rapid and enduring transition to virtual care in the COVID era. J Technol Behav Sci 2022 Dec 14:1-9. [doi: 10.1007/s41347-022-00286-y] [Medline: 36530382]
29. Hassani K, McElroy T, Coop M, et al. Rapid implementation and evaluation of virtual health training in a subspecialty hospital in British Columbia, in response to the COVID-19 pandemic. Front Pediatr 2021 May;9:638070. [doi: 10.3389/fped.2021.638070] [Medline: 34095023]
30. Hayden EM, Nash CJ, Farrell SE. Simulated video-based telehealth training for emergency physicians. Front Med (Lausanne) 2023 Aug;10:1223048. [doi: 10.3389/fmed.2023.1223048] [Medline: 37700768]
31. Khan S, Myers K, Busch B, et al. A national pediatric telepsychiatry curriculum for graduate medical education and continuing medical education. J Child Adolesc Psychopharmacol 2021 Sep;31(7):457-463. [doi: 10.1089/cap.2021.0024] [Medline: 34283939]
32. MDcme.ca home. URL: https://mdcme.ca/ [accessed 2024-07-29]
33. Curran V, Kirby F, Parsons E, Lockyer J. Short report: satisfaction with on-line CME. Evaluation of the ruralMDcme website. Can Fam Physician 2004 Feb;50:271-274. [Medline: 15000339]
34. Curran VR, Hollett A, Peddle E. Patient experiences with virtual care during the COVID-19 pandemic: phenomenological focus group study. JMIR Form Res 2023 May 1;7:e42966. [doi: 10.2196/42966] [Medline: 37036827]
35. Understanding Mainpro+ certification. The College of Family Physicians of Canada. 2021 Mar. URL: https://www.cfpc.ca/CFPC/media/PDF/Understanding-Mainpro-Certification-English-April15-2021.pdf [accessed 2022-04-15]
36. The Maintenance of Certification Program. Royal College of Physicians and Surgeons of Canada. URL: https://www.royalcollege.ca/rcsite/cpd/maintenance-of-certification-program-e [accessed 2022-04-15]
37. Van Hecke A, Duprez V, Pype P, Beeckman D, Verhaeghe S. Criteria for describing and evaluating training interventions in healthcare professions - CRe-DEPTH. Nurse Educ Today 2020 Jan;84:104254. [doi: 10.1016/j.nedt.2019.104254] [Medline: 31689586]
38. Cheung L. Using the ADDIE model of instructional design to teach chest radiograph interpretation. J Biomed Educ 2016 Jun 20;2016:1-6. [doi: 10.1155/2016/9502572]
39. Gagne R. The Conditions of Learning, 4th edition: Holt, Rinehart & Winston; 1985.
40. Herrington J, Oliver R. An instructional design framework for authentic learning environments. Educ Technol Res Dev 2000 Sep;48(3):23-48. [doi: 10.1007/BF02319856]
41. Hrastinski S. Asynchronous and synchronous e-learning. Educause Q 2008 Jan;31(4):51-55.
42. Pei L, Wu H. Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis. Med Educ Online 2019 Dec;24(1):1666538. [doi: 10.1080/10872981.2019.1666538] [Medline: 31526248]
43. Moreno R, Mayer R. Interactive multimodal learning environments. Educ Psychol Rev 2007 Sep 10;19(3):309-326. [doi: 10.1007/s10648-007-9047-2]
44. Brown AR, Voltz BD. Elements of effective e-learning design. Int Rev Res Open Distance Learn 2005 Mar;6(1). [doi: 10.19173/irrodl.v6i1.217]
45. Adedoyin OB, Soykan E. Covid-19 pandemic and online learning: the challenges and opportunities. Interact Learn Environ 2023 Feb 17;31(2):863-875. [doi: 10.1080/10494820.2020.1813180]

XSL•FO
RenderX

46. Curran VR, Fleet L. A review of evaluation outcomes of web-based continuing medical education. Med Educ 2005 Jun;39(6):561-567. [doi: 10.1111/j.1365-2929.2005.02173.x] [Medline: 15910431]
47. Calleja P, Wilkes S, Spencer M, Woodbridge S. Telehealth use in rural and remote health practitioner education: an integrative review. Rural Remote Health 2022 Jan;22(1):6467. [doi: 10.22605/RRH6467] [Medline: 35038387]
48. Kirkpatrick DL. Evaluating Training Programs: The Four Levels: Berrett-Koehler Publishers, Inc; 1994.
49. Moore DE, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]

Abbreviations

ADDIE: Analysis, Design, Development, Implementation, and Evaluation
AFMC: Association of Faculties of Medicine of Canada
CACME: Committee on Accreditation of Continuing Medical Education
CFPC: College of Family Physicians of Canada
CIHI: Canadian Institute for Health Information
CMA: Canadian Medical Association
CMPA: Canadian Medical Protective Association
CPD: continuing professional development
CRe-DEPTH: Criteria for Reporting on Development and Evaluation of Professional Training interventions in Healthcare
OPED: Office of Professional & Educational Development
PHP: Hypertext Preprocessor
RCPSC: Royal College of Physicians and Surgeons of Canada
SPC: scientific planning committee

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna; submitted 19.09.23; peer-reviewed by J Draper-Rodi, R Daynes-Kearney; revised version received 13.05.24; accepted 23.05.24; published 08.08.24.

Please cite as:
Curran V, Glynn R, Whitton C, Hollett A
An Approach to the Design and Development of an Accredited Continuing Professional Development e-Learning Module on Virtual Care
JMIR Med Educ 2024;10:e52906
URL: https://mededu.jmir.org/2024/1/e52906
doi:10.2196/52906

© Vernon Curran, Robert Glynn, Cindy Whitton, Ann Hollett. Originally published in JMIR Medical Education (https://mededu.jmir.org), 8.8.2024. 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.

XSL•FO
RenderX

Roles and Responsibilities of the Global Specialist Digital Health Workforce: Analysis of Global Census Data

Kerryn Butler-Henderson[1,2,*], PhD; Kathleen Gray[2,3,*], PhD; Salma Arabi[1,2,*], PhD

1

2

3

*all authors contributed equally

Corresponding Author:
Kerryn Butler-Henderson, PhD

Abstract

Background: The Global Specialist Digital Health Workforce Census is the largest workforce survey of the specialist roles that support the development, use, management, and governance of health data, health information, health knowledge, and health technology.

Objective: This paper aims to present an analysis of the roles and functions reported by respondents in the 2023 census.

Methods: The 2023 census was deployed using Qualtrics and was open from July 1 to August 13, 2023. A broad definition was provided to guide respondents about who is in the specialist digital health workforce. Anyone who self-identifies as being part of this workforce could undertake the survey. The data was analyzed using descriptive statistical analysis and thematic analysis of the functions respondents reported in their roles.

Results: A total of 1103 respondents completed the census, with data reported about their demographic information and their roles. The majority of respondents lived in Australia (n=870, 78.9%) or New Zealand (n=130, 11.8%), with most (n=620, 56.3%) aged 35 - 54 years and identifying as female (n=720, 65.3%). The top four occupational specialties were health informatics (n=179, 20.2%), health information management (n=175, 19.8%), health information technology (n=128, 14.4%), and health librarianship (n=104, 11.7%). Nearly all (n=797, 90%) participants identified as a manager or professional. Less than half (430/1019, 42.2%) had a formal qualification in a specialist digital health area, and only one-quarter (244/938, 26%) held a credential in a digital health area. While two-thirds (502/763, 65.7%) reported undertaking professional development in the last year, most were self-directed activities, such as seeking information or consuming online content. Work undertaken by specialist digital health workers could be classified as either leadership, functional, occupational, or technological.

Conclusions: Future specialist digital health workforce capability frameworks should include the aspects of leadership, function, occupation, and technology. This largely unqualified workforce is undertaking little formal professional development to upskill them to continue to support the safe delivery and management of health and care through the use of digital data and technology.

(JMIR Med Educ 2024;10:e54137) doi:10.2196/54137

KEYWORDS

workforce; functions; digital health; census; census data; workforce survey; survey; support; development; use; management; health data; health information; health knowledge; health technology; Australia; New Zealand; online content; digital data

Introduction

The importance of a specialist digital health workforce to support the development, use, management, and governance of health data, health information, health knowledge, and health technology has been well-documented [1], particularly through the transformation of digital health during the COVID-19 pandemic. This largely hidden workforce [2] supports the digital health needs for care delivery and management. They are the clinical coders, health informaticians, health information managers, health librarians, health technologists, and so many other occupational specialties who work behind the clinical scenes to ensure that care providers and health managers have the right data, information, and knowledge at the right time and right place [1]. However, there is a lack of accurate data about this specialist digital health workforce to understand their educational needs, their roles and functions, and their professional development needs. This gap in evidence creates challenges for workforce and education planning and forecasting.

The Global Digital Health Workforce Census was launched in 2018 following a rigorous development process [3]. The census stemmed from a collaborative effort between the University of Tasmania and the University of Melbourne using a Delphi

XSL•FO
RenderX

approach. A 10-member expert panel, comprising representatives from key stakeholders, identified issues during a focus group, forming the basis for a health information workforce minimum data set. The items in the census tool were based on existing workforce data items from other surveys and census data sets, and were initially developed with Australian and New Zealand experts. Based on the Health Workforce Australia Report [4], which called for improved data collection about the workforce, the census was referred to as the Australian Health Information Workforce Census. Following the 2018 census [1], the project undertook a validation study to globalize data items with the 2021 census, a smaller pilot with a more global group of participants [5]. The census was referred to as the Global Health Informatics, Digital, Data, Information, and Knowledge (HIDDIN) Workforce Census. The 2023 census was the first full census with global participants, renamed to the Global Specialist Digital Health Workforce, as defined by Butler-Henderson et al [6]. In addition, the census project team worked with Telstra Health to incorporate their Women and Digital Health [7] survey questions into the census. The purpose of this paper is to present the data from the 2023 census related to the roles and functions of the various specialist occupational groups in the specialist digital health workforce.

Methods

Ethical Considerations

The census was held online from July 1 to August 13, 2023. The census project was approved by the RMIT University Human Research Ethics Committee (#26607). No identifiable information is collected in the census and the survey system automatically allocated a unique identifier code to each response. For any questions with <5 responses, the number of responses is not reported to maintain confidentiality. Participants were not compensated for completing the census.

Survey Instrument

The census is a survey deployed through the Qualtrics survey system at RMIT University. It consists of 186 questions across 9 sections, as outlined in Textbox 1.

XSL•FO
RenderX

Textbox 1. Census sections and question topics in each section.

Demographic

- Country, state, and postcode of residence
- Country of birth and citizenship status
- Year of birth
- Gender
- Indigenous or ethnic group
- Disability

Professional membership and health practitioner registration

- What digital health memberships they hold
- If they are a registered health professional and field
- Hours worked in clinical role

Formal education

- Specialist digital health formal education at vocational or higher education level
- Clinical qualifications
- Other relevant qualifications

Credentials

- Relevant credentials

Occupation and paid employment information

- Discipline group
- Time worked in the specialist digital health workforce
- Seeking work
- Current digital health role(s)—for up to two roles, including country, state, postcode, role title, time in role, role intentions next 12 months, top 5 functions, permanency, organization type (both public/private and service type, eg, hospital, educational, department, not for profit), and renumeration
- How many different roles they have

Unpaid and voluntary work

- Voluntary roles and other unpaid related work

Professional development

- What professional development they have done in the last 12 months
- Needs and plans for next three years

Workforce intentions

- How much longer they plan to stay in the workforce
- Why they will leave
- If they will continue to volunteer or do unpaid specialist digital health work

Women and digital health

- Questions from the Women and Digital Health survey

Recruitment

The promotion of the census occurred in multiple ways. The 2023 census was supported by the Australian Digital Health Cooperative Research Centre, Australian Department of Health and Aged Care, Telstra Health, Australasian Institute of Digital Health, Australian Library and Information Association Health Libraries Australia, and the Health Information Management Association of Australia, all of which promoted the census to their networks. The census was launched at the 2023

XSL•FO
RenderX

international health and medical informatics conference, MedInfo. It was also promoted through other professional membership organizations, such as the International Federation of Health Information Management Associations and several other national organizations, such as ANDHealth, and through academic organizations. The census was advertised in several different publications, such as Pulse+ IT and What the Health. Several posts were shared on the census LinkedIn channel and X (formerly Twitter) account. Lastly, individuals could register for a distribution list, which received 2 alerts about the census.

Completion of the census was open to those who self-identified as part of the specialist digital health workforce. The following general guidance was provided:

> *You are part of the workforce if any part of your role (including volunteer or actively seeking) includes a function (listed below) related to health data, information, or knowledge. You may undertake a role that has both a Specialist Digital Health component and another component (for example, clinical or management). For this Census, only consider the Specialist Digital Health component. Functions could include analysing, designing, developing, implementing, maintaining, managing, operating, evaluating, or governing the data, technology, systems, and services for the health sector. You might not identify as part of the Specialist Digital Health workforce if the primary function of your role is limited to using health data, information, or knowledge but none of the other functions listed above.*

An information sheet was provided so that participants could make an informed decision about participation. At the start of the census, participants were reminded about the information sheet and asked to review the questions with regard to providing consent. If they did not consent to participate in the study, they were taken out of the survey. The census took an average of 14 minutes to complete; however, this varied depending on how much detail the participant chose to provide.

Statistical Analysis

Once the survey was closed, the data for all responses was cleaned and only responses that completed section 1 were included in the analysis. Most data items were analyzed using descriptive statistics, such as the number and percentage of responses. When there were fewer than 5 responses, the data is presented as "<5." Only items relevant to capabilities and skills were analyzed for this manuscript, and not all sections of the census are presented in this paper due to the relevance of the topic.

Participants were asked to provide up to 5 functions related to their primary specialist digital health role. All responses were grouped and, using NVivo 14.23 (Lumivero), analyzed for word frequency. The top 5% of the most frequently reported terms were then thematically analyzed, using a modification of the themes identified by Prommegger et al [8]. While Prommegger et al [8] examined occupational aspects, human aspects, and technological aspects, this study examined leadership aspects, functional aspects, occupational aspects, and technological aspects. The top 4 occupational specialties were identified, and the functions listed by respondents who identified with those occupational specialties were then extracted and thematically analyzed in the same fashion, where 5 or more participants identified the term.

Results

Overview

Complete responses for all of section 1 were received from 1103 participants. The majority of responses were from Australia (n=870, 78.9%). Countries with more than 5 responses are shown in Table 1. More than half (n=620, 56.3%) of participants were aged 35‑54 years, and two-thirds (n=720, 65.3%) identified as female. A total of 73 (7.1%) participants identified as Indigenous and 42 (3.8%) as living with a disability.

XSL•FO
RenderX

Table . Participants' characteristics of the 2023 Global Specialist Digital Health Workforce Census (N=1103).

Characteristic and selections		Participants, n (%)
Countries (>5 respondents)		
	Australia	870 (78.9)
	New Zealand	130 (11.8)
	United States	33 (3.0)
	England	9 (0.8)
	Nigeria	8 (0.7)
	Saudi Arabia	6 (0.5)
	Spain	5 (0.5)
	India	5 (0.5)
Age group (years)		
	<25	15 (1.4)
	25 - 34	123 (11.2)
	35 - 44	283 (25.7)
	45 - 54	337 (30.6)
	55 - 64	273 (24.8)
	≥65	72 (6.5)
Gender		
	Female	720 (65.3)
	Male	364 (33.0)
	Nonbinary, gender-fluid, agender	8 (0.7)
	Prefer not to say	11 (1.0)

Occupational Specialization

Respondents were asked to select which occupational specialty they identified with from a list of 16 occupation areas previously identified through the analysis of responses to the 2018 census. The top four occupational specialties were health informatics (n=179, 20.2%), health information management (n=175, 19.8%), health information technology (n=128, 14.4%), and health librarianship (n=104, 11.7%; Table 2). When asked how they classify their occupation against the major categories used by the Australian Bureau of Statistics (ABS) [9], 90% (n=797) of participants identified as managers or professionals. While these classifications are based on the Australian context, the census recognized the international nature of the digital health workforce. Respondents from other countries were encouraged to align their occupations with the provided categories, acknowledging that the ABS classifications served as a reference point for a standardized comparable analysis across diverse geographical contexts. This approach facilitated a more inclusive representation of the global specialist digital health workforce while maintaining a structured framework for analysis.

Respondents also were asked to review the definition of 8 digital health profiles developed by the Australian Digital Health Agency (ADHA) [10] and to select which one they identified as most aligning with their work. These 8 digital health profiles capture the diverse perspectives of the health workforce based on individual roles in the design, development, implementation, and adoption of digital technologies. The profiles include patient, carer, and consumer; frontline clinical; digital champion; clinical and technology bridging; education and research; technologist; leadership and executive; and business, administration, and clinical support [10]. There is no known analysis of the ADHA profiles previously published. In this 2023 census, there was a distribution across a range of profiles, with the top four being leadership and executive (n=174, 19.6%); education and research (n=162, 18.3%); business, administration, and clinical support (n=159, 17.9%); and clinical and technology bridging (n=136, 15.3%). Only 16.7% (n=148) of respondents identified as either a technologist or digital champion.

Table 2 summarizes respondents' categorization of their occupations.

XSL•FO
RenderX

Table . Occupational specializations and classifications in the 2023 Global Specialist Digital Health Workforce Census (n=886).

		Participants, n (%)
Occupation area		
	Biomedical engineering	<5
	Clinical coding	47 (5.3)
	Clinical documentation improvement	31 (3.5)
	Epidemiology	5 (0.6)
	Health artificial intelligence	7 (0.8)
	Health cybersecurity	<5
	Health data science/analytics	53 (6.0)
	Health informatics	179 (20.2)
	Health information management	175 (19.8)
	Health information technology	128 (14.4)
	Health innovation	56 (6.3)
	Health interoperability	28 (3.2)
	Health librarianship	104 (11.7)
	Health simulation	<5
	Health technology assessment	7 (0.8)
	Translational bioinformatics	<5
	Unable to classify	54 (6.1)
Occupation classification		
	Clerical or administrative worker	35 (4.0)
	Community or personal service worker	<5
	Laborer	<5
	Manager	323 (36.5)
	Professional	474 (53.5)
	Sales worker	<5
	Technician or trades worker	10 (1.1)
	Unable to classify	38 (4.3)
Australian Digital Health Agency classification		
	Business, administration, and clinical support	159 (17.9)
	Clinical and technology bridging	136 (15.3)
	Digital champion	54 (6.1)
	Education and research	162 (18.3)
	Frontline clinical	34 (3.8)
	Leadership and executive	174 (19.6)
	Patient, consumer, and carer	27 (3.0)
	Technologist	94 (10.6)
	Unable to classify	45 (5.1)

Qualifications

Participants were asked about their qualifications. With regard to a qualification in a specialist digital health area, the majority (589/1019, 57.8%) of respondents reported no formal educational qualification in a specialist digital health area. Further, only one-quarter (244/938, 26%) reported any industry-issued credential in a digital health area. Of 1033 responses, 30% (n=310) reported that they were a registered clinician.

XSL•FO
RenderX

With regard to professional development activities undertaken in the last year, 65.7% (502/763) reported undertaking some form of professional development. Participants were given the option to identify where they had undertaken the activity and could select more than one organization that delivered that professional development activity. Self-directed professional development activities, such as information seeking, reading/listening/watching blogs/podcasts/vodcasts, and other self-directed activities, were the most reported (676/2438, 27.7%) form of professional development activity (Table 3).

Table . Sources of professional development activities reported in the 2023 Global Specialist Digital Health Workforce Census.

Organization delivering activity	Participants, n (%[a])
Government	223 (9.1)
Industry organization	509 (20.9)
Membership organization	511 (21.0)
Self	676 (27.7)
Training provider	162 (6.6)
Workplace	357 (14.6)

[a]Percentages are based on the total number of reported professional development activities (N=2438).

Employment

With regard to their primary specialist digital health role, more than half (n=487, 55%) of respondents reported that they had worked in their current role for <10 years (Table 4). Three-quarters (n=607, 76.5%) of respondents reported that they were in a permanent specialist digital health role. Representing two-thirds (n=544, 68.6%) of respondents, the top four organizational types were hospital (n=300, 37.8%), health technology organization (n=96, 12.1%), state health department (n=83, 10.5%), and educational facility (n=65, 8.2%). Most (n=552, 69.6%) were public organizations.

XSL•FO
RenderX

Table . Employment characteristics of primary specialist digital health role in the 2023 Global Specialist Digital Health Workforce Census.

Characteristic and selections		Participants, n (%)
Years in current role (n=886)		
	<5	244 (27.5)
	5 - 9	184 (20.8)
	10 - 14	142 (16.0)
	15 - 19	84 (9.5)
	20 - 24	87 (9.8)
	≥25	145 (16.4)
Employment status (n=793)		
	Casual	28 (3.5)
	Contract	142 (17.9)
	Permanent	607 (76.5)
	Self-employed	16 (2.0)
Employment setting (n=793)		
	Community health care service	25 (3.2)
	Defense force/military	<5
	Educational facility	65 (8.2)
	Federal health organization	44 (5.5)
	Health technology organization	96 (12.1)
	Hospital	300 (37.8)
	Indigenous health service	9 (1.1)
	Local health service/district/network	57 (7.2)
	Other not-for-profit organization	29 (3.7)
	Other private organization	32 (4.0)
	Other public/government organization	26 (3.3)
	Primary care or primary health network	16 (2.0)
	Private practice	6 (0.8)
	Residential health care facility	<5
	State health department	83 (10.5)
Employer status (n=793)		
	Not-for-profit	73 (9.2)
	Private	147 (18.5)
	Public	552 (69.6)
	Public/private partnership	21 (2.6)

Functions

The Census asked respondents to list the top five functions of their primary specialist digital health role; 792 respondents provided between one to five functions. Thematic analysis of these functions (as described in the Methods section using a modified list of themes [8]) identified four broad ways of describing their work responsibilities, with example terms shown in Textbox 2.

1. Leadership aspects: these are functions related to leadership.
2. Functional aspects: these are functions related to the operational aspects of roles.
3. Occupational aspects: these are functions that describe the occupation.
4. Technological aspects: these are functions related to the technological aspects of the occupation.

The analysis identified that there was a broad range of functions across these themes, which is to be expected when analyzing the functions across 4 occupational specialist groups representing more than half of the workforce. There was a total of 1353 functions provided across these 4 groups. The functions

XSL•FO
RenderX

of health informatics (n=183 responses for functions), health information management (n=175), health information technology (n=135), and health librarian (n=104) were themed.

Textbox 2. Example terms for describing work responsibilities in the 2023 Global Specialist Digital Health Workforce Census.

Leadership aspects

Leadership, policy, strategic, strategy

Functional aspects

Advice, analysis, governance, manage, searching, teaching

Occupational aspects

Design, development, plans, research, support, service

Technological aspects

Applications, data, digital, software, systems, user

Discussion

Traditionally, throughout the world, capability and competency frameworks have been developed by experts based on their many years of experience. Thus the existing frameworks for digital health specialist occupational areas in many countries, including but not limited to those shown in Table 5, have been developed by industry and academic experts. However, it is crucial to acknowledge the limitation of our findings, as nearly 80% of responses came from Australia. This geographic concentration may limit the generalizability of the results, particularly for countries with single-digit responses. We recognize that while the census provides valuable insights, its predominantly Australian data set may impact the applicability of our conclusion globally. Therefore, it is imperative to interpret our framework recommendations within the context of this geographic bias. Nevertheless, this approach was once the only way to develop these frameworks; today, we have access to a large resource of data about the workforce to inform these frameworks. The Global Specialist Digital Health Workforce Census is one such source.

This paper shows how capability frameworks can be informed by data from those working in these roles. The insights from this analysis not only inform the types of roles and their functions and responsibilities but also help validate expert-originated frameworks and identify new emerging roles with the analysis of census data over time. The four themes identified in this review, leadership aspects, functional aspects, occupational aspects, and technological aspects, and associated functions within each theme, could guide future capability framework development for the specialist digital health workforce (Textbox 3).

Table . Distribution of work responsibilities by occupational group and theme in the 2023 Global Specialist Digital Health Workforce Census.

Occupational specialist	Responses, n	Functions listed, n	Functions included in theme analysis, n	Leadership aspects, n (%)	Functional aspects, n (%)	Occupational aspects, n (%)	Technological aspects, n (%)
Health informatics	183	584	141	7 (5.34)	44 (33.59)	39 (29.77)	41 (31.30)
Health information management	175	610	127	6 (4.88)	44 (35.77)	36 (29.27)	37 (30.08)
Health information technology	135	477	85	4 (4.88)	20 (24.39)	30 (36.59)	28 (34.15)
Health librarian	104	432	106	6 (5.88)	33 (32.35)	26 (25.49)	37 (36.27)

XSL•FO
RenderX

Textbox 3. Modified, with addenda, list of competency lists in specialist digital health occupational areas [11].

Health data scientist

- Canadian Institute of Health Information [12]

Health Informatics

- American Medical Informatics Association [13]
- Australasian Institute of Digital Health [14]
- Digital Health Canada [15]
- Gulf Cooperation Council Health Informatics Workforce Working Group [16]
- Faculty of Informatics United Kingdom [17]

Health information and communications technologists

- Health Information Technology Competencies (HITCOMP) [18]

Health information managers

- American Health Information Management Association [19]
- Canadian Health Information Management Association [20]
- Global Health Workforce Council [21]
- Health Information Management Association of Australia [22]

Health librarianship

- Australian Library & Information Association (Health Libraries Australia) [23]
- Medical Library Association [24]

Of critical concern, this census identified that the broad specialist digital health workforce is largely untrained in digital health capabilities, with more than half (589/1019, 57.8%) reporting that they did not have a specialist digital health qualification. Further, this workforce is not developing these skills consistently through a credentialing program (only 26% hold a credential) or through professional development activities (65.7% reported undertaking professional development in digital health in the past year).

While it could be assumed that most respondents were developing these skills on the job, most (55%) have only been in their role for <10 years, and one-quarter (27.5%) have been in their role for <5 years. On-the-job training is an important factor in improving the quality of health care [25], and the time it takes to become fully productive in a new job is significantly longer in the health workforce, varying depending on the complexity of the job, the individual's prior experience and skills, and the organization's orientation and induction process. While the first 90 days are important, it can take years for a new recruit to a role to be fully productive [26].

There is unquestioned recognition that qualifications to practice and continue professional development are critical for safe health care [27]. Yet amid the ever-increasing digital transformation of the health and care sector, this census shows that professional training and continuing professional development of digital health specialists is at least underreported or at worst absent [28-30].

This is the largest known analysis of the functions of the specialist digital health workforce; however, it is acknowledged that this analysis is of 792 respondents and is largely an Australian data set. It is important to note that the recruitment process may introduce response bias, as those who chose to participate may differ systematically from those who did not. The Australian-centric focus of this data set could limit the generalizability of findings to a broader global context. Future censuses, with a more diverse and extensive respondent pool, will be essential to mitigate potential biases and enhance the robustness and representativeness of the analysis.

The specialist digital health workforce has dedicated roles where their primary function is to support the development, use, management, and destruction of health data, health information, health knowledge, and health technology. The Global Specialist Digital Health Workforce Census is the only survey of its kind to capture critical information about this workforce, including the functions and the capabilities required for them to undertake their roles. However, to enhance the depth of this work, it is essential to provide greater granularity about the specific functions these roles entail. Understanding the intricacies of their daily tasks and responsibilities is crucial for a more comprehensive analysis. This overview emphasizes the largely unqualified nature of the workforce and their limited engagement in formal professional development. This underscores the need for a detailed exploration of the functions performed by these roles, which will not only shed light on the current state but also inform the creation of a more nuanced and informed capability framework. Future frameworks should encompass leadership, function, occupation, and technology aspects to offer a holistic perspective on the specialist digital health workforce.

XSL•FO
RenderX

Acknowledgments

The Global Specialist Digital Health Workforce Census project was funded through the Digital Health Cooperative Research Centre, with funding support from RMIT University, the Australian Department of Health and Aged Care, and Telstra Health, and in kind support from the University of Melbourne, Australian Digital Health Agency, Australasian Institute of Digital Health, Australian Library and Information Association Health Libraries Australia, and the Health Information Management Association of Australia. The authors wish to acknowledge the representatives of each of these organizations and their support as members of the Census Steering Committee: Allison Clarke, Clare Morgan, Gemma Siemensma, James Katte, Joycelyn Linh, Maureen McCarthy, Paul Creech, Sonya Hilberts, and Vickie Irving.

Conflicts of Interest

None declared.

References

1. Butler-Henderson K, Gray K. A glimpse at the Australian health information workforce: findings from the first Australian census. Stud Health Technol Inform 2019 Aug 21;264:1145-1149. [doi: 10.3233/SHTI190405] [Medline: 31438104]
2. Gray K, Gilbert C, Butler-Henderson K, Day K, Pritchard S. Ghosts in the machine: identifying the digital health information workforce. In: Lau F, Bartle-Clar JA, Bliss G, Borycki EM, Courtney KL, Kuo AMH, et al, editors. Improving Usability, Safety and Patient Outcomes With Health Information Technology 2019:146-151. [doi: 10.3233/978-1-61499-951-5-146]
3. Butler-Henderson K, Gray K, Greenfield D, et al. The development of a national census of the health information workforce: expert panel recommendations. Stud Health Technol Inform 2017;239:8-13. [doi: 10.3233/978-1-61499-783-2-8] [Medline: 28756430]
4. Health information workforce report. Australian Institute of Medical and Clinical Scientists. 2013 Oct. URL: https://www.aims.org.au/documents/item/401 [accessed 2024-07-11]
5. Butler-Henderson K, Gray K. 2021 Global HIDDIN Workforce Census. University of Tasmania. 2021. URL: https://www.utas.edu.au/health/projects/hiwcensus/2021-global-hiddin-workforce-census [accessed 2024-07-11]
6. Butler-Henderson K, Day K, Gray K, editors. The Health Information Workforce: Current and Future Developments: Springer; 2021.
7. Understanding gender diversity in Australia's digital health sector. Digital Health CRC. 2022. URL: https://digitalhealthcrc.com/future-thinking/understanding-gender-diversity-in-australias-digital-health-sector/ [accessed 2024-07-18]
8. Prommegger B, Wiesche M, Kremar H. What makes IT professionals special? A literature review on context-specific theorizing in IT workforce research. Presented at: SIGMIS-CPR '20: 2020 Computers and People Research Conference; Jun 19 to 21, 2020; Nuremberg, Germany. [doi: 10.1145/3378539.3393861]
9. Metadata. Australian Bureau of Statistics. 2023. URL: https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/metadata [accessed 2024-07-11]
10. National digital health workforce and education roadmap. Australian Digital Health Agency. 2020 Sep. URL: https://www.digitalhealth.gov.au/sites/default/files/2020-11/Workforce_and_Education-Roadmap.pdf [accessed 2024-07-11]
11. Ritchie A, Siemensma G, Fenton SH, Butler-Henderson K. Competencies, education, and accreditation of the health information workforce. In: Butler-Henderson K, Day K, Gray K, editors. The Health Information Workforce: Current and Future Developments: Springer; 2021. [doi: 10.1007/978-3-030-81850-0_5]
12. CIHI's health data and information governance and capability framework: toolkit. Canadian Institute of Health Information. 2020. URL: https://www.cihi.ca/sites/default/files/document/health-data-info-governance-capability-framework-toolkit-en.pdf [accessed 2024-07-11]
13. Valenta AL, Berner ES, Boren SA, et al. AMIA board white paper: AMIA 2017 core competencies for applied health Informatics education at the master's degree level. J Am Med Inform Assoc 2018 Dec 1;25(12):1657-1668. [doi: 10.1093/jamia/ocy132] [Medline: 30371862]
14. Australian health informatics competency framework for health informaticians: second edition. Australasian Institute of Digital Health. 2022 Feb. URL: https://digitalhealth.org.au/wp-content/uploads/2022/06/AHICFCompetencyFramework.pdf [accessed 2024-07-11]
15. Health informatics professional competencies. Digital Health Canada. 2022. URL: https://digitalhealthcanada.com/wp-content/uploads/2022/05/Health-Informatics-Professional-Competencies.pdf [accessed 2024-07-11]
16. Almalki M, Jamal AA, Elhassan O, Zakaria N, Alhefzi M. Toward the development of the GCC health informatics career paths and matrix. Computer Methods Programs Biomedicine 2021 Jun;205:105987. [doi: 10.1016/j.cmpb.2021.105987]
17. Davies A, Hassey A, Williams J, Moulton G. Creation of a core competency framework for clinical informatics: from genesis to maintaining relevance. Int J Med Inform 2022 Dec;168:104905. [doi: 10.1016/j.ijmedinf.2022.104905] [Medline: 36332519]
18. Competencies. HITCOMP. 2023. URL: http://hitcomp.org/competencies/ [accessed 2024-07-11]

XSL•FO
RenderX

19. 2018 AHIMA Health Information Management Curricula Competencies©. American Health Information Management Association. 2018. URL: https://www.ahima.org/him-curricula/ [accessed 2024-07-11]
20. Career matrix. Canadian Health Information Management Association. 2022. URL: https://www.echima.ca/careers/career-matrix/ [accessed 2024-07-11]
21. Global Health Workforce Council. Global academic curricula competencies for health information professionals. International Federation of Health Information Management Associations. 2015 Jun 30. URL: https://ifhimasitemedia.s3.us-east-2.amazonaws.com/wp-content/uploads/2018/01/20033722/AHIMA-GlobalCurricula_Final_6-30-15.pdf [accessed 2024-07-11]
22. Competency standards. Health Information Management Association of Australia. 2023. URL: https://www.himaa.org.au/our-work/competency-standards/#:~:text=HIMAA%20has%20developed%20the%20following%20competencies%20used%20for,development%29%20Clinical%20Documentation%20Improvement%20competency%20standards%20%28Under%20development%29 [accessed 2024-07-11]
23. ALIA HLA competencies. Australian Library and Information Association. 2018. URL: https://read.alia.org.au/alia-hla-competencies [accessed 2024-07-11]
24. Competencies. The Medical Library Association. 2017. URL: https://www.mlanet.org/professional-development/mla-competencies/ [accessed 2024-07-18]
25. Radeva S. On-the-job training as a model for adapting to the working environment. Int J 2019 Jun 5;31(5):1609-1614. [doi: 10.35120/kij31051609r]
26. Grek A, Stanton A, Monnig B, Whitman A, Chaney A. Advanced practice nurse and physician assistant orientation program: a critical piece in the onboarding process. J Nurse Pract 2022 Jun;18(6):653-659. [doi: 10.1016/j.nurpra.2022.02.028]
27. Mlambo M, Silén C, McGrath C. Lifelong learning and nurses' continuing professional development, a metasynthesis of the literature. BMC Nurs 2021 Apr 14;20(1):62. [doi: 10.1186/s12912-021-00579-2] [Medline: 33853599]
28. Crawford J, Butler-Henderson K. Professional learning and development for the health information workforce. In: Butler-Henderson K, Day K, Gray K, editors. The Health Information Workforce Health Informatics: Springer; 2021. [doi: 10.1007/978-3-030-81850-0_7]
29. Ramsden R, Colbran R, Christopher E, Edwards M. The role of digital technology in providing education, training, continuing professional development and support to the rural health workforce. Health Education 2022 Mar 9;122(2):126-149. [doi: 10.1108/HE-11-2020-0109]
30. Randhawa GK, Jackson M. The role of artificial intelligence in learning and professional development for healthcare professionals. Healthc Manage Forum 2020 Jan;33(1):19-24. [doi: 10.1177/0840470419869032] [Medline: 31802725]

Abbreviations

ABS: Australian Bureau of Statistics
ADHA: Australian Digital Health Agency
HIDDIN: Health Informatics, Digital, Data, Information, and Knowledge

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna; submitted 30.10.23; peer-reviewed by A Davies, C Eldredge; revised version received 13.03.24; accepted 31.05.24; published 25.07.24.

Please cite as:
Butler-Henderson K, Gray K, Arabi S
Roles and Responsibilities of the Global Specialist Digital Health Workforce: Analysis of Global Census Data
JMIR Med Educ 2024;10:e54137
URL: https://mededu.jmir.org/2024/1/e54137
doi:10.2196/54137

© Kerryn Butler-Henderson, Kathleen Gray, Salma Arabi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 25.7.2024. 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.

XSL•FO RenderX

Knowledge Transfer and Networking Upon Implementation of a Transdisciplinary Digital Health Curriculum in a Unique Digital Health Training Culture: Prospective Analysis

Juliane Kröplin[1], MBA, MD, DMD; Leonie Maier[1]; Jan-Hendrik Lenz[1,2], MME, MD, DMD; Bernd Romeike[2], MME, MD

1

2

Corresponding Author:
Juliane Kröplin, MBA, MD, DMD

Abstract

Background: Digital health has been taught at medical faculties for a few years. However, in general, the teaching of digital competencies in medical education and training is still underrepresented.

Objective: This study aims to analyze the objective acquisition of digital competencies through the implementation of a transdisciplinary digital health curriculum as a compulsory elective subject at a German university. The main subject areas of digital leadership and management, digital learning and didactics, digital communication, robotics, and generative artificial intelligence were developed and taught in a transdisciplinary manner over a period of 1 semester.

Methods: The participants evaluated the relevant content of the curriculum regarding the competencies already taught in advance during the study, using a Likert scale. The participants' increase in digital competencies were examined with a pre-post test consisting of 12 questions. Statistical analysis was performed using an unpaired 2-tailed Student *t* test. A *P* value of <.05 was considered statistically significant. Furthermore, an analysis of the acceptance of the transdisciplinary approach as well as the application of an alternative examination method (term paper instead of a test with closed and open questions) was carried out.

Results: In the first year after the introduction of the compulsory elective subject, students of human medicine (n=15), dentistry (n=3), and medical biotechnology (n=2) participated in the curriculum. In total, 13 participants were women (7 men), and 61.1% (n=11) of the participants in human medicine and dentistry were in the preclinical study stage (clinical: n=7, 38.9%). All the aforementioned learning objectives were largely absent in all study sections (preclinical: mean 4.2; clinical: mean 4.4; *P*=.02). The pre-post test comparison revealed a significant increase of 106% in knowledge (*P*<.001) among the participants.

Conclusions: The transdisciplinary teaching of a digital health curriculum, including digital teaching methods, considers perspectives and skills from different disciplines. Our new curriculum facilitates an objective increase in knowledge regarding the complex challenges of the digital transformation of our health care system. Of the 16 student term papers arising from the course, robotics and artificial intelligence attracted the most interest, accounting for 9 of the submissions.

(JMIR Med Educ 2024;10:e51389) doi:10.2196/51389

KEYWORDS

big data; digital didactics; digital health applications; digital leadership; digital literacy; generative artificial intelligence; mobile working; robotics; telemedicine; wearables

Introduction

Background

With the Digital Healthcare Act (German: Digitale-Versorgung-Gesetz), the spectrum of digitalization in the health care system was expanded in Germany in 2019. It includes, among others, the promotion of telemedicine and the expansion of the telematics infrastructure. In addition, a legal framework was created, which, for the first time, entitles insured persons to digital health applications. Digital health applications belong to low-risk medical devices and are primarily intended to support the detection, monitoring, treatment, or alleviation of diseases, injuries, or disabilities. Since January 2021, patients have also been entitled to have access to their data, which have generated during hospital treatment and stored in their electronic patient record. This facilitates electronic provision of medical information, in particular findings, diagnoses, treatment measures carried out and planned, as well as treatment reports for use across facilities, disciplines, and sectors [1,2].

These and further developments show that digital health is creating a new form of health care and is changing the way medicine is delivered and managed [3].

XSL•FO
RenderX

For medical educators, this evolution presents a 2-fold challenge: first, to understand and keep up with the rapidly evolving digital health landscape; and second, to effectively integrate this knowledge into medical curricula to prepare the next generation of health care professionals. Recognizing this gap and the opportunity it presents, the implementation of a comprehensive digital health curriculum is paramount.

Previous studies have suggested that digital health education should be integrated into medical school curricula, with a special emphasis on topics related to knowledge, skills, and attitudes [4].

Several other studies have emphasised the need for medical schools to prepare students for a future in digital health by incorporating digital health competencies into their curricula [4-7].

However, the transdisciplinary approach within university (digital) teaching is still not widespread. The need for such an approach arises from the potential for innovation [8] and is based on professional policy framework conditions such as the new dental licencing regulations [9]. Elective classes seem to be suitable formats for timely introduction, but a longitudinal implementation in mandatory curricula should be the goal [5].

The Implementation of a Transdisciplinary "Digital Health" Curriculum at Our University

The curriculum "Digital Health - Digitalisation and Digital Transformation of Medicine" was offered for the first time at our university in the winter semester of 2022-2023. Students from all faculties and all semesters of the university were eligible to participate.

The learning objectives were developed on the basis of existing literature [4-6,10] and interviews with transdisciplinary experts in the areas of human medicine, dental medicine, medical didactics, computer science, business administration, theology, and ethics. The curriculum is divided into the 4 subareas of digital didactics, namely digital communication, management and digital leadership, and robotics and generative artificial intelligence (AI), each with 14 weekly lessons as well as an introductory event and a final examination and evaluation event. The lessons particularly encompassed the following topics: augmented or virtual reality, big data or generative AI, data protection or information security, digital leadership, digital didactics, ethical aspects of digital health, new work, robotics, social media, open educational resources, digital health applications, wearables, simulation training, and telemedicine (Table 1).

Table . Digital health curriculum.

Topics	Goals, subareas, and time
Digital communication	• **Goal:** knowledge transfer regarding modern communication systems, consideration of legal framework conditions, and ethical aspects during transdisciplinary implementation and application • **Subareas:** telemedicine, digital patient files, ethics, messenger apps, digital health applications • **Time:** 3 lessons, each lasting 90 minutes
Digital didactics	• **Goal**: application of modern teaching and learning methods and creating a nondiscriminatory framework for studies • **Subareas**: open educational resources, virtual or augmented reality, simulation training • **Time**: 3 lessons, each lasting 90 minutes
Management and digital leadership	• **Goal:** knowledge transfer regarding digital transformation including economic aspects as well as the importance of innovative leadership styles • **Subareas:** leadership, information security, data protection, economy, social media, and mobile working • **Time:** 3 lessons, each lasting 90 minutes
Robotics and artificial intelligence	• **Goal:** knowledge transfer about possible applications of surgical robots, individualized medicine, and possible uses of generative artificial intelligence in teaching, research, and patient treatment • **Subareas:** robotics, generative artificial intelligence, wearables, and big data • **Time**: 3 lessons, each lasting 90 minutes

The aims of this digital health curriculum are as follows: (1) integrating basic digital health content into the curriculum of a university in northern Germany; in a transdisciplinary approach, students will be taught the necessary competencies to be able to apply digital health technologies in their later work; (2) considering the new licencing regulations for dentists; dental students, in particular, should be encouraged to use the newly implemented compulsory elective subject to gain knowledge in the field of digital health; and (3) to encourage students to critically engage with the topic of digital health within the framework of a scientific thesis; this also intended to reflect currently relevant digital health topics from the students' perspective as a basis for further curriculum development.

XSL•FO
RenderX

The curriculum contents were taught over a period of 1 semester within the framework of a compulsory elective subject.

Furthermore, this study aims to analyze the objective acquisition of digital competencies through the implementation of a transdisciplinary digital health curriculum at a German university.

Methods

Ethical Considerations

The study has been reviewed by the ethics committee of the Faculty of Medicine of the University of Rostock, Germany, and has been approved (A 2022-0137).

Demographics and Previous Teaching of Digital Health

Student-related data about educational level, gender distribution, and career goals were analyzed. At the beginning of the semester, students were asked whether digital health learning objectives had already been taught in previous courses, using a Likert scale (1=very well taught to 5=not taught at all).

Students' Assessment and a New Examination Approach for Further Development

To measure the allocation of knowledge of the participants, the participants' prior knowledge was assessed during the introductory lesson through a theoretical test (pretest) consisting of 12 questions. Ten questions were multiple-choice and 2 were open questions. The test was specifically related to the topics covered in the curriculum. Multiple-choice questions assessed knowledge on the topics of digital transformation, ethics, change management, data protection, robot-assisted surgery, digital patient files, video consultation, and simulation training. The didactics section was covered by 2 open questions and 1 multiple-choice question. At the final seminar, the theoretical test was repeated with similar questions (posttest).

In addition to the standardized questions, students were asked to write a scientific paper. The topic could be chosen independently. However, a prerequisite was a content-related reference to the overarching topic of digital health. The objectives of the examination are to (1) encourage students to critically engage with a digital health topic of their choice, (2) promote scientific work, and (3) obtain an insight into the topics of digital health perceived as relevant by the students as a basis for further curriculum development.

For further structuring of the curriculum, the scientific papers were assigned to one of the main topic areas based on the selected headings and abstract contents.

Statistical Analysis

The data were analyzed using SPSS (version 27; IBM Corp) software. The gender distribution, career goals, intended subject area, and scientific papers were analyzed descriptively. Statistical analysis of pre-post test results and previous teaching of learning objectives was performed using an unpaired (learning objectives) and paired (pre-post test results) 2-tailed Student *t* test. A *P* value of <.05 was considered statistically significant.

Results

Educational Level of the Participants

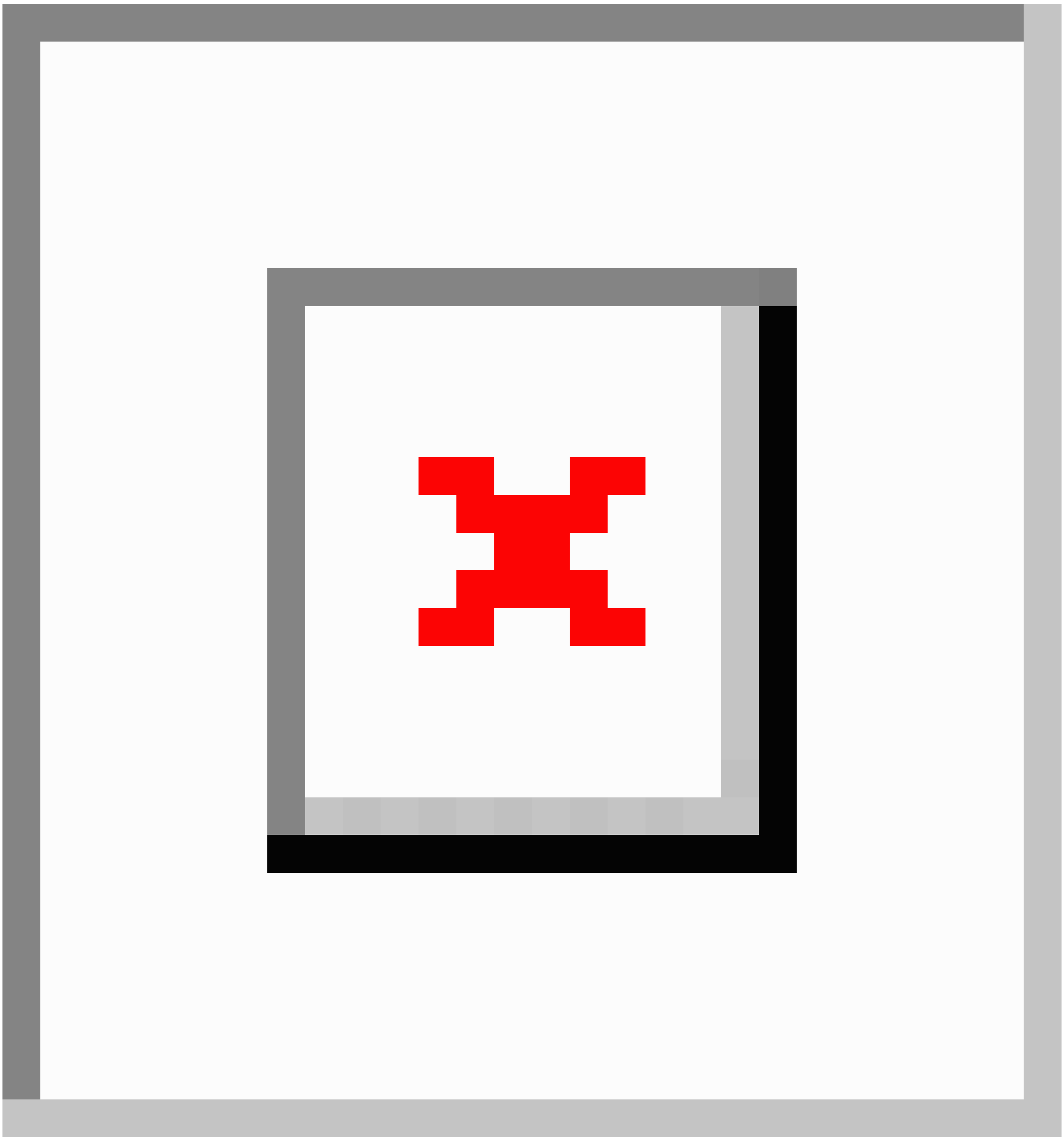
Within the first year, a total of 20 students (5 in the winter term and 15 in the summer term) participated in the digital health curriculum. The average age of the participants was 22.3 (range 19-30) years. At the time of participation, 15 participants studied human medicine, 3 participants were studying dentistry, and 2 participants were studying medical biotechnology. In total, 11 (61.1%) students in human and dental medicine were in the preclinical phase and 7 (38.9%) were in the clinical phase.

Gender Distribution

Figure 1 shows the gender ratio according to the subjects of study among the participants. In total, 13 participants were female and 7 were male. Among human medicine students, 10 were female and 5 were male. Two dentistry students were male and 1 was female. Both biotechnology students were female.

XSL•FO
RenderX

Figure 1. Gender distribution by study subject.



Career Goals

Two questions were aligned with the focus on future professional activities. The first question asked whether the respondents wanted to work in an inpatient or outpatient setting. The options “other” and “don’t know yet” could also be selected. Furthermore, the students were asked about their desired goal of becoming a specialist doctor. As shown in Multimedia Appendix 1, the majority of participants are still undecided on whether they want to work in the outpatient or inpatient sector in future. Multimedia Appendix 2 shows the answers to the question about the goal of becoming a medical specialist, which was answered by the participating human medicine students. According to this, most of the participants who already know their career goal intended to become a specialist in surgery (n=4).

Previous Teaching of Digital Health

During the first lesson, students were asked whether digital health learning objectives have already been taught in previous courses, using a Likert scale (1=very well taught to 5=not taught at all). Table 2 shows the corresponding evaluations.

XSL•FO
RenderX

Table . Evaluation of the learning objectives of previous teaching of the digital health curriculum.

Learning objectives	Clinical, mean	Preclinical, mean	*P* value
All	4.2	4.4	.02
Big data	4.9	4.7	.36
Artificial intelligence	4.7	4.6	.84
Digital health applications	4.1	4.3	.80
Messenger apps	4.3	4.7	.26
Wearables	4.4	4.8	.18
Telemedicine	4.1	4.7	.20
Data protection and information security	3.6	3.7	.82
Digital ethics	4.0	4.1	.87
Simulation training	3.9	4.5	.28
Virtual or augmented reality	3.8	4.6	.10
Economy	3.7	4.6	.09
Digital didactics	3.0	3.6	.41
Robotics	4.6	4.5	.83
Digital leadership	4.6	4.8	.46
Mobile working	4.7	4.7	.97
Social media	4.6	4.4	.69
Open educational resources	3.9	4.3	.55

Among clinical students, significantly better overall coverage of the digital health learning objectives is evident.

Pre-Post Test Results and Term Paper Evaluation

In the pretest, the participants scored an average of 4 points compared to 8.3 points in the posttest. Consequently, there was a significant increase of 106% in knowledge ($P<.001$; Table 3).

Table . Increase in knowledge determined via a pre-post test (maximum achievable score 12; 106% increase in knowledge by 4.3 points; $P<.001$).

Increase in knowledge	Pretest[a]	Posttest[b]
Total score	4.0	8.3
Clinical	4.1	8.7
Preclinical	4.2	8.2
Female participants	3.6	8.3
Male participants	4.7	8.1

[a]Difference in pretest scores between clinical and preclinical participants: $P=.96$; differences in posttest scores between clinical and preclinical participants: $P=.38$.

[b]Difference in pretest scores between male and female participants: $P=.11$; difference in posttest scores between male and female participants: $P=.17$.

Neither gender nor study phase affected pre- or posttest results. As shown in Table 4, the most frequently selected main topic was robotics and AI.

XSL•FO
RenderX

Table . Digital health topics selected by students for their term papers.

Titles of the students' term papers	Digital health main topic
Progress of computer-assisted procedures and robotics in implantology	Robotics and AI[a]
Mind reading with functional magnetic resonance imaging and AI	Robotics and AI
To what extent can the Da Vinci Robot help reduce postoperative complications?	Robotics and AI
Algorithms against prejudice? The role of AI in combating gender discrimination in the health sector	Robotics and AI
What opportunities arise from the use of AI in medicine and what are the associated problems?	Robotics and AI
Applications of AI in Radiology	Robotics and AI
AI and robotics in Orthopaedics and Trauma Surgery	Robotics and AI
Opportunities and limits of AI in the health sector	Robotics and AI
Use of AI for early detection of dementia	Robotics and AI
Data ethics in the digital world	Management and leadership
Does digitalisation in medicine lead to a loss of skills and knowledge among medical staff?	Digital didactics
"Flipped Classroom": Possibilities of redesigning of an accompanying seminar on the study of human medicine.	Digital didactics
Implementation of an interdisciplinary elective subject "Digital Health"	Digital didactics
Aspects of discrimination against older people in digital medicine	Digital communication
What role do chatbots play in medical studies	Digital communication

[a]AI: artificial intelligence.

Discussion

Overview

Current social, political, and economic developments in Germany require a reorientation of university teaching, considering digital learning and teaching strategies. The necessity is also reflected in the restructuring of established framework conditions, such as the amendment of dental and medical licencing regulations [9,11].

This study aimed to analyze the objective acquisition of digital competencies through the implementation of a transdisciplinary digital health curriculum at a German university.

The learning objectives were imparted on the main topics of management and digital leadership, robotics and AI, digital communication, and digital didactics within the framework of a 1-semester curriculum. Objective knowledge gain was determined using a pre-post test design. In addition, the extent to which the approach of transdisciplinary networking could be implemented was analyzed. This was quantified by the disciplines and the number of clinical and preclinical participants. Overall, the results were analysed over 1 year (2 cohorts). In the second run, the number of participants has already tripled.

Characterization of the Participants

According to the Federal Statistical Office, 64.8% of students in human medicine in 2021 were female [12]. This corresponds to the distribution of participants in our curriculum, even when considering the isolated subject group of human medicine being the most frequently represented. Consequently, it can be assumed that the topic is not gender-specific and is of equal interest to male and female students. This cannot be confirmed for participants from the fields of dentistry and medical biotechnology. However, the small number of participants must be considered here.

Previous Teaching of Digital Health

Evaluation of the students at the beginning of the semester revealed that all the content of the curriculum has not been taught at all or only to a very limited extent. Even though there was a significant difference in knowledge between the clinical and preclinical sections, this concerns all participants. Consequently, it can be assumed that this deficit will not be sufficiently compensated for in higher semesters with regard to the clinical phase.

The results also indicate that most participants are still open about their career goals. This applies both to the future field of work (outpatient vs inpatient) and to the intended specialization. Therefore, the general approach to teaching content can be considered suitable.

Assessment of the Increase in Knowledge

As reported by studies with a similar study design, a significant objective increase in knowledge could be achieved among participants through the curricular dissemination of knowledge on relevant digital health topics. It should be noted that some students participated in the curriculum out of interest in the content but without aiming to achieve a good grade.

XSL•FO
RenderX

Consequently, it can be assumed that some students did not prepare for the posttest. The fact that summative assessment of the intended learning objectives at the beginning of the curriculum increases learning success has previously been described [13].

Regarding the current evidence in the development of digital literacy, the focus is increasingly on social interaction and lifelong learning skills in an innovative teaching and learning culture, in addition to subject knowledge [5].

Term Paper Evaluation

When analyzing the selected term papers, it quickly became clear that the topic of AI is of outstanding importance among digital health topics. This seems to be explained, in particular, by the strong media presence of the topic. The rapid development of generative AI has received special media attention with the launch of ChatGPT in 2022 [14-16]. Two challenges arise, in particular, for the curriculum. Although the special importance of flexible and adaptive teaching formats to be able to integrate innovations into teaching without delay is becoming apparent, the establishment of framework conditions for the application of generative AI in teaching, research, and clinical practice is coming into focus. Both focal points and associated challenges were already considered and will be further developed for our future digital health curriculum.

The Role of Leadership in a Digital Health Training Culture

Digital transformation is a continuous process that is better accepted by those who perceive digitalization as relevant to their own work. Digital leadership describes the special role of managers in the implementation of digital transformation. It is up to managers in the health care sector to align the strategic orientation to digital transformation with the company's goals and needs and to create an appropriate digital culture. Regarding the provision of early access to the necessary knowledge on topics related to digital health, managers in the field of education have a special responsibility [17].

The transdisciplinary approach of the digital health curriculum acknowledges the current evidence for the success of digital transformation. In particular, evidence from economic evaluations has shown that in a networked environment, the opening of boundaries is necessary to create innovation and exploit synergies [8].

With an average value of 5 on the Likert scale, the results of the initial evaluation show that this knowledge has not yet been imparted in the participants' previous curricula. Consideration of the transdisciplinary digital health curriculum is, therefore, of particular importance.

Digitalization Connects: the Necessity of a Transdisciplinary Digital Health Curriculum

The goal of opening of the curriculum to all faculties is to expand the transdisciplinary network to promote an innovation-driven teaching and learning environment. This basic idea represents a unique selling point for previously established digital health curricula.

Our results indicate that this opportunity was already realized in the first year by students from 3 different disciplines, such as human medicine, dentistry, and medical biotechnology. The distribution of clinical and preclinical students also shows cross-semester interest.

In the future, an increase in the participation of dental medicine students is expected. This is due to the new orientation of the dental licencing regulations, which mandate participating in curricula by choosing from among the elective subject areas (to which the compulsory digital health elective subject is assigned), both for the preclinical and clinical study phases [9].

It should be noted, however, that only 5 out of 20 students did not belong to the field of human medicine. These results suggest that the transdisciplinary approach needs to be further promoted, addressed, and implemented to achieve an even better transdisciplinary exchange.

Social media use may present an opportunity for increasing the visibility of our transdisciplinary curriculum and its learning objectives. The curriculum is currently already accompanied by a social media channel. The importance of social media in teaching and research is currently the focus of social debates and scientific studies [18,19]. For better assessment of the importance of social media in a modern academic teaching and learning culture, the authors believe that further studies are needed.

Emerging Technologies in a Transforming Health Care System

The use of modern technologies has enormous potential for optimizing patient treatment [20,21]. In surgery, in particular, there is a wide range of applications in the operating theater and perioperative management.

A recent editorial describes current emerging innovations with particular potential, which are also included in the digital health curriculum [20]. In particular, this involves the contents of machine learning–enabled clinical decision-making support, computer vision and augmented reality, as well as wearable devices and remote patient monitoring. The dynamic nature of these developments, among others, shows the particular importance of a flexible and adaptive curriculum to be able to integrate emerging technologies into teaching without delay.

Robot-assisted surgery, including approaches to telesurgery, is of particular importance, especially in surgery. The special importance of robotics for patient care has already been described several times and is now an integral part of numerous hospitals [22,23].

The special importance of robotics is also reflected in the selection of homework topics. Three of the 16 papers submitted focus on robotics in medicine.

However, the increasing use of robotics in the operating theater also requires special skills that can and must be practised extensively in a simulation-based setting [24]. This requires time and financial resources, as well as training in a supervised setting [25]. In teaching and further education, these prerequisites represent a hurdle. In particular, cost-intensive virtual and augmented reality simulators are often only rarely

XSL•FO
RenderX

available; their use in teaching is generally yet not structured [26]. User acceptance is indisputably high and can increase satisfaction in addition to learning success [27]. However, the topic requires economic reflection and a basic understanding of project management—an aspect that was addressed in the curriculum section of Management and Digital Leadership.

In addition to the implementation and continuous further development of technical innovations in clinical applications, achievements with disruptive innovation power also play a special role in future teaching and research. The disruptive potential of digital transformation is currently manifesting itself in particular in the launch of generative AI, such as ChatGPT [14].

Generative AI, Web-Based Meetings, and the Challenge of Flexible Adaptive Training

The examination of digital teaching methods has experienced a surge in innovation, particularly in the context of the COVID-19 pandemic [28]. Experience in the field of telemedicine has provided a blueprint for web-based teaching with simultaneous integration of knowledge content in telemedicine. Thus, knowledge transfer could be extended by the achievement of local flexibility [28].

But approaches that account for time flexibility are also described: the "flipped classroom" model, for example, is an approach to active self-directed learning in which students acquire the basic concepts themselves before class—for example, through recorded lectures or interactional learning modules provided by a learning management system—so that class time can be used for active learning activities such as exercises, projects, or discussions. Valuable time spent in presence is used for the application, rather than acquisition, of knowledge. This can increase both student performance and student satisfaction [29,30].

In addition to the flexibility of location and time, there are often limits to accessing real-world working environments. To be able to train practical and theoretical skills in a realistic setting, such as an operating theater, teaching using virtual and augmented reality offers promising potential.

Virtual reality refers to complete visual immersion in an artificial, computer-generated environment. In augmented reality, holograms, which often also enable interaction, appear projected into the room through semitransparent glasses. Mixed reality is the combination of digital screens with projected interactional holograms. The user sees the real world while simultaneously manipulating the digital content generated by the device [31].

Both technologies are increasingly being integrated in the clinical setting, but also in teaching, such as the visualization of organs. In clinical applications, augmented reality enables the simulation of patient encounters to train communication skills or intraoperative decision-making to increase safety during surgery [32].

Limitations

This study's limitations particularly include its single-center design and the small number of participants at the time of analysis. In addition, the final test only examined excerpts from topics that cannot represent the full scope of the curriculum. The choice of term paper is also subject to numerous influencing factors, so the motivation for choosing the topic cannot be clearly identified.

Conclusions

This study aims to analyze the objective acquisition of digital competencies through the implementation of a transdisciplinary digital health curriculum at a German university. The results show that relevant content on digital health topics has not been taught sufficiently at the university outside our new digital health curriculum. The objective increase in the knowledge on these topics within the framework of the digital health curriculum could be verified as significant via a pre-post test design.

The approach of transdisciplinary development of a digital health curriculum seems especially promising. We provided dentistry students a platform to complete their recently compulsory elective subject. We observed that dentistry students could complete their recently compulsory elective subject when using an appropriate digital platform.

The integration of written assignments as a special examination element can promote critical engagement with digital health content. This also facilitates gaining insight into digital health topics and issues that are relevant to students. We can harness these insights in further developing our curriculum.

Together with the current literature, our data indicate that the content of digital health curricula must be transferred into standard teaching for all health science students.

Acknowledgments

We would like to thank all the speakers who supported the digital health curriculum with their presentations and expert knowledge.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Career goals.
[PNG File, 37 KB - mededu_v10i1e51389_app1.png]

XSL•FO
RenderX

Multimedia Appendix 2
Aspired specialist of the human medicine participants.
[PNG File, 25 KB - mededu_v10i1e51389_app2.png]

References

1. Gerlinger G, Mangiapane N, Sander J. Digital health applications (Diga) in medical and psychotherapeutic care. opportunities and challenges from the perspective of the healthcare providers. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Oct;64(10):1213-1219. [doi: 10.1007/s00103-021-03408-8] [Medline: 34550412]
2. Digitale Versorgung-Gesetz: Vorstellung des Gesetzgebers Zur Digitalisierung des Gesundheitswesens. Fortschr Röntgenstr 2019 Aug;191(8):769-771. [doi: 10.1055/a-0875-9009] [Medline: 31344725]
3. Lowery C. What is digital health and what do I need to know about it? Obstet Gynecol Clin North Am 2020 Jun;47(2):215-225. [doi: 10.1016/j.ogc.2020.02.011] [Medline: 32451013]
4. Kröplin J, Huber T, Geis C, Braun B, Fritz T. eSurgery—digital transformation in surgery, surgical education and training: survey analysis of the status quo in Germany. Eur Surg 2022 Oct;54(5):249-258. [doi: 10.1007/s10353-022-00747-x]
5. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019 Dec 11;19(1):460. [doi: 10.1186/s12909-019-1891-5] [Medline: 31829208]
6. Kuhn S, Huettl F, Deutsch K, Kirchgässner E, Huber T, Kneist W. Surgical education in the digital age - virtual reality, augmented reality and robotics in the medical school [Article in German]. Zentralbl Chir 2021 Feb;146(1):37-43. [doi: 10.1055/a-1265-7259] [Medline: 33588501]
7. Machleid F, Kaczmarczyk R, Johann D, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827. [doi: 10.2196/19827] [Medline: 32667899]
8. Grafström M, Falkman LL. Everyday narratives: CEO rhetoric on Twitter. JOCM 2017 May 8;30(3):312-322. [doi: 10.1108/JOCM-10-2016-0197]
9. Approbationsordnung für Zahnärzte und Zahnärztinnen [Article in German]. Gesetze im Internet. 2023. URL: https://www.gesetze-im-internet.de/zappro/index.html
10. Khurana MP, Raaschou-Pedersen DE, Kurtzhals J, Bardram JE, Ostrowski SR, Bundgaard JS. Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Med Educ 2022 Feb 26;22(1):129. [doi: 10.1186/s12909-022-03163-7] [Medline: 35216611]
11. Kuhlmann E. Ärztliche Approbationsordnung: Neuer Anlauf für überfällige Reform [Article in German]. Dtsch Arztebl 2023;120(20):A-906.
12. Studierende Insgesamt und Studierende Deutsche Im Studienfach Medizin (Allgemein-Medizin) Nach Geschlecht [Article in German]. Statistisches Bundesamt. 2021. URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/lrbil05.html#242500
13. Raupach T, Brown J, Anders S, Hasenfuss G, Harendza S. Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. BMC Med 2013 Mar 5;11:61. [doi: 10.1186/1741-7015-11-61] [Medline: 23497243]
14. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
15. Moritz S, Romeike B, Stosch C, Tolks D. Generative AI (gAI) in medical education: Chat-GPT and co. GMS J Med Educ 2023;40(4):Doc54. [doi: 10.3205/zma001636] [Medline: 37560050]
16. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023 Mar;5(3):e102. [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
17. Cortellazzo L, Bruni E, Zampieri R. The role of leadership in a digitalized world: a review. Front Psychol 2019;10:1938. [doi: 10.3389/fpsyg.2019.01938] [Medline: 31507494]
18. Lima DL, Viscarret V, Velasco J, Lima R, Malcher F. Social media as a tool for surgical education: a qualitative systematic review. Surg Endosc 2022 Jul;36(7):4674-4684. [doi: 10.1007/s00464-022-09150-9] [Medline: 35230534]
19. Huber T, Hüttl F, Braun B, et al. Fridays for future! - All days for surgery!: Thoughts of young surgeons on a modern promotion of the next generation [Article in German]. Chirurg 2022 Mar;93(3):250-255. [doi: 10.1007/s00104-022-01577-z] [Medline: 35132445]
20. Marwaha JS, Raza MM, Kvedar JC. The digital transformation of surgery. NPJ Digit Med 2023 May 31;6(1):103. [doi: 10.1038/s41746-023-00846-3] [Medline: 37258642]
21. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219. [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
22. Williamson T, Song SE. Robotic surgery techniques to improve traditional laparoscopy. JSLS 2022;26(2):e2022.00002. [doi: 10.4293/JSLS.2022.00002] [Medline: 35655469]
23. Reinisch A, Liese J, Padberg W, Ulrich F. Robotic operations in urgent general surgery: a systematic review. J Robot Surg 2023 Apr;17(2):275-290. [doi: 10.1007/s11701-022-01425-6] [Medline: 35727485]

XSL•FO
RenderX

24. Kiely DJ, Gotlieb WH, Lau S, et al. Virtual reality robotic surgery simulation curriculum to teach robotic suturing: a randomized controlled trial. J Robot Surg 2015 Sep;9(3):179-186. [doi: 10.1007/s11701-015-0513-4] [Medline: 26531197]
25. Sridhar AN, Briggs TP, Kelly JD, Nathan S. Training in robotic surgery-an overview. Curr Urol Rep 2017 Aug;18(8):58. [doi: 10.1007/s11934-017-0710-y] [Medline: 28647793]
26. Brunner S, Kröplin J, Meyer HJ, Schmitz-Rixen T, Fritz T. Use of surgical simulators in further education-a nationwide analysis in Germany. Chirurg 2021 Nov;92(11):1040-1049. [doi: 10.1007/s00104-020-01332-2] [Medline: 33399900]
27. Kröplin J, Zauner EU, Dopp H, et al. Training strategies for a sustainable medical care: a survey among assistant and chief physicians in a tertiary care hospital in Germany. Innov Surg Sci 2020 Dec;5(3-4):20200024. [doi: 10.1515/iss-2020-0024] [Medline: 33506099]
28. Jumreornvong O, Yang E, Race J, Appel J. Telemedicine and medical education in the age of COVID-19. Acad Med 2020 Dec;95(12):1838-1843. [doi: 10.1097/ACM.0000000000003711] [Medline: 32889946]
29. Street SE, Gilliland KO, McNeil C, Royal K. The flipped classroom improved medical student performance and satisfaction in a pre-clinical physiology course. MedSciEduc 2015 Mar;25(1):35-43. [doi: 10.1007/s40670-014-0092-4]
30. Hew KF, Lo CK. Flipped classroom improves student learning in health professions education: a meta-analysis. BMC Med Educ 2018 Mar 15;18(1):38. [doi: 10.1186/s12909-018-1144-z] [Medline: 29544495]
31. Verhey JT, Haglin JM, Verhey EM, Hartigan DE. Virtual, augmented, and mixed reality applications in orthopedic surgery. Int J Med Robot 2020 Apr;16(2):e2067. [doi: 10.1002/rcs.2067] [Medline: 31867864]
32. Cho B, Geng E, Arvind V, et al. Understanding artificial intelligence and predictive analytics: a clinically focused review of machine learning techniques. JBJS Rev 2022 Mar 18;10(3). [doi: e21.00142] [Medline: 35302963]

Abbreviations

AI: artificial intelligence

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna; submitted 30.07.23; peer-reviewed by C Tsou, F Chiabrando; revised version received 08.02.24; accepted 13.02.24; published 15.04.24.

Please cite as:
Kröplin J, Maier L, Lenz JH, Romeike B
Knowledge Transfer and Networking Upon Implementation of a Transdisciplinary Digital Health Curriculum in a Unique Digital Health Training Culture: Prospective Analysis
JMIR Med Educ 2024;10:e51389
URL: https://mededu.jmir.org/2024/1/e51389
doi:10.2196/51389

© Juliane Kröplin, Leonie Maier, Jan-Hendrik Lenz, Bernd Romeike. Originally published in JMIR Medical Education (https://mededu.jmir.org), 15.4.2024. 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.

XSL•FO
RenderX

Hospital Use of a Web-Based Clinical Knowledge Support System and In-Training Examination Performance Among Postgraduate Resident Physicians in Japan: Nationwide Observational Study

Koshi Kataoka[1], MMSc; Yuji Nishizaki[1], MPH, MD, PhD; Taro Shimizu[2], MSc, MBA, MPH, MD, PhD; Yu Yamamoto[3], MD; Kiyoshi Shikino[4], MHPE, MD, PhD; Masanori Nojima[5], MPH, MD, PhD; Kazuya Nagasaki[6], MD, PhD; Sho Fukui[7], MPH, MD; Sho Nishiguchi[8], MD; Kohta Katayama[9,10], MD, PhD; Masaru Kurihara[11], MD, PhD; Rieko Ueda[12], MS, PhD; Hiroyuki Kobayashi[6], MD, PhD; Yasuharu Tokuda[13,14], MPH, MD, PhD

1
2
3
4
5
6
7
8
9
10
11
12
13
14

Corresponding Author:
Yuji Nishizaki, MPH, MD, PhD

Abstract

Background: The relationship between educational outcomes and the use of web-based clinical knowledge support systems in teaching hospitals remains unknown in Japan. A previous study on this topic could have been affected by recall bias because of the use of a self-reported questionnaire.

Objective: We aimed to explore the relationship between the use of the Wolters Kluwer UpToDate clinical knowledge support system in teaching hospitals and residents' General Medicine In-Training Examination (GM-ITE) scores. In this study, we objectively evaluated the relationship between the total number of UpToDate hospital use logs and the GM-ITE scores.

Methods: This nationwide cross-sectional study included postgraduate year–1 and –2 residents who had taken the examination in the 2020 academic year. Hospital-level information was obtained from published web pages, and UpToDate hospital use logs were provided by Wolters Kluwer. We evaluated the relationship between the total number of UpToDate hospital use logs and residents' GM-ITE scores. We analyzed 215 teaching hospitals with at least 5 GM-ITE examinees and hospital use logs from 2017 to 2019.

Results: The study population consisted of 3013 residents from 215 teaching hospitals with at least 5 GM-ITE examinees and web-based resource use log data from 2017 to 2019. High-use hospital residents had significantly higher GM-ITE scores than low-use hospital residents (mean 26.9, SD 2.0 vs mean 26.2, SD 2.3; *P*=.009; Cohen *d*=0.35, 95% CI 0.08-0.62). The GM-ITE scores were significantly correlated with the total number of hospital use logs (Pearson *r*=0.28; *P*<.001). The multilevel analysis revealed a positive association between the total number of logs divided by the number of hospital physicians and the GM-ITE scores (estimated coefficient=0.36, 95% CI 0.14-0.59; *P*=.001).

Conclusions: The findings suggest that the development of residents' clinical reasoning abilities through UpToDate is associated with high GM-ITE scores. Thus, higher use of UpToDate may lead physicians and residents in high-use hospitals to increase the implementation of evidence-based medicine, leading to high educational outcomes.

(JMIR Med Educ 2024;10:e52207) doi:10.2196/52207

XSL•FO
RenderX

KEYWORDS

clinical knowledge support system; GM-ITE; postgraduate clinical resident; in-training examination performance; exam; exams; examination; examinations; resident; residents; cross-sectional; national; nationwide; postgraduate; decision support; point-of-care; UpToDate; DynaMed; knowledge support; medical education; performance; information behavior; information behaviour; information seeking; teaching; pedagogy; pedagogical; log; logs; usage; evidence-based medicine; EBM; educational; decision support system; clinical decision support; Japan; General Medicine In-Training Examination

Introduction

Sir William Osler [1] stated that "to study the phenomena of disease without books is to sail in an uncharted sea, while to study books without patients is not to go to sea at all." Self-learning is known to develop basic clinical skills [2-4], and several studies have demonstrated the effectiveness of web-based clinical knowledge support systems. For example, a study examining the US Residency Internal Medicine In-Training Examination (IM-ITE) score reports a 3.7% increase in the IM-ITE score per 100 hours of UpToDate use [5]. In addition, UpToDate users are more satisfied with their answer accuracy, interaction, and overall satisfaction than PubMed Clinical Queries users [6]. Thus, UpToDate may be effective at the hospital level because hospitals using UpToDate have been reported to show a significantly shorter length of stay for patients [7-9]. UpToDate is already the most widely used web-based clinical knowledge support system among residents (65.5%) and the third most used system among physicians (40.4%) [9].

The General Medicine In-Training Examination (GM-ITE) is an in-training examination developed to provide residents and training program directors with an objective, reliable, and valid assessment of clinical knowledge during training. It uses the same methodology as the IM-ITE [10-12] and comprises the following 4 domains: medical interview/professionalism, symptomatology/clinical reasoning, clinical procedures, and disease knowledge. The examinations consist of 60 questions (6 on medical interview/professionalism, 15 on symptomatology/clinical reasoning, 15 on clinical procedure, and 24 on disease knowledge) and include video- and audio-format questions. The GM-ITE was first introduced in 2011 by the Japan Institute for Advancement of the Medical Education Program (JAMEP), a nonprofit organization, and is administered annually. The questions are prepared annually by a committee of experienced physicians, and peers are reviewed by an independent committee. The examinations are open to the residents of teaching hospitals that have applied to offer the examinations [13,14].

We previously reported that self-study time and use of UpToDate had positive relationships with GM-ITE scores [4]. However, those findings could have been affected by recall bias because of the use of a self-reported questionnaire, which meant that objectivity could not be guaranteed. In this study, therefore, we objectively evaluated the relationship between the total number of hospital use logs in UpToDate and the GM-ITE scores of hospital residents.

Hospital use logs were used because residents have several opportunities to acquire second-hand knowledge from their supervisors, reflecting the evidence-based medicine (EBM) culture of teaching hospitals. The introduction of clinical knowledge support systems has recently increased among resident and senior doctors, although the frequency of use is low because of language barriers and is far from the global standard [9]. The postgraduate 2-year residency system was established in 2004 in Japan. The use of the *Yanegawara* ("tiled roof" in Japanese) style of education, in which senior doctors teach resident physicians and postgraduate year (PGY)–2 residents teach PGY-1 residents based on EBM using web-based medical resources, such as UpToDate, has also become widespread [15]. The merit of the *Yanegawara*-style education is the aspect of teaching among residents with close grade levels. Internationally, peer teaching or peer tutor systems have been shown to be effective in medical education [16,17].

The aim of this study was to evaluate the correlation between the total number of UpToDate hospital use logs and the GM-ITE scores of resident physicians objectively.

Methods

Study Design and Population

We conducted a nationwide observational study of postgraduate residents in Japan using both mean GM-ITE scores and the total number of UpToDate hospital use logs to examine their relationship. The 2020 GM-ITE and self-reported questionnaire were conducted between January 13 and 31, 2021, and the data were collected during the same period. We accessed the data set for research purposes on June 16, 2021.

In Japan, postgraduate resident physicians are required to undergo at minimum a 2-year postgraduate residency program after 6 years of undergraduate medical school. In the program, the resident physicians rotate around 7 clinical departments: internal medicine, surgery, emergency medicine, pediatrics, obstetrics and gynecology, psychiatry, and community-based medicine. The Ministry of Health, Labour and Welfare has established guidelines for postgraduate clinical training programs to teach communication skills, professionalism, and ethics, in addition to basic clinical knowledge and skills, to resident physicians. Medical students in their final year of an undergraduate medical program can apply for the postgraduate residency program at more than approximately 1000 clinical teaching hospitals in Japan using a web-based matching system [18].

Measurements

We collected hospital-level information (number of physicians, monthly salary, number of ambulances, number of permitted beds, type of tertiary emergency care, location, and type of hospital) from published web pages. The hospital use logs of the web-based clinical knowledge support system (UpToDate) in the 3 years from 2017 to 2019 were provided by Wolters

XSL•FO
RenderX

Kluwer. UpToDate log data were defined as the number of topic review page views. We also collected GM-ITE scores. We hypothesized that supervisors' use of UpToDate reflects the culture of EBM resident education at each teaching hospital. Furthermore, we decided to use UpToDate hospital use logs from 2017 to 2019 to examine their association with the 2020 GM-ITE scores because educational effects are not immediately reflected after an intervention. Resident-level information (sex, grade, number of monthly emergency department duties, average number of patients in charge, general medicine department rotation, self-study time, and weekly duty hours) were obtained using a self-reported questionnaire administered immediately after the GM-ITE. These variables were selected based on previous studies [19-21].

Statistical Analyses

Hospitals were classified as low or high use according to their UpToDate hospital use logs. The total number of use logs was divided by the number of physicians and was log-transformed into base 2. The monthly salary, number of ambulances, and number of permitted beds were also log-transformed into base 2. Low-use hospitals were defined as those with fewer than the median log-transformed number of hospital use logs, whereas high-use hospitals were defined as those with greater than or equal to the median log-transformed number of hospital use logs. Differences between low-use and high-use hospitals were examined for statistical significance using the Student 2-tailed *t* test. Categorical variables were compared using the χ^2 test and presented as frequencies with percentages. The effect size (Cohen *d*) was estimated using the median difference between low- and high-use hospitals divided by the pooled SD—a value of 0.2 was considered a small effect, 0.5 was considered a medium effect, and 0.8 was considered a large effect [22]. Hospital-level analysis was performed using scatter plots to examine the association between the mean GM-ITE score and the number of UpToDate hospital use logs aggregated at the hospital level. We analyzed the association between the GM-ITE scores and the total number of UpToDate hospital use logs in each hospital over 3 years, using a linear multilevel regression model. The multilevel analysis was adjusted for sex, location, and type of hospital in addition to statistically significant factors in the univariate analysis. In those analyses, the domain of medical interview/professionalism in the GM-ITE was excluded from the analysis because we believed that it was not a clinical skill that could be improved using UpToDate. All analyses were performed using SAS software (version 9.4; SAS Institute Inc).

Ethical Considerations

This study was performed in accordance with the principles of the Declaration of Helsinki and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines. All the methods followed the *Ethical Guidelines for Medical and Health Research Involving Human Subjects*. Informed consent was obtained from each participant after clarifying the explanatory research document, including data anonymization and voluntary participation. Only participants who provided consent were included in this study, and they were also provided an opportunity to opt out. The study was approved by the Ethics Review Board of JAMEP (approval 21-1).

Results

The 2020 GM-ITE was offered at 593 teaching hospitals nationwide, and 7669 residents took the exams. A total of 6816 residents from 588 teaching hospitals participated in the survey on the training environment. The study population consisted of 3013 residents from 215 teaching hospitals with at least 5 GM-ITE examinees and web-based resource use log data from 2017 to 2019. Hospitals in all regions of Japan, namely Hokkaido, Tohoku, Kanto, Chubu, Kinki, Chugoku, Shikoku, Kyushu, and Okinawa, were included. The mean number of GM-ITE examinees per hospital was 14.1 (SD 8.6).

The hospital-level information is presented in Table 1. The mean GM-ITE score of all the hospitals was 26.5 (SD 2.2); of the 215 hospitals, 115 (53.5%) were secondary care hospitals, 159 (74%) were located in rural areas, and 204 (94.9%) were community-based hospitals. Residents of high-use hospitals achieved significantly higher GM-ITE scores than those of low-use hospitals (mean 26.9, SD 2.0 vs mean 26.2, SD 2.3; *P*=.009; Cohen *d*=0.35, 95% CI 0.08-0.62). Monthly salary (in JPY ¥100,000; JPY ¥100=US $0.64) was significantly higher in low-use hospitals than high-use hospitals (mean 3.7, SD 0.8 vs mean 3.3, SD 0.7; *P*<.001). The resident-level information is presented in Multimedia Appendix 1; 68.5% (2076/3031) were male and 50.5% (1531/3031) were PGY-2 residents.

Correlations between total use in 3 years divided by the number of physicians and GM-ITE scores were analyzed (Figure 1). The mean GM-ITE hospital score was significantly correlated with the total number of UpToDate hospital use logs (Pearson *r*=0.28, *P*<.001; Spearman *r*=0.27, *P*<.001). The linear regression function was $y = 24.13 + 0.66 \times \log_2(\text{total use/number of physicians})$; therefore, the difference in the mean GM-ITE score between the total use divided by the number of physicians at values of 8 and 128 was 2.64 (Figure 1). Multimedia Appendix 2 shows the relationship between GM-ITE scores and hospital- and resident-level information using an univariate analysis. The statistically significant factors were log-transformed total number of hospital use logs in 3 years divided by the number of physicians (*P*<.001), log-transformed number of ambulances (*P*<.001), log-transformed number of permitted beds (*P*=.005), type of tertiary emergency care (*P*=.01), grade (*P*<.001), number of monthly emergency department duty (*P*=.004-.046), average number of patients in charge (from *P*<.001 to *P*=.01), general medicine department rotation (*P*=.004), self-study time (*P*=.02-.04), and weekly duty hours (*P*<.001). The multilevel analysis was adjusted for all these factors in addition to sex, location, and type of hospital. Table 2 shows the relationship between GM-ITE scores and hospital- and resident-level information using a multilevel analysis. The multilevel analysis revealed a positive association between 3-year total hospital use logs and GM-ITE scores (estimated coefficient=0.36, 95% CI 0.14-0.59; *P*=.001). Multimedia Appendix 3 shows the results of the analysis of all 4 domains (medical interview/professionalism, symptomatology/clinical reasoning, clinical procedures, and disease knowledge). The result also revealed a positive association between the use of UpToDate and GM-ITE scores (estimated coefficient=0.41, 95% CI 0.18-0.65; *P*<.001).

XSL•FO
RenderX

Table . Background characteristics of the teaching hospitals.

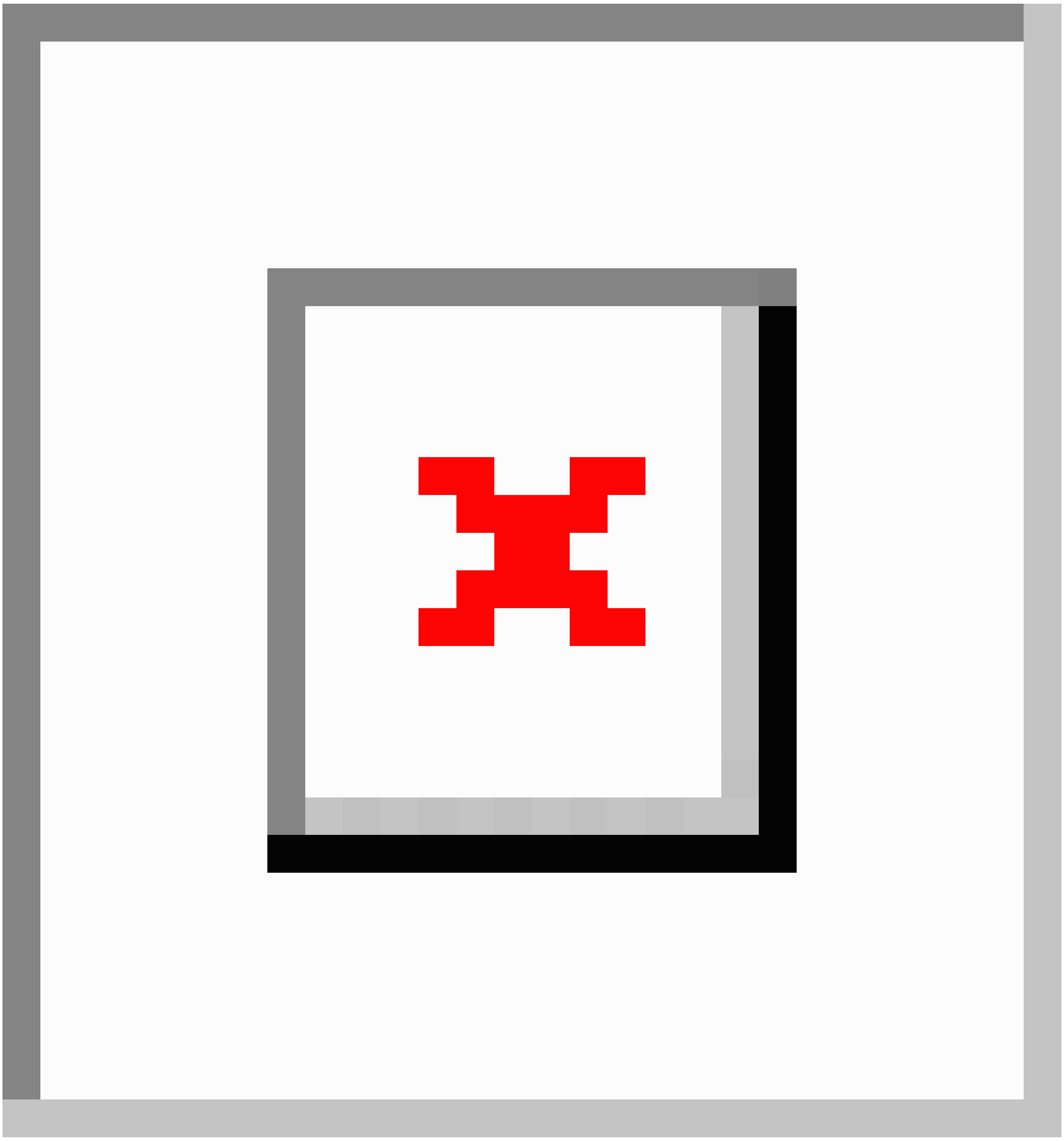
Hospital-level information		Total (N=215)	Low-use hospitals (n=107)	High-use hospitals (n=108)	*P* value
Total number of use logs of UpToDate, mean (SD)		10,485.1 (20,231.4)	2578.7 (2,278.2)	18,318.2 (26,249.5)	<.001
Number of physicians, mean (SD)		144.2 (91.4)	116.3 (61.9)	171.8 (106.6)	<.001
Total use in 3 years/number of physicians, mean (SD)		56.5 (62.2)	20.2 (8.7)	92.4 (71.1)	<.001
Log-transformed total use in 3 years/number of physicians, mean (SD)		5.2 (1.3)	4.1 (0.9)	6.3 (0.8)	<.001
Monthly salary (in JPY ¥100,000[a]), mean (SD)		3.5 (0.8)	3.7 (0.8)	3.3 (0.7)	<.001
Log-transformed monthly salary (in JPY ¥100,000), mean (SD)		1.8 (0.3)	1.9 (0.3)	1.7 (0.3)	<.001
Number of ambulances, mean (SD)		4882.6 (2399.2)	4462.0 (2183.9)	5299.4 (2536.8)	.01
Log-transformed number of ambulances, mean (SD)		12.1 (0.8)	11.9 (0.7)	12.2 (0.8)	.02
Number of permitted beds, mean (SD)		497.8 (166.8)	461.2 (158.9)	534.0 (167.3)	.001
Log-transformed number of permitted beds, mean (SD)		8.9 (0.5)	8.8 (0.5)	9.0 (0.4)	<.001
Type of tertiary emergency care, n (%)					.02
	Tertiary medical care	100 (46.5)	41 (38.3)	59 (54.6)	
	Secondary care	115 (53.5)	66 (61.7)	49 (45.4)	
Location, n (%)					.13
	Urban area	56 (26)	23 (21.5)	33 (30.6)	
	Rural area	159 (74)	84 (78.5)	75 (69.4)	
Type of hospital, n (%)					.03
	University hospital	11 (5.1)	2 (1.9)	9 (8.3)	
	Community-based hospital	204 (94.9)	105 (98.1)	99 (91.7)	
GM-ITE[b] score, mean (SD)		26.5 (2.2)	26.2 (2.3)	26.9 (2.0)	.009

[a]JPY ¥100=US $0.64.

[b]GM-ITE: General Medicine In-Training Examination.

XSL•FO
RenderX

Figure 1. Correlation between total use of UpToDate and mean General Medicine In-Training Examination (GM-ITE) scores.



XSL•FO
RenderX

Table . Factors related to General Medicine In-Training Examination (GM-ITE) scores (multilevel analysis).

Factors			Estimated coefficient (95% CI)	*P* value
Hospital-level information				
	Log-transformed total use of UpToDate in 3 years/number of physicians		0.36 (0.14 to 0.29)	.001
	Log-transformed number of ambulances		0.34 (−0.08 to 0.77)	.11
	Log-transformed number of permitted beds		0.36 (−0.39 to 1.12)	.35
	Type of tertiary emergency care			
		Tertiary medical care	Reference	Reference
		Secondary care	−0.20 (−0.80 to 0.40)	.51
	Location			
		Urban area	Reference	Reference
		Rural area	0.72 (0.09 to 1.35)	.02
	Type of hospital			
		University hospital	Reference	Reference
		Community-based hospital	0.52 (−0.82 to 1.88)	.44
Resident-level information				
	Sex			
		Male	Reference	Reference
		Female	0.08 (−0.28 to 0.45)	.66
	Grade			
		PGY[a]-1	Reference	Reference
		PGY-2	0.81 (0.45 to 1.17)	<.001
	Number of monthly emergency department duties			
		0	Reference	Reference
		1-2	0.46 (−0.64 to 1.58)	.41
		3-5	0.81 (−0.28 to 1.92)	.15
		>6	0.48 (−0.75 to 1.72)	.45
		Unknown	−0.42 (−3.38 to 2.54)	.78
	Average number of patients in charge			
		0-4	Reference	Reference
		5-9	0.76 (0.30 to 1.21)	.001
		10-14	0.62 (−0.10 to 1.35)	.09
		>15	1.20 (−0.07 to 2.47)	.06
		Unknown	−1.03 (−2.27 to 0.20)	.10
	General medicine department rotation			
		Yes	Reference	Reference
		No	−0.12 (−0.54 to 0.29)	.56
	Self-study time (min/d)			
		None	Reference	Reference
		0-30	−0.10 (−1.13 to 0.92)	.84
		31-60	0.31 (−0.70 to 1.33)	.54

XSL•FO
RenderX

Factors		Estimated coefficient (95% CI)	*P* value
	61-90	0.94 (−0.12 to 2.01)	.08
	>91	1.03 (−0.25 to 2.32)	.12
Weekly duty hours (h/wk)			
	0-59	Reference	Reference
	60-79	0.67 (0.27 to 1.07)	<.001
	>80	−0.10 (−0.60 to 0.38)	.67

[a]PGY: postgraduate year.

Discussion

We objectively evaluated the relationship between hospital use logs of the web-based clinical knowledge support system, UpToDate, at teaching hospitals and residents' GM-ITE scores. Residents of high-use hospitals achieved significantly higher GM-ITE scores, an objective index of the basic clinical ability of residents, than those of low-use hospitals. There are 2 main situations in which residents use web-based clinical knowledge support systems such as UpToDate: "actual clinical sittings such as bedside and outpatient care" and "self-improvement." UpToDate is useful in situations where immediate evidence-based care must be provided to the patient in front of a resident [23]. In terms of self-improvement, among both residents and senior doctors, web-based clinical knowledge support systems could lead to the development of basic clinical abilities because they can collect the latest information more efficiently than from textbooks [9].

The use of a web-based clinical knowledge support system is associated with high GM-ITE scores owing to the residents' knowledge of clinical reasoning. The clinical training guidelines of the Ministry of Health, Labour and Welfare highlight the importance of studying clinical reasoning and problem-solving abilities, and residents are required to have the ability to (1) make a differential diagnosis and initial response to high-frequency symptoms through an appropriate clinical reasoning process and (2) collect patient information and make clinical decisions in consideration of the patient's intentions and quality of life based on the latest medical knowledge. Residents constantly acquire the latest medical knowledge and use evidence-based and their own experiences to solve clinical problems. The questions in the GM-ITE include clinical reasoning questions in accordance with the guidelines of the Ministry of Health, Labour and Welfare. As UpToDate contains a series of flows that "list differential diagnoses from symptoms and link them to accurate diagnoses," we believe the use of UpToDate would help residents develop their clinical reasoning abilities. Therefore, we speculate that the development of residents' clinical reasoning abilities through UpToDate is associated with high GM-ITE scores.

Japanese residents are required to gain greater outpatient clinical experience to acquire basic clinical skills, including communication and clinical reasoning, during this 2-year training period. The postgraduate clinical residency system has been revised regularly, and the latest revision in 2020 requires a 1-month general outpatient training rotation for residents. Therefore, outpatient training is becoming increasingly important in Japanese clinical resident education programs.

Previous studies have demonstrated the usefulness of web-based clinical knowledge support systems in outpatient care. A comparison of outpatient diagnostic errors made by physicians with and without the use of UpToDate shows that diagnostic errors were significantly reduced in the case of physicians who used UpToDate [24]. Internal medicine residents' responses to patient-specific questions encountered in outpatient settings have been known to improve their clinical skills and patient care decisions. UpToDate has been reported to be the second most commonly used tool for gathering information after MEDLINE and is an extremely helpful information source [25]. We speculate that the GM-ITE includes questions regarding outpatient care that are associated with the development of clinical residents' outpatient care abilities and high GM-ITE scores.

Factors significantly and positively associated with GM-ITE scores in the multilevel analysis results, besides the use of UpToDate, were location, PGY-2 grade, average number of patients in charge, and weekly duty hours. Residents of rural teaching hospitals may have the opportunity to examine more patients, because the number of physicians affiliated with rural teaching hospitals is lower than that with urban teaching hospitals. Consequently, they may acquire greater clinical experience and knowledge [26]. Regarding the difference in grades, we believe that the difference in clinical experience is directly reflected in GM-ITE scores. This finding is consistent with the results of our previous study [4]. We recommend that residents develop a proactive attitude toward patient care because basic clinical skills tend to develop with daily patient management. The results of the multilevel analysis showed that 60-79 weekly duty hours were significantly and positively associated with GM-ITE scores. This finding supports our previous results [20]; we believe there are optimal working hours for improving clinical competency.

The development of residents' basic clinical skills does not require many supervisors; however, high-quality and highly productive education is necessary. A previous Japanese study showed that education delivered by a limited number of supervisors was more likely to develop residents' basic clinical skills [27]. Furthermore, residents who rotated in general medicine achieved higher GM-ITE scores [4]. We believe the following factors are required for future residency education: generalist residency education by general medicine specialists; use of productive web-based clinical knowledge support

XSL•FO
RenderX

systems, such as UpToDate; EBM culture; and the *Yanegawara*-style educational system.

This study has a few limitations. First, the scores were examined among a limited number of GM-ITE examinees. Although there are approximately 18,000 PGY-1 and -2 residents in Japan, only 3013 residents were analyzed in this study, accounting for approximately one-sixth (16.7%) of the total population. In addition, as the GM-ITE is a voluntary examination, a bias toward highly motivated residents taking the exam may exist. Therefore, the generalizability of this study is not ensured. Second, causal relationships could not be guaranteed because the study design was cross-sectional. To control for selection bias and to assess causality, we believe that planning a randomized controlled trial targeting nationwide resident physicians is necessary. In this randomized controlled trial, the GM-ITE scores would be the primary end point, and the intervention would control for the presence or absence of web-based clinical knowledge support systems. Third, we did not assess the baseline clinical skills of the GM-ITE examinees in this study, and differences in undergraduate medical school education could have impacted the study results. Fourth, the hospital use logs did not include detailed information, such as user information and access time. It was not possible to identify user-specific information, such as residents, physicians, and co-medical professionals (eg, nurses), from the log data. Fifth, the results came from a single web-based clinical knowledge support system. Although there are other web-based clinical knowledge support systems that aid residents, physicians, and paramedical workers, we did not compare UpToDate with them. Some resident physicians in Japan may use web-based clinical knowledge support systems other than UpToDate. Although we could not obtain data on systems other than UpToDate for this study, we aim to include them in our next research project to validate the current results.

In conclusion, residents in high-use hospitals had significantly higher GM-ITE scores than those in low-use hospitals, indicating that GM-ITE scores are associated with web-based resource use logs. A previous study showed an association between web-based resource use and resident GM-ITE scores using data from a self-reported survey of clinical residents [4]. Our findings are consistent with those of previous studies and include data that ensure objectivity. Frequent use of web-based clinical knowledge support systems will increase the likelihood of physicians, including faculty, senior, and junior residents, implementing EBM and senior physicians teaching juniors using the *Yanegawara*-style education, which may lead to higher educational outcomes.

Acknowledgments

We thank the members of Japan Institute for Advancement of Medical Education Program (JAMEP) for their assistance. We would like to thank Editage for English language editing and Wolters Kluwer for providing log data on web-based resource use. This work was supported by the Health, Labour and Welfare Policy Grants from the Ministry of Health, Labour and Welfare's Research on Region Medical (21IA2004).

Data Availability

The data are not available for sharing because we did not obtain relevant consent from the participants to publish them. The corresponding author will respond to inquiries on data analyses.

Authors' Contributions

K Kataoka and MN had full access to all data in the study and take responsibility for the integrity of the data and accuracy of the data analysis. YN and YT contributed to study concept and design. K Kataoka, YN, TS, YY, KS, MN, KN, SF, SN, K Katayama, MK, RU, HK, and YT contributed to the acquisition, analysis, and interpretation of data. K Kataoka and YN contributed to manuscript drafting. YN, HK, and YT contributed to critical revision of the manuscript for important intellectual content. K Kataoka and MN contribute to statistical analysis. RU and YN contributed to administrative, technical, or material support. YN, HK, and YT contributed to supervision.

Conflicts of Interest

YN received an honorarium from the Japan Institute for Advancement of Medical Education Program (JAMEP) as a General Medicine In-Training Examination (GM-ITE) project manager. YT is a director of JAMEP. YT, HK, and KS received honoraria for delivering lectures for JAMEP. TS, YY, KS, and SF received honoraria from JAMEP as exam preparers for the GM-ITE. YN and YT received honoraria from Wolters Kluwer for delivering the Wolters Kluwer lecture. K Kataoka, MN, KN, SN, K Katayama, MK, and RU declare no competing interests.

Multimedia Appendix 1

Background characteristics of the residents.

[DOCX File, 29 KB - mededu_v10i1e52207_app1.docx]

Multimedia Appendix 2

XSL•FO
RenderX

Factors related to General Medicine In-Training Examination scores (univariate analysis).
[DOCX File, 30 KB - mededu_v10i1e52207_app2.docx]

Multimedia Appendix 3
Factors related to General Medicine In-Training Examination scores, including the 4 domains (multilevel analysis).
[DOCX File, 31 KB - mededu_v10i1e52207_app3.docx]

References

1. Osler W. Address on the dedication of the new building. Boston Med Surg J 1901 Jan 17;144(3):60-61. [doi: 10.1056/NEJM190101171440304]
2. Bull DA, Stringham JC, Karwande SV, Neumayer LA. Effect of a resident self-study and presentation program on performance on the Thoracic Surgery In-Training Examination. Am J Surg 2001 Feb;181(2):142-144. [doi: 10.1016/s0002-9610(00)00567-5] [Medline: 11425055]
3. Philip J, Whitten CW, Johnston WE. Independent study and performance on the Anesthesiology In-Training Examination. J Clin Anesth 2006 Sep;18(6):471-473. [doi: 10.1016/j.jclinane.2006.01.003] [Medline: 16980169]
4. Nishizaki Y, Shimizu T, Shinozaki T, et al. Impact of general medicine rotation training on the In-Training Examination scores of 11, 244 Japanese resident physicians: a nationwide multi-center cross-sectional study. BMC Med Educ 2020 Nov 13;20(1):426. [doi: 10.1186/s12909-020-02334-8] [Medline: 33187497]
5. McDonald FS, Zeger SL, Kolars JC. Factors associated with medical knowledge acquisition during internal medicine residency. J Gen Intern Med 2007 Jul;22(7):962-968. [doi: 10.1007/s11606-007-0206-4] [Medline: 17468889]
6. Sayyah Ensan L, Faghankhani M, Javanbakht A, Ahmadi SF, Baradaran HR. To compare PubMed Clinical Queries and UpToDate in teaching information mastery to clinical residents: a crossover randomized controlled trial. PLoS One 2011;6(8):e23487. [doi: 10.1371/journal.pone.0023487] [Medline: 21858142]
7. Bonis PA, Pickens GT, Rind DM, Foster DA. Association of a clinical knowledge support system with improved patient safety, reduced complications and shorter length of stay among Medicare beneficiaries in acute care hospitals in the United States. Int J Med Inform 2008 Nov;77(11):745-753. [doi: 10.1016/j.ijmedinf.2008.04.002] [Medline: 18565788]
8. Isaac T, Zheng J, Jha A. Use of UpToDate and outcomes in US hospitals. J Hosp Med 2012 Feb;7(2):85-90. [doi: 10.1002/jhm.944] [Medline: 22095750]
9. Sakai Y, Sato Y, Sato M, Watanabe M. Clinical usefulness of library and information services in Japan: the detailed use and value of information in clinical settings. PLoS One 2018 Jun 28;13(6):e0199944. [doi: 10.1371/journal.pone.0199944] [Medline: 29953527]
10. Garibaldi RA, Subhiyah R, Moore ME, Waxman H. The In-Training Examination in Internal Medicine: an analysis of resident performance over time. Ann Intern Med 2002 Sep 17;137(6):505-510. [doi: 10.7326/0003-4819-137-6-200209170-00011] [Medline: 12230352]
11. Kanna B, Gu Y, Akhuetie J, Dimitrov V. Predicting performance using background characteristics of international medical graduates in an inner-city university-affiliated internal medicine residency training program. BMC Med Educ 2009 Jul 13;9:42. [doi: 10.1186/1472-6920-9-42] [Medline: 19594918]
12. Perez JA, Greer S. Correlation of United States Medical Licensing Examination and Internal Medicine In-Training Examination performance. Adv Health Sci Educ Theory Pract 2009 Dec;14(5):753-758. [doi: 10.1007/s10459-009-9158-2] [Medline: 19283500]
13. Nagasaki K, Nishizaki Y, Nojima M, et al. Validation of the General Medicine In-Training Examination using the Professional and Linguistic Assessments Board examination among postgraduate residents in Japan. Int J Gen Med 2021 Oct 7;14:6487-6495. [doi: 10.2147/IJGM.S331173] [Medline: 34675616]
14. Nishizaki Y, Nozawa K, Shinozaki T, et al. Difference in the General Medicine In-Training Examination score between community-based hospitals and university hospitals: a cross-sectional study based on 15,188 Japanese resident physicians. BMC Med Educ 2021 Apr 15;21(1):214. [doi: 10.1186/s12909-021-02649-0] [Medline: 33858403]
15. Yano R. Yanegawara style PBL tutorial education in Kinjo Gakuin University [Article in Japanese]. Jpn J Pharm Educ 2018 Sep 28;2. [doi: 10.24489/jjphe.2018-008]
16. Ten Cate O, Durning S. Peer teaching in medical education: twelve reasons to move from theory to practice. Med Teach 2007 Sep;29(6):591-599. [doi: 10.1080/01421590701606799] [Medline: 17922354]
17. Burgess A, McGregor D, Mellis C. Medical students as peer tutors: a systematic review. BMC Med Educ 2014 Jun 9;14:115. [doi: 10.1186/1472-6920-14-115] [Medline: 24912500]
18. Japan residency matching program interim report [Article in Japanese]. Japan Residency Matching Program. 2021. URL: https://jrmp2.s3.ap-northeast-1.amazonaws.com/chukan/2021chukan.pdf [accessed 2024-05-14]
19. Nishizaki Y, Shinozaki T, Kinoshita K, Shimizu T, Tokuda Y. Awareness of diagnostic error among Japanese residents: a nationwide study. J Gen Intern Med 2018 Apr;33(4):445-448. [doi: 10.1007/s11606-017-4248-y] [Medline: 29256086]
20. Nagasaki K, Nishizaki Y, Shinozaki T, et al. Impact of the resident duty hours on In-Training Examination score: a nationwide study in Japan. Med Teach 2022 Apr;44(4):433-440. [doi: 10.1080/0142159X.2021.2003764] [Medline: 34818129]

XSL•FO
RenderX

21. Kinoshita K, Tsugawa Y, Shimizu T, et al. Impact of inpatient caseload, emergency department duties, and online learning resource on General Medicine In-Training Examination scores in Japan. Int J Gen Med 2015;8:355-360. [doi: 10.2147/IJGM.S81920] [Medline: 26586961]
22. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Lawrence Erlbaum Associates, Inc; 1988.
23. Phua J, See KC, Khalizah HJ, Low SP, Lim TK. Utility of the electronic information resource UpToDate for clinical decision-making at bedside rounds. Singapore Med J 2012 Feb;53(2):116-120. [Medline: 22337186]
24. Shimizu T, Nemoto T, Tokuda Y. Effectiveness of a clinical knowledge support system for reducing diagnostic errors in outpatient care in Japan: a retrospective study. Int J Med Inform 2018 Jan;109:1-4. [doi: 10.1016/j.ijmedinf.2017.09.010] [Medline: 29195700]
25. Schilling LM, Steiner JF, Lundahl K, Anderson RJ. Residents' patient-specific clinical questions: opportunities for evidence-based learning. Acad Med 2005 Jan;80(1):51-56. [doi: 10.1097/00001888-200501000-00013] [Medline: 15618093]
26. Shimizu T, Tsugawa Y, Tanoue Y, et al. The hospital educational environment and performance of residents in the General Medicine In-Training Examination: a multicenter study in Japan. Int J Gen Med 2013 Jul 29;6:637-640. [doi: 10.2147/IJGM.S45336] [Medline: 23930077]
27. Mizuno A, Tsugawa Y, Shimizu T, et al. The impact of the hospital volume on the performance of residents on the General Medicine In-Training Examination: a multicenter study in Japan. Intern Med 2016;55(12):1553-1558. [doi: 10.2169/internalmedicine.55.6293] [Medline: 27301504]

Abbreviations

EBM: evidence-based medicine
GM-ITE: General Medicine In-Training Examination
IM-ITE: Internal Medicine In-Training Examination
JAMEP: Japan Institute for Advancement of the Medical Education Program
PGY: postgraduate year
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna; submitted 28.08.23; peer-reviewed by J Walsh, MD Pumpo, M Zahmatkeshan; revised version received 02.05.24; accepted 02.05.24; published 30.05.24.

Please cite as:
Kataoka K, Nishizaki Y, Shimizu T, Yamamoto Y, Shikino K, Nojima M, Nagasaki K, Fukui S, Nishiguchi S, Katayama K, Kurihara M, Ueda R, Kobayashi H, Tokuda Y
Hospital Use of a Web-Based Clinical Knowledge Support System and In-Training Examination Performance Among Postgraduate Resident Physicians in Japan: Nationwide Observational Study
JMIR Med Educ 2024;10:e52207
URL: https://mededu.jmir.org/2024/1/e52207
doi:10.2196/52207

© Koshi Kataoka, Yuji Nishizaki, Taro Shimizu, Yu Yamamoto, Kiyoshi Shikino, Masanori Nojima, Kazuya Nagasaki, Sho Fukui, Sho Nishiguchi, Kohta Katayama, Masaru Kurihara, Rieko Ueda, Hiroyuki Kobayashi, Yasuharu Tokuda. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.5.2024. 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.

XSL•FO
RenderX

A Call for a Health Data–Informed Workforce Among Clinicians

Joy Doll[1], OTD, OTR/L; A Jerrod Anzalone[2], PhD; Martina Clarke[3], PhD; Kathryn Cooper[3], PhD; Ann Polich[4], MD; Jacob Siedlik[5], PhD

1
2
3
4
5

Corresponding Author:
Joy Doll, OTD, OTR/L

Abstract

A momentous amount of health data has been and is being collected. Across all levels of health care, data are driving decision-making and impacting patient care. A new field of knowledge and role for those in health care is emerging—the need for a health data–informed workforce. In this viewpoint, we describe the approaches needed to build a health data–informed workforce, a new and critical skill for the health care ecosystem.

(JMIR Med Educ 2024;10:e52290) doi:10.2196/52290

KEYWORDS

health data–informed workforce; health data; health informaticist; data literacy; workforce development

Background

Health care has become a data-driven business. It is no longer acceptable that both incoming and current health care professionals and business leaders lack an understanding of the influence data has on health care delivery. The clinician coauthors listed here represent this sphere and are still learning every day. We represent the diverse background of professionals that exist in the health data space, with a wide variety of journeys into this arena [1]. “Health data” is a broad term, often referring to data collected and exchanged in electronic systems. Everyday health data are entered, exchanged, and used to make important decisions from the patient level to the systems level. Health care professionals today need an understanding of the utilization and impact of health data to optimize care delivery and interact with the many systems they encounter daily.

When we entered the health care industry over 20 years ago, we were hopeful clinicians excited to impact patients’ lives. For many of us, we quickly became disillusioned by a system that was driven not by patient outcomes but by reimbursement. Yet, we regained hope with pivotal moments, including when Don Berwick challenged health care organizations to promote quality and evidence-based medicine with the Institute for Healthcare Improvement; the proliferation of electronic health record (EHR) usage leading to the potential to share patient information across systems [2-4]; and the opportunity to move from fee-for-service to value-based payment [5]. We continue to grow in hope, as many openly discuss health equity and social determinants and drivers of health. In addition, the conversations and investments in workforce related to health data knowledge and expertise are ongoing and receiving national attention. Opportunities abound with the expansive growth of artificial intelligence and machine learning.

However, none of these impending innovations can grow and disseminate without understanding data. Gaining an understanding about health data and their use by clinicians is critical to promote the key structural aspects necessary to improve health care delivery, including interoperability, data standards, quality measures, and reimbursement for health outcomes. When we started in health care, the understanding of the impact of health data did not truly and widely exist. In our experience, we find that many clinicians are unconsciously incompetent—lacking a basic understanding of how health data are used, what health data consist of, and where data flow [6]. Unconscious incompetence occurs when the decision makers lack the true information and expertise needed to make an informed decision [6]. This lack of competence causes uninformed decision-making in the health care ecosystem, which causes more challenge. Technology and data become a burden and not a solution.

In health care, a health data–informed workforce is needed to remedy the gaps and make the important connections for positive change. In our experiences and those of our peers, we interact with clinicians who learned about informatics and health data on the job [1]. Many stories start with an interest in data and technology or some savviness with technology. These are individuals willing to lean into innovation and learn through failure. Yet, their learning curve is steep and lacks the efficiency

XSL•FO
RenderX

that a health data–informed workforce could address. The understanding of health data has become a shared team value critical to growing and expanding the evidence to support interprofessional practice. Now is the time to move beyond the early adopters and explore how we can expand the health data–informed workforce. We acknowledge previous authors that have called for this momentum to grow and call for ongoing and widespread engagement. In this viewpoint, we attempt to define the health data–informed workforce at the micro-, meso-, and macrolevels. We then offer suggestions for clinicians wanting to level up their competence in health data.

What Is a Health Data–Informed Workforce?

A health data–informed workforce includes clinicians with a basic understanding of data along with their exchange and influence on decision-making. The ideal would be to move clinicians from being unconsciously incompetent to consciously competent. However, the amount of knowledge expected is overwhelming. The complexity of health data has evolved into the field of health informatics. Multiple studies have indicated that the field of health informatics is diverse, with a wide variety of education and workplace requirements [1,7,8]. Health informatics is a field that explores the use of health data for “scientific inquiry, problem-solving, decision making” with the intent to improve health care delivery and impact [9]. Yet, health data impact every level of health care, from the micro- to macrolevel, calling upon all clinicians to hold a basic understanding.

For the purposes of this viewpoint, we consider the microlevel to be interactions with patients and clinicians; the mesolevel focuses on the infrastructure and systems in place for health data sharing; and the macrolevel addresses the impact of policy on health data. Clinicians who are data informed at each of these levels will improve the impact of health data utility and ensure that decisions made around health data and technology will facilitate positive change.

At the clinician and patient level (ie, the microlevel), data are used to make clinical decisions. The widespread adoption of EHR systems supported by the 21st Century Cures Act and provisions against information blocking in the Office of the National Coordinator of Health Information Technology’s Final Rule place a premium on data, and data literacy, in health care delivery [10]. The availability of data and the ability for patients to access their health data through patient portals and other digital applications can advance shared decision-making, promoting improved health outcomes while empowering patient’s involvement in their care. Yet, EHRs have introduced burden, and many clinicians are under information overload, which can result in health care errors [9,11]. A recent piece in the *Journal of the American Medical Association* titled “Death by Patient Portal” illustrates the love-hate relationship that occurs with much of health technology and data [12]. Data are flowing and being shared, but questions remain on how much and how to make information usable for patients and clinicians. Despite these challenges, data and technology are reported to only continue to grow in health care. This calls on clinicians to know how to access patient data in their EHRs, understand where patients track and record data, and feel comfortable translating health information to multiple levels of digital and health literacy. A health data–informed clinician knows to use tools such as health information exchanges (HIEs) to ensure that they are making clinical decisions with comprehensive patient data beyond the EHR [13]. An HIE extracts data from multiple EHRs and matches that data into a comprehensive patient record. In some health care organizations, HIEs are integrated into the EHR. They can provide quick and comprehensive patient data for clinical decision-making [14]. Due to HIEs, health care becomes more proactive and less reactive when clinicians are aware of a recent emergency department visit, for example. At the same time, HIEs can lead to information overload for providers. In addition, clinicians improve their patient experience when they have information about the patient journey and history, which an HIE can provide [15]. Patients also report a better patient experience when they are not forced to “repeat their story” or re-enter information they have already reported.

Many health care organizations use data for various reasons, including use by health care delivery systems, payers, and academic researchers. When it comes to the mesolevel, the health data–informed workforce needs to understand data governance, including understanding how, why, and when data are shared and recognizing the importance of privacy and security. Patient consent remains important to ensure patients know when and where their data are being shared and how they are being used. In addition, health technology selection and vetting, along with vendor management, is critical. Vendors can offer solutions, yet at the same time, these tools can have unintended consequences from data being entered into multiple systems, causing burden on clinicians and a lack of data completeness in a patient’s record. Clinicians need to recognize the importance of interoperability as it impacts data access and use between systems. Interoperability refers to the ability to exchange data in a useful manner. An interoperable approach reduces double documentation and siloed health data [16]. At the same time, health data have extensive protections under the Health Insurance Portability and Accountability Act (HIPAA), which requires thoughtfulness to the exchange and use of data across systems. We have witnessed too many clinicians enamored with a piece of technology without vetting its ability to further health care. Great technology that further siloes data into multiple systems and lacks expanded adoption can cause more burden and potential patient harm. We need a workforce that questions the benefits and challenges how additional technology and data can actually improve health care delivery along with their interoperability.

Health data–informed clinicians also recognize the importance and value of data standards [17]. Data standards provide a critical foundation for data exchange. Decisions are being made daily in health care organizations without the recognition or use of data standards. One example is choosing to create a health-related social need screener without considerations of existing tools or data standards work, such as that led by the Gravity Project [18]. These approaches further denigrate the system and cause a myriad of challenges to interoperability.

XSL•FO
RenderX

For the macrolevel, understanding and advocating for local and federal policies that support the proliferation of growing workforce expertise is critical for the health data–informed clinician. Clinicians need a basic recognition and understanding of how policy drives health data utility [19]. The gaps in the workforce around health informatics have been identified and acknowledged [7,20,21]. Efforts in this area have been made by the American Medical Informatics Association's 10×10 program and the federal funding to support the Public Health Informatics and Technology Workforce Program by the Office of National Coordinator of Health Information Technology. These policies and investments provide opportunities to support both current professionals and those entering the workforce, representing examples of the impact of policy on health data.

How Do Clinicians Level Up?

Overview

If this sparks something inside you, the next step is to be curious about how to develop into a health data–informed clinician. All clinicians should be on a journey as lifelong learners. Health and health care constantly change, not to mention technology and data use. Becoming more data informed does not mean getting a new degree, even though that is an option. In this next section, we share some basic aspects for those desiring to become a health data–informed clinician. Certainly, we cannot go into extensive depth, but we hope this plants seeds to grow the health data–informed workforce. Some strategies to level up are as follows.

Get to Know a Health Informaticist

No one can or is expected to know everything, which is why health care is a team sport. One strategy to help build a health data–informed workforce is for clinicians to learn the role of health informaticists. Health informatics "is the interprofessional field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem-solving, decision making, motivated by efforts to improve human health" [9]. In other words, health informatics is a wide field focused on health data and their utility to impact health care outcomes. Health informaticists hold expertise in data management, security, privacy, and governance requirements to support safe handling of protected health information [1]. They are also challenged to ensure that health data are interpreted and presented meaningfully to stakeholders, including clinicians; health care leaders; and most importantly, patients [22]. Health informaticists go by different names in different organizations, including clinical informaticist, data analyst, business analyst, etc [23,24]. Their roles and demands may vary based on where they work. However, many organizations have informatics expertise in their organization. The next step would be to include an informaticist as part of the team. They can be invaluable in selecting health technology, vendor management, training and implementation, and project implementation, not to mention data handling! They offer a wide variety of skills to a team interacting with health technology and data, including data extraction, quality metrics, data analysis, dashboard builds, etc [1].

Use an HIE

HIEs are state-based or regional infrastructures that match data across multiple EHRs to provide a comprehensive patient record [25]. The sophistication of HIEs vary, yet they are a tool available to clinicians in multiple health systems that often go underutilized. In some cases, HIEs can be queried for information on the patient. They can also be used to send and receive information on a patient to allow for more comprehensive decision-making [14,15]. Clinicians can find out if their health care organization is part of their local HIE to gain access and training on how to use an HIE to improve clinical decision-making.

Recognize the Importance of Data Standards

Clinicians have an important role in entering health data, which impacts the ability to analyze data from health data utilities. Recognizing the importance of the use of appropriate data standards is important for clinicians. In the informatics field, you often hear the term "garbage in, garbage out," and much effort has been made to extract and clean data to show the impact of quality payment programs, which have induced new health care system burden. Clinicians can work with their informatics team to ensure documentation is structured in a meaningful way. The United States Core Data for Interoperability offers guidance around common ways to document that can promote data sharing [26].

Get Some Training

There is a variety of professional organizations that can support the learning and growing of professionals targeted at clinicians. These organizations offer conferences, web-based trainings, and certifications. Some federal resources also exist, including the Office of the National Coordinator for Health Information Technology, who offers webinars and other valuable resources. Table 1 calls out some of these resources. Each organization offers a variety of training opportunities. Another option is to seek a mentor in health informatics, partnering with someone with experience to learn from.

Table . Organizations and resources.

Resource	Website
American Health Information Management Association (AHIMA)	[27]
American Medical Informatics Association (AMIA)	[28]
Civitas Networks for Health	[29]
Healthcare Information and Management Systems Society (HIMSS)	[30]
Office of the National Coordinator for Health Information Technology	[31]

XSL•FO
RenderX

For some, upskilling may be entering the field of health informatics. Academic programs exist in health information management and health informatics across the country. Many professional organizations offer discipline-tailored programming in health informatics specifically in medicine and nursing. Many programs offer web-based options and teach core skills.

It is normal to feel intimidated by the terminology and concepts. However, it is important to remember that health data are being used to drive lots of decisions. Garnering a basic understanding will improve clinical skills and help with patient advocacy to improve care delivery. Everyone can take some simple steps to become more health data informed.

What Can Educators Do?

Overview

It is impossible to know everything about the field of informatics and health data. Instead, the intent should not be about teaching all the skills but instead the critical thinking skills necessary to consider how and why technology and data can be used in health care. As a society, we need to cultivate minds that can think and problem solve for a future we do not yet exist in. As educators, we need to encourage the ability to embrace ambiguity and innovation while recognizing that human beings approach these elements in different ways that can cultivate adoption at different rates of speed. Educating a health data–informed workforce requires educators to recognize that technical and technology skills are important but not enough. The focus should include the following.

Promotion of Data Literacy

Basic data literacy involves understanding how data can be used to effect positive change in patient outcomes, cost reduction, and mitigation of caregiver burnout, among other applications. Data literacy is the ability to read and understand data. For those advanced in this area, data literacy includes communicating and sharing data in ways appropriate to the audience. Health data literacy in an informed health care workforce includes training on effective data management throughout the health data life cycle and how to traverse the knowledge discovery process, from data to information to knowledge and, ultimately, wisdom and actionable insights. The Data, Information, Knowledge and Wisdom Model provides a theoretical framework that spans from reviewing data to applying data in impactful ways [32]. A significant amount of health data is collected, and deciding what to do with it requires a deeper understanding. Educators should push learners to move beyond reviewing data to deeply engaging with them in meaningful ways to improve health care.

Ethical Use of Health Data

Data, especially health data, require a high level of care and stewardship. Educators need to focus on the ethics of data use; data governance, including privacy and security along with appropriate data-sharing strategies; and the importance of recognizing data literacy for key stakeholders, including patients, policy makers, payers, clinicians, and health care executives. Data brokering and its impact on health care continue to evolve.

The infusion of artificial intelligence will continue to generate new ethical questions, opportunities, and concerns [33]. In addition, gaps in data and new data areas such as social determinants and drivers of health offer new and interesting challenges to consider [18]. Furthermore, innovation should always be grounded in asking the "what if" questions to ensure that ethical considerations are always an aspect of data use.

Focus on Data Utility

Health data are being collected at a momentous rate. Educators must focus on preparing a health data–informed workforce to recognize the utility of data for the audience. This must also be considered in implementing health information technology mechanisms focused on user experience and human-centered design to ensure that health data are used thoughtfully and ethically. Data standards are also critical to utility, such as the United States Core Data for Interoperability [26]. We have witnessed many implementations without the consideration of data standards, causing barriers to interoperability that can produce harm in patient care. We can name multiple examples where technology is purchased without even considering how systems will share or integrate data, causing myriad other challenges in health care.

Recognize the Impact of the System

Health care is a large system within systems. Health technology and data are driven by systems, whether they be legal, policy, or reimbursement. Implementing data and technology without a strong understanding of the mechanisms and systems thinking is problematic. A health data–informed workforce recognizes the many layered systems impacting health information technology and data use implementation. Ensuring that the workforce engages in systems thinking and searching for "the why" in implementation and data use is a critical skill. In addition, clinicians should not feel disempowered and instead recognize the role they can play at the microlevel in patient interactions to improve the use of health data for improved outcomes.

Conclusion

Our hope is to promote a conversation and spark innovation around the need to expand and grow the health data–informed workforce. We certainly cannot provide every piece of advice or suggestion here. Yet, we hope to spark a revolution to grow the cadre of passionate advocates for the proliferation of health data and technology in ways that truly support equity, reduce burden, and improve health care delivery. Additionally, we are not saying that data skills are not critical—they are. We recognize that we need more than that. We need a workforce that asks questions about where data go and how they are used and that becomes more informed on the data tools of their patients. We need a health data–informed workforce now and into the future.

XSL•FO
RenderX

Acknowledgments

The project described is supported by the National Institute of General Medical Sciences (U54 GM115458), which funds the Great Plains IDeA-CTR (Institutional Development Award–Clinical & Translational Research) Network. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH).

Conflicts of Interest

JD owns and operates Hello Better Healthcare, LCC, where she collaborates with a variety of organizations as a strategic advisor or evaluator. These currently include CHI Health, the Iowa Community Care HUB, Family Room, Matter Health, and the Centers for Disease Control and Prevention (CDC). The other authors have no further interests to declare.

References

1. Bossen C, Bertelsen PS. Digital health care and data work: who are the data professionals? Health Inf Manag 2023 Jul 25:18333583231183083. [doi: 10.1177/18333583231183083] [Medline: 37491822]
2. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
3. Tripathi M. EHR evolution: policy and legislation forces changing the EHR. J AHIMA 2012 Oct;83(10):24-29, quiz 30. [Medline: 23061349]
4. Lin YK, Lin M, Chen H. Do electronic health records affect quality of care? evidence from the HITECH Act. Inf Syst Res 2019 Mar 12;30(1):306-318. [doi: 10.1287/isre.2018.0813]
5. Teisberg E, Wallace S, O'Hara S. Defining and implementing value-based health care: a strategic framework. Acad Med 2020 May;95(5):682-685. [doi: 10.1097/ACM.0000000000003122] [Medline: 31833857]
6. Lynch D, Christensen UJ, Howe NJ. AI technology and personalized learning design—uncovering unconscious incompetence. In: Burgos D, editor. Radical Solutions and Learning Analytics: Personalised Learning and Teaching Through Big Data: Springer; 2020:157-172. [doi: 10.1007/978-981-15-4526-9_10]
7. Desai S, Mostaghimi A, Nambudiri VE. Clinical informatics subspecialists: characterizing a novel evolving workforce. J Am Med Inform Assoc 2020 Nov 1;27(11):1711-1715. [doi: 10.1093/jamia/ocaa173] [Medline: 32951031]
8. Patel JS, Vo H, Nguyen A, Dzomba B, Wu H. A data-driven assessment of the U.S. health informatics programs and job market. Appl Clin Inform 2022 Mar;13(2):327-338. [doi: 10.1055/s-0042-1743242] [Medline: 35354210]
9. Jen MY, Mechanic OJ, Teoli D. Informatics. In: StatPearls: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK470564/ [accessed 2024-06-07]
10. Office of the National Coordinator for Health Information Technology, Department of Health and Human Services. 21st Century Cures Act: interoperability, information blocking, and the ONC Health IT Certification Program. Federal Register. 2020 May 1. URL: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2023-07-31]
11. Nijor S, Rallis G, Lad N, Gokcen E. Patient safety issues from information overload in electronic medical records. J Patient Saf 2022 Sep 1;18(6):e999-e1003. [doi: 10.1097/PTS.0000000000001002] [Medline: 35985047]
12. Stillman M. Death by patient portal. JAMA 2023 Jul 18;330(3):223-224. [doi: 10.1001/jama.2023.11629] [Medline: 37389857]
13. Chen M, Esmaeilzadeh P. Adoption and use of various health information exchange methods for sending inside health information in US hospitals. Int J Med Inform 2023 Sep;177:105156. [doi: 10.1016/j.ijmedinf.2023.105156] [Medline: 37487455]
14. Dixon BE, Holmgren AJ, Adler-Milstein J, Grannis SJ. Health information exchange and interoperability. In: Finnell JT, Dixon BE, editors. Clinical Informatics Study Guide: Text and Review: Springer; 2022:203-219. [doi: 10.1007/978-3-030-93765-2_14]
15. Janakiraman R, Park E, Demirezen EM, Kumar S. The effects of health information exchange access on healthcare quality and efficiency: an empirical investigation. Manag Sci 2023 Feb;69(2):791-811. [doi: 10.1287/mnsc.2022.4378]
16. Vorisek CN, Lehne M, Klopfenstein SAI, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724. [doi: 10.2196/35724] [Medline: 35852842]
17. Schulz S, Stegwee R, Chronaki C. Standards in healthcare data. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science: Springer; 2019:19-36. [doi: 10.1007/978-3-319-99713-1_3]
18. Rousseau JF, Oliveira E, Tierney WM, Khurshid A. Methods for development and application of data standards in an ontology-driven information model for measuring, managing, and computing social determinants of health for individuals, households, and communities evaluated through an example of asthma. J Biomed Inform 2022 Dec;136:104241. [doi: 10.1016/j.jbi.2022.104241] [Medline: 36375772]
19. Health data utility framework — a guide to implementation. Civitas Networks for Health. 2023 Mar. URL: https://www.civitasforhealth.org/wp-content/uploads/2023/03/Civitas-HDU-Framework-Final-2023-03-26.pdf [accessed 2023-10-20]

XSL•FO
RenderX

20. Klinedinst J. Preparing the health Informatics workforce for the future. In: Hübner UH, Mustata Wilson G, Morawski TS, et al, editors. Nursing Informatics: A Health Informatics, Interprofessional and Global Perspective: Springer; 2022:603-626. [doi: 10.1007/978-3-030-91237-6_39]
21. Dixon BE, McFarlane TD, Grannis SJ, Gibson PJ. Public health informatics workforce skills and needs: a descriptive analysis using the 2017 PH WINS. Eur J Public Health 2020 Sep 30;30(Supplement_5):ckaa165.027. [doi: 10.1093/eurpub/ckaa165.027]
22. Gadd CS, Steen EB, Caro CM, Greenberg S, Williamson JJ, Fridsma DB. Domains, tasks, and knowledge for health Informatics practice: results of a practice analysis. J Am Med Inform Assoc 2020 Jun 1;27(6):845-852. [doi: 10.1093/jamia/ocaa018] [Medline: 32421829]
23. McFarlane TD, Dixon BE, Grannis SJ, Gibson PJ. Public health informatics in local and state health agencies: an update from the public health workforce interests and needs survey. J Public Health Manag Pract 2019;25(2 Suppl):S67-S77. [doi: 10.1097/PHH.0000000000000918] [Medline: 30720619]
24. Brommeyer M, Whittaker M, Mackay M, Ng F, Liang Z. Building health service management workforce capacity in the era of health informatics and digital health - a scoping review. Int J Med Inform 2023 Jan;169:104909. [doi: 10.1016/j.ijmedinf.2022.104909] [Medline: 36347141]
25. Dixon BE, Rahurkar S, Apathy NC. Interoperability and health information exchange for public health. In: Magnuson J, Dixon B, editors. Public Health Informatics and Information Systems: Springer; 2020:307-324. [doi: 10.1007/978-3-030-41215-9_18]
26. United States Core Data for Interoperability (USCDI). Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi [accessed 2023-07-31]
27. American Health Information Management Association (AHIMA). URL: https://www.ahima.org/ [accessed 2023-10-24]
28. American Medical Informatics Association (AMIA). URL: https://amia.org/ [accessed 2023-10-24]
29. Civitas Networks for Health. URL: https://www.civitasforhealth.org/ [accessed 2023-10-24]
30. Healthcare Information and Management Systems Society (HIMSS). URL: https://www.himss.org/ [accessed 2024-10-24]
31. Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/ [accessed 2023-10-24]
32. Nelson R. Informatics: evolution of the Nelson Data, Information, Knowledge and Wisdom Model: part 2. Online J Issues Nurs 2020 Jul 21;25(3). [doi: 10.3912/OJIN.Vol25No03InfoCol01]
33. Gray K, Slavotinek J, Dimaguila GL, Choo D. Artificial intelligence education for the health workforce: expert survey of approaches and needs. JMIR Med Educ 2022 Apr 4;8(2):e35223. [doi: 10.2196/35223] [Medline: 35249885]

Abbreviations

EHR: electronic health record
HIE: health information exchange
HIPAA: Health Insurance Portability and Accountability Act

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna, TDA Cardoso; submitted 29.08.23; peer-reviewed by J McClay, S Helou; revised version received 26.03.24; accepted 09.05.24; published 17.06.24.

Please cite as:
Doll J, Anzalone AJ, Clarke M, Cooper K, Polich A, Siedlik J
A Call for a Health Data–Informed Workforce Among Clinicians
JMIR Med Educ 2024;10:e52290
URL: https://mededu.jmir.org/2024/1/e52290
doi:10.2196/52290

© Joy Doll, A Jerrod Anzalone, Martina Clarke, Kathryn Cooper, Ann Polich, Jacob Siedlik. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.6.2024. 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.

XSL•FO
RenderX

A Proposed Decision-Making Framework for the Translation of In-Person Clinical Care to Digital Care: Tutorial

Anna DeLaRosby[1], PT, DPT; Julie Mulcahy[2], PT, DPT; Todd Norwood[2], PT, DPT

1

2

Corresponding Author:
Anna DeLaRosby, PT, DPT

Abstract

The continued demand for digital health requires that providers adapt thought processes to enable sound clinical decision-making in digital settings. Providers report that lack of training is a barrier to providing digital health care. Physical examination techniques and hands-on interventions must be adjusted in safe, reliable, and feasible ways to provide digital care, and decision-making may be impacted by modifications made to these techniques. We have proposed a framework to determine whether a procedure can be modified to obtain a comparable result in a digital environment or whether a referral to in-person care is required. The decision-making framework was developed using program outcomes of a digital physical therapy platform and aims to alleviate barriers to delivering digital care that providers may experience. This paper describes the unique considerations a provider must make when collecting background information, selecting and executing procedures, assessing results, and determining whether they can proceed with clinical care in digital settings.

(JMIR Med Educ 2024;10:e52993) doi:10.2196/52993

KEYWORDS

clinical decision-making; digital health; telehealth; telerehab; framework; digital medicine; cognitive process; telemedicine; clinical training

Introduction

Background

Digital health is revolutionizing health care, and the COVID-19 pandemic has led to rapid acceleration of the use of digital health technologies, particularly the adoption of telehealth. Digital health, including the use of telehealth or telemedicine, allows health care practitioners to provide services without being in the same physical location as the patient. Telehealth can include synchronous or asynchronous messaging with providers, video calls, audio-only calls, and the secure transmission of information over the internet between patients and their providers [1]. Digital health can also include information gathered by medical devices, wearable sensors, apps, or other software [2]. The application of technology in health care has a vast potential to increase access to care and improve quality.

Research indicates that telehealth outcomes are equivalent to in-person care in rehabilitation [3-5] and can be an effective intervention for addressing pain and function limitations in a variety of musculoskeletal conditions [6]. Clinical outcomes from telehealth episodes of care are comparable with in-person rehabilitation for conditions such as osteoarthritis, low-back pain, hip and knee replacement, multiple sclerosis, and cardiac and pulmonary rehabilitation [3]. Increasing evidence supports that telehealth physical therapy delivered by a mobile app provides clinical outcomes comparable with those of in-person care [3,4,7]. Research also reveals that telehealth decreases travel time and costs [8]. It is well documented that patients recognize the benefits of telehealth as well, demonstrating high engagement [9-11] and high levels of satisfaction across multiple metrics, including quality of care, convenient access to multiple specialists, improved care and coordination with digital care, and outcomes similar to in-person care [12-17].

Despite evidence of the benefits of telehealth, there are barriers to the integration of telehealth into traditional health care models. For example, physical therapists (PTs) report apprehension toward utilizing telehealth in their practice, reporting insufficient preparation and inadequate knowledge about how to implement telehealth visits, influencing providers' acceptance, preferences, and outcomes [12,18,19]. Further, less than half (42%) of health care providers surveyed believed telehealth was as effective as face-to-face care, and 21% reported insufficient training [18,19]. Another significant barrier to digital health adoption is the belief that lack of physical contact hampers accurate diagnosis and management [12,18,19]. Successful integration of telehealth into traditional health care models will only be achieved through addressing provider beliefs about the efficacy of telehealth and instruction in providing equivalent care through a new model.

Telehealth requires the translation of traditional clinical skills to a new medium [20,21]. Remote patient care is characterized by dynamic patient environments, unique safety concerns, and

XSL•FO
RenderX

a lack of traditional patient care tools, forcing the provider to act in new and dynamic ways to provide effective care. When encountering new clinical scenarios, many providers look for guidance through decision-making frameworks. Frameworks outline a structured and systematic approach to problem-solving that incorporates evidence and specific context, and promotes informed decisions [22]. When used in health care, decision-making frameworks can ensure consistency, reduce bias, and enhance the quality of decisions and quality of care [22-24]. A standardized process assists health care professionals in assessing risks and benefits, improves outcomes, and provides patient-centered evidence-based care [24].

Delivering effective care in a digital health setting requires that health care providers adapt their thought processes to account for the nuance of the interactions between technology and the patient to enable sound clinical decision-making in the digital health setting. This paper introduces a decision-making framework to determine whether a clinical procedure is feasible in a telehealth setting with similar quality, accuracy, and reliability as in-person encounters, or when the use of an equivalent but alternative procedure is most appropriate. We propose that utilizing a clinical decision-making framework can alleviate clinicians' concerns about the efficacy of digital health and assist the implementation of clinical best practices in a digital setting. The purpose of this paper is threefold: (1) to propose a decision-making framework to train and inform health care providers that increases provider efficacy with the translation of skills to this new medium; (2) to propose a thought model that allows quantitative testing through implementation research; and (3) to realize the potential for telehealth for patients and providers to improve access to care independent of geography.

Development of the Framework

This framework was the result of a review of the current literature and the authors' combined expertise in providing telehealth physical therapy. The authors have a combined 18 years of experience in telehealth, including providing patient care, designing and implementing training for providers, as well as managing a nationwide network of telehealth PTs. This framework has been applied to clinical practice and refined based on the outcomes of over 10,000 patient cases.

Analysis of program outcomes and the identification of PT behaviors that lead to positive clinical outcomes influenced the development of this framework. Program data confirmed that provider behavior during telehealth episodes directly impacts clinical outcomes in an app-based telehealth physical therapy program [4] and that when interventions provide high value, patients will be highly engaged [11] resulting in cost savings [25]. Prior literature describes how to translate specific evidence-based evaluation techniques for the application of telehealth and how to utilize established clinical practice guidelines for telehealth episodes [26-32]. However, procedure-specific training cannot prepare providers for the dynamic nature of telehealth encounters that include variations in the patient's environment, health status, caregiver support, digital literacy, equipment availability, and other factors. In response to the ever-changing context of telehealth visits and to fully equip health care providers working in a digital environment, a decision-making framework was developed. This framework was designed to help providers identify the relevant factors in the clinical picture, assess possible actions, and make decisions that lead to positive clinical outcomes. The process of defining this framework was iterative, data-driven, and emphasized patient-centered design. We incorporated an understanding of the users on our platform, the tasks they completed, and the digital environment; development was driven and refined by patient surveys, feedback, and outcomes. We believe this framework will assist clinicians in translating their clinical skills to digital practice to enable optimal clinical outcomes, convenience, and satisfaction. Initially, learning to leverage the steps of the framework may increase time in decision-making but as the clinician becomes experienced the process will become efficient and give more options for the telehealth environment.

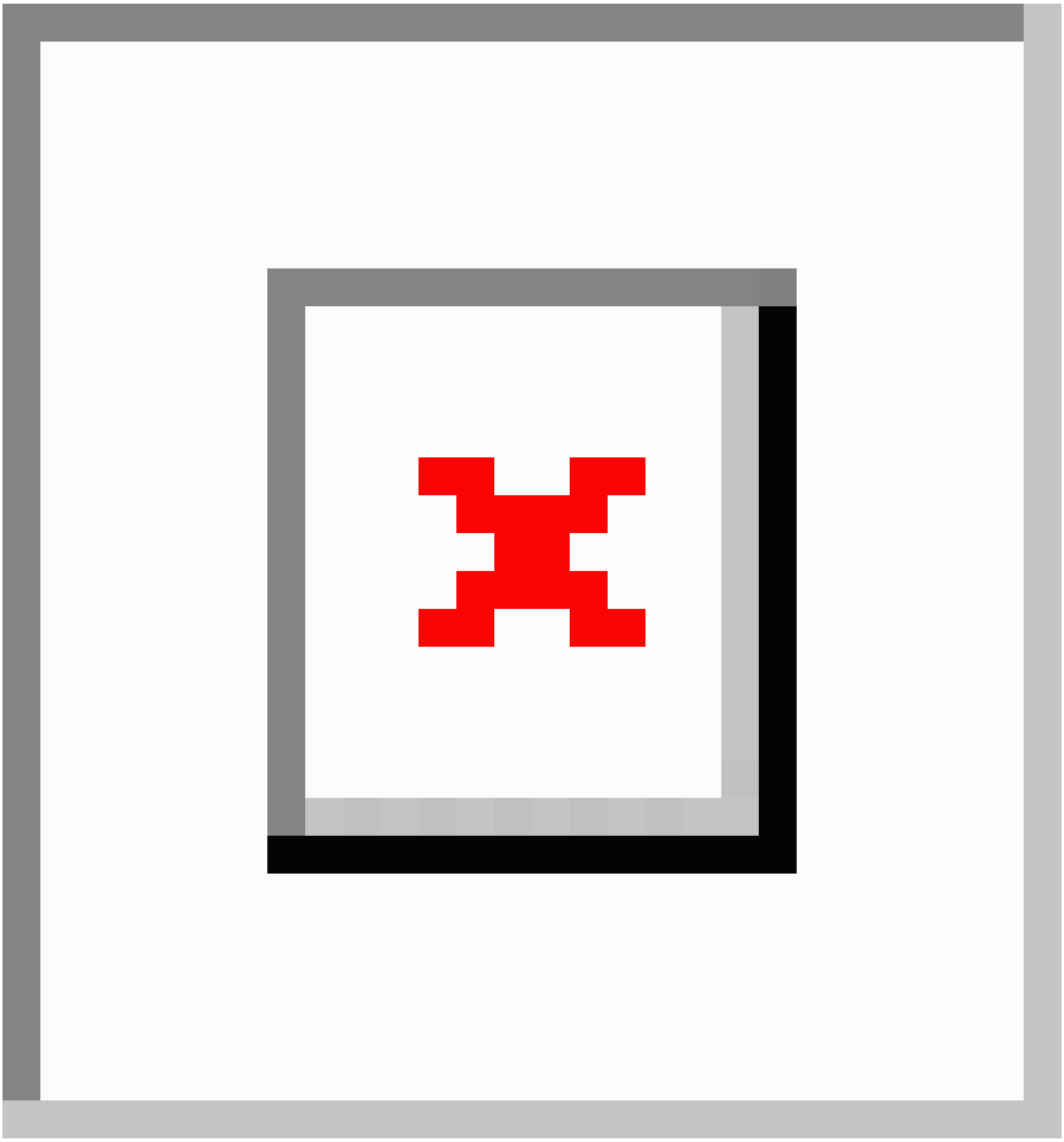
Utilizing the Framework

Appropriate application of a decision-making framework in a clinical setting requires that certain conditions are met. First, the clinical problem must be within the scope of the clinician's practice. This ensures the clinician is appropriately trained and licensed to provide care and make clinical decisions. In the case of digital health, appropriate training includes proficiency with digital tools, technology, and website manner in addition to medical or clinical training [33,34]. Second, the patient must be appropriate for digital care. Appropriateness for care requires that the patient's cognition level, medical status, digital literacy, communication abilities, technology access, physical environment, and preference all support safe digital care interactions. Finally, providers must consider the security and regulatory implications of digital care, including ensuring compliance with HIPAA (Health Insurance Portability and Accountability Act), local and federal privacy regulations, and data security requirements. If the provider, patient, and technology all meet these conditions, the application of this framework is appropriate.

At each step of the process, the provider must determine whether telehealth is the most appropriate method of providing care. When a provider determines that telehealth is not appropriate for the patient, they should inform the patient of the next steps, which may include activation of emergency services, coordination of care to facilitate referral to a specialist, in-person visit, or obtaining labs or imaging. Figure 1 provides a visual representation of the steps included in this decision making framework.

XSL•FO
RenderX

Figure 1. The decision-making path. At every step of the patient encounter, providers must determine whether telehealth is the best option for the clinical scenario. The determination process should be the same whether the provider is using a traditional procedure or a procedure that has been modified for the patient's environment. At each step, the provider must determine whether they can continue down the decision-making path, or if they need to return to the start of the decision-making process using an alternative procedure. If no acceptable digital option exists at any step, they must refer to in-person care.



Description of the Decision-Making Framework

Step 1: Collect Background Information

Clinicians may collect relevant clinical information using data from chart review and review of a digital intake form. The subjective interview of a telehealth visit should proceed as it does in an in-person visit, with emphasis on the chief complaint, relevant health history, current and past medical conditions, and social history. The subjective portion of the examination may also include a visual assessment of the patient's environment, inquiry about equipment availability, and availability of caregiver support, which are factors unique to telehealth but enhance the clinical picture. If at the conclusion of the subjective interview, the provider has identified an urgent medical need, or that telehealth is not appropriate then the patient may be referred to in-person care at this time. If the provider is confident that they have collected the information needed to inform the objective examination and that it is safe and appropriate to continue with a telehealth objective examination, they will move to the next step.

XSL•FO
RenderX

Step 2: Select an Examination Procedure

Providers will select the examination procedures based on the information gathered in the subjective examination. Procedures should be evidence-based and relevant to the differential diagnosis process. Once a procedure has been selected, the provider must consider the feasibility, reliability, and validity of the procedure when performed in a digital setting.

To evaluate feasibility, we consider whether the patient has the resources, space, ability, and knowledge necessary to complete the procedure safely. The provider will consider information gathered in the subjective portion regarding the patient's cognitive status, physical ability, social support, environment and technological resources, and time available to determine if the procedure can be accurately performed. If the setup for a test is complicated or the instructions are lengthy, the time constraints of a patient visit may make a test not feasible.

Reliability is the quality of a measure that produces reproducible scores on repeat administrations of a test. Reliability is thus a prerequisite for test validity [35]. Validity is the measure of how accurately a test measures the underlying trait of interest [35,36]. When assessing patients in-person, reliability is supported by a clinical environment standardized for all sessions. In digital health settings, tests are performed in the patient's environment and providers must look for alternative ways to ensure results are reliable and valid. If a traditional procedure cannot be performed with acceptable feasibility and reliability, then providers should consider if an alternative procedure can provide the same clinical information. Alternative methods will be unique to the patient's resources, abilities, and environment, but alternatives should be assessed for feasibility and reliability. Functional testing is often an acceptable alternative for traditional tests when the equipment or environment is standardized.

The reliability of functional tests can be increased if the same equipment in the home is used for subsequent testing. For example, a 30-second sit-to-stand test [37] using the same chair in the patient's home will give a clinician reliable data for each assessment. Further, measurements such as joint range of motion, can be tracked by having the patient reach to low, medium, or high shelves in their home and reassessed using the same shelves. This technique allows the provider to monitor and document progress in an easily accessible, functional and standardized way within the patient's environment.

Selecting a procedure means that the provider will make dynamic decisions unique to the patient they are seeing. For example, during an in-person visit, manual muscle testing of internal rotation of the shoulder is often used to indicate subscapularis muscle rupture or dysfunction. In digital settings, the provider cannot provide manual resistance, but the same information can be obtained using the Gerber test [38]. If the patient is unable to achieve the testing position for a Gerber test, a provider could consider functional strength testing such as lifting canned goods. In this scenario, the provider will ensure reliability by using the same number of cans at each assessment. To ensure validity, the provider must ensure that the patient is performing the test correctly; in this example, a patient lifting the canned goods with a straight arm would provide an invalid result but lifting with a bent elbow would appropriately stress the biceps and give a valid result.

If there is no procedure that can be performed that is feasible and reliable in the digital setting, and this information is required for clinical decision-making, then a referral to in-person care would be indicated. For example, if a clinician suspects rupture of the anterior cruciate ligament and determines that a Lachman test is necessary, but is not feasible via telehealth, then a referral for in-person assessment is required.

Step 3: Execute the Clincial Procedure

Performing the clinical procedures in digital settings requires different skills than in in-person settings. Digital settings require the provider to assist the patient in managing their environment and any relevant equipment needed during the visit. Therefore, it is incumbent on the provider to communicate with the patient explicitly about the procedure prior to execution and ensure they have the relevant equipment and can use it appropriately.

The provider should communicate what equipment is needed (eg, a sturdy chair with arms). Providers should give clear directions to the patient on how to set up any equipment and where the patient should be positioned. Additionally, the provider must describe how to utilize technology during the procedure. Appropriate audio, video, and lighting setup ensures the provider can see and hear the patient adequately while they perform the tasks. The provider should review each step of the procedure with the patient prior to performing it and allow the patient to ask questions or clarify instructions. The patient should have a good understanding of what information the procedure is gathering so that the patient can monitor and report the appropriate variable during the procedure. For example, during a balance assessment, the patient should understand if they are balancing for as long as they can without toe touches, or if they should count the number of toe touches within the given time frame. The provider should document the method used for the procedure, equipment, setup, and outcome to ensure subsequent tests can be performed in a standard way. If the patient is unable to perform the procedure as directed by the provider, then the provider should consider alternative procedures or referral to in-person care.

Step 4: Assess Results

Once the procedure has been performed, the provider determines whether the result answers the original clinical question and their confidence level in the result. Confidence will be affected by how accurately the patient was able to follow the provider's instructions, and if technology worked as expected. If the patient performed the test incorrectly or if there was video or audio lag or poor clarity available, the provider may have low confidence in the result. A procedure that was performed as instructed in an environment that was reliably standardized using the same equipment and set up with technology that worked without disruption will provide high confidence.

Step 5: Proceed With Clinical Care, Repeat, or Refer

High confidence in the outcome allows the provider to continue care in the digital setting. If the provider has low confidence in the result, they can repeat steps 1 through 4 again using an

XSL•FO
RenderX

alternative procedure to achieve a result that provides high confidence. If the provider is seeking information that is essential to the care of the patient and no procedure can be performed in a manner that provides a result that is reliable, reproducible, and yields high confidence, then a referral to in-person care is needed. Table 1 provides a list of the factors that should be considered when making clinical decisions in digital settings.

Table . The relevant factors the provider should consider as they progress through the decision-making process. At each stage, the provider must determine whether telehealth is appropriate for this clinical scenario.

Factors	Key points

Collect background information	• Subjective history may include chief complaint and health history as well as:
	  • Cognition level
	  • Digital literacy
	  • Communication abilities
	  • Technology access
	  • Features of physical environment
	  • Patient preference for digital health tools
	• If each criterion is not met, then the patient must be referred to in-person care

Select procedures	• Traditional procedures, digital alternatives, or functional tests may be used if they are:
	  • Necessary for clinical reasoning
	  • Evidence-based
	  • Feasible
	  • Reliable
	• If no procedure meets these criteria, then the patient must be referred to in-person care

Execute procedures	• Instruct the patient about:
	  • Equipment required
	  • Technology settings
	  • Environment set up
	  • Performance of the procedure
	  • Outcome reporting
	• If execution of the procedure is impeded by any of these factors, the provider will consider alternative procedures or refer to in-person care

Assess results	• Determine if the reliability of the result was affected by:
	  • Procedure performance
	  • Technology
	  • Reporting accuracy
	• Does the provider have confidence in the result of the procedure?

Proceed with clinical care, repeat, or refer	• Do you need more clinical information?
	  • If no:
	    • Proceed with clinical care
	  • If yes:
	    • Repeat decision-making steps
	    • Refer if no alternative exists

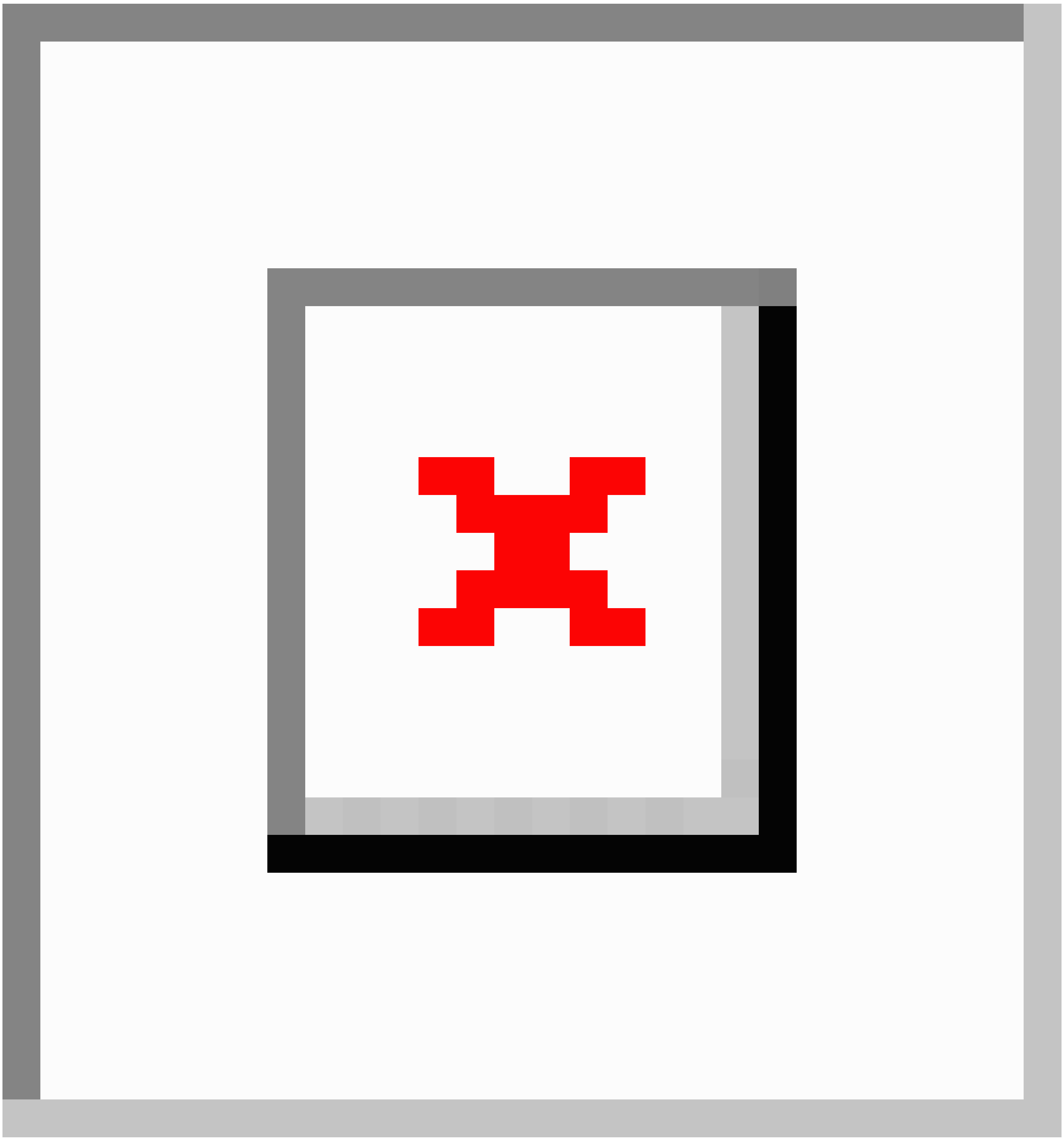
Clinical Application

Overview

The application of this decision-making framework can be illustrated through clinical examples. This example provides descriptions of how procedures can be modified but provides high-value clinical information when feasibility, reliability, and reproducibility are considered. Assessment of confidence allows providers to determine the value of the result prior to proceeding with clinical care or referring to in-person care. Figure 2 provides a visual representation of the decision-making process used in the patient scenario.

XSL•FO
RenderX

Figure 2. Description of clinical application of proposed decision-making framework using the timed up and go test and modified test. The provider proceeded through the first process but had low confidence in performance. They then repeated the decision-making process with modifications made to the test environment and procedure. The modified test produced a high-confidence result and allowed clinical care to proceed.



Patient Scenario

Consider a hypothetical case of a 79-year-old woman living in a rural community who scheduled a telehealth visit with her primary care provider (PCP) to discuss concerns regarding mobility. Mobility assessment is within the scope of the provider in this example, who has the appropriate training and expertise to perform telehealth visits. The visit will take place on the platform provided by the health system and meet applicable HIPAA and data security requirements. The provider has access to the patient's medical history as a part of the software platform.

Step 1: Collect Background Information

The patient's chief complaint is frequent stumbling, often the result of catching her toe while walking, which has resulted in loss of balance, frequently holding onto furniture or walls while walking, and avoiding walking in the community due to fear of falling. She reports no falls to the ground and no other health status changes but is concerned that her balance will continue to decline. The provider assesses the patient's cognitive status, communication ability, and preference for digital health during the subjective assessment. As part of the telehealth visit, the provider completes red flag screening and review of systems and finds no neurological deficits, no indication of cardiac

XSL•FO
RenderX

impairment, and no history to suggest that the mobility concerns are the result of sinister pathology. Her PCP would like to quantify the mobility impairments in a standardized way during the telehealth visit and the patient agrees to this. The patient reports that her husband is available during the telehealth visit to assist if needed. As the patient has no current history of falls, health history is clear, and the patient has a caregiver present, the PCP feels confident that they can complete a mobility assessment safely via telehealth.

Step 2: Select Procedure

The PCP chooses the Timed Up and Go (TUG) test [39] as it is evidenced-based and recommended by the Center for Disease Control STEADI protocol [40]. TUG is a timed mobility test in which patients rise from a standard chair, walk to a line on the floor 10 feet away, turn, return to the chair, and sit. Patients are instructed to wear their regular footwear and can use a walking aid during the test if needed.

The PCP assesses feasibility by asking if the patient has access to the equipment needed: a sturdy chair such as a dining chair, stopwatch, tape measure, and tape. The PCP describes the test to the patient and husband and asks if they feel able to achieve the setup and execute the test. The PCP will be able to gather qualitative information about gait during the test by having the patient face their device's camera toward the test area. The outcome of the TUG is time-based, which the PCP determines to be reliable through digital means. The PCP decides that the caregiver will manage the stopwatch to mitigate any lag in the internet connection during the test. Using a tape measure to define distance and using the same chair in the same location will ensure that the test setup is reproducible for subsequent testing. The provider educates the patient's caregiver on the start or stop timing procedure of the TUG, further ensuring reliability. The PCP will assess qualitative mobility by visually assessing movement during the test using the camera of the mobile device. The PCP determines that the TUG is feasible and reliable in a digital setting and provides the clinical information required to make clinical decisions about treatments for this patient, so no alternative is necessary.

Step 3: Execute Procedure

The PCP instructs the patient's husband to gather a sturdy chair and stopwatch, measure 10 feet on the floor, and mark it with a line of tape. The provider instructs the patient and caregiver to arrange the camera of their mobile device in a manner that allows the PCP to observe the test. The caregiver is instructed on starting or stopping the stopwatch. The patient is instructed on the test procedure according to the standard TUG instructions. The caregiver is instructed to report the time to completion of the procedure to the PCP. The provider answers clarifying questions for the patient and caregiver, and they perform the test. During the test the provider can hear that the caregiver fumbles with the stopwatch, and the patient leaves the video frame.

Step 4: Assess Results

While the environment setup was standardized supporting reliability, the caregiver reported difficulty with starting or stopping the timer, impacting the accuracy of the timed result. The patient left the visual frame during the test, impacting the ability to assess qualitative aspects of gait such as stopping and changing directions. The provider determines they have low confidence in the result and is unable to determine if the patient exceeded the recommended time of <12 seconds for test completion, or if there are mobility deficits that prompt recommendations for assistive device use.

Step 5: Proceed With Clinical Care, Repeat or Refer

The provider has low confidence in the result of the test and does not feel they can proceed with clinical care based on the results. The need for mobility assessment remains, and the provider feels that modifications of the testing scenario may allow them to gain the clinical information they need, so a referral to in-person care is not necessary. The home environment had only one area where a 10-foot space was available to complete the TUG, however, the family was unable to position the camera in a manner that allowed the whole area to be seen by the provider. Additionally, the caregiver had difficulty starting and stopping the timer, decreasing the accuracy of the result. The provider determines that the variables measured by the TUG test appropriately provide the clinical information they need, but he will need to utilize an alternative testing method to enable him to address the limitations. He will repeat decision-making steps using a digital alternative to gain the information he needs from the mobility assessment.

Background information remains the same, so the provider can proceed to select an alternative procedure. They decide to address the limitations of the first attempt by choosing a new testing location where they can standardize the test using landmarks in the patient's home. The caregiver is instructed to position the front legs of the chair even with a door frame and will have the patient walk to the end of the hallway, touch the wall, and return to the chair. The distance walked is less than the 10 feet required of the TUG, but the patient is visible to the provider the whole distance. Additionally, the provider will give audio cues to start and stop the test while he manages the timer remotely. The provider and patient determine together that this setup is feasible and easily reproducible for subsequent testing. The modifications will allow the provider to assess movement quality as well as ensure timed results are accurate, which addresses the limitations of the prior test.

Execution of the modified procedure requires instruction regarding chair location and placement of the mobile device so the camera captures the whole testing area. The patient is instructed on how to perform the modified test procedure. The performance of the modified test proceeds without audio or video lag or distortion. After the second test provider feels confident that the timed result was successful. The provider was able to assess the quality of mobility throughout the whole test. Because the provider has high confidence in the clinical information they obtained through the alternative test, they can proceed with clinical care. The provider determines that the patient would benefit from using a single-point cane to improve balance with changing directions when walking. The PCP also prescribes physical therapy to address balance, gait, and lower extremity strength. The patient will schedule a follow-up telehealth visit with the PCP in 4 weeks and they will repeat

XSL•FO
RenderX

the modified mobility test at that time using the same setup to assess the effect of these interventions.

Discussion

Principal Findings

Providing a decision-making framework for clinicians to utilize in digital care can alleviate clinicians' concerns about implementing digital care in their practice. To our knowledge, a framework that assists providers in translating in-person clinical skills to digital care does not exist. This framework enables clinicians to practice effectively in the most accessible environment for the patient while prioritizing evidence-based practice, assessing risks, and providing patient-centered care. As digital care is increasingly desired by patients [19,23,41,42], it is imperative that providers are confident in decision-making in digital settings so telehealth remains safe, efficient, and equivalent to in-person care.

The value of a clinical procedure is reliant upon the feasibility, reliability, and clinician confidence, as well as the interaction of those variables with digital technology. A procedure that is feasible, reliable, and reproducible, but is performed poorly and provides low confidence has less value in clinical decision-making than an alternative digital procedure that deviates from standard performance but instills high confidence in clinical decision-making. This improves patient safety by determining whether a patient can remain in a digital care environment or should be referred to in-person care. Additionally, the framework encourages clinicians to use evidence-based practice guidelines as the basis for care, modifying procedures in a feasible and reliable manner to improve outcomes. This will ensure consistency, reduce bias, and enhance the quality of decisions in digital care [22-24]. The application example demonstrated that modifications made based on the patient's environment and technology limitations enabled the provider to proceed with digital care in a manner consistent with clinical best practices and supported the provision of safe, effective, and quality care.

Time is a valuable resource in medical care, and providers must be confident in decisions made during clinical encounters. In situations where decisions must be made quickly, utilizing a framework can assist with decision-making efficiency [22-24]. Novice clinicians or providers who are transitioning to digital care may benefit from a framework to help them determine the best course of action in a timely manner. With increased provider experience and repetition, the decision-making process will be more efficient and timelier. For example, experienced telehealth clinicians become proficient in scanning the patient environment, determining feasibility based on available resources, as well as becoming efficient at modifying traditional procedures based on the patient's environment, and instructing patients on camera setup and how to utilize technology efficiently. In scenarios like the clinical application described above, an experienced provider may identify potential barriers prior to execution and decide to utilize a modified procedure from the start to save time.

This framework builds on the existing literature that shows similar diagnostic accuracy between in-person and digital examination techniques [26,29-31]. Lack of physical contact when working through telehealth was perceived to hamper accurate and effective diagnosis and management [18]. However, many commonly performed physical examination techniques have poor sensitivity and interrater reliability. This is evident in the poor interrater reliability scores of techniques such as palpation of lumbar structures [43] and assessment of breath sounds [44]. Decision-making tools that enable providers to evaluate alternative methods for gathering clinical information help to overcome these barriers and increase confidence that practitioners are providing effective, safe care. Additionally, adapting procedures allows patients the full benefit of telehealth, including convenience, cost-savings, better adherence, higher engagement, and improved access to care in rural or underserved areas [20].

Future Research

Avenues for further research should include randomized control trials comparing trained versus untrained providers to determine whether the utilization of this framework leads to improved clinical outcomes, provider self-efficacy, and patient satisfaction scores, and would provide insight to overcoming the barriers to digital health that providers may experience. Research is needed in implementation science to determine if training clinicians in using a framework will increase treatment fidelity. Similarly, this framework can be considered in future studies to provide further evidence of the efficacy of digital care and enable the full potential of telehealth for all stakeholders.

Further understanding of how providers make decisions to include digital tools in patient care is needed. Understanding provider confidence in modifying in-person techniques and clinical problem-solving in digital settings may improve providers' willingness to utilize digital care with their patients. Provider training about how to modify traditional procedures, evaluating the efficacy of modified procedures, and assessing confidence in results may increase provider self-efficacy in digital settings. Best practices and standardized education for health care providers on how to effectively use digital tools should be established.

Limitations

There are limitations of this framework as it is broad in scope and cannot address every situation. Independent tests should be performed to evaluate the usability of the framework and its effectiveness in improving guideline implementation. We recognize that no single framework can be used for all guidelines or contexts. Provider behavior will be influenced by environment, resources, technology, and other factors despite training in using a decision-making framework.

Conclusion

We created a framework for clinicians to determine whether a particular procedure can be performed feasibly in a digital health setting with the same quality, accuracy, and reliability as in a traditional setting. Utilizing a framework to assist in clinical decision-making is important to alleviate clinicians' concerns

XSL•FO
RenderX

about using digital tools and help guide the translation of the best available evidence from traditional care to digital care. The increased demand by patients for digital care requires a new set of clinical skills, and this framework enables providers to comply with clinical best practices and offer high-quality care for patients who want to receive their care via telehealth.

Conflicts of Interest

JM and TN are employed shareholders in Omada Health Inc.

References

1. What is telehealth? Telehealth.HHS.gov. URL: https://telehealth.hhs.gov/patients/understanding-telehealth [accessed 2024-01-26]
2. What is digital health? U.S. Food and Drug Administration. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health [accessed 2024-01-26]
3. Seron P, Oliveros MJ, Gutierrez-Arias R, et al. Effectiveness of telerehabilitation in physical therapy: a rapid overview. Phys Ther 2021 Jun 1;101(6):pzab053. [doi: 10.1093/ptj/pzab053] [Medline: 33561280]
4. Beresford L, Norwood T. The effect of mobile care delivery on clinically meaningful outcomes, satisfaction, and engagement among physical therapy patients: observational retrospective study. JMIR Rehabil Assist Technol 2022 Feb 2;9(1):e31349. [doi: 10.2196/31349] [Medline: 35107436]
5. Levy CE, Silverman E, Jia H, Geiss M, Omura D. Effects of physical therapy delivery via home video telerehabilitation on functional and health-related quality of life outcomes. J Rehabil Res Dev 2015;52(3):361-370. [doi: 10.1682/JRRD.2014.10.0239] [Medline: 26230650]
6. Cottrell MA, Galea OA, O'Leary SP, Hill AJ, Russell TG. Real-time telerehabilitation for the treatment of musculoskeletal conditions is effective and comparable to standard practice: a systematic review and meta-analysis. Clin Rehabil 2017 May;31(5):625-638. [doi: 10.1177/0269215516645148] [Medline: 27141087]
7. Callaghan T, McCord C, Washburn D, et al. The changing nature of telehealth use by primary care physicians in the United States. J Prim Care Community Health 2022 Jan;13:21501319221110418. [doi: 10.1177/21501319221110418] [Medline: 35795898]
8. Kosterink SM, Huis in 't Veld R, Cagnie B, Hasenbring M, Vollenbroek-Hutten MMR. The clinical effectiveness of a myofeedback-based teletreatment service in patients with non-specific neck and shoulder pain: a randomized controlled trial. J Telemed Telecare 2010;16(6):316-321. [doi: 10.1258/jtt.2010.006005] [Medline: 20798425]
9. Batsis JA, DiMilia PR, Seo LM, et al. Effectiveness of ambulatory telemedicine care in older adults: a systematic review. J Am Geriatr Soc 2019 Aug;67(8):1737-1749. [doi: 10.1111/jgs.15959] [Medline: 31066916]
10. Argent R, Daly A, Caulfield B. Patient involvement with home-based exercise programs: can connected health interventions influence adherence? JMIR Mhealth Uhealth 2018 Mar 1;6(3):e47. [doi: 10.2196/mhealth.8518] [Medline: 29496655]
11. Mulcahy J, Beresford LS, DeLaRosby A. Defying stereotypes. Top Geriatr Rehabil 2023;39(4):307-311. [doi: 10.1097/TGR.0000000000000414]
12. Fernandes LG, Oliveira RFF, Barros PM, Fagundes FRC, Soares RJ, Saragiotto BT. Physical therapists and public perceptions of telerehabilitation: an online open survey on acceptability, preferences, and needs. Braz J Phys Ther 2022 Nov;26(6):100464. [doi: 10.1016/j.bjpt.2022.100464] [Medline: 36410257]
13. Moo LR, Gately ME, Jafri Z, Shirk SD. Home-based video telemedicine for dementia management. Clin Gerontol 2020 Mar;43(2):193-203. [doi: 10.1080/07317115.2019.1655510] [Medline: 31431147]
14. Record JD, Ziegelstein RC, Christmas C, Rand CS, Hanyok LA. Delivering personalized care at a distance: how telemedicine can foster getting to know the patient as a person. J Pers Med 2021 Feb 17;11(2):137. [doi: 10.3390/jpm11020137] [Medline: 33671324]
15. Appleman ER, O'Connor MK, Rockefeller W, Morin P, Moo LR. Using video telehealth to deliver patient-centered collaborative care: the G-IMPACT pilot. Clin Gerontol 2022 Jul;45(4):1010-1019. [doi: 10.1080/07317115.2020.1738000] [Medline: 32228299]
16. Essery R, Geraghty AWA, Kirby S, Yardley L. Predictors of adherence to home-based physical therapies: a systematic review. Disabil Rehabil 2017 Mar;39(6):519-534. [doi: 10.3109/09638288.2016.1153160] [Medline: 27097761]
17. Slightam C, Gregory AJ, Hu J, et al. Patient perceptions of video visits using veterans affairs telehealth tablets: survey study. J Med Internet Res 2020 Apr 15;22(4):e15682. [doi: 10.2196/15682] [Medline: 32293573]
18. Malliaras P, Merolli M, Williams CM, Caneiro JP, Haines T, Barton C. "It's not hands-on therapy, so it's very limited": telehealth use and views among allied health clinicians during the coronavirus pandemic. Musculoskelet Sci Pract 2021 Apr;52:102340. [doi: 10.1016/j.msksp.2021.102340] [Medline: 33571900]
19. Almathami HKY, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407. [doi: 10.2196/16407] [Medline: 32130131]

XSL•FO
RenderX

20. The digitally enabled physical therapist: an APTA foundational paper. American Physical Therapy Association. 2022 Nov 21. URL: https://www.apta.org/your-practice/practice-models-and-settings/digital-health-technology/digitally-enabled-physical-therapist [accessed 2023-08-16]
21. Jeffries PR, Bushardt RL, DuBose-Morris R, et al. The role of technology in health professions education during the COVID-19 pandemic. Acad Med 2022 Mar 1;97(3S):S104-S109. [doi: 10.1097/ACM.0000000000004523] [Medline: 34789662]
22. Mandelblatt JS, Ramsey SD, Lieu TA, Phelps CE. Evaluating frameworks that provide value measures for health care interventions. Value Health 2017 Feb;20(2):185-192. [doi: 10.1016/j.jval.2016.11.013] [Medline: 28237193]
23. Little LM, Pickett KA, Proffitt R, Cason J. Keeping pace with 21st century healthcare: a framework for telehealth research, practice, and program evaluation in occupational therapy. Int J Telerehabil 2021 Jun;13(1):e6379. [doi: 10.5195/ijt.2021.6379] [Medline: 34345350]
24. Moberg J, Oxman AD, Rosenbaum S, et al. The GRADE evidence to decision (EtD) framework for health system and public health decisions. Health Res Policy Syst 2018 May 29;16(1):45. [doi: 10.1186/s12961-018-0320-2] [Medline: 29843743]
25. Chen F, Siego CV, Jasik CB, et al. The value of virtual physical therapy for musculoskeletal care. Am J Manag Care 2023 Jun 1;29(6):e169-e175. [doi: 10.37765/ajmc.2023.89375] [Medline: 37341981]
26. Lamplot JD, Pinnamaneni S, Swensen-Buza S, et al. The knee examination for video telemedicine encounters. HSS J 2021 Feb;17(1):80-84. [doi: 10.1177/1556331620975039] [Medline: 33967647]
27. Lawton CD, Swensen-Buza S, Awender JF, et al. The elbow physical examination for telemedicine encounters. HSS J 2021 Feb;17(1):65-69. [doi: 10.1177/1556331620975040] [Medline: 33967644]
28. Lamplot JD, Pinnamaneni S, Swensen-Buza S, et al. The virtual shoulder and knee physical examination. Orthop J Sports Med 2020 Oct;8(10):2325967120962869. [doi: 10.1177/2325967120962869] [Medline: 33614791]
29. Wright-Chisem J, Trehan S. The hand and wrist examination for video telehealth encounters. HSS J 2021 Feb;17(1):70-74. [doi: 10.1177/1556331620975341] [Medline: 33967645]
30. Iyer S, Shafi K, Lovecchio F, et al. The spine telehealth physical examination: strategies for success. HSS J 2021 Feb;17(1):14-17. [doi: 10.1177/1556331620974954] [Medline: 33967636]
31. Laskowski ER, Johnson SE, Shelerud RA, et al. The telemedicine musculoskeletal examination. Mayo Clin Proc 2020 Aug;95(8):1715-1731. [doi: 10.1016/j.mayocp.2020.05.026] [Medline: 32753146]
32. Kotnour J. Forward motion: clinical practice guidelines in telehealth physical therapy: a case report. Orthop Phys Thery Practice 2024;36(1):39-43.
33. McConnochie KM. Webside manner: a key to high-quality primary care telemedicine for all. Telemed J E Health 2019 Nov;25(11):1007-1011. [doi: 10.1089/tmj.2018.0274] [Medline: 30648924]
34. Modic MB, Neuendorf K, Windover AK. Enhancing your webside manner: optimizing opportunities for relationship-centered care in virtual visits. J Patient Exp 2020 Dec;7(6):869-877. [doi: 10.1177/2374373520968975] [Medline: 33457513]
35. Patino CM, Ferreira JC. Internal and external validity: can you apply research study results to your patients? J Bras Pneumol 2018 May;44(3):183. [doi: 10.1590/S1806-37562018000000164] [Medline: 30043882]
36. Lachin JM. The role of measurement reliability in clinical trials. Clin Trials 2004;1(6):553-566. [doi: 10.1191/1740774504cn057oa] [Medline: 16279296]
37. Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. Res Q Exerc Sport 1999 Jun;70(2):113-119. [doi: 10.1080/02701367.1999.10608028] [Medline: 10380242]
38. Yoon JP, Chung SW, Kim SH, Oh JH. Diagnostic value of four clinical tests for the evaluation of subscapularis integrity. J Shoulder Elbow Surg 2013 Sep;22(9):1186-1192. [doi: 10.1016/j.jse.2012.12.002] [Medline: 23434234]
39. Podsiadlo D, Richardson S. The timed “Up & Go”: a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991 Feb;39(2):142-148. [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]
40. Bergen G, Shakya I. CDC STEADI: evaluation guide for older adult clinical fall prevention programs. Centers for Disease Control and Prevention. 2019. URL: https://www.cdc.gov/steadi/pdf/Steadi-Evaluation-Guide_Final_4_30_19.pdf [accessed 2024-06-18]
41. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022 Jun;2022. [doi: 10.31478/202206e] [Medline: 36177208]
42. Gajarawala SN, Pelkowski JN. Telehealth benefits and barriers. J Nurse Pract 2021 Feb;17(2):218-221. [doi: 10.1016/j.nurpra.2020.09.013] [Medline: 33106751]
43. Nolet PS, Yu H, Côté P, et al. Reliability and validity of manual palpation for the assessment of patients with low back pain: a systematic and critical review. Chiropr Man Therap 2021 Aug 26;29(1):33. [doi: 10.1186/s12998-021-00384-3] [Medline: 34446040]
44. Benbassat J, Baumal R. Narrative review: should teaching of the respiratory physical examination be restricted only to signs with proven reliability and validity? J Gen Intern Med 2010 Aug;25(8):865-872. [doi: 10.1007/s11606-010-1327-8] [Medline: 20349154]

XSL•FO
RenderX

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act
PCP: primary care provider
PT: physical therapist
TUG: Timed Up and Go

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna; submitted 21.09.23; peer-reviewed by M Avdagovska, S Hinder; revised version received 12.02.24; accepted 09.05.24; published 26.06.24.

Please cite as:
DeLaRosby A, Mulcahy J, Norwood T
A Proposed Decision-Making Framework for the Translation of In-Person Clinical Care to Digital Care: Tutorial
JMIR Med Educ 2024;10:e52993
URL: https://mededu.jmir.org/2024/1/e52993
doi:10.2196/52993

© Anna DeLaRosby, Julie Mulcahy, Todd Norwood. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.6.2024. 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.

XSL•FO
RenderX

Digital Health Education for the Future: The SaNuRN (Santé Numérique Rouen-Nice) Consortium's Journey

Julien Grosjean[1,2,*], PhD; Frank Dufour[3,*], PhD; Arriel Benis[4,*], PhD; Jean-Marie Januel[1], RN, PhD; Pascal Staccini[3], MD, PhD; Stéfan Jacques Darmoni[1,2], MD, PhD

1

2

3

4

*these authors contributed equally

Corresponding Author:
Arriel Benis, PhD

Abstract

Santé Numérique Rouen-Nice (SaNuRN; "Digital Health Rouen-Nice" in English) is a 5-year project by the University of Rouen Normandy (URN) and Côte d'Azur University (CAU) consortium to optimize digital health education for medical and paramedical students, professionals, and administrators. The project includes a skills framework, training modules, and teaching resources. In 2027, SaNuRN is expected to train a significant portion of the 400,000 health and paramedical students at the French national level. Our purpose is to give a synopsis of the SaNuRN initiative, emphasizing its novel educational methods and how they will enhance the delivery of digital health education. Our goals include showcasing SaNuRN as a comprehensive program consisting of a proficiency framework, instructional modules, and educational materials and explaining how SaNuRN is implemented in the participating academic institutions. SaNuRN is aimed at educating and training health and paramedical students in digital health. The project is a cooperative effort between URN and CAU, covering 4 French departments. It is based on the *French National Referential on Digital Health* (*FNRDH*), which defines the skills and competencies to be acquired and validated by every student in the health, paramedical, and social professions curricula. The SaNuRN team is currently adapting the existing URN and CAU syllabi to *FNRDH* and developing short-duration video capsules of 20-30 minutes to teach all the relevant material. The project aims to ensure that the largest student population earns the necessary skills, and it has developed a 2-tier system involving facilitators who will enable the efficient expansion of the project's educational outreach and support the students in learning the needed material efficiently. With a focus on real-world scenarios and innovative teaching activities integrating telemedicine devices and virtual professionals, SaNuRN is committed to enabling continuous learning for health care professionals in clinical practice. The SaNuRN team introduced new ways of evaluating health care professionals by shifting from a knowledge-based to a competencies-based evaluation, aligning with the Miller teaching pyramid and using the Objective Structured Clinical Examination and Script Concordance Test in digital health education. Drawing on the expertise of URN, CAU, and their public health and digital research laboratories and partners, SaNuRN represents a platform for continuous innovation, including telemedicine training and living labs with virtual and interactive professional activities. SaNuRN provides a comprehensive, personalized, 30-hour training package for health and paramedical students, addressing all 70 *FNRDH* competencies. The project is enhanced using artificial intelligence and natural language processing to create virtual patients and professionals for digital health care simulation. SaNuRN teaching materials are open access. It collaborates with academic institutions worldwide to develop educational material on digital health in English and multilingual formats. SaNuRN offers a practical and persuasive training approach to meet the current digital health education requirements.

(JMIR Med Educ 2024;10:e53997) doi:10.2196/53997

KEYWORDS

digital health; medical informatics; education; health education; curriculum; students; teaching materials; hybrid learning; program development; capacity building; access to information; e-learning; open access; open data; skills framework; competency-based learning; telemedicine training; medical simulation; objective structured clinical examination; OSCE; script concordance test; SCT; virtual patient

XSL•FO
RenderX

Introduction and Background

Digital health and health informatics are at the crossroads of medicine and health sciences, computer science and engineering, information and communication sciences, mathematics, statistics, technology, and innovation management [1]. Digital health has been a component of regular training in medical schools for 40 years [2-4] under different labels, such as medical informatics [5], medical computing (in the United States) [6], and e-health [7], with high heterogeneity in content at the national level. In France, digital health is a subdomain of public health, which is also part of the training in medical and paramedical schools [8].

In 2022, the French Ministry of Health, and in particular its Delegation of Digital Health, published an open call for project proposals to support innovative approaches to develop initial academic and continuing professional education in digital health to health-related students, professionals and administrators; law specialists; computer scientists; and data protection officers. "Health-related students and professionals" was mainly referring to students enrolled in health-related programs, including medicine; odontology; pharmacy; midwifery; and paramedical fields such as nursing, physiotherapy, speech therapy, and hearing-aid technician, as well as training programs for social workers [9]. A budget of €71 million (US $75.73 million) has been secured to achieve this specific call to deal with the expected need to train over 400,000 health and paramedical professions students in 2027 at the national level.

The University of Rouen Normandy (URN) and the Côte d'Azur University (CAU), as a consortium, have successfully answered this call by getting a 5-year grant for their joint project, *Santé Numérique Rouen-Nice* (SaNuRN; "Digital Health Rouen-Nice" in English) [10]. SaNuRN began on September 1, 2022, with a cost estimate of €6,891,923 (US $7,351,441) and a grant contribution of €3,951,200 (US $4,214,646), with the goal of training around 30,000 students by 2027.

Before the initiation of this national project in France, there was a notable deficiency in digital health training for health students and practically none in paramedical schools. The primary focus was on health students pursuing master's degrees, such as medicine, pharmacy, dentistry, and midwifery. For instance, a national master's program in medical informatics has been established at Sorbonne University for the past 25 years. However, up until 2020, there was no existing digital health training curriculum for health students at the bachelor's degree level. Consequently, a comprehensive curriculum in digital health had to be developed from scratch for both health and paramedical students at the bachelor's degree level.

Before the SaNuRN project, a 10-hour module was introduced for all first-year medical students in CAU in 2020, and in URN, a 20-hour module was implemented for some first-year medical students in 2021. The open call from the Delegation of Digital Health at the French Ministry of Health emphasized allocating 80% of the training effort to the bachelor's degree level. One of the challenges of the SaNuRN project was assembling a team of digital health specialists to train all health-related students. The initial 2 years of the SaNuRN project (2022-2024) were dedicated to implementing a digital health teaching module for all health-related students, including both health and paramedical programs, at the bachelor's degree level.

This paper aims to provide an overview of the SaNuRN project, highlighting its pedagogical innovations and how its implementation will optimize digital health education. Our objectives are to present SaNuRN as a whole, comprising a skills framework, training modules, and teaching resources, and to describe how SaNuRN is and will be deployed in the consortium institutions. Below, we describe the SaNuRN project and its objectives. Next, we detail the skills framework and training modules, explaining the teaching resources and how they are deployed. Finally, we discuss the pedagogical innovations and expected impact of the SaNuRN project on digital health education and the quality of care and patient outcomes.

Building a Digital Health Education Lifelong Platform (SaNuRN) as a Cooperation Achievement

Overview

The SaNuRN project emerged in the context of a long-lasting cooperation between URN and CAU in *digital health* (SJD and PS), *medical simulation* (Professors Louis Sibert and Jean-Paul Fournier), and *general practice* (Professors Matthieu Schuers and David Darmon), as a primary use case for teaching digital health during postgraduate studies and residency. From our perspective, this extensive cooperation was a decisive factor in the grant application's success. URN is located in the northwest of France and CAU is located in the southeast. The distance between them is around 1000 km, and this points out the challenges related to the SaNuRN consortium, which is well managed by using as much dematerialized infrastructure as possible to deliver digital health teaching and learning content. Thus, the first-stage objective of SaNuRN is to educate and train all students in health-related fields in digital health at 4 French departments (Seine-Maritime and Eure in Normandy for URN, and Alpes-Maritimes and Var in Provence-Alpes-Côte d'Azur for CAU), to cover a population of 4 million inhabitants; the target of SaNuRN is to train about 2800 health-related students each year.

Targeted Skills and Competencies

To support this effort, the SaNuRN team activities are based on the *French National Referential on Digital Health* (*FNRDH*) [11]. Created and published in 2021, this referential gives a framework and defines the skills and competencies to be acquired and validated by every student in the health, paramedical, and social professions curricula [8]. The exhaustive list of skills and competencies of the *FNRDH* is detailed in Multimedia Appendix 1. The skills defined in the *FNRDH* are organized into five competency categories: (1) security, (2) health data, (3) communication in health, (4) digital tools in health, and (5) telehealth and teleactivities. *FNRDH* is built around a three-level hierarchy: (1) the 5 competencies as introduced above, (2) a total of 25 subcompetencies (eg, to identify an end user or a health professional and to characterize

XSL•FO
RenderX

and manage nominative data, applying the European rules such as the General Data Protection Regulation [GPDR]), and (3) a total of 70 different abilities (eg, to understand the life cycle of the digital health data and to take actions against virus and malware). In June 2023, *FNRDH* was integrated [12] into the HeTOP terminology server [13] to create a Catalog and Index of Health Digital Teaching Resources (CIDHR) [14-16] to be usable by all French health and paramedical students.

Adapting Existing Resources to the FNRDH Framework

Since the beginning of the project in October 2022, the primary need has consisted of adapting the existing URN and CAU syllabi to *FNRDH*. The SaNuRN project builds in a matrix format to adapt *FNRDH* for each degree (bachelor', master's, doctorate, or residency) and each field of study (eg, medicine, nursing, or physiotherapy). Furthermore, the course focuses on digital health in each field of study and, at each degree level, is limited to 30 hours of lectures and practices to address the 70 competencies of the *FNRDH* skills framework. To manage this challenge, the SaNuRN team is developing short-duration video capsules of 20-30 minutes to teach all the relevant material adapted to the degree levels and fields of study by taking into account the expectations within each degree, discipline, and potential learning sites (URN, CAU, and their partners).

It is essential to notice that the pedagogical components of the SaNuRN project are derived from existing teaching resources previously developed by the 2 Departments of Digital Health at URN and CAU. For example, West Normandy has 7 nursing schools (partners of URN), and the SaNuRN project has adapted its training to each of them specifically.

Since the inception of the SaNuRN project, additional teaching modules have been introduced to address the list of *FNRDH* skills and competencies. Among these modules, one is dedicated to cybersecurity and another to health data. This supplementary course emphasizes practical applications, featuring instructional videos on the use of specific tools aligned with the skills and competencies of the *FNRDH*. These tools include (1) a secure email tool, (2) guidance on accessing Mon Espace Santé—a digital platform designed for citizens and patients to manage their digital documents actively, (3) Health French National Identification, and (4) ethics in health (refer to Figure 1). The SaNuRN consortium used existing videos from the French National Digital Health Agency to develop these new teaching resources.

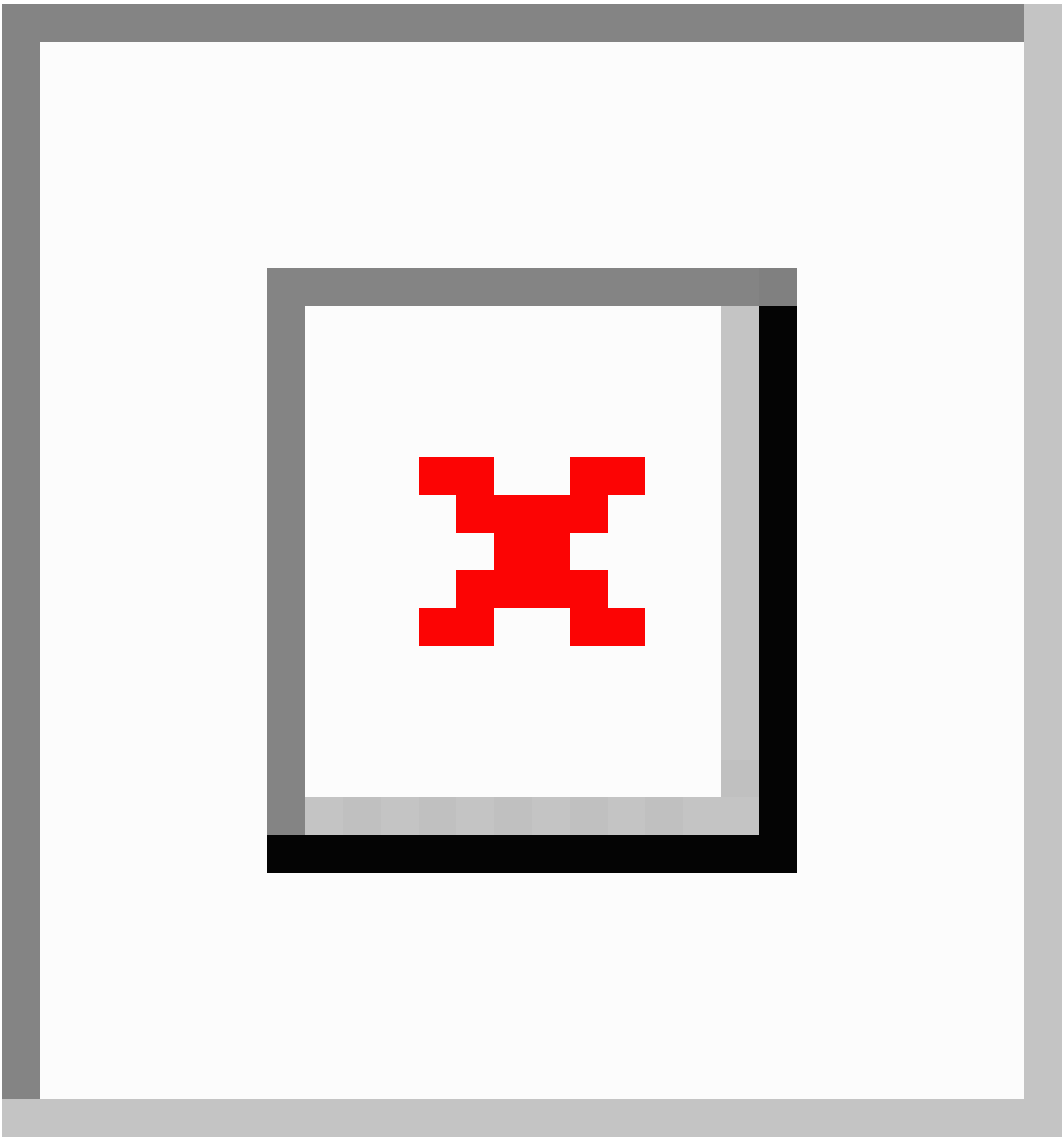
As of January 2024, the personalized 30-hour module is accessible in 2 modes: as freely available teaching resources for any health-related student through an open data website [17] and as specific video capsules within the URN and CAU private teaching environments. A total of 24 hours of preexisting resources, predating the SaNuRN project, were adapted to cater to various audiences, focusing on nurses and pharmacists. Overall, 80% (56/70) of the *FNRDH* competencies are covered by at least 1 SaNuRN teaching resource [17].

In line with the strategic decision made by the SaNuRN consortium in response to the French Delegation of Digital Health, all teaching materials generated during the project will be openly accessible on a website ("teaching open data") [17]. Furthermore, from these teaching materials (eg, cybersecurity), several short-duration capsules were developed to suit the specific needs of students in various specialties, including medicine, pharmacy, and nursing.

A significant advantage of SaNuRN is that all these resources are freely accessible to everyone, aiming to benefit health and paramedical students and professionals.

XSL•FO
RenderX

Figure 1. Overall schema of the SaNuRN project timeline. DES: Diplôme d'Etudes Spécialisées (Residency Program); OSCE: Objective Structured Clinical Examination; SaNuRN: Santé Numérique Rouen-Nice (Digital Health Rouen-Nice).



Facilitating Digital Health Education Adoption and Improvement

To ensure that the largest student population earns the necessary skills, the SaNuRN project has developed a 2-tier system. The first tier involves selecting teaching staff in each professional specialty field involved in the project that will act as facilitators. They will enable the efficient expansion of the project's educational outreach and facilitate periodic updating of the skills framework within each professional specialty. Additionally, these facilitators will support the students in learning the needed material efficiently. The training of these facilitators began in May 2023 for a year.

Based on the outcomes of the first tier, the second tier will consist of adjusting the educational resources that were initially only based on the *FNRDH*. This will facilitate the deployment, student and teaching staff engagement, and adoption of the digital health teaching modules in all health and paramedical specialties.

The SaNuRN Approach to Digital Health Education Innovation

Overview

As a part of the requirements of the SaNuRN's grant, 80% of the funding is dedicated to first-degree students (bachelor's), corresponding to most of the health and paramedical students

XSL•FO
RenderX

that are enrolled. The SaNuRN consortium has already largely fulfilled this objective by massively educating and training in traditional classroom settings or through self-training.

In the next 3 and half years (see Figure 1), the SaNuRN consortium will focus on pedagogical digital health innovations for the second and third degrees of all health-related fields of study. For example, the consortium has already planned pluriprofessional training sessions for the first semester of 2024 (eg, medicine residents with nursing students, both involved in specific teleconsultations). The first scheduled training session is about clinical data warehouses from various health and paramedical perspectives.

New paradigms are already present in the SaNuRN digital health syllabus and have been introduced to public health residents, particularly the paradigm of “One Digital Health” [18], defined as the intersection of *one health* and *digital health*. The SaNuRN team has developed two other innovations: (1) modification of the evaluation process with a shift from a knowledge-based to a competencies-based evaluation, as proposed by the Miller teaching pyramid [19], such as the Objective Structured Clinical Examination, and (2) Script Concordance Test in digital health, using a Health Professional Connected Office (see Figure 1).

Furthermore, several aspects will be mainly at the heart of the innovation of the SaNuRN project, as presented below.

Interactive and Innovative Components of Courses

The characteristics of the SaNuRN project are primarily the combination of knowledge and expertise from URN, CAU, and their public health and digital research laboratories and partners.

Integrated Telemedicine Devices

Two medical simulation centers, at URN [20] and at CAU [21], have already established living labs. These labs include simulated professional offices and patient apartments, providing a platform to test various software in different health situations, especially in general practice. Soon, 2 integrated telemedicine units will be available for health and paramedical students to test different health situations using scenarios of simulated patients.

By 2027, the SaNuRN project aims to implement several teaching modules that will be financially self-sustainable (ie, that will run in the future without the financial support available for the grant period). These modules will offer digital health training for continuous learning in general practice and private companies, including big pharma and health technology (or “medtech”) companies. For instance, a full-day teaching module has been developed to help private companies handle clinical data warehouses. The medical simulation centers in Rouen (URN) and Nice (CAU) will be used to conduct most of these training sessions.

Living Labs With Virtual and Interactive Professional Activities

The SaNuRN program includes a conversational virtual clinical simulator using artificial intelligence techniques combined with natural language processing, with the modeling of clinical situations defined for the training of all health care professionals (eg, physicians, pharmacists, nurses, and physiotherapists).

Analogous to expert systems, this tool, on the one hand, will be able to play the role of a professional (ie, backward chaining), asking questions to a patient while adjusting to the patient’s responses (similarly to a computer-assisted diagnostic aid). On the other hand, the system will play the role of the patient (ie, forward chaining), answering a professional’s questions (simulated clinical examination of the virtual patient).

A Comprehensive Overview of SaNuRN’s First Achievements

Digital Health Education Before SaNuRN (September 2022)

Before SaNuRN (September 2022), digital health was already taught in URN and CAU. Indeed, for example, first-year students at the health schools at URN and CAU received an initial and primary education on digital health. Specifically, since 2021, a total of 15% (150/1000) of the students at URN took a 20-hour course as a part of a minor in health digital science, and since 2020, a total of 100% (1000/1000) at CAU received a 10-hour mandatory course. These digital health courses are performed in a traditional large classroom setting at URN and CAU.

Adaptations in Nursing Schools

In West Normandy, a teaching self-learning module was provided to the 7 nursing schools, representing 600 nursing students. In the first semester of 2022, the teaching module was directly derived from the one provided for health students, with an identical duration of 20 hours. Very quickly, in response to the feedback from nursing students, the SaNuRN consortium analyzed the teaching discrepancies between the real needs of these nursing students and the content of the digital health teaching module. Therefore, a 6-hour training was reorganized for the second semester. Furthermore, for each teaching module, all the examples provided were modified and adapted to the nursing student’s needs (eg, to demonstrate the need for health smart cards in their specific practice).

Implementation and Expansion During SaNuRN’s First Year

During the first academic year of the SaNuRN project (2022-2023), around 2000 students specializing in health care and paramedical specialties were trained. According to our knowledge and the various national agencies involved in the program, this number is significantly higher than those at other French universities. The West Normandy nursing schools provided only 20 hours of digital health training. These schools have requested an additional 10-hour course in the third year of the curriculum to fulfill the 30-hour teaching requirement.

Expansion and Hybrid Learning

In the ongoing academic year (2023-2024), the SaNuRN consortium has engaged several new student cohorts. At URN, 200 second-year medical students (ie, students who passed the first highly selective year at the school of health and chose to

XSL•FO
RenderX

study medicine) have participated in hybrid training, including 4 hours of face-to-face organized courses and 25 hours of self-training (see Figure 1). Additionally, around 100 second-year pharmacy students (ie, similar to students who chose to study medicine, except they chose to study pharmacy instead) may opt for this digital health teaching module. A total of 100 physiotherapy and ergotherapy students are also involved in the project, dedicating 24 hours to self-training. For all these new URN students, the SaNuRN team proposes the following module: a 2-hour introduction teaching (see Figure 1).

At CAU, 400 nursing students from 4 nursing schools will participate in the SaNuRN project in 2024. Lastly, 20 public health residents from URN and CAU will have access to advanced teaching resources. Thus, 100% of medical and paramedical students will be trained in digital health at both universities in 2025.

The Delegation of Digital Health of the French Ministry of Health aims to educate and train 400,000 students in digital health by 2027 using a 30-hour module based on the *FNRDH* guideline. The SaNuRN project plans to teach 13,200 students over 5 years.

The overall SaNuRN project during the 5 years is summarized in Figure 1; most of the effort is made for students in the first degree of their studies to attain the 100% rate of trained students in digital health. The SaNuRN consortium will fit this goal in June 2025. Then, specific contents are available for second and third degrees to improve the knowledge and competencies in specific situations and medical and paramedical disciplines (eg, videos and live demonstrations of teleconsulting with nurses, physicians, or physiotherapists).

Training Trainers

Goals and Framework

Since May 2023, specific training sessions have been performed for digital health trainers. The "Training the Trainers" component of the SaNuRN project is instrumental to attaining the project's primary goal, that is, providing education on digital health for undergraduate students of all medical, paramedical, and social disciplines in the academic year 2024-2025. The goals of this component are to obtain from the trainees—who all are faculty members actively teaching in the various academic programs and institutions responsible for undergraduate education in medicine, paramedicine, and social work—the most accurate information about their students such as their profiles, schedules and course works, and preparation for the discipline of digital health.

Program Structure and Implementation

This information is further used to:

- Design the most appropriate pedagogical resources in terms of format, depth of knowledge, types of learning activities, and modes of assessment.
- Evaluate and train the faculty members in the discipline of digital health.
- Train them in the design of e-learning curricula and the use of digital pedagogical resources.
- Prepare them for the integration of the 30 hours of education to digital health in their respective programs.

Core Activities and Learning Objectives

The "Training the Trainers" program has been organized as a yearlong, ongoing, asynchronous, and remote training activity primarily to respond to the significant disparity regarding the trainees' availability, who, for the most part, could not commit to a fixed time slot, and to allow for an extensive immersion within the discipline itself and consistent exposure to digital technologies. All the collaborative and remote tools used in the development and course of this program were unknown to the trainees, and it took time and practice for all of them to attain a good level of proficiency and confidence.

This model allowed the program to admit new trainees at different stages and moments of its development.

Development of Pedagogical Resources

The program started in May 2023 with 15 faculty members enrolled. It is scheduled to last until the end of May 2024, with, as of today, 34 members representing all the disciplines concerned with the integration of new courses in digital health.

The core activity of the program consists of the complete understanding of the reference framework (*FNRDH*; Multimedia Appendix 1) and the planning and design of its integration within existing courses and academic programs. Through a collective explication of all capacities included in the *FNRDH*, the group of trainees has identified 5 learning topics forming the core common foundation shared by all health-related disciplines: "cybersecurity," the "digital health system," "digital communication," "digital professional communication," and "further developments of digital health." In order to accommodate the various specificities of the existing academic programs, each of these topics has been divided into teaching modules of roughly 20 minutes, allowing easy customization and integration into existing curricula.

All program trainees contribute to designing and producing these 30 modules, representing the first 10 hours of education in digital health for first-year undergraduate students in health-related disciplines. These modules are designed to be delivered as autonomous self-teaching, asynchronous modules, thus allowing all faculty members to monitor students' activity and progress with their own methods and tools.

Together with these modules, the trainees are producing web-based interactive resources with the help of a partner of the SaNuRN program, IKIGAI, a nonprofit game design company. Two types of such interactive resources are currently being produced: a gamified quiz and a set of flashcards for practice and memorization.

Innovative Pedagogical Tools: Introduction of Learning and Assessment Scenario

The "Training the Trainers" program provides trainees with a fully immersive experience in digital communication and education and an in-depth analysis of the *FNRDH*, allowing them to clearly envision the multiple ramifications of the new discipline of e-health.

XSL•FO
RenderX

With the "Training the Trainers" program (see Figure 1), the SaNuRN project has introduced, at the undergraduate level, one major innovative pedagogical tool, the Learning and Assessment Scenario. This tool consists of a detailed outline of a complex professional situation involving digital tools and technologies and the collaboration of professionals from other disciplines. The students presented with this situation must engage in collaborative activities to assess the situation's multiple dimensions and propose a coordinated plan of action to solve the issue. This teaching tool prefigures tools used at the graduate and postgraduate levels, such as the Objective Structured Clinical Examination and Script Concordance Test. The Learning and Assessment Scenario also serves as an efficient tool to teach the much-needed interprofessional collaboration skills that are brought to higher levels of complexity and depth by digital technologies. With this progressive strategy, the educational program created by SaNuRN, covering the 3 cycles of medical, paramedical, and social work studies, creates a consistent continuum of educational engagement for faculty members and students in meaningful interactions with digital technologies.

Evaluation Plans

Currently, no formal (qualitative or quantitative) evaluation has been performed in the SaNuRN project; 2 qualitative evaluations have already been planned: in URN and CAU, 1 for medical students and 1 for nurse students. One indirect positive measure is the presence of health-related students in URN and CAU during the first year's training sessions. However, the presence was not mandatory, and over 90% of health-related students were present in the 20-hour training module in URN and 10-hour training module in CAU.

Discussion

Overview

During the first academic year of the SaNuRN project (2022-2023), around 2000 students specializing in health care and paramedical specialties received training, a significantly higher number than other French universities. In the ongoing academic year (2023-2024), various new student cohorts are participating in the project, including medical, pharmacy, physiotherapy, and ergotherapy students and public health residents. The SaNuRN project aims to educate 13,200 students over 5 years, contributing to the Delegation of Digital Health's goal of training 400,000 students by 2027.

The project primarily focuses on first-degree students in the initial years (bachelor's), with specific content for second- and third-degree students (master's and PhD or residency) to enhance knowledge and competencies in various medical and paramedical disciplines.

The SaNuRN consortium plans to introduce innovative teaching methods, including interprofessional training sessions, competency-based evaluations, and the use of telemedicine devices. Interactive and innovative course components, combined with living labs and virtual clinical simulators, form the core of the project's innovations.

The key strengths and limitations of the SaNuRN project rely on (1) the fulfillment of the French Ministry of Health's aim to make digital health learning mandatory and (2) compliance with professional international recommendations, even when the specificities for the French higher education system make it challenging.

Strengths

Fulfilling National Commitments With the SaNuRN Project

By September 2024, learning digital health will be mandatory for all health and paramedical students in France. At that time, the SaNuRN project will be able to fulfill the national commitment to teaching digital health by addressing the 70 *FNRDH* competencies in a 30-hour training package.

Fitting Global Trends in Digital Health Education

In addition to France, several countries are proposing digital health training at the national level. However, only a handful of countries have established such competencies for clinical practice in their core medical school curriculum [22]. In England, the National Health Service has launched the Digital Readiness Education program [23]. It aims to improve digital skills, understanding, knowledge, and awareness across the multidisciplinary health and care workforce to support new working methods. This program focuses on continuous training. A qualitative study evaluated digital competencies in Singapore for its national medical school curriculum (which included 4 medical schools) [22]. One of the main conclusions was the need to enhance the sharing of educational resources and expertise. This point is also crucial in the French program, so the SaNuRN project has decided to create "open access" and "open data" teaching resources that are shareable with all French health and paramedical schools. An experiment was also conducted in Italy, using Petri-Nets to improve digital health literacy [24].

Looking at Internationalizing the SaNuRN's Concept

An essential advantage of SaNuRN is that all teaching materials created during the project are freely available to anyone, even explicitly targeting all health and paramedical students and professionals [17]. Most of the teaching material is created in French. However, thanks to international collaborations with institutions such as the Holon Institute of Technology (HIT) in Israel, a large part of the SaNuRN material is coproduced and available in English. This approach allows these partners, URN and CAU, to use the relevant SaNuRN resources in their curricula. For example, some SaNuRN resources (lessons) have been cocreated or coenhanced with lecturers in charge of health data science courses of the Department of Digital Medical Technologies at the HIT in Israel [25].

Complying With the International Medical Informatics Association Recommendations

The International Medical Informatics Association has published 2 versions of its international recommendations in biomedical and health informatics education, initially in 2000 and revised in 2010 and 2023 [26]. The International Medical Informatics Association recommendations are a framework for national

XSL•FO
RenderX

initiatives in biomedical and health informatics education and for constituting international programs and exchange of students and teachers in this digital health field. Zainal et al [27] have proposed a scoping review on clinical informatics training in medical school education curricula; these authors proposed 4 main recommendations that are very similar to those used in the SaNuRN project: situating digital health curriculum within specific contexts, developing evidence-based guidelines for robust digital health education, developing validated assessment techniques to evaluate curriculum effectiveness, and equipping educators with relevant digital health training.

Limitations

Needing to Align With Other International Standards

The teaching model may not be entirely compatible with other international approaches; drawing inspiration from experiences in other countries and attempting to fit within a shared framework would be advisable.

Temporarily Focusing on Undergraduate Students

Because the project was principally focusing its efforts on undergraduate degrees for its first 2 years, no international collaboration was initiated, apart from a cooperation with the HIT in Israel, as such collaboration usually targets graduate and postgraduate levels. At the national level, for the master's and PhD degrees, the SaNuRN consortium is planning to cooperate in 2024 with several French universities (Sorbonne Université, Paris Cité, Besançon Université, and Rennes Université), as well as European universities, in particular University for Health Sciences, Medical Informatics and Technology in Austria and continuing its cooperation with the HIT in Israel.

Conclusion

The SaNuRN project addresses France's national commitment to teaching digital health. SaNuRN addresses all 70 *FNRDH* skills and competencies (Multimedia Appendix 1) with a comprehensive, personalized, 30-hour training package for each health or paramedical student according to their degree level, field of study, and university curriculum. This innovative approach is enhanced by using artificial intelligence and natural language processing to create virtual patients and professionals for digital health care simulation, allowing each student to replay and practice various clinical situations. SaNuRN teaching materials are openly accessible. Moreover, SaNuRN, aiming to answer new needs of the French health schools and paramedical professions, is collaborating with academic institutions worldwide to develop educational material in digital health in English and multilingual formats. SaNuRN offers an enhanced training approach that is both effective and persuasive, making it a challenging solution to the current digital health education requirements in France and potentially Europe and worldwide.

Acknowledgments

This work was supported by the Santé Numérique Rouen-Nice (SaNuRN) project (ANR_22-CMAS-0014 3.951.200), granted by the Delegation of Digital Health of the French Ministry of Health and the French National Research Agency.

Data Availability

Data sharing does not apply to this paper, as no data sets were generated or analyzed during this study.

Authors' Contributions

JG contributed to conceptualization, funding acquisition, investigation, methodology, project administration, supervision, visualization, and writing (original draft, review, and editing). FD contributed to methodology, supervision, and writing (review and editing). AB contributed to conceptualization, investigation, visualization, and writing (original draft, review, and editing). J-MJ contributed to methodology, supervision, and writing (review and editing). PS contributed to conceptualization, funding acquisition, investigation, methodology, project administration, supervision, visualization, and writing (original draft, review, and editing). SJD contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation, visualization, and writing (original draft, review, and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of skills and competencies of the *French National Referential on Digital Health* (*FNRDH*).
[PDF File, 145 KB - mededu_v10i1e53997_app1.pdf]

References

1. Benis A, Grosjean J, Billey K, et al. Medical Informatics and Digital Health Multilingual Ontology (MIMO): a tool to improve international collaborations. Int J Med Inform 2022 Nov;167:104860. [doi: 10.1016/j.ijmedinf.2022.104860] [Medline: 36084537]

XSL•FO
RenderX

2. Chen D, Gorla J. The need to develop digital health competencies for medical learners. Med Teach 2023 Jul;45(7):790-791. [doi: 10.1080/0142159X.2023.2178886] [Medline: 36787406]
3. Rachmani E, Haikal H, Rimawati E. Development and validation of Digital Health Literacy Competencies for Citizens (DHLC), an instrument for measuring digital health literacy in the community. Comput Methods Programs Biomed Update 2022;2:100082. [doi: 10.1016/j.cmpbup.2022.100082] [Medline: 36407680]
4. Kleib M, Arnaert A, Nagle LM, et al. Digital health education and training for undergraduate and graduate nursing students: a scoping review protocol. JBI Evid Synth 2023 Jul 1;21(7):1469-1476. [doi: 10.11124/JBIES-22-00266] [Medline: 36728743]
5. Collen MF. Origins of medical informatics. West J Med 1986 Dec;145(6):778-785. [Medline: 3544507]
6. Kaplan B. The medical computing “lag”: perceptions of barriers to the application of computers to medicine. Int J Technol Assess Health Care 1987;3(1):123-136. [doi: 10.1017/s026646230001179x] [Medline: 10282217]
7. Eysenbach G. What is e-health? J Med Internet Res 2001;3(2):E20. [doi: 10.2196/jmir.3.2.e20] [Medline: 11720962]
8. Ministère de l’Enseignement supérieur et de la Recherche, Ministère de la Santé et de la Prévention. Arrêté du 10 novembre 2022 relatif à la formation socle au numérique en santé des étudiants en santé. Légifrance. 2022 Nov 10. URL: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000046548689 [accessed 2023-09-29]
9. Appel à manifestation d’intérêt (AMI) « compétences et métiers d’avenir » France 2030 - adapter l’appareil de formation des jeunes et des salariés aux métiers de demain. Ministère du Travail, de la Santé et des Solidartiés. 2023. URL: https://travail-emploi.gouv.fr/actualites/l-actualite-du-ministere/article/appel-a-manifestation-d-interet-ami-competences-et-metiers-d-avenir [accessed 2024-04-17]
10. Université de Rouen, Université Côte d’Azur. Sante Numerique Rouen Nice: digital health training program. SaNuRN. 2023. URL: https://sanurn.eu/ [accessed 2024-04-17]
11. Délégation Ministérielle au Numérique en Santé. Numérique en santé: référentiel socle et transversal de compétences. G_NIUS. 2022. URL: https://gnius.esante.gouv.fr/sites/default/files/2022-03/Référentiel de compétences numérique en santé.pdf [accessed 2024-04-17]
12. RCSN: référentiel socle et transversal de compétences en santé numérique. HeTOP. 2022. URL: https://www.hetop.eu/hetop/rep/fr/TER_RCSN/ [accessed 2024-04-17]
13. HeTOP. 2023. URL: https://www.hetop.eu/hetop/en/ [accessed 2022-04-29]
14. CISMeF. CIDHR: Catalog and Index of Digital Health Teaching Resources on the internet. CHU de Rouen. 2023. URL: https://doccismef.chu-rouen.fr/dc/#env=cidhr [accessed 2023-09-23]
15. Darmoni S, Benis A, Lejeune E, et al. Digital health multilingual ontology to index teaching resources. Stud Health Technol Inform 2022 Aug 31;298:19-23. [doi: 10.3233/SHTI220900] [Medline: 36073449]
16. Grosjean J, Benis A, Dufour JC, et al. Sharing digital health educational resources in a one-stop shop portal: tutorial on the Catalog and Index of Digital Health Teaching Resources (CIDHR) semantic search engine. JMIR Med Educ 2024 Mar 4;10:e48393. [doi: 10.2196/48393] [Medline: 38437007]
17. Cours de santé numérique (digital health). CISMeF. 2023. URL: https://www.cismef.org/cismef/d2im/cours/ [accessed 2024-01-29]
18. Benis A, Tamburis O, Chronaki C, Moen A. One Digital Health: a unified framework for future health ecosystems. J Med Internet Res 2021 Feb 5;23(2):e22189. [doi: 10.2196/22189] [Medline: 33492240]
19. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
20. Medical training center. Rouen University Hospital. 2023. URL: https://www.mtc-rouen.com/ [accessed 2024-04-17]
21. Centre de simulation médicale Nice - Harvard. Canal-U. 2008 Oct 15. URL: https://www.canal-u.tv/chaines/univcotedazur/medecine/centre-de-simulation-medicale-nice-harvard [accessed 2024-04-17]
22. Zainal H, Xiaohui X, Thumboo J, Kok Yong F. Digital competencies for Singapore’s national medical school curriculum: a qualitative study. Med Educ Online 2023 Dec;28(1):2211820. [doi: 10.1080/10872981.2023.2211820] [Medline: 37186901]
23. About digital readiness education. National Health Service. 2023 Apr 4. URL: https://digital-transformation.hee.nhs.uk/about-digital-readiness-education/about-digital-readiness-education [accessed 2023-10-03]
24. Ricci FL, Consorti F, Pecoraro F, Luzi D, Tamburis O. A Petri-Net-based approach for enhancing clinical reasoning in medical education. IEEE Trans Learning Technol 2022 Apr 1;15(2):167-178. [doi: 10.1109/TLT.2022.3157391]
25. Digital medical technologies - overview. Holon Institute of Technology. 2023. URL: https://www.hit.ac.il/en/DMT/Overview [accessed 2024-04-17]
26. Bichel-Findlay J, Koch S, Mantas J, et al. Recommendations of the International Medical Informatics Association (IMIA) on education in biomedical and health informatics: second revision. Int J Med Inform 2023 Feb;170:104908. [doi: 10.1016/j.ijmedinf.2022.104908] [Medline: 36502741]
27. Zainal H, Tan JK, Xiaohui X, Thumboo J, Yong FK. Clinical informatics training in medical school education curricula: a scoping review. J Am Med Inform Assoc 2023 Feb 16;30(3):604-616. [doi: 10.1093/jamia/ocac245] [Medline: 36545751]

XSL•FO
RenderX

Abbreviations

CAU: Côte d'Azur University
CIDHR: Catalog and Index of Health Digital Teaching Resources
FNRDH: French National Referential on Digital Health
GPDR: General Data Protection Regulation
HIT: Holon Institute of Technology
SaNuRN: Santé Numérique Rouen-Nice (Digital Health Rouen-Nice)
URN: University of Rouen Normandy

Edited by F Pietrantonio, I Said-Criado, JL Castro, M Montagna, TDA Cardoso; submitted 27.10.23; peer-reviewed by A Arbabisarjou, JJ Beunza, M Wolfien, O Tamburis; revised version received 16.03.24; accepted 21.03.24; published 30.04.24.

Please cite as:
Grosjean J, Dufour F, Benis A, Januel JM, Staccini P, Darmoni SJ
Digital Health Education for the Future: The SaNuRN (Santé Numérique Rouen-Nice) Consortium's Journey
JMIR Med Educ 2024;10:e53997
URL: https://mededu.jmir.org/2024/1/e53997
doi:10.2196/53997

© Julien Grosjean, Frank Dufour, Arriel Benis, Jean-Marie Januel, Pascal Staccini, Stéfan Jacques Darmoni. Originally published in JMIR Medical Education (https://mededu.jmir.org), 30.4.2024. 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.

XSL•FO
RenderX

Original Paper

Multidisciplinary Design–Based Multimodal Virtual Reality Simulation in Nursing Education: Mixed Methods Study

Ji-Young Yeo[1], PhD; Hyeongil Nam[2], PhD; Jong-Il Park[2], PhD; Soo-Yeon Han[3], PhD

[1]College of Nursing, Hanyang University, Seoul, Republic of Korea

[2]Department of Computer Science, Hanyang University, Seoul, Republic of Korea

[3]Department of Nursing, Bucheon University, Bucheon, Republic of Korea

Corresponding Author:
Soo-Yeon Han, PhD
Department of Nursing
Bucheon University
sosa-ro 56
Bucheon, 14774
Republic of Korea
Phone: 82 326108312
Fax: 82 326108300
Email: sooyeonhan@bc.ac.kr

Abstract

Background: The COVID-19 pandemic underscored the necessity for innovative educational methods in nursing. Our study takes a unique approach using a multidisciplinary simulation design, which offers a systematic and comprehensive strategy for developing virtual reality (VR) simulations in nursing education.

Objective: The aim of this study is to develop VR simulation content for a pediatric nursing module based on a multidisciplinary simulation design and to evaluate its feasibility for nursing education.

Methods: This study used a 1-group, posttest-only design. VR content for pediatric nursing practice was developed by integrating the technological characteristics of a multimodal VR system with the learning elements of traditional nursing simulation, combining various disciplines, including education, engineering, and nursing. A user test was conducted with 12 nursing graduates (preservice nurses) followed by post hoc surveys (assessing presence, VR systems, VR sickness, and simulation satisfaction) and in-depth, one-on-one interviews.

Results: User tests showed mean scores of 4.01 (SD 1.43) for presence, 4.91 (SD 0.81) for the VR system, 0.64 (SD 0.35) for VR sickness, and 5.00 (SD 1.00) for simulation satisfaction. In-depth interviews revealed that the main strengths of the immersive VR simulation for pediatric pneumonia nursing were effective visualization and direct experience through hands-on manipulation; the drawback was keyword-based voice interaction. To improve VR simulation quality, participants suggested increasing the number of nursing techniques and refining them in more detail.

Conclusions: This VR simulation content for a pediatric nursing practice using a multidisciplinary educational design model was confirmed to have positive educational potential. Further research is needed to confirm the specific learning effects of immersive nursing content based on multidisciplinary design models.

(JMIR Med Educ 2024;10:e53106) doi:10.2196/53106

KEYWORDS

multidisciplinary; multimodal; nursing; simulation; virtual reality; VR; education; allied health; educational; simulations; pediatric; pediatrics; paediatric; paediatrics; feasibility; nurse; nurses; qualitative; interview; interviews; development; develop; teaching; educator; educators; user test; user testing; module; modules; usability; satisfaction

XSL•FO
RenderX

Introduction

Overview

Virtual reality simulation (VRS) education, which integrates the latest information and communications technology innovations into simulation education, is a cost-effective solution for nursing practice education. It allows for attainable and predictable results without limitations of time and place. Consequently, VRS is expected to be extensively used in nursing curricula [1]. The demand for immersive education is growing rapidly following the clinical practice challenges experienced during the COVID-19 pandemic [2]. Various studies have been conducted to enhance immersion, a major advantage of virtual reality (VR)–based educational content; increase the educational efficacy of VR-based content; and minimize the side effects of the technology [3].

VR technology for on-site nursing simulation education is in its early stages and is centered on 2D web-based simulations, primarily using computers and monitors [4]. Even in realistic 3D simulation education, most studies are based on fragmentary skill-oriented scenarios using handheld devices [5].

It is essential to clarify the desired direction of immersive education, learning situational elements, and skill levels required for students to use immersive simulations in practical nursing training effectively. This approach should be based on an optimal theoretical framework and educational design [6]. Since immersive simulation is an educational method that integrates elements of traditional nursing simulation education and engineering elements, to maximize the effect of education using immersive simulation, the learning design must consider various elements of each area.

Despite significant advancements in IT and a growing interest in immersive content, little research exists on multidisciplinary design models for immersive education and the feasibility of design model-based content [7-9].

This study aimed to systematically develop 3D nursing simulation content based on a VR educational design model and the National League for Nursing (NLN) Jeffries Simulation Theory. It also conducted user tests to evaluate its feasibility in nursing education. We propose a multidisciplinary VRS content development model and provide the basis for future VR-based nursing education.

Theoretical Framework

The VR-based nursing simulation content was developed using the NLN Jeffries Simulation Theory [10] and Han’s [11] VR-based Educational Simulation Model. The selection of these 2 theories as a framework for this study was a strategic approach to systematically consider the characteristics of VR media and educational elements to maximize learning effectiveness. Rather than simply borrowing VR technical elements for nursing education, we sought to develop the most effective educational content in a multidisciplinary manner by integrating technology, educational elements, and nursing. The NLN Jeffries Simulation Theory provides a well-established foundation in nursing education, while Han’s [11] VR-based Educational Simulation Model offers specific guidelines for creating immersive and user-centered VR experiences.

The NLN Jeffries Simulation Theory is widely used in nursing education to guide the development of simulation-based learning experiences. This theory presents components as participant factors, facilitator factors, educational strategies factors, and expected outcomes, offering versatility in their application in nursing education. By outlining specific components that can be tailored to the needs and objectives of learners, the theory provides a robust framework for educators. It maximizes the effectiveness of simulation as a teaching strategy, ensuring that educational goals are met and that learners are better prepared for clinical practice. Especially as educational strategies factors, this includes the design and execution aspects of simulation activities, such as the complexity of simulations, feedback mechanisms, and debriefing methods. These factors are key to optimizing the learning environment and supporting the achievement of educational goals. Endorsed by the NLN, this theory has become a cornerstone in the field, guiding educators on how to integrate simulation into their teaching practices effectively.

Han’s [11] research aims to develop and validate design principles for optimizing VR-based educational simulations. It explores ways to use VR to extend users’ learning experiences into realistic contexts. Han’s [11] model comprises 12 design principles based on the 3 categories of contextual scenario, affordance in the simulation, and user activity and response. These categories emphasize creating immersive learning experiences that more closely resemble real-life situations, thereby enhancing the learner’s engagement and facilitating a deeper understanding of the subject matter. The 12 design principles encompass creating realistic scenarios, engaging user actions, and reflective activities closely mirroring real-life contexts. Applying these principles to nursing education simulations is justified as they enhance experiential learning, improve critical thinking and decision-making skills, and provide a safe environment for clinical practice. The principles ensure that the simulations are relevant, technologically apt, and realistically mimic health care settings, thereby making theoretical knowledge tangible and facilitating embodied learning.

This theoretical framework can provide a structured approach to developing simulations that are both educational and reflective of actual nursing practice. Thus, Han’s [11] model has 2 main advantages when applied to develop this nursing simulation content. First, it comprehensively considers educational engineering elements and the technical characteristics of VR. Second, it is subdivided into 3 clear categories and 12 subcategories, specifying the characteristics of each area. This structure allows for the provision of specific and clear guidelines for application in developing this nursing simulation content. We comprehensively considered the characteristics of VR, a technical element, as well as nursing and educational engineering elements, and selected an appropriate model to develop optimal immersive content based on a multidisciplinary perspective (Textbox 1).

XSL•FO
RenderX

Textbox 1. Immersive content development based on virtual reality (VR) education engineering design principles.

Principle of replicating real-life problems:

- To reflect the nature and importance of real-life problems, a contextual scenario was constructed using the clinical pathway of a disease to allow the learner to experience the entire flow of actual clinical practice for a specific disease and the corresponding nursing care.

Principle of adequacy of VR technology:

- Technique to perform nursing skills with bare hands without a hand device was applied through a variety of technologies, allowing the learner to have experiences similar to clinical practice.
- Using deep learning technology, the virtual caregiver was configured to give feedback to the learner with a verbal response.

Principle of similarity to real environment:

- To construct a practice environment as similar to real-life situations as possible; actual photos were provided as reference materials when creating virtual objects, which was refined through several revisions.

Principle of structural planning:

- Based on the Jeffries Simulation Theory template, learning goals for each module were set, upon which simulation elements and specific learning contents were organized.
- A flowchart was designed to show the VR simulation (VRS) content deployment order according to the user's activities.

Principle of implementing a professional approach:

- The content was algorithmically designed to recognize appropriate or inappropriate nursing procedures performed by learners and to provide feedback according to their level of performance, allowing them to ultimately embody the knowledge, skills, attitudes, and so forth, that was required of a nurse.

Principle of structured activity deployment:

- Contents and screens were designed according to the expected progress of the learner's activities, and a storyboard including the main contents was constructed according to the progress of the simulation activities.

Principle of simple-to-complex process:

- The modules were structured to align with the sequential order of patient care from admission to discharge, ensuring that activities are carried out in a systematic and sequential manner.

Principle of virtual recognition:

- To enhance the learner's sense of presence and awareness of activity direction in the VR space, the simulation modules were designed to display the learner's hands or body parts on the screen.

Principle of the reality of operation and selection:

- To enable realistic exploration, manipulation, and selection activities, real-life activities and voice responses were presented.
- The nursing outcomes were algorithmically configured to vary depending on the learner's VRS activities.

Principle of providing relevant information:

- The content was designed to assist the learner with the initial orientation and facilitate nursing activities by providing information on the envisaged simulation activities.
- In the introduction stage, instructional materials and preinstructional videos were provided.
- In the activity stage, pop-up quizzes or audio formats were used to provide information on the elements and procedures that the learner needed to consider during simulation performance.

Principle of promoting critical thinking:

- The problem situations related to pediatric patients hospitalized with pneumonia, learning objectives, and framework for required nursing interventions were presented.
- Critical thinking was promoted by allowing for decision-making and implementation within the presented situation.
- Use of tablet devices was included in the content to allow the learner to retrieve information on the patient's conditions and related information whenever necessary.

Principle of encouraging critical reflection:

XSL•FO
RenderX

- Feedback was provided to the learner during the simulation process through quizzes or a virtual agent's responses in a manner to give them the opportunity to reflect on the appropriateness of their own nursing actions.
- A performance checklist was displayed on the screen just before the end of the VRS, and feedback videos were provided after the simulation ended to allow the learner to reflect on their nursing performance again.

Objectives

Therefore, this study's theoretical framework, which integrates nursing theory and educational technology models, can provide a structured approach to developing VRSs for educational purposes that reflect actual nursing practice situations.

Methods

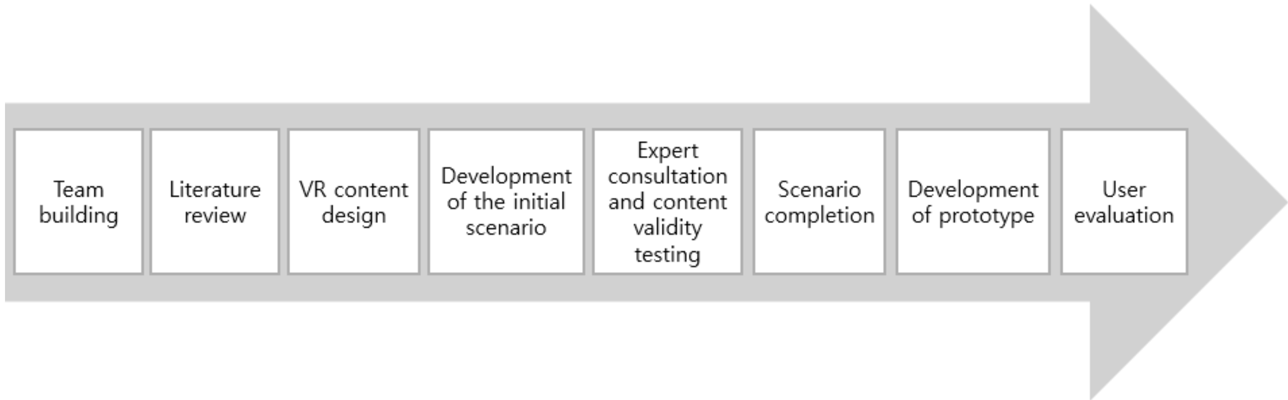
Study Design

The study used a 1-group, posttest-only design to develop and evaluate immersive simulation content in nursing education.

Procedures

The VRS content was developed via (1) team building, (2) literature review, (3) VR content design, (4) development of the initial scenario, (5) expert consultation and content validity testing, (6) scenario completion, (7) development of prototype content, and (8) user evaluation (Figure 1). This process can be broadly divided into 2 parts—development of content and user evaluation.

Figure 1. Research overview. VR: virtual reality.



Part 1: Development of Immersive Simulation Content

Step1: Team Building

The 5 developers (2 nursing experts, 2 IT developers, and 1 3D-modeling expert) produced the immersive content throughout the study from October 2020 to April 2021. A total of 8 nursing professionals (3 professors and 5 clinicians) supported the scenario development process.

First, the project team drafted a pediatric pneumonia nursing scenario using VR-based educational design principles and the Jeffries Simulation Theory. The 8 experts evaluated the validity of the scenario's content.

The project team followed the design process throughout the content development and technical work. The final prototype content was iteratively refined via expert group meetings. It has been modified several times to better reflect possible clinical situations at a level that can be implemented in a virtual environment. The content validity of the developed scenario was confirmed by experts. The prototypes were categorized into 2 groups—low-fidelity prototypes using 2D materials, such as printed matter, and high-fidelity prototypes using content development programming techniques to reflect the actual environment accurately [12]. In this study, high-fidelity prototypes were developed to provide a training experience similar to that in a real clinical environment to evaluate the content applicability to nursing education.

Step 2: Literature Review

The project team conducted an extensive literature review to gather and analyze existing research on VR-based educational simulations. This review provided a foundation for developing the scenario and identifying best practices and potential challenges in VRS.

Step 3: VR Content Design

This design phase included defining the learning objectives, creating detailed storyboards for each scenario, and outlining the technical requirements for the VR environment. The design ensured that the content was pedagogically sound and technically feasible. The design phase also involved specifying the interactive elements and feedback mechanisms that would be integrated into the VR experience to enhance engagement and learning efficacy.

Step 4: Development of the Initial Scenario

Detailed flowcharts were created to visualize the learner's journey through the simulation, ensuring that each step aligned with the intended learning outcomes. The initial scenario for pediatric pneumonia was developed, integrating the educational design principles and the Jeffries simulation theory. This

XSL•FO
RenderX

scenario included specific nursing tasks and clinical decisions that learners would need to perform and make during the simulation.

Step 5: Expert Consultation and Content Validity Testing

An expert group of 8 nursing professionals with practical clinical training in pediatric nursing, including 3 pediatric nursing professors, evaluated the adequacy of the assessment data, nursing diagnoses, nursing interventions, and content relevance of the nursing evaluation algorithms. Scenario content validity testing resulted in an overall content validity index of 0.99 and 1 of 32 items was removed as "not relevant" or "somewhat relevant." In module 4, among the assessment data to make a nursing diagnosis of skin integrity disorder, unnecessary assessment data were deleted according to expert opinion. The final version of the scenario was developed after modifications and refinements based on expert opinions.

Step 6: Scenario Completion

The simulation scenario included 6 modules, covering the period from admission to discharge for a child with pneumonia (Table 1). First, a disease clinical path for pneumonia was created, and based on this, 6 modules were constructed focusing on the medical treatment and nursing intervention that a 7-year-old child diagnosed with pneumonia would receive from the first day of hospitalization to the day of discharge. The scenario consisted of 6 modules following the clinical pathway of pediatric pneumonia; the nursing process was applied to each module based on detailed learning goals. The scenarios were configured to provide feedback on each learner's performance.

Table 1. Virtual reality simulation modules.

Module	Day	Nursing task
Module 1	Admission	Admission assessment
Module 2	Hospital day 1-1	Fever management and medication
Module 3	Hospital day 1-2	Oxygen therapy and monitoring
Module 4	Hospital day 2	Antibiotic skin test and intravenous injection
Module 5	Hospital day 3	Respiratory therapy (nebulizer application)
Module 6	Discharge	Discharge nursing care

Step 7: Development of Prototype Content

The final prototype content was developed using Unreal Engine 4.25 (Epic Games) and includes 2 unique technical features. First, learners performed all procedures using their hands directly without handheld devices. Second, learners could use voice interaction with virtual agents to enhance communication competencies. To implement the VR experience, Vive Pro HMD (HTC) and a hand-tracking software development kit was used to obtain the user's natural hand position information. Voice communication was implemented using Google's speech kit and application programming interface.

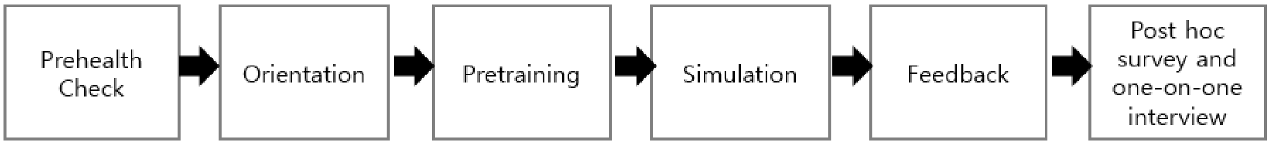
Each module took 20-30 minutes to complete and included key nursing procedures. After completing the scenario simulation, the learner received system feedback using a visual checklist. In summary, the developed prototype provides an immersive and interactive VR experience for enhancing nursing competencies through hands-on practice and voice communication with virtual objects.

Specifically, we tested this prototype content based on feedback from the nursing expert. They indicated difficulties with voice recognition, as the system did not always accurately recognize spoken keywords. To address this, we incorporated a synonym learning algorithm using Google's synonym crawling technology to enhance the system's ability to understand and respond to varied inputs. This improvement aimed to provide a more seamless and intuitive user experience.

Part 2: User Testing

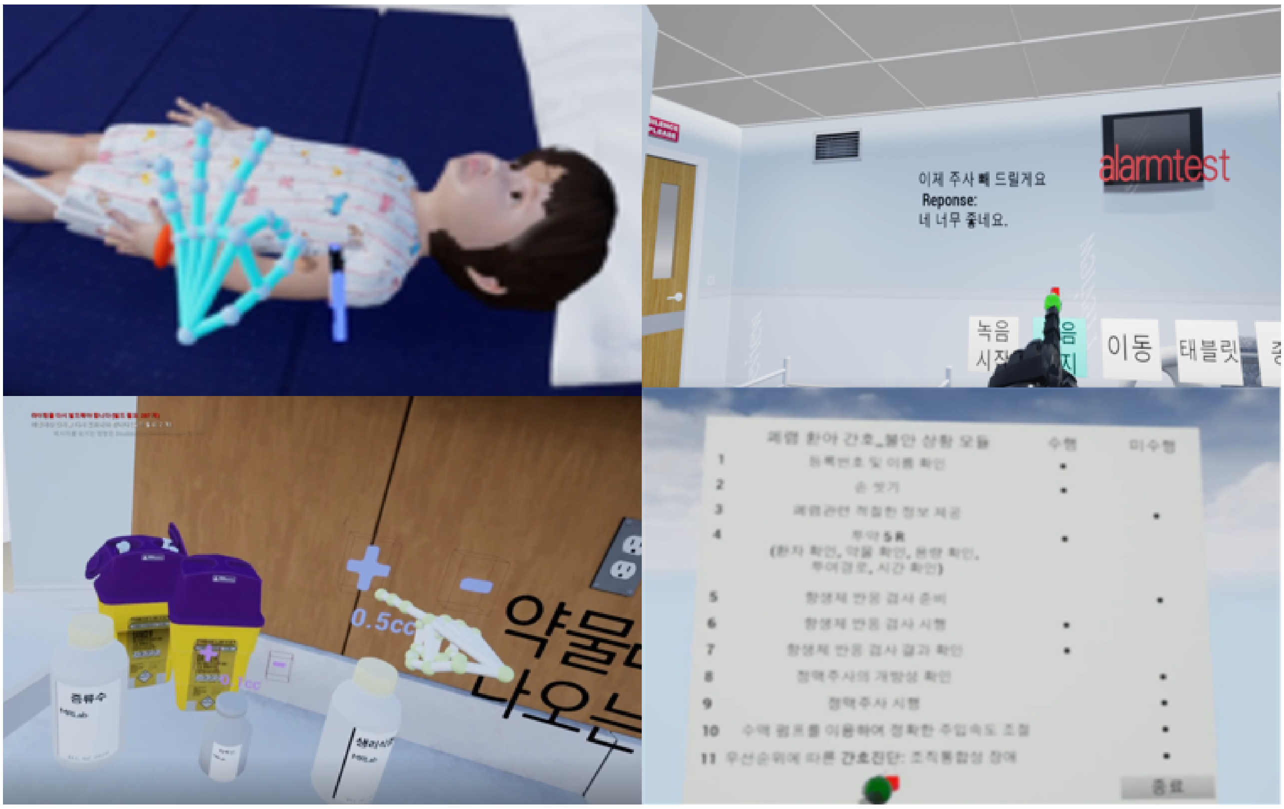
The feasibility of this VRS content was evaluated using a survey (quantitative method) and an interview (qualitative method). Before user evaluation, the participant's health was assessed. Immediately following the VRS experience, participants completed a survey and semistructured, in-depth, one-on-one interviews (Figure 2). Part 2 began in April 2021 (Figure 3).

Figure 2. Flow of the user test.



XSL•FO
RenderX

Figure 3. Example of screens from the developed virtual reality content.



Participants

Participants were recruited by convenience sampling of nursing graduates of a university in Hanyang. Selection criteria included preservice nurses with clinical practice and simulation training experience. After obtaining approval from the institutional review board, a recruitment notice was posted in the university's KakaoTalk chat room for nursing graduates. Those who volunteered to participate were given explanations of the study's purpose and procedures, and all participants provided written consent. A total of 12 nursing students were selected for participation to assess the feasibility of, and satisfaction with, the developed VRS content and identify areas for improvement [13].

Measurement

Overview

Participants completed an 85-item questionnaire based on previous studies; items included presence, VR technological elements, cybersickness, learning satisfaction, and participant characteristics (age, sex, general health status, experience of motion sickness, VR experience, practical experience, and satisfaction with practical training). With permission from their developers, individual survey tools were modified to suit the study characteristics and translated and back-translated by 2 bilingual translators.

Checklist for Participant Health Status

A total of 3 items were used to assess the participants' current health, including 2 items from the Screening Questionnaire by Costa et al [14] and 1 item related to dizziness.

Postexperience Evaluation Questionnaire

Presence Questionnaire

The Presence Questionnaire (PQ) was developed by Witmer and Singer [15] to measure presence in a virtual environment. It consists of 5 subscales—realism, the possibility of acting, interface quality, the possibility of examination, and self-evaluation of performance. The PQ comprises 32 items, each rated on a 7-point Likert scale (1=not at all and 7=completely).

Virtual Reality Systems Questionnaire

The Virtual Reality Systems Questionnaire (VRSQ) is a 22-item tool developed by Norman [16] to evaluate VR games. It consists of 7 technical elements, each rated on a 9-point Likert scale. The original VRSQ was modified to suit this study and consisted of 20 items.

Simulator Sickness Questionnaire

The Simulator Sickness Questionnaire (SSQ) is a cybersickness scale developed by Kennedy et al [17] to measure symptoms of discomfort in VR environments using 16 items rated on a 4-point Likert scale ranging from 0 (none) to 3 (severe). The scores for each item were summed to obtain a total score indicating the overall level of discomfort experienced by individuals in the VR environment.

Simulation Satisfaction Questionnaire

User satisfaction was measured using a simulation satisfaction tool developed by Wotton et al [18]. It consists of 8 items rated on a 7-point Likert scale (1=not at all and 7=very much).

XSL•FO
RenderX

In-Depth Interviews: Core Questions

The in-depth interviews consisted of 4 open-ended questions (1-4) first used by Servotte et al [19] in their VRS study and 2 additional questions (5 and 6). The questions were (1) what was your first impression and feeling about the VRS experience? (2) What did you find enjoyable during the VRS experience? (3) What did you find uncomfortable during the VRS experience? (4) What improvements would you suggest? (5) Compared with your previous clinical practice experience, what aspects were useful? and (6) Compared with your previous clinical practice experience, what aspects were lacking?

Data Collection

Data collection was conducted on April 15 and 16, 2021. Led by the principal researcher, a professor in the Department of Nursing, a total of 3 researchers conducted a user test of this VRS content. Of the 3 researchers, 2 were nursing majors and 1 was an engineering major. Strict COVID-19 infection control guidelines, such as fever checks and social distancing, were followed. Using the health checklist, only participants without nausea, vomiting, or physical discomfort in the 48 hours prior to the VRS experience took part.

In our study, to minimize cybersickness, the duration of wearing the equipment per person was limited to less than 1 hour at a time. Based on these internal test guidelines, each participant was allowed to experience only 1 module of 6 pediatric pneumonia contents through random assignment using a computer in advance. Participants received the scenario via email the day before the simulation. On the day of the VRS experience, participants watched an orientation video about all 6 modules and prepracticed for approximately 40 minutes. After a 20-minute break in an open space, they participated in a user test. When the simulation experience was over, participants filled out a questionnaire and had a one-on-one personal interview with the principal researcher.

Data Analysis

Quantitative data were analyzed using SPSS Statistics software (version 25.0; IBM Corp) with descriptive statistics, such as means, SDs, and frequencies. Qualitative interview data were analyzed using thematic analysis by Braun and Clarke [20]. Interview transcripts were coded using NVivo 12.0 Pro software (Lumivero), and member checking, continuous comparison, and interrater verification processes were implemented to enhance the reliability and validity of the analyzed qualitative data. All researchers participated in the iterative analysis process.

Ethical Considerations

This study was conducted after obtaining approval from the institutional review board of Hanyang University (HYUIRB-202103-017). To ensure anonymity and to protect participants' personal information, all data collected were immediately encoded in compliance with the Declaration of Helsinki.

Results

Participant Demographics

A total of 12 licensed nurses, 3 (25%) male nurses and 9 (75%) female nurses with a mean age of 24.3 (SD 1.23) years, participated in the VRS training (Table 2). Most respondents reported good health on a regular basis (n=11, 92%) and did not experience motion sickness in their daily lives (n=9, 75%). Only 1 person subjectively assessed her health status as unhealthy due to her underlying disease, and there were no problems during the preliminary health check. One-third of participants (n=4, 33%) had experience with VR through games or other means, and all had previous clinical training with high-fidelity simulation (HFS) using mannequins and web-based simulation with vSim for Nursing. All participants had no problems checking the preliminary health checklist.

XSL•FO
RenderX

Table 2. General characteristics of participants (N=12).

Characteristics and categories	Values
Sex, n (%)	
Female	9 (75)
Male	3 (25)
Age (years), mean (SD)	24.3 (1.23)
General health status, n (%)	
Very unhealthy	0 (0)
Unhealthy	1 (8)
Healthy	7 (58)
Very healthy	4 (33)
Experience of motion sickness when traveling in a car, n (%)	
Always	0 (0)
Mostly	3 (25)
Rarely	9 (75)
Never	0 (0)
VR[a] experience such as games, n (%)	
No	8 (67)
Yes	4 (33)
Clinical practice experience, n (%)	
1 semester	1 (8)
2 semesters	7 (58)
3 or more semesters	4 (33)
Practice experience outside of hospitals (one or both), n (%)	
vSim for Nursing	12 (100)
HFS[b]	2 (17)
Satisfaction with previous practical training, n (%)	
Very dissatisfied	0 (0)
Dissatisfied	0 (0)
Neutral	3 (25)
Satisfied	9 (75)
Very satisfied	0 (0)

[a]VR: virtual reality.

[b]HFS: high-fidelity simulation.

Presence

The mean presence score during the VRS training was 4.01 (SD 1.43) out of 7 points, which is moderate. High scores were achieved for the following items: "How well could you identify sounds?" (mean 5.50, SD 1.09), "How well could you localize sounds?" (mean 5.42, SD 1.68), "How proficient in moving and interacting with the virtual environment did you feel at the end of the experience?" (mean 5.17, SD 1.53), and "To what extent did the visual aspects of the environment produce user involvement?" (mean 5.10, SD 1.20). Items with low scores were "How natural was the mechanism that controlled movement through the environment?" (mean 2.40, SD 0.90) and "How compelling was your sense of moving around inside the virtual environment?" (mean 2.50, SD 1.24; Table 3).

XSL•FO
RenderX

Table 3. Presence (N=12)[a].

Item	Values, mean (SD)
How much were you able to control events?	3.30 (1.20)
How responsive was the environment to actions that you initiated (or performed)?	3.10 (0.80)
How natural did your interactions with the environment seem?	3.10 (1.10)
How completely were all of your senses engaged?	4.80 (1.40)
How much did the visual aspects of the environment involve you?	5.10 (1.20)
How much did the auditory aspects of the environment involve you?	4.00 (1.30)
How natural was the mechanism which controlled movement through the environment?	2.40 (0.90)
How aware were you of events occurring in the real world around you?	3.80 (1.60)
How aware were you of your display and control devices?	4.80 (1.50)
How compelling was your sense of objects moving through space?	3.75 (1.55)
How inconsistent or disconnected was the information coming from your various senses?	3.83 (1.03)
How much did your experiences in the virtual environment seem consistent with your real-world experiences?	3.75 (1.42)
Were you able to anticipate what would happen next in response to the actions that you performed?	4.67 (1.50)
How completely were you able to actively survey or search the environment using vision?	4.42 (1.78)
How well could you identify sounds?	5.50 (1.09)
How well could you localize sounds?	5.42 (1.68)
How well could you actively survey or search the virtual environment using touch?	3.00 (1.41)
How compelling was your sense of moving around inside the virtual environment?	2.50 (1.24)
How closely were you able to examine objects?	3.75 (1.60)
How well could you examine objects from multiple viewpoints?	4.00 (1.71)
How well could you move or manipulate objects in the virtual environment?	3.67 (0.98)
To what degree did you feel confused or disoriented at the beginning of breaks or at the end of the experimental session?	3.00 (1.54)
How involved were you in the virtual environment experience?	4.92 (1.68)
How distracting was the control mechanism?	4.17 (1.70)
How much delay did you experience between your actions and expected outcomes?	3.75 (1.96)
How quickly did you adjust to the virtual environment experience?	4.58 (1.24)
How proficient in moving and interacting with the virtual environment did you feel at the end of the experience?	5.17 (1.53)
How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?	3.33 (1.23)
How much did the control devices interfere with the performance of assigned tasks or with other activities?	4.92 (1.31)
How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities?	2.75 (1.66)
Did you learn new techniques that enabled you to improve your performance?	3.33 (1.78)
Were you involved in the experimental task to the extent that you lost track of time?	4.67 (2.06)

[a]Total: mean 4.01 (SD 1.43).

VR Systems

The mean score for technical elements of the VR system was 4.91 (SD 0.81) out of 9. High scores were achieved for “Auditory glitches” (mean 7.30, SD 1.70), “Trying to locate the source of sounds” (mean 6.92, SD 1.73), “Trying to turn and see what is to the left or right”(mean 6.58, SD 1.93), and “Trying to turn and see what is behind” (mean 6.42, SD 2.02), which indicates a few technical problems. Low scores were achieved for “Trying to aim or point targets with skeletal hands” (mean 3.83, SD 2.29), “Calibrating the system and tracking” (mean 3.80, SD 1.80), and “Trying to aim or point targets with robotic hands” (mean 3.42, SD 1.93).

VR Sickness (Simulator Sickness)

The mean score for VR sickness during the training was 0.64 (SD 0.35) out of 3, indicating that most participants did not experience cybersickness. Only minimal discomfort (1 point)

XSL•FO
RenderX

was reported for most items. However, scores for eye strain (mean 1.60, SD 1.20), fatigue (mean 1.40, SD 0.80), and head fullness (mean 1.0, SD 0.95) were ≥1 point, indicating some symptoms (Table 4).

Table 4. Virtual reality sickness (N=12)[a].

Item	Values, mean (SD)
General discomfort	0.40 (0.70)
Fatigue	1.40 (0.80)
Headache	0.80 (1.00)
Eyestrain	1.60 (1.20)
Difficulty focusing	0.80 (0.80)
Increased salivation	0.00 (0.00)
Sweating	0.40 (0.90)
Nausea	0.30 (0.90)
Difficulty concentrating	0.40 (0.70)
Fullness of head	1.00 (0.95)
Blurred vision	0.75 (1.06)
Dizzy (eyes open)	0.92 (0.67)
Dizzy (eyes closed)	0.83 (0.94)
Vertigo	0.50 (0.67)
Stomach awareness	0.00 (0.00)
Burping	0.00 (0.00)

[a]Total: mean 0.64 (SD 0.35).

Simulation Satisfaction

The overall mean score (8 items) for user satisfaction with the VRS training program was 5.00 (SD 1.00) out of 7. The highest scores (mean 5.30, SD 1.20) were related to experience enjoyment and problem-solving opportunities provided by the VRS training program. Items that received lower scores were those asking whether the participants could concentrate during the VRS training (mean 4.80, SD 1.40), whether the challenges were appropriate (mean 4.80, SD 1.40), and whether feedback provided during the VRS training helped them improve nursing knowledge (mean 4.80, SD 1.70; Table 5).

Table 5. Simulation satisfaction[a].

Item	Values, mean (SD)
Enjoyment of the virtual reality simulation practice	5.30 (1.20)
Ability to concentrate during the virtual reality simulation	4.80 (2.00)
Provision of appropriate challenges	4.80 (1.40)
Opportunity for problem-solving	5.30 (1.20)
Usefulness in learning on hospitalized children with pneumonia	5.10 (1.20)
Usefulness in improving pediatric nursing practice skills	4.90 (1.40)
Usefulness of feedback in patient care	4.90 (1.50)
Usefulness of feedback in enhancing nursing knowledge	4.80 (1.70)

[a]Total: mean 5.00 (SD 1.00).

Qualitative Evaluation

Overview

By analyzing the one-on-one in-depth interview data of the 12 participants, the strengths, weaknesses, improvement requirements, and comparison points of the content developed in this study were confirmed.

Strengths of VRS Content

Of the 12 participants, 9 (75%) responded that the visual features were well implemented, and they enjoyed manipulating and moving objects with their hands in the virtual environment. Some participants also found the virtual interactive experiences fascinating and were pleased to apply what they learned in practice.

XSL•FO
RenderX

> *I enjoyed the feeling of interaction and communication in the virtual environment as if I were observing the patient's reactions during real practice.* [Participant 7]

Weaknesses of VRS Content

Participants reported difficulties with voice recognition because communication centered on specific keywords and with technical aspects of virtual object manipulation. Additionally, some participants experienced disruptions during the nursing procedure owing to instability in certain virtual spaces.

> *It was frustrating when my words were not recognized properly, resulting in a generic response such as "Yes, I understand" to whatever I said.* [Participant 6]

Comparison With Existing Practical Training

Strengths of VRS Content

Participants noted that traditional hospital–based clinical training often involves fragmented training based on situational needs, whereas they mentioned the most significant advantage of VRS is the ability to experience independent nursing practice from beginning to end. They also found it highly beneficial to experience the entire continuum of nursing care, from admission to discharge, for specific diseases, and they highly appreciated the absence of time and space constraints in VRS.

> *In the actual clinical practice setting, we surround the patient like a folding screen, and there are many things we cannot see due to patient privacy. This makes us feel like we are gathering skills bit by bit rather than acquiring them systematically. However, with VR, we can perform the entire procedure from start to finish. In that sense, it is better than clinical practice.* [Comparison with clinical practice: participant 11]

> *Our group consisted of only 4 members, with some taking on the roles of nurse or doctor, while others were responsible for taking pictures or did not have the opportunity to perform nursing skills, which was disappointing. However, doing everything by myself allowed me to focus on improving my nursing skills.* [Comparison with HFS: participant 4]

> *Other programs like vSim for Nursing do not provide two-way communication. However, with this program, I felt like there was some two-way communication depending on what I said or did. It was fun and felt like I was a real nurse in the virtual environment, communicating and receiving reactions from a caregiver in my presence.* [Comparison with web-based simulations: participant 10]

Weaknesses of VRS Content

Compared with actual clinical practice, the limited manipulation of nursing supplies was evaluated as disappointing in that students were unable to perform detailed movements. Some pointed out that they had difficulty immersing themselves in the virtual clinical situation because they were concentrating on manipulating objects.

> *The simulations were limited to predetermined scenarios, whereas unexpected situations and various patient reactions may be encountered in a hospital setting. The tension that accompanies these situations may be lacking in simulations because all students are required to complete the same simulation from start to finish.* [Comparison with clinical practice: participant 2]

> *During the simulation, I was able to practice communication while working with my friends as a team. Unfortunately, I was doing everything alone in this program.* [Comparison with HFS: participant 5]

> *Communicating as if it were reality was helpful, but it was uncomfortable when my words were not accurately recognized.* [Comparison with a simple web-based simulation: participant 4]

Discussion

Principal Findings

VRS can provide realistic learning experiences through lifelike 3D environments, enhancing immersion and presence [21]. In this study, we sought to develop effective immersive nursing content by maximizing the advantages of VR and the learning effect in nursing through a multidisciplinary approach.

The content was developed using VR-based simulation design to create a realistic environment, considering clinical issues and contextual factors, while incorporating learning elements, such as clues and hints for learning and evaluation. Emphasis was placed on technical features, including visual implementation, to enhance immersion and presence and the educational design of the content [9]. Immersion relates to the technical aspect of VRS as an objective characteristic of the development environment, whereas presence refers to the user's psychological and subjective perception that they are present in the immersive virtual environment [19]. Immersion can evoke a feeling of physical "presence" by providing an experience similar to real-world ones [22].

The research team held numerous discussions and participated in refinement with modeling experts to enhance immersion by increasing visual fidelity and creating a more realistic environment. Because vision is the primary sense for perceiving information, high immersion and presence are needed in educational content design [23]. Most respondents rated the visual implementation as excellent, which helped them maintain high levels of immersion. However, some participants noted that engagement was hampered in part by the unsophisticated nursing supplies and the racial inconsistency of the caregiver avatars. This mismatch partially disrupted their immersion [24]. In future VR content production, it will be essential to consider not only the color, brightness, and spatial perception of virtual objects but also learners' cultural characteristics.

The main goal of the content development in this study was to enhance the sense of presence and provide a practice environment that would allow learners to perform nursing

XSL•FO RenderX

procedures with their own hands and provide emotional support to a virtual patient through verbal communication. According to user evaluations, multimodal interaction with voice communication and implementation of direct actions in the environment were viewed as positive and provided a moderate level of presence. However, communication accuracy and technical features related to virtual object manipulation required improvement. To the best of our knowledge, this is the first study to combine voice interactions with direct hand manipulations in practical nursing training. Multimodal interaction technology can increase resemblance to reality and provide a deeper sense of immersion, enhancing learning effectiveness [25,26]. Future studies should aim to improve the accuracy of multimodal interactions.

Despite the demand for improvement of virtual object manipulation and communication accuracy, satisfaction with the VRS program was relatively high, rated as moderate or better. The 2 most highly rated items were related to joyful experiences and problem-solving opportunities. These results align with the work by Lin et al [27] on multimodal interactions in VR for thoracic diagnostic scenarios, underscoring the effectiveness of such technologies in creating immersive educational experiences. This indicates the strong educational potential of immersive VRS. If immersive simulation content is developed by focusing on the strengths of VRS compared with existing training methods, it will be more effective in future nursing practice education.

The participants stated that the biggest advantage of VRS compared with traditional clinical practice was the opportunity to directly experience nursing procedures independently from start to finish. This finding is consistent with a previous study on multiuser virtual environment simulation experiences for nursing students [28], which highlighted building a “frame of thinking like a nurse” as a major theme. This self-directed nursing experience can facilitate learners’ metacognition and enhance competence [28,29]. The participants suggested that this immersive content could improve nursing students’ critical thinking and problem-solving skills. In VRS training, learners make clinical decisions and perform nursing activities based on their own judgments, which can effectively enhance critical thinking and problem-solving skills [5]. A previous study on nursing students [30] also confirmed the usefulness of VRS for developing clinical decision-making skills. However, this preliminary study was limited by time, and each participant experienced only 1 module. Follow-up studies are required to confirm the effectiveness of VR-based training.

Another strength of VRS pointed out by participants was the ability of students to experience the entire nursing process for patient care based on theoretical knowledge. In this study, participants were unable to experience all 6 modules due to time constraints while wearing VR, but the entire flow of all 6 modules was shown through video material during orientation on the day of the experience. Participants experienced the entire nursing process on a specific day on their own within 1 randomly selected module. This highlights the value of the full-cycle scenario design based on the clinical pathway of a corresponding illness. Traditional clinical training often limits nursing students to fragmented or random opportunities situations rather than structured, comprehensive practice sessions for a given condition [31]. Although traditional clinical training allows for an experience of unexpected situations, it lacks structure. Well-structured VRS practical training content design allows learners to apply their theoretical knowledge to practical situations systematically. Thus, VRS will be useful for reducing gaps between theoretical knowledge and practical skills.

Because VRS has some drawbacks, that is, lack of actual clinical practice and teamwork scenarios, it may be more effective to integrate VRS into traditional clinical training, allowing the two training methods to complement one another. Immersive VRS content can expand learning experiences into previously inaccessible quantitative and qualitative areas [32]. To improve the quality of practical nursing training in the post–COVID-19 era, a systematic education design model that leverages the advantages of VR and real-world training to optimize nursing education is needed.

Limitations and Future Research

The COVID-19 pandemic posed certain challenges for this study, particularly regarding participant recruitment and experiment execution. These constraints resulted in a smaller sample size, affecting the findings’ generalizability. Additionally, the duration of the VR simulation was limited to minimize the cybersickness risk, and each participant experienced only 1 module instead of the entire hospitalization scenario.

Despite these limitations, the study significantly contributes by systematically developing immersive VRS educational content for nursing. Using the Jeffries Simulation Theory template and VRS education design principles, the study confirmed the feasibility of this content in practical nursing training. Importantly, the study implemented multimodal interaction, enabling learners to use their hands directly and engage in voice communication, thereby closely simulating real clinical nursing practices.

Future research should address these limitations by recruiting a larger and more diverse sample population. Allowing participants to experience full-cycle clinical pathways of diseases will facilitate a more comprehensive verification of learning effects, enhancing the validity and applicability of the results. Furthermore, expanding the scope of multimodal interactions and refining the VR environment will improve the realism and effectiveness of the education, establishing it as an even more robust tool for nursing education.

Conclusions

We developed multidisciplinary VRS educational content that integrates the characteristics of nursing, education, and engineering and confirmed its feasibility as effective simulation education. Compared with conventional practical nursing training, this VRS content was valuable, allowing students to perform tasks independently and to experience the overall flow of nursing. The physical training environment was well-implemented visually, and the multimodal interaction increased immersion and presence. The VRS was also useful for enhancing individual nursing competencies through one-on-one training. However, keyword-based voice interactions

XSL•FO
RenderX

were a major obstacle to immersion, highlighting the need for additional research. Future research should explore the detailed learning effects and educational possibilities of immersive content using multidisciplinary design models. In conjunction with conventional clinical practice and HFS, VRS content based on a multidisciplinary educational model can be used in practical nursing training in the post–pandemic era to optimize clinical competency.

Conflicts of Interest

None declared.

References

1. Giordano NA, Whitney CE, Axson SA, Cassidy K, Rosado E, Hoyt-Brennan AM. A pilot study to compare virtual reality to hybrid simulation for opioid-related overdose and naloxone training. Nurse Educ Today 2020;88:104365. [doi: 10.1016/j.nedt.2020.104365] [Medline: 32088524]
2. Shea KL, Rovera EJ. Preparing for the Covid-19 pandemic and its impact on a nursing simulation curriculum. J Nurs Educ 2021;60(1):52-55. [doi: 10.3928/01484834-20201217-12] [Medline: 33400810]
3. Foronda CL, Fernandez-Burgos M, Nadeau C, Kelley CN, Henry MN. Virtual simulation in nursing education: a systematic review spanning 1996 to 2018. Simul Healthc 2020;15(1):46-54. [doi: 10.1097/SIH.0000000000000411] [Medline: 32028447]
4. Kim KA, Choi DW. The effect of virtual simulation in nursing education: an application of care for acute heart disease patients. J Korean Soc Simul Nurs 2018;6(2):1-13. [doi: 10.17333/jkssn.6.2.1]
5. Tolarba JEL. Virtual simulation in nursing education: a systematic review. Int J Nurs Educ 2021;13(3):48-54. [doi: 10.37506/ijone.v13i3.16310]
6. Suh E. Development of a conceptual framework for nursing simulation education utilizing human patient simulators and standardized patients. J Korean Acad Soc Nurs Educ 2012;18(2):206-219. [doi: 10.5977/jkasne.2012.18.2.206]
7. Farra SL, Smith SJ, Ulrich DL. The student experience with varying immersion levels of virtual reality simulation. Nurs Educ Perspect 2018;39(2):99-101. [doi: 10.1097/01.nep.0000000000000258]
8. Han H, Lim C. A developmental study on design principles for virtual reality based educational simulation. J Educ Technol 2020;36(2):221-264. [doi: 10.17232/kset.36.2.221]
9. Radianti J, Majchrzak TA, Fromm J, Wohlgenannt I. A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. Comput Educ 2020;147:103778. [doi: 10.1016/j.compedu.2019.103778]
10. Jeffries P, Rodgers B, Adamson K. NLN Jeffries simulation theory: brief narrative description. Nurs Educ Perspect 2015;36(5):292-293. [doi: 10.5480/1536-5026-36.5.292] [Medline: 26521496]
11. Han H. Development of a design model for virtual reality based educational simulation. Seoul National University. Seoul; 2019. URL: http://www.riss.kr/link?id=T15399066 [accessed 2024-06-18]
12. Sonderegger A, Sauer J. The influence of design aesthetics in usability testing: effects on user performance and perceived usability. Appl Ergon 2010;41(3):403-410. [doi: 10.1016/j.apergo.2009.09.002] [Medline: 19892317]
13. Sandelowski M. Sample size in qualitative research. Res Nurs Health 1995;18(2):179-183. [doi: 10.1002/nur.4770180211] [Medline: 7899572]
14. Costa IKF, Tibúrcio MP, Costa IKF, Dantas RAN, Galvão RN, Torres GDV. Development of a virtual simulation game on basic life support. Rev Esc Enferm USP 2018;52:e03382 [FREE Full text] [doi: 10.1590/S1980-220X2017047903382] [Medline: 30403269]
15. Witmer BG, Singer MJ. Measuring presence in virtual environments: a presence questionnaire. Presence (Camb) 1998;7(3):225-240. [doi: 10.1162/105474698565686]
16. Norman KL. Evaluation of virtual reality games: simulator sickness and human factors. 2018 Presented at: International Working Conference on Advanced Visual Interfaces; June 29, 2018; Grosseto, Italy.
17. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int J Aviat Psychol 1993;3(3):203-220. [doi: 10.1207/s15327108ijap0303_3]
18. Wotton K, Davis J, Button D, Kelton M. Third-year undergraduate nursing students' perceptions of high-fidelity simulation. J Nurs Educ 2010;49(11):632-639. [doi: 10.3928/01484834-20100831-01] [Medline: 20795614]
19. Servotte JC, Goosse M, Campbell SH, Dardenne N, Pilote B, Simoneau IL, et al. Virtual reality experience: immersion, sense of presence, and cybersickness. Clin Simul Nurs 2020;38:35-43. [doi: 10.1016/j.ecns.2019.09.006]
20. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
21. Liu Z, Fan X, Liu Y, Ye XD. Effects of immersive virtual reality cardiopulmonary resuscitation training on prospective kindergarten teachers' learning achievements, attitudes and self‐efficacy. Br J Educ Technol 2022;53(6):2050-2070. [doi: 10.1111/bjet.13237]

XSL•FO
RenderX

22. Arp L, Woodard BS, Mestre L. Accommodating diverse learning styles in an online environment. Ref User Serv Q 2006;46(2):27-32. [doi: 10.5860/rusq.46n2.27]
23. Sherman WR, Craig AB. Understanding Virtual Reality: Interface, Application, and Design 2nd ed. London: Morgan Kaufmann; 2018.
24. Chen VHH, Ibasco GC, Leow VJX, Lew JYY. The effect of VR avatar embodiment on improving attitudes and closeness toward immigrants. Front Psychol 2021;12:705574 [FREE Full text] [doi: 10.3389/fpsyg.2021.705574] [Medline: 34721153]
25. Capece N, Gruosso M, Erra U, Catena R, Manfredi G. A preliminary investigation on a multimodal controller and freehand based interaction in virtual reality. 2021 Presented at: AVR 2021, 8th International Conference on Augmented Reality, Virtual Reality and Computer Graphics, Virtual event; September 16, 2021; India p. 53-65. [doi: 10.1007/978-3-030-87595-4_5]
26. Uchino S, Abe N, Tabuchi Y, Taki H, He S. VR interactive dialog system with verbal and nonverbal communication. Artif Life Robotics 2009;13(2):512-516. [doi: 10.1007/s10015-008-0595-4]
27. Lin WH, Liou WK, Chen PJ, Chen S. Using multimodal interaction in a virtual reality thoracic diagnostic scenario. Int J Nurs Educ 2024;16(2):45-50. [doi: 10.37506/5rbjt116]
28. Rim D, Shin H. Development and assessment of a multi-user virtual environment nursing simulation program: a mixed methods research study. Clin Simul Nurs 2022;62:31-41. [doi: 10.1016/j.ecns.2021.10.004]
29. Brown KM, Swoboda SM, Gilbert GE, Horvath C, Sullivan N. Integrating virtual simulation into nursing education: a roadmap. Clin Simul Nurs 2022;72:21-29. [doi: 10.1016/j.ecns.2021.08.002]
30. Smith SJ, Farra S, Ulrich DL, Hodgson E, Nicely S, Matcham W. Learning and retention using virtual reality in a decontamination simulation. Nurs Educ Perspect 2016;37(4):210-214. [doi: 10.1097/01.NEP.0000000000000035] [Medline: 27740579]
31. Zhang J, Shields L, Ma B, Yin Y, Wang J, Zhang R, et al. The clinical learning environment, supervision and future intention to work as a nurse in nursing students: a cross-sectional and descriptive study. BMC Med Educ 2022;22(1):548 [FREE Full text] [doi: 10.1186/s12909-022-03609-y] [Medline: 35841091]
32. Elliman J, Loizou M, Loizides F. Virtual reality simulation training for student nurse education. 2016 Presented at: VS-Games 2016: the 8th International Conference on Games and Virtual Worlds for Serious Applications; September 7, 2016; Barcelona, Spain p. 1-2. [doi: 10.1109/vs-games.2016.7590377]

Abbreviations

HFS: high-fidelity simulation
NLN: National League for Nursing
PQ: Presence Questionnaire
SSQ: Simulator Sickness Questionnaire
VR: virtual reality
VRS: virtual reality simulation
VRSQ: Virtual Reality Systems Questionnaire

Edited by J López Castro, M Montagna, I Said-Criado, F Pietrantonio; submitted 26.09.23; peer-reviewed by M Brown, E Jeong; comments to author 18.12.23; revised version received 31.01.24; accepted 11.06.24; published 26.07.24.

Please cite as:
Yeo JY, Nam H, Park JI, Han SY
Multidisciplinary Design–Based Multimodal Virtual Reality Simulation in Nursing Education: Mixed Methods Study
JMIR Med Educ 2024;10:e53106
URL: https://mededu.jmir.org/2024/1/e53106
doi:10.2196/53106
PMID:

©Ji-Young Yeo, Hyeongil Nam, Jong-Il Park, Soo-Yeon Han. Originally published in JMIR Medical Education (https://mededu.jmir.org), 26.07.2024. 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.

XSL•FO
RenderX

Original Paper

Resources to Support Canadian Nurses to Deliver Virtual Care: Environmental Scan

Manal Kleib[1], BSN, MSN, MBA, PhD; Antonia Arnaert[2], PhD; Lynn M Nagle[3], PhD; Elizabeth Mirekuwaa Darko[1], BSN, MSN; Sobia Idrees[1], BSN, MSN; Daniel da Costa[2], BSN; Shamsa Ali[1], BSN, MSN

[1]Faculty of Nursing, University of Alberta, Edmonton, AB, Canada

[2]Ingram School of Nursing, McGill University, Montreal, QC, Canada

[3]Faculty of Nursing, University of New Brunswick, Fredericton, NB, Canada

Corresponding Author:
Manal Kleib, BSN, MSN, MBA, PhD
Faculty of Nursing
University of Alberta
5-112 Edmonton Clinic Health Academy
11405 - 87 Avenue NW
Edmonton, AB, T6G 1C9
Canada
Phone: 1 780 248 1422
Fax: 1 780 492 2551
Email: manal.kleib@ualberta.ca

Abstract

Background: Regulatory and professional nursing associations have an important role in ensuring that nurses provide safe, competent, and ethical care and are capable of adapting to emerging phenomena that influence society and population health needs. Telehealth and more recently virtual care are 2 digital health modalities that have gained momentum during the COVID-19 pandemic. Telehealth refers to telecommunications and digital communication technologies used to deliver health care, support health care provider and patient education, and facilitate self-care. Virtual care facilitates the delivery of health care services via any remote communication between patients and health care providers and among health care providers, either synchronously or asynchronously, through information and communication technologies. Despite nurses' adaptability to delivering virtual care, many have also reported challenges.

Objective: This study aims to describe resources about virtual care, digital health, and nursing informatics (ie, practice guidelines and fact sheets) available to Canadian nurses through their regulatory and professional associations.

Methods: An environmental scan was conducted between March and July 2023. The websites of nursing regulatory bodies across 13 Canadian provinces and territories and relevant nursing and a few nonnursing professional associations were searched. Data were extracted from the websites of these organizations to map out educational materials, training opportunities, and guidelines made available for nurses to learn and adapt to the ongoing digitalization of the health care system. Information from each source was summarized and analyzed using an inductive content analysis approach to identify categories and themes. The Virtual Health Competency Framework was applied to support the analysis process.

Results: Seven themes were identified: (1) types of resources available about virtual care, (2) terminologies used in virtual care resources, (3) currency of virtual care resources identified, (4) requirements for providing virtual care between provinces, (5) resources through professional nursing associations and other relevant organizations, (6) regulatory guidance versus competency in virtual care, and (7) resources about digital health and nursing informatics. Results also revealed that practice guidance for delivering telehealth existed before the COVID-19 pandemic, but it was further expanded during the pandemic. Differences were noted across available resources with respect to terms used (eg, telenursing, telehealth, or virtual care), types of documents (eg, guideline vs fact sheet), and the depth of information shared. Only 2 associations provided comprehensive telenursing practice guidelines. Resources relative to digital health and nursing informatics exist, but variations between provinces were also noted.

Conclusions: The use of telehealth and virtual care services is becoming mainstream in Canadian health care. Despite variations across jurisdictions, the existing nursing practice guidance resources for delivering telehealth and virtual care are substantial and

XSL•FO
RenderX

can serve as a beginning step for developing a standardized set of practice requirements or competencies to inform nursing practice and the education of future nurses.

(JMIR Med Educ 2024;10:e53254) doi:10.2196/53254

KEYWORDS

virtual care; digital health; nursing practice; environmental scan; telehealth; nurses; Canada; health care

Introduction

Background

Telehealth refers to the “delivery and facilitation of health and health-related services including medical care, provider and patient education, health information services, and self-care via telecommunications and digital communication technologies. Examples of the technologies used in telehealth include, but are not limited to, live video conferencing, mobile health apps, ‘store and forward’ electronic transmission, and remote patient monitoring.” [1]. During the COVID-19 pandemic, a transition to virtual care became necessary to protect the public and ensure the continuity of health services. Virtual health denotes the facilitation of the delivery of care services through any remote interactions between patients and health care providers and among health care providers themselves, whether synchronous or asynchronous, using information and communication technologies (ICTs) [2]. Various technologies were applied to facilitate virtual care such as SMS text messaging and email; phone; mobile health applications; electronic medical records; chatbots; remote monitoring technologies; and telecommunication applications such as Zoom (Zoom Video Communications), Skype (Skype Communications), FaceTime (Apple Inc), and WhatsApp (WhatsApp LLC) for video consultations [3-9]. Both modalities are subsumed under the umbrella term of digital health, which denotes the proper use of technology to improve the health and wellbeing of people at all levels.

Nurses providing care across a diversity of health care settings had to adapt to virtual care delivery. Despite their engagement and adaptability to this new form of care, many nurses also reported challenges. Those who have not previously used digital health technologies to deliver care had to quickly adapt and learn new skills, and some reported receiving limited guidance or best practice guidelines on how to provide care through these new modalities [10-13]. Before the COVID-19 pandemic, the most recent survey of practicing Canadian nurses’ use of technology identified that only 60% of nurses surveyed about the use of virtual care reported having adequate knowledge and skills to use these technologies [14]. A secondary analysis comparing data from the 2017 and 2020 versions of this survey identified that virtual care was predominantly delivered by nurse practitioners. Several factors were found to predict Canadian nurses’ use of virtual care, including their professional designation, perceived quality of care in the facility where they worked, the type of electronic record used, their perceptions of the quality of care they delivered through virtual care technologies, and their perceptions of the skills and knowledge needed to use these technologies [15]. In addition to gaps among practicing nurses, research has also identified that while new graduate nurses have high digital literacy skills, they struggle to understand the broad spectrum of digital health technologies and their applications in health care [16]. Other studies identified that digital competence, organizational support, prior education about technology, and ongoing training and support are essential factors for effective and safe use and adoption of health information technologies [17-20].

Nurses are integral to the successful implementation and use of technology in health care. Currently, there are >459,000 regulated nurses represented by the Canadian Nurses Association (CNA) at the national level [21]. Provincial regulatory associations also exist across all 13 Canadian provinces and territories. As part of their mandate, nursing regulatory bodies address the registration requirements, standards for education and practice, code of ethics, and continuing competency to ensure nurses are qualified and competent to provide safe, ethical, and evidence-informed practice as well as are prepared to respond to emerging phenomena that influence population health needs and nursing professional practice roles [22,23]. They also develop documents (eg, practice guidelines and practice directives) that outline the activities the regulated nurses are authorized to perform [23]. Professional nursing associations are available to all nurses, nursing students, and nurse retirees, enabling these groups to have a voice in the policy and advocacy work related to profession-wide and societal issues that may have broader impacts on health and health care. In addition to these associations, nursing specialty practice organizations and interest groups serve to provide resources and support for professional development in a specialty practice area such as critical care [24]. Considering the expanded and ongoing use of virtual care and for nurses to thrive when providing care in this increasingly digitalized health care environment, it is important to understand how professional and regulatory nursing associations are contributing to supporting nurses in this area and what knowledge gaps exist in offering necessary support, training, and professional guidance for optimal nursing practice with technology. Addressing this knowledge gap could serve to encourage nursing regulatory and professional bodies to update existing resources or develop new ones, which may consequently contribute to motivating nurses to develop their competency in digital health and virtual care.

Objective

This study aims to describe resources about virtual care, digital health, and nursing informatics (ie, practice guidelines and fact sheets) available to Canadian nurses through their regulatory and professional associations.

XSL•FO
RenderX

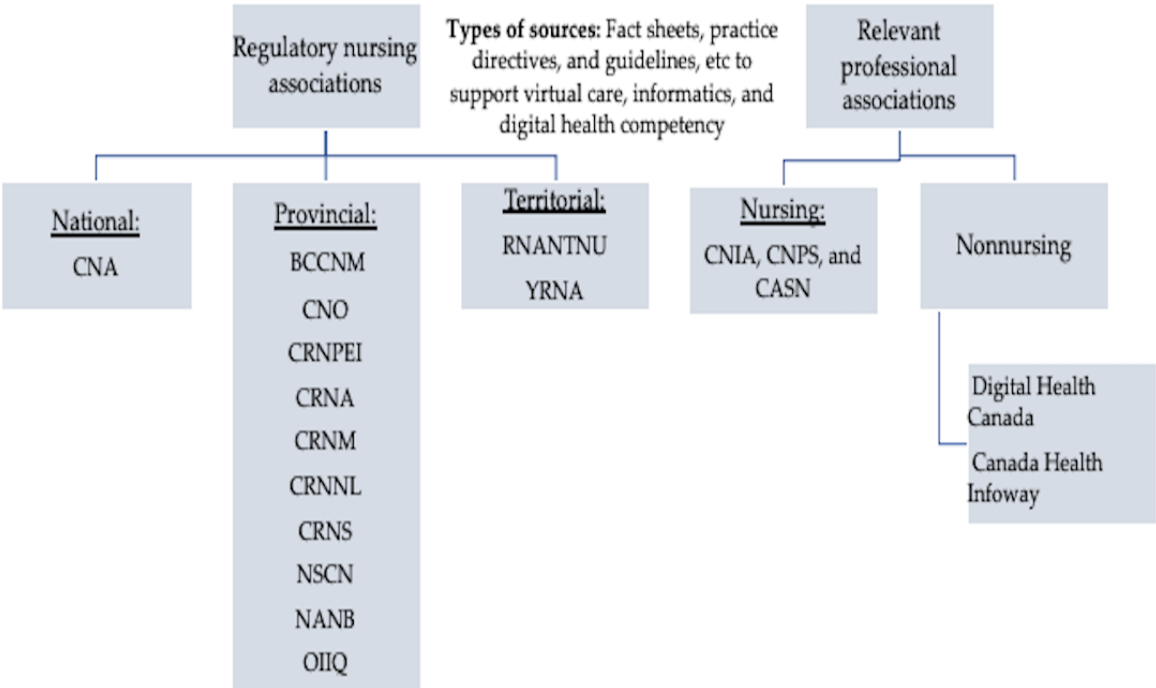
Methods

Overview

This study used an environmental scan method [25,26]. An environmental scan is often undertaken when there are emerging issues that necessitate an immediate investigation to determine their potential impact. Environmental scans originated in the business industry as a tool to inform strategic planning and policy decision-making by gathering, interpreting, and using information about the internal and external environment of an organization to guide future actions. Frameworks such as Strengths, Weakness, Opportunities, and Threats analysis or Political, Economic, Social, Technological, Legal, and Environmental are often used to guide this process [25,26]. In health care, environmental scans serve to collect, organize, and analyze information on issues in health care to make evidence-informed decisions, guide program planning, and inform public policy decisions [25]. Environmental scans are also different from other systematic literature searches such as scoping or systematic reviews and may incorporate different and diverse data sources and approaches for collecting information to answer questions of interest. Although the use of environmental scanning has increased in health service research, there is limited methodological guidance on its application. However, because scanning the environment can be extensive, determining the most relevant information sources and outlining the scope of the scan by identifying clear objectives is recommended [25]. In this project, the environmental scan was guided by the following steps: (1) determining the scope of the environmental scan and data sources to be included, (2) scanning resources identified for information that is relevant to the focus of the research, and (3) analyzing and summarizing the information to identify strengths and limitations.

The environmental scan was conducted between March 2023 and July 2023 to identify and retrieve pertinent information and resources available to guide nursing practice or education about virtual care, digital health, and informatics for Canadian nurses. A search of the websites of national, provincial, and territorial Canadian nursing regulatory bodies and the websites of selected professional nursing associations, nursing specialty practice organizations, and some nonnursing organizations and associations was conducted. To locate the relevant available data on the websites of each organization and association, we used different search terms such as *digital health*, *nursing informatics (NI)*, *virtual care*, *telehealth*, *telenursing*, *telemedicine*, and *nurses*. The authors then examined what was reported in each record identified on each association website by reviewing the content, date, relevance, and the type of the resource. Nursing unions were excluded from this scan because their mandate is not related to professional development or regulation. Only information published in English was reviewed. Figure 1 summarizes the sources of data used in this scan.

Figure 1. Visual representation of the data sources used in the environmental scan. BCCNM: British Columbia College of Nurses and Midwives; CASN: Canadian Association of Schools of Nursing; CNA: Canadian Nurses Association; CNIA: Canadian Nursing Informatics Association; CNO: College of Nurses of Ontario; CNPS: Canadian Nurses Protective Society; CRNA: College of Registered Nurses of Alberta; CRNM: College of Registered Nurses of Manitoba; CRNNL: College of Registered Nurses of Newfoundland and Labrador; CRNPEI: College of Registered Nurses and Midwives of Prince Edward Island; CRNS: College of Registered Nurses of Saskatchewan; NANB: Nurses Association of New Brunswick; NSCN: Nova Scotia College of Nursing; OIIQ: Ordre des Infirmières et Infirmiers du Québec; RNANTNU: The Registered Nurses Association of the Northwest Territories and Nunavut; YRNA: Yukon Registered Nurses Association.



XSL•FO
RenderX

Data Analysis

We applied an inductive qualitative content analysis approach to categorize data extracted from the retrieved documents [27,28]. Inductive content analysis is an appropriate approach for descriptive qualitative analysis. It focuses on identifying easily observable items or data without the need to discern intent or identify deeper meaning. It uses a systematic approach to transform large amounts of data into a highly organized and concise summary of results that can be grouped into codes, categories, and themes. The Provincial Health Service Authority (PHSA) Virtual Health Competency Framework was also used to support the analysis and interpretation of the data [29]. Two reviewers independently completed the process of data abstraction and coding, and data were compared for accuracy to ensure consistency and rigor in the abstraction process.

Ethical Considerations

As per the ethics requirements at the University of Alberta [30], we did not seek ethics clearance because the research did not involve collecting input or information from human participants. We primarily relied on publicly available information that is legally accessible to the public.

Results

Overview

A total of 17 websites were included in this scan, of which 13 (76%) were for regulatory bodies (n=1, 8% national; n=10, 77% provincial; and n=2, 15% territorial), 2 (12%) were professional associations, and 2 (12%) were part of the relevant nonnursing organizations. On the basis of the nature of the information provided in the retrieved documents or resources, seven overarching themes were identified: (1) types of resources available about virtual care, (2) terminologies used in virtual care resources, (3) currency of virtual care resources identified, (4) requirements for providing virtual care between provinces, (5) resources through professional nursing associations and other relevant organizations, (6) regulatory guidance versus competency in virtual care, and (7) resources about digital health and NI. These are discussed in detail in the subsequent sections.

Types of Resources Available About Virtual Care

Different types of guidance documents such as fact sheets, practice directives, guidance, practice guidelines, and general guidelines were identified on the web pages of nursing regulatory bodies (Table 1). The CNA issued a telehealth fact sheet in 2017 and did not provide national level guidance pertaining to telenursing practice or virtual care; however, it has identified virtual care as a priority advocacy area. Some associations provided documents in the form of fact sheets, guidance, and practice directives (College of Nurses of Ontario [CNO], College of Registered Nurses of Manitoba [CRNM], and College of Registered Nurses and Midwives of Prince Edward Island [CRNPEI]). Only 2 associations published telenursing practice guidelines for registered nurses and nurse practitioners (Nurses Association of New Brunswick [NANB] and the Nova Scotia College of Nursing [NSCN]). Some associations supplemented their guidelines with educational resources such as practice scenarios, case studies, and YouTube videos (NSCN and British Columbia College of Nurses and Midwives [BCCNM]), and some associations provided resources to support nurses' mental health and well-being (CNO). Some regulatory bodies provided external links to organizations, such as the Canadian Nurses Protective Society (CNPS), on their website.

XSL•FO
RenderX

Table 1. Websites of nursing and related organizations offering virtual care resources.

Organization	Virtual care	Concepts emphasized and general assessment
National and provincial regulatory nursing associations		
CNA[a]	• Virtual care as an advocacy priority (2023) • Telehealth fact sheet (2017)	• Virtual care has been identified as an advocacy priority • Telepractice fact sheet has not been updated since 2017 • No further guidance provided during or after the pandemic
BCCNM[b]	• Virtual care learning resource	• A unique web page providing a definition, FAQs[c], case study about providing virtual care to clients outside British Columbia, links to related standards, policies, guidelines, professional standards, and additional resources on virtual health (a handbook and toolkit)
CNO[d]	• Telepractice fact sheet (2023) • COVID-19 practice resources	• The document provides brief information on the expectations of nurses providing care both within and outside Ontario. In addition, it is stated that the telepractice guideline is under review. Furthermore, there are links to resources, such as the scope of practice, code of conduct, therapeutic nurse-client relationship, documentation, and confidentiality and privacy: personal health information • Specific resources for COVID-19, such as self-care fact sheets and other information and tips to support mental health for nurses
CRNPEI[e]	• Practice directive—telehealth nursing practice (2019) • Practice directive—technology in practice (2020)	• Telehealth practice directives: address regulatory requirements • Technology in nursing practice: provide guidance on aspects of virtual care • Concepts addressed: therapeutic nurse-patient relationship, competencies beyond basic nursing program, consent, privacy, confidentiality, and ethical and legal considerations • There are links to external resources (eg, CNPS[f] infoLAW page that has information on relevant topics such as confidentiality of health information and social media)
CRNA[g]	• Telepractice and virtual care (2020)	• Regulatory guidance to RNs[h] and NPs[i] on licensing requirements, indicating that they must be licensed in the jurisdiction they are providing care, regardless of the practice setting • RN entry-level competencies include indicators about nursing informatics
CRNM[j]	• Practice expectations spotlight: telepractice or virtual care (2020) • Guidance on telepractice (2020) • Telepractice (2021)	• A news item providing generic guidance regarding risks associated, focusing on miscommunication, privacy breaches, and poor coordination of care • Additional resources on the page link to CNPS Telephone Advice (2009), but no link or listing of more current CNPS resources about virtual care • A guide providing information on general performance expectations, informed consent, privacy, safety considerations and adverse events, safety expectations, competency, and documentation and billing for health care providers including nurses • An information sheet of telepractice on professional practice both within and outside of Manitoba

XSL•FO
RenderX

Organization	Virtual care	Concepts emphasized and general assessment
CRNNL[k]	• Virtual nursing practice (2020) • Fact sheet: virtual nursing licensure requirements (2022)	• A document articulating the practice expectations required of RNs and NPs participating in virtual nursing practice and examples of virtual care technologies • A fact sheet that addresses licensure requirements for engaging in virtual nursing practice.
CRNS[l]	• Tag: telehealth—nursing use of ICTs[m] (2020)	• General and regulatory guidance on telehealth and nursing use of ICT technologies. Concepts briefly addressed the following: upholding standards, competencies, code of ethics, clinical knowledge, clinical judgment, communication, documentation skills, and nurse-client relationship
NSCN[n]	• Telenursing practice guidelines for nurses (2019, last revised in February 2023) • Practice scenarios and FAQs for nurses working with virtual MDs[o] or NPs (2023) • YouTube video (12 minutes) • Telenursing online education module • Fact sheet: LPN[p] insurance program—providing telehealth • Glossary (2023) • COVID-19 information and resources including social media use	• NSCN has the most recent and updated detailed guidelines on telenursing practice for RNs and NPs. It has also recently published a glossary of associated terms related to virtual care • Core concepts addressed the following: telenursing; professional practice; competencies; risk management; informed consent; confidentiality; therapeutic nurse-client relationship; communication; documentation; virtual care; and telehealth services including video-conferencing, phone call, web-based patient portals, email and SMS text messaging, and other electronic communications • Following the outbreak of the pandemic, it provided detailed information about social media use
NANB[q]	• Telenursing practice guidelines (February 2023) • FAQ: what are the registration requirements to provide telenursing care in NB?	• The toolkit provides guidance and resources to assist nurses and employers in providing safe, competent, and ethical telenursing care • Core concepts addressed the following: telenursing practice, accountability, competency, client-centered care, legal requirements, confidentiality, client safety, custodian of record, competency, evidence-informed practice, communication, documentation, nurse-client relationship, informed consent, and risk management • Additional guidance on regulatory requirements is explored in the FAQ component
OIIQ[r]	—[s]	• Information on this website is in French and therefore was not verified
RNANTNU[t]	• FAQs: telehealth • News page provides hyperlinks to external resources	• FAQs provide general guidance on licensure and providing care in different jurisdictions. • News page provides hyperlinks to information through other organizations, such as CASN[u] and CNPS, that provide information about virtual care
YRNA[v]	• None	• No virtual care guidelines. Documentation guidelines have not been updated since 2013
Relevant professional nursing associations		

XSL•FO
RenderX

Organization	Virtual care	Concepts emphasized and general assessment
CNIA[w]	• None	• The 2023 conference (Accelerating Digital Care Capacity in Nursing) focused on how virtual care influenced nursing practice
CNPS	• InfoLAW: Telepractice (2020) • FAQ providing information about 12 things to consider before joining a virtual care practice (2020, updated in 2022)	• The telepractice InfoLAW guide provides a general overview of telepractice addressing aspects related to privacy, risk management, jurisdictional considerations with links to regulatory bodies' telepractice resources, and a case study supplemented with a quiz. It does not replace regulatory requirements. • Some resources are accessible publicly, but membership is needed to access resources such as case studies and quizzes, which require registration with an access code.
Relevant nonnursing organizations		
CHI[x]	• Clinician change virtual care toolkit (2022)	• CHI has substantive resources about digital health—refer to the main web page. A detailed guide is available for new and experienced users or clinicians on how to provide safe and high-quality virtual care. • The toolkit provides guidance to new and experienced users on how to provide safe and high-quality virtual care.
Digital Health Canada	• Virtual care in Canada: Lexicon (2021) • Virtual care in Canada: maturity model framework (2021)	• This organization has dedicated and extensive resources and certification for health informatics professionals.

[a]CNA: Canadian Nurses Association.
[b]BCCNM: British Columbia College of Nurses and Midwives.
[c]FAQ: frequently asked question.
[d]CNO: College of Nurses of Ontario.
[e]CRNPEI: College of Registered Nurses and Midwives of Prince Edward Islan.
[f]CNPS: Canadian Nurses Protective Society.
[g]CRNA: College of Registered Nurses of Alberta.
[h]RN: registered nurse.
[i]NP: nurse practitioner.
[j]CRNM: College of Registered Nurses of Manitoba.
[k]CRNNL: College of Registered Nurses of Newfoundland and Labrador.
[l]CRNS: College of Registered Nurses of Saskatchewan.
[m]ICT: information and communication technology.
[n]NSCN: Nova Scotia College of Nursing.
[o]MD: medical doctor.
[p]LPN: licensed practical nurse.
[q]NANB: Nurses Association of New Brunswick.
[r]OIIQ: Ordre des Infirmières et Infirmiers du Québec.
[s]Not available.
[t]RNANTNU: The Registered Nurses Association of the Northwest Territories and Nunavut.
[u]CASN: Canadian Association of Schools of Nursing.
[v]YRNA: Yukon Registered Nurses Association.
[w]CNIA: Canadian Nursing Informatics Association.
[x]CHI: Canada Health Infoway.

Terminologies Used in Virtual Care Resources

Across the available resources, regulatory bodies used different terms and to some extent interchangeably, including telehealth, telepractice, telepractice nursing, virtual care, and technology. In their fact sheet, the CNA used the term telehealth; similarly, the Registered Nurses Association of the Northwest Territories and Nunavut (RNANTNU) and the College of Registered Nurses of Saskatchewan (CRNS) used telehealth on their web pages to describe the care provided using ICTs. The CRNPEI, in their

XSL•FO
RenderX

practice directive document, used the terms telehealth nursing practice and technology in practice, whereas the CNO, in their fact sheet, used the term telepractice. In the CRNM guidance on telepractice, the term telepractice was used, but both the terms telepractice and virtual care were used on the web page. In describing telepractice, the CRNM also indicated that telepractice is also known as virtual care. Furthermore, the College of Registered Nurses of Newfoundland and Labrador (CRNNL), in a document provided on their website, used the term virtual nursing practice. The practice guidelines and guidelines provided by the NSCN and NANB, respectively, used the term telenursing and telenursing practice. At the same time, the College of Registered Nurses of Alberta (CRNA), on their web page, used the terms telepractice and virtual care simultaneously.

Some associations, including CNO, BCCNM, NANB, and CRNPEI, provided links to nurses' scope of practice and practice standards of practice documents on their web pages and in the documents. This implies that when providing virtual care or telehealth, nurses are also expected to abide by these overarching practice requirements; however, these documents do not explicitly discuss or guide nurses in understanding virtual care. Associations such as CNO, NANB, and NSCN had links to other registered nurse professional practice resources (eg, scope of practice, privacy of health information, code of ethics, code of conduct, and documentation practices) that are included in these documents, but these also did not have specific guidance on telehealth or virtual care.

Currency of Virtual Care Resources Identified

The information provided in the resources screened was evaluated using the Currency, Relevance, Authority, Accuracy and Purpose checklist [31]. This method evaluates different dimensions to ensure the credibility of a source of information. This scan used the description under the Currency, which examines the date of publication and the last date updated as well as the relevance of the topic or information, that is, determining if the information addresses current events. Relating to the description provided by the Currency, this scan focused on the COVID-19 pandemic, virtual care and nursing regulatory bodies, and professional associations' websites, as these sources are more likely to provide up-to-date and accurate information to their membership.

In 2023, the NSCN and NANB provided practice guidelines and guidelines to direct the provision of care in a safe, competent, and ethical manner. The current nature of these documents suggests that the NSCN and NANB noticed the increased use of virtual care and the importance of providing detailed guidelines to assist nurses in increasing their knowledge and capacity in telenursing. In addition, the CNO provided an updated version of their fact sheet on telepractice in 2023 while the guideline was being reviewed. Some of the associations have not updated their documents and information in recent times. The CNA last updated their telehealth fact sheet in 2017; CRNPEI updated their practice directive for telehealth nursing practice in 2019 and technology in practice in 2020; CRNA had information on telepractice and virtual care information on their web page dated 2020; CRNM guidance on telepractice was published in 2020, an informational document on telepractice was also published in 2021, and the practice expectations spotlight—telepractice or virtual care—was posted in 2020; CRNNL published virtual nursing practice guidelines in 2020; and CRNS made the tag telehealth—nursing use of information and communication technologies available in 2020. As noted in the publication date of these documents, the surge for developing guidance about telehealth and virtual care occurred mainly during the COVID-19 pandemic.

Requirements for Providing Virtual Care Between Provinces

Regulatory bodies in 12 provinces provided a statement on the legislative requirement of practicing telehealth or virtual care in and out of the nurse's jurisdiction. Information from the Quebec's nursing regulatory body, which had the information in French, was not verified. The fundamental requirement, according to the regulatory bodies, is that nurses must have a valid registration in their jurisdiction. Most of the regulatory bodies, BCCNM, CNO, CRNPEI, CRNA, CRNM, CRNNL, NSCN, and NANB, described the legislative requirements around providing telehealth or virtual care when care is provided in the nurses' jurisdictions and across jurisdictions. The CRNNL, CRNM, CRNA, and CNO also provided additional information on the expectations for an out-of-province nurse in providing telepractice to residents outside their jurisdiction. Because of the variation in the registration requirement and the legal aspects involved, especially for out-of-province nurses, it is important for nurses to contact the appropriate authorities or organizations, such as the CNPS and regulatory bodies, before starting to practice in a virtual environment.

Resources Through Professional Nursing Associations and Other Relevant Organizations

The CNPS is a not-for-profit society that offers legal advice, risk management services, legal assistance, and professional liability protection to >140,000 nurses registered with it [32]. During the course of the pandemic, the CNPS took the initiative to look up regulatory and legal guidance available about virtual and telehealth practice for nurses in Canada and made these resources more accessible by providing the URL links to information already posted on the provincial nursing regulatory associations' websites. They also developed some educational resources in the form of case studies for nurses and nursing students to understand different scenarios and possible legal risks associated with virtual nursing practice. Most of these resources were freely accessible in the first 2 years of the pandemic, but now the learning portion of these resources is accessible to CNPS members only. The Canadian Nursing Informatics Association (CNIA) annual conference in 2023 focused on highlighting nurses' innovations and successes in delivering virtual care and directions toward accelerating nursing digital capacity. Digital Health Canada and Canada Health Infoway published up-to-date key resources related to virtual care and both organizations continue to hold regular webinars to facilitate and guide virtual care delivery.

XSL•FO
RenderX

Regulatory Guidance Versus Competency in Virtual Care

To ascertain whether current resources could be leveraged to further enhance nurses' competency in delivering virtual care, we compared the information provided by the regulatory nursing associations in the form of guidelines, practice directives, and fact sheets on telehealth or virtual care with the PHSA Virtual Health Competency Framework for health care providers delivering virtual health, which was developed based on a joint collaboration between the Office of Virtual Health in British Columbia, clinical partners, patients, and family partners. It also applied a comprehensive search of existing literature to identify the domains that reflect the foundational and functional competencies needed to deliver safe care in a virtual environment [29]. As shown in Table 2, only the NANB and NSCN provided comprehensive guidelines, which to a large extent were congruent with most competency indicators proposed in the PHSA framework, although both associations used the term telenursing in their guidelines, as opposed to virtual health. The CRNPEI, in their practice directive, provided some competency expectations for nurses under the section virtual care modalities; however, the competencies identified were related to general technology use and policies without explicitly mentioning virtual care or telehealth. Similarly, the CRNM, in their guidance on telepractice, identified competency in relation to the use of technology and telepractice.

XSL•FO
RenderX

Table 2. Comparison between Provincial Health Service Authority (PHSA) virtual health competencies and the existing regulatory nursing guidance.

PHSA virtual health competencies	Nursing associations		
Domains and competencies	NANB[a]	NSCN[b]	CRNM[c]
Domain 1: virtual health practice requirements—encompasses awareness and understanding of legal, regulatory, and organizational virtual health policies			
• Demonstrates an awareness of the legal and regulatory requirements and practice standards that inform and guide delivery of virtual health	✓	✓	
• Applies relevant organizational policies and decision support tools for safe and effective virtual health	✓	✓	
Domain 2: technology for virtual health—encompasses an understanding of how to use organizational virtual health technologies appropriately and safely			
• Demonstrates the knowledge and skills needed to use virtual health tools	✓	✓	✓
• Demonstrates an awareness and understanding of the privacy, security, and safety features of virtual health tools	✓	✓	
Domain 3: equity-oriented care for virtual health—encompasses the ability to use equity-oriented care in virtual health practice			
• Applies principles of equity-oriented care to determine if virtual health can improve access or exacerbate barriers to care	✓		
• Advocates for and leverages resources to ensure access to virtual health	✓		
Domain 4: delivery of virtual health—encompasses knowledge and capacity to deliver safe, high-quality virtual health			
• Incorporate virtual health safely and appropriately into clinical practice	✓	✓	
• Apply trauma awareness, cultural humility and sensitivity, and harm reduction in virtual health practice	✓		
• Support patient's and family's informed decision-making on the risks and benefits of virtual health	✓	✓	
• Determine what technological supports patients and families need when using virtual health tools	✓	✓	
• Communicate clearly and respectfully in the virtual health environment	✓	✓	
• Demonstrate the skills and judgment needed to safely and effectively complete a virtual health clinical interaction	✓	✓	
• Recognize and respond appropriately to the patients' emotional, psychological, social, and physical needs in the virtual health environment	✓	✓	
• Provide effective patient and family support and share education and follow-up recommendations in the virtual health environment	✓	✓	
• Integrate into virtual health practice the appropriate referral process and documentation standards to ensure quality care	✓	✓	

[a]NANB: Nurses Association of New Brunswick.

[b]NSCN: Nova Scotia College of Nursing.

[c]CRNM: College of Registered Nurses of Manitoba.

Resources About Digital Health and NI

Within the Canadian context, NI competency involves the use of digital health tools to support information synthesis in accordance with professional and regulatory standards in care delivery [16]. As shown in Table 3, various resources are available on regulatory websites to strengthen nursing capacity in digital health and informatics. The CNA and CNIA in 2017 developed a joint position statement on NI, emphasizing the importance of embracing NI competencies to advance nursing practice and knowledge development. At the time of this writing, the development of a revised CNA and CNIA position statement is underway. Across the 13 provinces and territories, only 4 provinces had an NI group presence (Ontario, Alberta, Saskatchewan, and Nova Scotia). Some information could not be verified. For example, the New Brunswick NI group has a

XSL•FO
RenderX

Facebook account, but the last event posted on this page was in 2017. In addition, there is a web link on the Facebook page with a link to the NI group website, which is inactive. At the time of this writing, the Atlantic Canada Nursing Informatics Chapter has been created as part of CNIA in 2023, including the provinces of New Brunswick, Nova Scotia, Newfoundland and Labrador, and Prince Edward Island. The other association is the Association of Registered Nurses of Manitoba, which includes the Manitoba Nursing Informatics Association as a specialty group; however, the link to the Manitoba Nursing Informatics Association is inactive.

Most regulatory nursing associations included information related to the privacy of personal health information and electronic or paper-based documentation as part of standard practice documents such as the entry-to-practice competencies; however, there is limited discussion of NI or digital health concepts. Several associations, including the CRNS, CRNA, and NANB, provided guidelines for social media and e-professionalism for nurses in 2021 and 2022, including the ethical and professional obligations of nurses. Although the Registered Nurses Association of the Northwest Territories and Nunavut also had a document on social media, this was before the pandemic, and it has not been updated recently. Professional nursing associations, including the CNIA and the Canadian Association of Schools of Nursing, have publicly available educational resources about digital health and NI standards such as "the Nursing Informatics Entry-to-Practice Competencies for Registered Nurses" to guide nursing practice and education. At the time of this writing, the development of a revised NI competencies is underway. Nonnursing organizations, such as including Digital Health Canada and Canada Health Infoway, offer extensive resources related to digital health in Canadian health care as well as health informatics competencies for health care providers. However, currently, it is not known how many nurses are actually using these resources.

XSL•FO
RenderX

Table 3. Websites of nursing and related organizations offering digital health and nursing informatics (NI) resources.

Organization	Informatics and digital health	Comments
National and provincial regulatory nursing associations		
CNA[a]	• Position statement: NI (2017) • Infographic: advancing an essential clinical data set in Canada (2019-present) • Fact sheet: privacy of personal health information (2011)	• A joint position statement by CNA and CNIA[b] on the development of NI and competencies • A fact sheet on the privacy of personal health information, providing information on federal, provincial, and territorial privacy laws. In addition, information on resources that support the protection of the privacy of personal health information, including the Pan-Canadian Health Information Privacy and Confidentiality Framework, PIPEDA[c] Awareness Raising Tools Initiative, and CNA's Code of Ethics for Registered Nurses, is provided.
BCCNM[d]	• Not available	—[e]
CNO[f]	• Ontario Nursing Informatics Group	• A web page containing information on the activities of the Ontario Nursing Informatics Group in the development and promotion of NI among nurses in Ontario.
CRNPEI[g]	• Documentation standards (2021)	• A practice directive document on documentation standards that describes the accountability and expectations of nurses concerning documentation in all practice settings, irrespective of the documentation method or storage. The major concepts addressed in the document are principles of documentation, confidentiality, importance of documentation, purpose, who has a role in documentation, cosigning and countersigning entries, key elements of nursing documentation, and timing of documentation. In addition, resources on legislation at the federal and provincial levels that affect nursing documentation are provided.
CRNA[h]	• Social media and e-professionalism: guidelines for nurses (2021) • Privacy and management of health information standards (2022) • Entry-level competencies for the practice of registered nurses (2019) • Documentation standards (2022) • Nursing Informatics Association of Alberta	• A document providing guidelines on professional and ethical obligations with online presence/social media and competencies expected of entry-level and experienced nurses • The responsibilities of managing health information and the requirements of the Health Information Act for custodians and affiliates are also addressed • The core concepts addressed in the documentation process are communication, accountability, legal implications, and expected standards for documentation • A blogspot providing information on activities of the Nursing Informatics Association
CRNM[i]	• Manitoba Nursing Informatics Association (link is inactive)	—
CRNNL[j]	• None	—
CRNS[k]	• Social media guidelines for RNs[l] (2021) • Documentation guidelines for RNs (2021) • General overview of NI to nurse managers of RNs • SNIA[m]	• A dedicated website on Saskatchewan Nursing Informatics Association with links to associated nursing informatics journals, digital health resources, relevant external resources, professional practice, and CNIA • The SNIA provides additional resources and links (publications and links to CNPS and CNO telepractice guidelines), but these do not include their own guidance on virtual care • General guidelines for NI practice on social media and documentation and a specific description of NI to nurse managers supervising RNs
NSCN[n]	• Nova Scotia Nursing Informatics Group	• NSCN has established an informatics group that holds monthly meetings and guides nurses on the use of ICT[o] in health care. • The web page provides links to other international, national, and provincial organizations that align with digital health and informatics.

XSL•FO
RenderX

Organization	Informatics and digital health	Comments
NANB[p]	• Fact sheet: misinformation and disinformation (April 2023) • FAQ[q]: nursing documentation • Practice guidelines: social media (2022) • New Brunswick Nursing Informatics Group	• A practice guide providing information on the overview of social media and the general expected responsibilities of nurses when using the platform • Additional information is provided on general nursing documentation with additional guidance on considerations on electronic documentation, how to document telepractice, or nursing care provided virtually • The New Brunswick NI group Facebook page has little detail about the group and has no current activity or presence. A new Atlantic chapter (the Atlantic Canada Nursing Informatics Chapter) has recently been created.
OIIQ[r]	—	• The information on the website is written in French.
RNANTNU[s]	• Documentation guidelines (2015) • Position statement—social media (2015)	• Detailed information on general documentation guidelines, which have not been updated since 2015. Concepts addressed are standards related to responsibility and accountability, knowledge-based practice, client-centered service, and public trust • Has a document on regulatory expectations for RNs and NPs[t] and the benefits and risks of using social media
YRNA[u]	• Documentation guidelines (2013)	• Generic information about the documentation process and details with aspects of electronic documentation
Professional nursing associations		
CNIA	• 2023 conference—Accelerating Digital Care Capacity in Nursing • General resources	• One of the conference themes focused on how virtual care influenced nursing practice • The resource page provides various informational resources on digital health, informatics within education and practice, NI teaching toolkits, CASN[v] NI entry-to-practice competencies for RNs, and CNA e-Nursing strategy document
CASN	• NI entry-to-practice competencies for RNs • Dedicated web page for digital health and informatics education	• A range of resources, including webinars, toolkits, whiteboard animation videos, research, and web-based self-paced 5 modules. These are primarily targeted at nurse educators and nursing programs. NI competencies update is currently underway.

[a]CNA: Canadian Nurses Association.
[b]CNIA: Canadian Nursing Informatics Association.
[c]PIPEDA: Personal Information and Electronic Documents Act.
[d]BCCNM: British Columbia College of Nurses and Midwives.
[e]Not available.
[f]CNO: College of Nurses of Ontario.
[g]CRNPEI: College of Registered Nurses and Midwives of Prince Edward Islan.
[h]CRNA: College of Registered Nurses of Alberta.
[i]CRNM: College of Registered Nurses of Manitoba.
[j]CRNNL: College of Registered Nurses of Newfoundland and Labrador.
[k]CRNS: College of Registered Nurses of Saskatchewan.
[l]RN: registered nurse.
[m]SNIA: Saskatchewan Nursing Informatics Association.
[n]NSCN: Nova Scotia College of Nursing.
[o]ICT: information and communication technology.
[p]NANB: Nurses Association of New Brunswick.
[q]FAQ: frequently asked question.
[r]OIIQ: Ordre des Infirmières et Infirmiers du Québec.
[s]RNANTNU: The Registered Nurses Association of the Northwest Territories and Nunavut.
[t]NP: nurse practitioner.
[u]YRNA: Yukon Registered Nurses Association.
[v]CASN: Canadian Association of Schools of Nursing.

XSL•FO
RenderX

Discussion

Principal Findings

This environmental scan revealed several strengths and opportunities for improvement. The existing resources available to nurses for providing virtual care are mostly current and have evolved during the course of the COVID-19 pandemic. Despite variations in the types, depth of the information shared, and terms used to describe virtual care and guide nursing practice as described in the existing resources, the availability of such resources indicates that regulatory nursing associations were responsive to the needs of their members and to the evolving phenomena affecting health care and nursing practice in Canada.

Considering the ongoing use of virtual care and the potential risk of reverting to a large-scale use of this form of care delivery should events similar to the COVID-19 pandemic occur in the future, there might be benefits in adopting consistent terminologies as well as practice requirements and competencies. By having formalized standards of practice and competencies related to telehealth, telenursing, or virtual care and constantly updating them to reflect current needs across provinces, this will likely reduce uncertainties among nurses and minimize the tendency for taking a reactive approach to virtual care delivery. Such an approach will also better enable nurses to comprehensively develop their knowledge and competency in providing virtual care, as opposed to seeking information on a need-to-know basis. The PHSA Virtual Health Competency Framework and its compatibility with the existing comprehensive telenursing practice guidelines developed by the NANB and NSCN and the currency of these resources is promising. This framework can be used by regulatory nursing association to complement telenursing practice guidelines as well as by schools of nursing to teach nursing students about virtual health.

Having consensus on the scope of telenursing practice and virtual care across jurisdictions not only enables nurses to better respond to potential future events similar to the scale of the COVID-19 pandemic but also enables them to successfully adapt to the ongoing digital transformation taking place in Canadian health care [33-36]. This is important because upon reviewing available resources aimed at guiding and developing NI competencies as well as supporting nursing practice in digital health, it was found that these resources also varied. For example, some associations provided documentation guidelines to their members, but these were not consistently updated. In addition, they did not specifically address documentation in electronic health records or link to NI competencies.

Nurses represent the largest group of Canadian care providers, and regardless of their professional designation, they are at crossroads with digitalization [33,36]. As such, engaging the nursing workforce and supporting the development of new skills is vitally important to realize the full potential of digital innovations [35]. This will enable nurses to leverage digital technologies to further enhance nursing practice, facilitate accessibility to care, and improve overall health outcomes [33,36]. As part of their professional responsibility and accountability, nurses must also pursue professional development opportunities in digital health and NI offered through their professional associations or other sources as well as to keep abreast with issues and phenomena that influence nursing practice and health care. It is also important that health systems critically reflect on lessons learned from the pandemic, including the benefits and challenges associated with virtual care delivery and proactively planning for what should be done differently going forward. Furthermore, it is incumbent upon health systems investing in digital health technologies to support nurses in adapting to this changing context of health care and to work collaboratively with regulatory and professional nursing associations and schools of nursing to develop digital care capacity among practicing and future nurses. An evidence of the commitment of health care governing authorities at the federal, provincial, and territorial levels to support efforts to increase technology adoption and virtual care is the report released by the Canadian Institute of Health Information, in which they developed a series of case studies based on semistructured interviews with representatives from provinces and territories [37]. These case studies addressed various aspects, including strategy, governance and direction setting, programs, and initiatives, providing an excellent resource to facilitate learning and policy direction regarding virtual care [37]. Similarly, Canadian Association of Schools of Nursing has recently secured funding from Health Canada to sponsor the project “Essential COVID-19 Skills for Graduating and New Nurses” [38]. Throughout the COVID-19 pandemic and continuing to the present, the CNA has engaged in focused advocacy work. This includes recommendations for implementing a pan-Canadian digital health strategy; investing in virtual care to support populations considered vulnerable; ensuring that health workers receive appropriate training; and providing financial support to help jurisdictions deploy essential infrastructure and access reliable, high-speed internet services [39].

Conclusions

Despite challenges during the COVID-19 pandemic, it also served as a catalyst for expanding access to care through digital health modalities such as telehealth and virtual care. Nursing associations play an important role in regulating and ensuring that nurses provide safe, competent, and ethical care. Considering that nursing practice with telehealth and virtual care has become mainstream, there are opportunities to build on these successes and mitigate potential risks in the future. Work completed to date to inform Canadian nursing practice in a virtual context is substantial and can serve as a beginning step for developing a standardized set of requirements to inform nursing practice in telehealth and virtual care and in the education of future nurses. This environmental scan followed a systematic search and provided important insights into the current state of resources available to nurses in Canada regarding virtual care through their professional and regulatory associations. To our knowledge, no prior work has been done on this topic. This scan may not be applicable to nurses outside Canada.

XSL•FO
RenderX

Acknowledgments

MK and AA received funding through the Social Sciences and Humanities Research Council Insight Development Grant, and a portion of this funding was used to pay for research assistantship costs.

Authors' Contributions

MK, AA, and LMN conceptualized the project idea, developed the initial draft, and led the discussion of findings. EMD, SI, SA, and DdC contributed to data searches, abstraction, and reviewing and editing the final paper.

Conflicts of Interest

None declared.

References

1. NEJM Catalyst. What is telehealth? NEJM Catalyst. 2018 Feb 1. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0268 [accessed 2024-07-19]
2. Shaw J, Jamieson T, Agarwal P, Griffin B, Wong I, Bhatia RS. Virtual care policy recommendations for patient-centred primary care: findings of a consensus policy dialogue using a nominal group technique. J Telemed Telecare 2018 Oct;24(9):608-615. [doi: 10.1177/1357633X17730444] [Medline: 28945161]
3. Ndayishimiye C, Lopes H, Middleton J. A systematic scoping review of digital health technologies during COVID-19: a new normal in primary health care delivery. Health Technol (Berl) 2023;13(2):273-284 [FREE Full text] [doi: 10.1007/s12553-023-00725-7] [Medline: 36628261]
4. Mbunge E, Batani J, Gaobotse G, Muchemwa B. Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19) pandemic in South Africa: a systematic review. Glob Health J 2022 Jun;6(2):102-113 [FREE Full text] [doi: 10.1016/j.glohj.2022.03.001] [Medline: 35282399]
5. Breton M, Deville-Stoetzel N, Gaboury I, Smithman MA, Kaczorowski J, Lussier MT, et al. Telehealth in primary healthcare: a portrait of its rapid implementation during the COVID-19 pandemic. Healthc Policy 2021 Aug;17(1):73-90 [FREE Full text] [doi: 10.12927/hcpol.2021.26576] [Medline: 34543178]
6. Regragui S, Abou Malham S, Gaboury I, Bois C, Deville-Stoetzel N, Maillet L, et al. Nursing practice and teleconsultations in a pandemic context: a mixed-methods study. J Clin Nurs 2023 Sep;32(17-18):6339-6353. [doi: 10.1111/jocn.16756] [Medline: 37202866]
7. Mohammed HT, Hyseni L, Bui V, Gerritsen B, Fuller K, Sung J, et al. Exploring the use and challenges of implementing virtual visits during COVID-19 in primary care and lessons for sustained use. PLoS One 2021 Jun 24;16(6):e0253665 [FREE Full text] [doi: 10.1371/journal.pone.0253665] [Medline: 34166441]
8. Health care workers' experiences providing virtual care during the COVID-19 pandemic. Statistics Canada. 2022 Nov 18. URL: https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2022073-eng.htm#shr-pg0 [accessed 2024-02-01]
9. Mahoney MF. Telehealth, telemedicine, and related technologic platforms: current practice and response to the COVID-19 pandemic. J Wound Ostomy Continence Nurs 2020;47(5):439-444 [FREE Full text] [doi: 10.1097/WON.0000000000000694] [Medline: 32970029]
10. Hughes L, Petrella A, Phillips N, Taylor RM. Virtual care and the impact of COVID-19 on nursing: a single centre evaluation. J Adv Nurs 2022 Feb;78(2):498-509 [FREE Full text] [doi: 10.1111/jan.15050] [Medline: 34590738]
11. Lee C, Scime S, Simon S, Ng F, Fitch MI. Exploring nursing engagement in providing virtual care to cancer patients in Canada. Can Oncol Nurs J 2022 Oct 1;32(4):580-585 [FREE Full text] [Medline: 38919778]
12. Lee JY, Lee S, Choi H, Oh EG. Exploring the experiences of frontline nurses caring for COVID-19 patients. Int Nurs Rev 2023 Mar;70(1):50-58 [FREE Full text] [doi: 10.1111/inr.12801] [Medline: 36018881]
13. Abdolkhani R, Petersen S, Walter R, Zhao L, Butler-Henderson K, Livesay K. The impact of digital health transformation driven by COVID-19 on nursing practice: systematic literature review. JMIR Nurs 2022 Aug 30;5(1):e40348 [FREE Full text] [doi: 10.2196/40348] [Medline: 35867838]
14. 2020 National Survey of Canadian Nurses: use of digital health technology in practice. Canada Health Infoway. 2020 May 11. URL: https://www.infoway-inforoute.ca/en/component/edocman/resources/reports/benefits-evaluation/3812-2020-national-survey-of-canadian-nurses-use-of-digital-health-technology-in-practice [accessed 2024-02-01]
15. Beauséjour W, Hagens S. Uncovering important drivers of the increase in the use of virtual care technologies in nursing care: quantitative analysis from the 2020 National Survey of Canadian Nurses. JMIR Nurs 2022 Mar 31;5(1):e33586 [FREE Full text] [doi: 10.2196/33586] [Medline: 35357326]
16. Kleib M, Nagle LM, Furlong KE, Paul P, Duarte Wisnesky U, Ali S. Are future nurses ready for digital health?: informatics competency baseline assessment. Nurse Educ 2022;47(5):E98-104 [FREE Full text] [doi: 10.1097/NNE.0000000000001199] [Medline: 35324499]

XSL•FO
RenderX

17. Kaihlanen AM, Elovainio M, Virtanen L, Kinnunen UM, Vehko T, Saranto K, et al. Nursing informatics competence profiles and perceptions of health information system usefulness among registered nurses: a latent profile analysis. J Adv Nurs 2023 Oct;79(10):4022-4033. [doi: 10.1111/jan.15718] [Medline: 37243421]
18. De Leeuw JA, Woltjer H, Kool RB. Identification of factors influencing the adoption of health information technology by nurses who are digitally lagging: in-depth interview study. J Med Internet Res 2020 Aug 14;22(8):e15630 [FREE Full text] [doi: 10.2196/15630] [Medline: 32663142]
19. Kleib M, Nagle L. Factors associated with Canadian nurses' informatics competency. Comput Inform Nurs 2018 Aug;36(8):406-415. [doi: 10.1097/CIN.0000000000000434] [Medline: 29596068]
20. Jarva E, Oikarinen A, Andersson J, Tuomikoski AM, Kääriäinen M, Meriläinen M, et al. Healthcare professionals' perceptions of digital health competence: a qualitative descriptive study. Nurs Open 2022 Mar;9(2):1379-1393 [FREE Full text] [doi: 10.1002/nop2.1184] [Medline: 35094493]
21. Nursing statistics. Canadian Nurses Association. URL: https://www.cna-aiic.ca/en/nursing/regulated-nursing-in-canada/nursing-statistics [accessed 2024-02-01]
22. Villeneuve M, Betker C. Nurses, nursing associations, and health systems evolution in Canada. Online J Issues Nurs 2020;25(1) [FREE Full text] [doi: 10.3912/OJIN.Vol25No01Man06]
23. Almost J. Regulated nursing in Canada: the landscape in 2021. Canadian Nurses Association. 2021 Feb. URL: https://hl-prod-ca-oc-download.s3-ca-central-1.amazonaws.com/CNA/2f975e7e-4a40-45ca-863c-5ebf0a138d5e/UploadedImages/documents/Regulated-Nursing-in-Canada_e_Copy.pdf [accessed 2024-02-01]
24. Why do professional associations matter? Canadian Nurses Association. URL: https://www.cna-aiic.ca/en/about-us/who-we-are/why-professional-associations-matter [accessed 2024-02-01]
25. Charlton P, Doucet S, Azar R, Nagel DA, Boulos L, Luke A, et al. The use of the environmental scan in health services delivery research: a scoping review protocol. BMJ Open 2019 Sep 06;9(9):e029805 [FREE Full text] [doi: 10.1136/bmjopen-2019-029805] [Medline: 31494613]
26. Graham P, Evitts T, Thomas-MacLean R. Environmental scans: how useful are they for primary care research? Can Fam Physician 2008 Jul;54(7):1022-1023 [FREE Full text] [Medline: 18625830]
27. Kleinheksel AJ, Rockich-Winston N, Tawfik H, Wyatt TR. Demystifying content analysis. Am J Pharm Educ 2020 Jan;84(1):7113 [FREE Full text] [doi: 10.5688/ajpe7113] [Medline: 32292185]
28. Erlingsson C, Brysiewicz P. A hands-on guide to doing content analysis. Afr J Emerg Med 2017 Sep;7(3):93-99 [FREE Full text] [doi: 10.1016/j.afjem.2017.08.001] [Medline: 30456117]
29. MacDougall M, Jiwa S, Lee B, Underhill D. Virtual health competency framework: for health-care providers delivering virtual health. Provincial Health Services Authority. 2024 Jan 23. URL: http://www.phsa.ca/health-professionals-site/Documents/Office%20of%20Virtual%20Health/OVH%20Virtual%20Health%20Competency%20Framework.pdf [accessed 2024-02-01]
30. Do I need research ethics approval? University of Alberta. URL: https://www.ualberta.ca/en/research/services/research-ethics/do-i-need-research-ethics-approval.html [accessed 2024-08-12]
31. Fielding JA. Rethinking CRAAP: getting students thinking like fact-checkers in evaluating web sources. C&RL News 2019 Dec 05;80(11):620 [FREE Full text] [doi: 10.5860/crln.80.11.620]
32. Canadian Nurses Protective Society. URL: https://cnps.ca/ [accessed 2024-02-01]
33. Booth RG, Strudwick G, McBride S, O'Connor S, Solano López AL. How the nursing profession should adapt for a digital future. BMJ 2021 Jun 14;373:n1190 [FREE Full text] [doi: 10.1136/bmj.n1190]
34. Sensmeier J. Virtual care: the time is now. Nurs Manage 2019 Feb;50(2):22-26. [doi: 10.1097/01.NUMA.0000552738.79448.3e] [Medline: 30695010]
35. Engaging and transforming the health workforce. In: Health in the 21st Century: Putting Data to Work for Stronger Health Systems. Paris, France: OECD Publishing; 2019.
36. Booth RG, Strudwick G. Preparing nursing for the virtual care realities of a post-pandemic future. Nurs Leadersh (Tor Ont) 2021 Dec;34(4):86-96. [doi: 10.12927/cjnl.2021.26685] [Medline: 35039123]
37. The expansion of virtual care in Canada: new data and information. Canadian Institute for Health Information. 2023. URL: https://www.cihi.ca/sites/default/files/document/expansion-of-virtual-care-in-canada-report-en.pdf [accessed 2024-02-01]
38. CASN secures funding for an essential COVID-19 virtual simulation education series for nurses. Canadian Association of Schools of Nursing. URL: https://www.casn.ca/2020/07/casn-secures-funding-for-an-essential-covid-19-virtual-simulation-education-series-for-nurses/ [accessed 2024-02-01]
39. Virtual care. Canadian Nurses Association. URL: https://www.cna-aiic.ca/en/policy-advocacy/advocacy-priorities/virtual-care [accessed 2024-02-01]

Abbreviations

BCCNM: British Columbia College of Nurses and Midwives
CNA: Canadian Nurses Association
CNIA: Canadian Nursing Informatics Association

XSL•FO
RenderX

CNO: College of Nurses of Ontario
CNPS: Canadian Nurses Protective Society
CRNA: College of Registered Nurses of Alberta
CRNM: College of Registered Nurses of Manitoba
CRNNL: College of Registered Nurses of Newfoundland and Labrador
CRNS: College of Registered Nurses of Saskatchewan
NANB: Nurses Association of New Brunswick
NI: nursing informatics
NSCN: Nova Scotia College of Nursing
PHSA: Provincial Health Service Authority

Edited by B Lesselroth, F Pietrantonio, J López Castro, M Montagna, I Said-Criado; submitted 30.09.23; peer-reviewed by K Butler-Henderson, G Strudwick, P Khorasani; comments to author 18.12.23; revised version received 11.02.24; accepted 20.06.24; published 13.08.24.

Please cite as:
Kleib M, Arnaert A, Nagle LM, Darko EM, Idrees S, da Costa D, Ali S
Resources to Support Canadian Nurses to Deliver Virtual Care: Environmental Scan
JMIR Med Educ 2024;10:e53254
URL: https://mededu.jmir.org/2024/1/e53254
doi:10.2196/53254
PMID:39137026

©Manal Kleib, Antonia Arnaert, Lynn M Nagle, Elizabeth Mirekuwaa Darko, Sobia Idrees, Daniel da Costa, Shamsa Ali. Originally published in JMIR Medical Education (https://mededu.jmir.org), 13.08.2024. 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.

XSL•FO
RenderX

Original Paper

Newly Qualified Canadian Nurses' Experiences With Digital Health in the Workplace: Comparative Qualitative Analysis

Manal Kleib[1], BSN, MSN, MBA, PhD; Antonia Arnaert[2], PhD; Lynn M Nagle[3], PhD; Rebecca Sugars[1], BSN; Daniel da Costa[2], BSN

[1]Faculty of Nursing, University of Alberta, Edmonton, AB, Canada
[2]Ingram School of Nursing, McGill University, Montreal, QC, Canada
[3]Faculty of Nursing, University of New Brunswick, Fredericton, NB, Canada

Corresponding Author:
Manal Kleib, BSN, MSN, MBA, PhD
Faculty of Nursing
University of Alberta
5-112 Edmonton Clinic Health Academy
Edmonton, AB, T6G1C9
Canada
Phone: 1 7802481422
Email: manal.kleib@ualberta.ca

Abstract

Background: Clinical practice settings have increasingly become dependent on the use of digital or eHealth technologies such as electronic health records. It is vitally important to support nurses in adapting to digitalized health care systems; however, little is known about nursing graduates' experiences as they transition to the workplace.

Objective: This study aims to (1) describe newly qualified nurses' experiences with digital health in the workplace, and (2) identify strategies that could help support new graduates' transition and practice with digital health.

Methods: An exploratory descriptive qualitative design was used. A total of 14 nurses from Eastern and Western Canada participated in semistructured interviews and data were analyzed using inductive content analysis.

Results: Three themes were identified: (1) experiences before becoming a registered nurse, (2) experiences upon joining the workplace, and (3) suggestions for bridging the gap in transition to digital health practice. Findings revealed more similarities than differences between participants with respect to gaps in digital health education, technology-related challenges, and their influence on nursing practice.

Conclusions: Digital health is the foundation of contemporary health care; therefore, comprehensive education during nursing school and throughout professional nursing practice, as well as organizational support and policy, are critical pillars. Health systems investing in digital health technologies must create supportive work environments for nurses to thrive in technologically rich environments and increase their capacity to deliver the digital health future.

(JMIR Med Educ 2024;10:e53258) doi:10.2196/53258

KEYWORDS
digital health; new graduate nurses; nursing practice; workplace; informatics

Introduction

Clinical practice settings have increasingly become dependent on the use of digital or eHealth technologies such as electronic health records, telehealth, mobile health, and medical devices to name a few [1,2]. The COVID-19 pandemic has also contributed to the increased use of digital technologies to facilitate virtual care delivery, creating both opportunities and challenges for care providers and patients [3-5]. Digital health refers to the "proper use of technology for improving the health and well-being of people at the individual and population levels" [6]. According to the World Health Organization, digital health expands the concept of eHealth, with a wide range of smart devices and connected equipment. It also encompasses other uses of digital technologies for health such as the Internet of Things, artificial intelligence (AI), big data, and robotics [7].

Recognizing the increased use of digital health across health systems, the International Council of Nurses [8] has recently

XSL•FO
RenderX

released a position statement affirming the central role of digital health in contemporary nursing practice and the importance of developing the skills and competencies of nurses through the integration of digital health content in formal undergraduate and postgraduate curricula and participation in continuing professional development. These recommendations have also recently been echoed by the Canadian government [9] identifying digital preparedness as one dimension of actionable strategies that should be put into place to mitigate the administrative burden associated with the use of digital tools by nurses as primary users of these technologies.

Digital capabilities or informatics competencies are critical core requirements for safe nursing practice with technology. In Canadian health care, nurses must be able to use digital health technologies to support information synthesis and patient care in accordance with their professional and regulatory standards [10]. Despite efforts to enhance digital preparedness among future nurses, research involving senior-level nursing students identified that most learning about digital health is taking place in clinical settings with limited theoretical education in the classroom [11,12]. In a practice-based profession, such as nursing, it is inherent that nursing students are socialized into their professional roles by observing and interacting with nurses [13] in health care settings that they go to for their clinical education; however, practicing nurses involved in mentoring nursing students also have limited digital capabilities and report challenges with technology use, mostly among older nurses [14,15]. Assumptions also prevail that younger nurses are tech-savvy because they were mostly born in the digital age; however, despite being described as digital natives, research suggests that they do not naturally have positive views on using technology for care provision [16] nor do their digital skills easily transfer to the clinical context [13,17-19].

While nursing programs acknowledge the importance of integrating digital health content at all levels of nursing education, research involving nurse educators [20] and academic administrators [21] from across schools of nursing in Canada revealed that nurse educators have limited knowledge and capacity to teach informatics, limited awareness about informatics competencies, and limited resources to teach nursing students about digital health applications in clinical practice. For example, 65% of these schools indicated they do not have access to a training version of any electronic health record system to teach students about this technology. Although these observations are focused on Canadian nurses, similar gaps have been reported in other countries [18,19,22,23]. Potentially, these gaps can have a negative impact on nursing graduates as they transition to the workplace. According to a recent study [24] involving clinical managers and newly registered nurses (RNs) in the United Kingdom, researchers identified several factors impacting these nurses in the workplace including technology infrastructure, time, skills, digital literacy training, support, leadership, familiarity, and confidence: creating barriers to optimal nursing practice with technology. It is vitally important to support nurses in adapting to more digitalized health care systems; however, little is known about Canadian nursing graduates' experiences. This study sought to (1) describe newly qualified nurses' experiences with digital health in the workplace, and (2) identify strategies that could help support their transition and clinical practice with digital health technologies.

Methods

Design

An exploratory descriptive qualitative design was most appropriate considering the limited research available [25]. This design aligns with the philosophical perspectives of pragmatism and constructionism and is concerned with understanding and describing the human experience within its unique context [25]. Through inductive and dynamic research processes, the researcher explores the subjective experiences of participants and their perceptions through a collection of data that describes the "who, what, and where of the events or experiences" [25]. In this study, qualitative data was collected from semistructured interviews with nurses, a technique suitable for collecting data that is open-ended and explores participants' thoughts, feelings, and beliefs about a particular topic [26].

Participants

Qualitative studies recommend 1-30 informants [25-28]. In this study, the number of participants was guided by the concept of information power to interview sampling, which indicates that if the sample holds more information that is relevant to the actual study, a lower number of participants is needed [27]. The five aspects of information power further supporting this approach were (1) the specific or narrow study aim, (2) participants holding characteristics that are specific to the study aim and identification supported by suitable recruitment strategies, (3) theoretical support, (4) the quality of dialogue with participants, and (5) a thematic cross-case analysis approach [27]. In addition, the concept of thematic saturation, which means researchers would stop data collection when there is no new information shared by participants [26], was also considered. Participants were newly qualified nurses who graduated from 2 undergraduate nursing programs in Eastern and Western Canada (WC) in the last 2 years and have been working in clinical settings that have digital health technologies, such as electronic health records, regardless of the stage of implementation. Purposive and snowball sampling techniques were applied to enroll interested participants with consideration of various sociodemographic backgrounds [28].

Data Collection

To recruit participants, administrative staff in the selected nursing programs circulated a recruitment poster invitation via a listserv email to graduates from these programs, and potential participants were asked to contact the researchers to express interest in participating in the research. An interview guide (Textbox 1) was developed by the researchers based on their expertise in this area and the general literature using open-ended questions and prompts to facilitate the discussion regarding aspects related to new graduates' experiences with digital health [26]. Pretesting of the interview schedule was accomplished by engaging 2 research assistants (RAs; RS and DC) involved in this project in a demo interview, as both were recent graduates. No further revisions were introduced to the initial interview

XSL•FO
RenderX

guide. In addition, sociodemographic data, including age, gender, and work setting, were collected at the beginning of interviews to understand the characteristics of participants and the context of their practice environments. Interviews were facilitated by RAs (RS and DC) without the presence of the researchers who were also faculty members and may have interacted with these participants when they were students, as such minimize undue discomfort to participants. The interviews were held via Zoom (Zoom Video Communications, Inc), each lasting between 45 and 60 minutes, and participants had the option to turn their cameras off during the interview. Although options for in-person interviews were offered, participants preferred a digital approach considering their busy clinical schedules, and this mode did not impact the quality of dialogue between the participants and the interviewers. The audio-recorded interviews were then uploaded to a secured SharePoint (Microsoft Corp) drive and later transcribed verbatim by a professional transcriptionist service provider and records.

Textbox 1. Interview guide.

- Could you share what types of technologies do you currently use in your day-to-day practice as a registered nurse? Prompt: Could you share some examples?
- Could you describe your experiences when working with these technologies? Prompt: what makes your work easy/difficult?
- Could you describe your experiences of learning on how to use these technologies? Prompt: In the nursing school/in the workplace?
- What suggestions do you have to enhance newly graduated nurses' experiences of using digital health technologies? Prompt: in the nursing school, in the workplace, other?
- Is there anything else you would like to share?

Data Analysis

RAs (RS and DC) received training in qualitative data analysis and an orientation to the data management software. Data were uploaded into the NVivo software (Lumivero) in a deidentified format to facilitate data management and analysis. An inductive manifest content analysis was used considering the research is exploratory. In this approach, the goal is to "describe what the informants actually say, stay close to the text, use the words themselves, and describe the visible and obvious in the text" [28]. The unit of analysis (ie, the sample) involved analyzing data from all the participants in its entirety [28]. The two RAs were involved in data analysis, which began by having each one independently read the interview transcripts line by line, several times, to make sense of the data and whole and achieve familiarity [28-30] Open coding (ie, identification of meaning units and labeling each with a code) was then done by having each RA independently analyze 2 transcripts. These pilot transcripts were then compared and discussed to enhance reliability before coding the remainder of the data. At this stage, we also created a coding list, including all the meaning units identified along with an explanation of the codes to enhance intercoder reliability and reduce cognitive fatigue. The initial coding was then done by the research team without additional changes introduced. Condensing of the meaning units was then done to identify categories and subcategories without losing the content of the unit, which was followed by grouping of categories with similar events into main categories. The researchers then constructed the themes based on the analysis of these data (Multimedia Appendix 1). The resulting analysis was reported in a narrative format. Rigor was enhanced by reviewing the coding scheme against the interview data and the involvement of all team members in the discussion and interpretation of the findings [28-30].

Ethical Considerations

Ethics approval was obtained from an Ethics Research Review Board in the WC research site (Pro00112596), and secondary approval from an ethics board at the university in Eastern Canada (EC; A12-E69-21B). Participants were provided with an information sheet along with the consent form and were given the opportunity to ask questions and verify any aspect related to the research before providing their consent for participation. Participants were also assured that they could opt out at any time without penalty. To protect privacy and confidentiality, the interview data were deidentified by using a code (eg, participant 01-E and participant 01-W). Each participant received a small gift certificate to convey appreciation of their time.

Results

Demographic Characteristics of Participants

A total of 14 participants were interviewed, 6 participants from WC and 8 participants from EC. Of those 14 participants, 8 participants were younger than 25 years old, 4 participants were in the age category of 26-30 years, and 2 participants were in the age category of 31-41 years. Participants were mostly females, except for 1 male participant. They worked in large urban hospitals in a variety of units including emergency room, intensive care, internal medicine, trauma surgery, and operating room.

Digital Health Technologies in the Workplace

In their current practice environments as RNs, participants indicated that they used different hardware (eg, laptops, mobile devices, workstations on wheels, and portable computers), clinical information systems (CIS), and medical devices and equipment (Table 1).

XSL•FO
RenderX

Table 1. Types of digital health technologies currently used in practice.

Technology	Western Canada site	Eastern Canada site
CIS[a]: new, old, and hybrid systems	• EPIC CIS/EMR[b] EDIS or Millennium NetCare Electronic Health Record Pyxis Vax Meditech	• OACIS (EMR used universally) • MedUrge (EMR used in acute care only) • VSign (mobile app to access Oacis remotely) • PANDAWebRx (formerly known as GESPHARxLite; eMAR used at some sites)
Medical devices and equipment (some devices are connected to newer CIS systems, and others are stand-alone with paper charting)	• Sphygmomanometer for blood pressure • Cardiac monitors • Glucometers • Bladder scanners • Transcutaneous bilirubin meter • Dopplers • ECG[c] machines • Pulse oximeter for O2 Saturation • Temperature probe	• Philips monitors (syncs to Oacis) • Nova glucometers (syncs to Oacis) • BBraun IV pumps • Apple watches (in terms of patient education and telemonitoring) • Continuous glucose monitors used by patients

[a]CIS: clinical information system.

[b]EMR: electronic medical record.

[c]ECG: electrocardiogram.

Key Findings

Overview

Data analysis revealed three themes: (1) experiences before becoming an RN, (2) experiences upon joining the workplace, and (3) suggestions for bridging the gap in the transition to digital health practice. These themes are discussed below with illustrative quotes from the participants.

Theme 1: Experiences Before Becoming an RN

This theme includes the categories: (1) academic, (2) clinical, and (3) personal experiences. Most participants stated having familiarity with technology before joining their schools, and perhaps because of that, it was assumed that they were highly literate with technology since they grew up with it and have used it frequently in their daily lives. In a way, this made learning about technology easier, but in other cases, it also created barriers. According to one participant:

> *...sometimes I feel like there's an assumption that maybe we know more than we do, just because of the age we are ... we know how to use things like a phone, or the computer, or whatever, but it's used in a different way in practice.* [P5, WC]

This sentiment was shared by another participant who noted that individual differences should also be considered:

> *I think it's individual, some people are just a little bit more inclined to be using it [technology] and have an easier time.* [P2, EC]

At the time participants were in school, they indicated being aware of the relevance of digital health technology (DHT) to nursing practice. Their education involved varied exposure to theoretical and laboratory training, with most learning taking place in clinical settings, but this was not without challenges. A participant described it as follows:

> *It was very much like somebody handed you a cheat sheet that just told you like, Type in these numbers, and then you will find your [Laughs] ... your medication page for this patient, and you can enter your medications. So, we got that from our clinical instructor, and then different nurses had different workflows. They would show us, this is how to print out my MAR, or this is the time of day when I'll go and log that I've given my medications.* [P1, WC]

The depth and breadth of these experiences were also noted by another participant who further elaborated:

> *At the beginning of every clinical, we had somewhat of an orientation on the first day where we would go into the unit where we'd be doing our stage and we would get a little session on how to use the different platforms that the unit uses ... They would show us a brief overview of the different tabs that are available and what kind of information you can get and that kind of stuff.* [P3, EC]

Another participant added:

> *Even just things of like having to do an ECG on someone ... Where do I put the other ECG tab? Because that's an actual thing that happens ... nobody comes in looking perfectly healthy and pristine as like the students that are in the program.* [P7, WC]

With respect to learning experiences in the classroom, the majority indicated receiving limited and broad education about DHT, with some having mixed feelings about these gaps. As described by one participant:

> *We would talk about it in lectures ... I don't really remember ever doing anything like digital in labs.* [P5, WC]

Another participant agreed adding:

> *During my studies, it was less obvious ... it wasn't through my classes that I learned about these technologies, it was more when I was using it in practice.* [P7, EC]

XSL•FO
RenderX

One participant went on saying:

> *There are so many unspoken things in nursing ... I always found myself, a lot of times, getting feedback from the instructor saying, "You should know this," and me being like, "I didn't know that I had to ... I didn't know that I didn't know that I was supposed to know this.* [P7, WC]

Some participants did not make sense of DHT and its relevance to nursing practice until they had experienced its expanded use; somewhat creating a sense of urgency that they should learn more about it. A participant explained:

> *It wasn't until the pandemic, really, that I understood. Now, this is what telehealth is used for; this is the function, this is what it can do, it can bridge the gaps ... it was mostly the pandemic that really changed my mindset there.* [P8, EC]

The large-scale transition to a new electronic record system was also mentioned:

> *...like not built into our curriculum, for example, when we would have guest speakers or informal discussions, we would always talk about electronic charting system because it was like a pretty big deal for the implementation of this provincial system in phases...* [P4: WC]

Theme 2: Experiences Upon Joining the Workplace

This theme includes two categories: (1) facilitating and (2) hindering experiences regarding the use of DHT in the workplace.

Hindering experiences discussed by participants included aspects related to training on CIS and medical devices and support resources, the IT infrastructure, and the potential impact of these challenges on their practice. Upon starting their practice as novice nurses, most participants indicated receiving either limited or condensed DHT training over a short period, which despite their tech-savviness, found to be a little overwhelming. This thought was shared by a participant who mentioned:

> *Our orientation for [__] was fairly short ... it was a lot of self-learning on how the program works ... you just have to play around with the program to kind of understand how it works.* [P1, EC]

Another participant added:

> *We never received any kind of formal training on how to interact with patients using telehealth. Not even in formal training, you are kind of thrown into that world and expected to pick up the phone and have these conversations.* [P3, EC]

Another participant shared

> *They'll teach us if we're the nurse for that patient on how to use it [a medical device]. But it's not like a full unit thing ... it's kind of limited for 30 minutes ... and then the one nurse that has the patient that day. I mean, when that nurse leaves, another nurse comes in, how are they going to help the patient with this device?* [P1, EC]

Participants also mentioned fragmentation of digital health across the health care system, exacerbated by a lack of interoperability, with less technology integration in rural settings. When working with hybrid or outdated systems, new graduates felt more prone to making errors while transferring patient information between different platforms. As one participant described:

> *I'm not looking to make my job easier. I'm looking to make myself more effective, and I would be significantly more effective if I had access to everything and could spend less time fumbling through papers and, you know, clicking on screens, and could do things faster, then, I can spend more time actually with my patients...* [P1, WC]

Participants also discussed that patient care became less efficient when the system was not user-friendly, the technology was outdated, the system lagged, IT glitches were not solved quickly, and there were not enough computers on the unit. All these issues increased concerns regarding patient privacy and confidentiality, safety, and potential legal liabilities to the nurse and the organization. A participant explained:

> *Because sometimes that happens, you know, one computer at the nursing station is not working, you go on another one, everything is down. And it came back within like 15-20 minutes. That's a long time when I need to call a doctor to see a critical value. Maybe it doesn't sound like a lot, but 15 minutes for the system, the entire system to be down is a lot. And when that was going to happen, it was a near change of shift. So, we weren't sure if the care plans were going to print for the next shift. So, we had to start writing them out. This is not a feasible way to be doing this. In that case, we were all lost. Even the people who had been there for 20 years didn't know what to do. I was asking people, "What can I do here?" Nobody knew. So, not really good.* [P8, EC]

These practices were perceived to consume valuable nursing time that could have been spent in the clinical care of patients as a participant explained:

> *When we take time trying to fix a machine that wasn't working so that we could take patients to monitor their cardiac function, then it takes away from our ability to get to know patients and to treat them more holistically.* [P5, EC]

Multiple participants also voiced concerns regarding inconsistencies in how nurses documented electronically, and how these differences although sometimes helped them learn different strategies, also caused confusion and could potentially result in legal or patient safety issues as one participant described:

> *I think there are some discrepancies that could lead to some actual challenges for patient safety.* [P5, EC]

Some also indicated redundant charting or copying electronic information from one platform to another, sometimes including an analog step requiring a physical document to be scanned. According to one participant:

XSL•FO
RenderX

> *Well, I know something happened that day, but nothing got charted about it, or there is not very complete charting about that incident, because that nurse prefers to chart everything in her flowsheets and doesn't write a lot in her note.* [P2, WC]

These issues were further compounded by the limited support and resources available to new graduates during night shifts and on weekends such as access to clinical nurse educators (CNEs) and tech support as one participant explained:

> *When you are on a dayshift, there's lots of extra support. It is a little bit tougher if you are on evenings or nights because those extra supports, like the managers and educators, aren't around.* [P2, WC]

On the other hand, participants discussed factors that facilitated their experiences with DHT including having access to a wide range of technologies to support care, being adaptable, and positive workplace culture. Having access to technology that worked well enhanced efficiency including time-saving actions such as the convenience of having prompt access to patient information in one system, which enabled prioritization of care, clinical decision-making, standardization, and continuity of care. According to one participant:

> *I think it [CIS] has definitely made me more detail-oriented, especially now that I've been using them for a little bit over a year ... there are so many things that you're thinking about day-to-day with every single patient, ... okay, what are the priorities for my patient, and what am I looking for? What could go wrong? What do I need to keep in mind?* [P2, WC]

Some participants also talked about the technology being helpful in giving them prompts about things they have learned about but could not easily remember. For instance, some of the newer CIS systems provided help pages with information about managing patient care and relevant nursing actions, which facilitated their practice in busy clinical environments as they were also adapting to their new roles as independent RNs:

> *It's been a great facilitator in my learning experience as a new grad nurse because I'm able to stay on top of everything more easily ... it gives us the opportunity to focus on patient interaction and on what's important.* [P5, EC]

Participants acknowledged that learning about technology is a learning curve and therefore being adaptable is key. In addition, ongoing learning about technology is important—not only upon orientation to the unit. As explained by participant:

> *...there was really just like a different learning curve ... now it's been like over two years since the system has been implemented, everyone is pretty much onboard ... there's still some of our staff that, you know, like need help with certain things, every now and then, something will surprise you and like you'll realize you're still learning, or like they'll update the software.* [P4, WC]

New graduates' practice with DHT became easier through the support of nurses in the unit, accessibility to other resource persons, such as CNEs and superusers, having the opportunity to learn from more experienced nurses and in exchange teaching them technology-related skills, and the mentoring behaviors of nurses toward the new hires. According to one participant:

> *The culture of the unit is so supportive in so many ways, I think that it exceeded all my expectations for the support that I feel in being able to ask questions and share knowledge with other people.* [P5, EC]

Other factors contributing to positive experiences, as discussed by some participants who have transitioned to a new system, included working with integrated CIS systems and having access to a playground or a sandbox during their initial training. In addition, those who have had positive training and education experiences with DHT as nursing students were also more likely to support new incoming nurses as they transitioned to their new workplace as described by one participant:

> *Because I was a student when I started using [___] in practice, I got a little bit more of a transition into how the system worked ... when we've new hires come in who've never used electronic charting before, we always try to go around and see patients together, and we'll say like, "Okay, you'll chart with so-and-so today, so that, you know, you can get used to the system and play around with it, and just get comfortable."* [P4, WC]

Theme 3: Suggestions for Bridging Gaps in Transitioning to Digital Health Practice

Participants shared different strategies on how to improve newly qualified nurses' practice with DHT discussing aspects related to improving the education of nursing students, and strategies that can be introduced to enhance new graduate nurses' experiences once they join the workplace.

With respect to nursing education and looking back at their own experiences as nursing students and recognizing the expanded use of technology in the health care system, participants would have appreciated if they had had adequate education and exposure to DHT during their school to help them make better sense of how DH relates to nursing practice and the clinical care of patients. One participant shared:

> *I feel like a lot of nursing school time was spent on different things that don't necessarily translate to current practice, nor does it translate to digital health information. I've never made a digital care plan in my time as a registered nurse, but we did spend a lot of time on care plans in school.* [P3, WC]

They felt nursing education could be improved by providing relevant hands-on learning that goes beyond knowing how to use the system to critically think about why a nurse is doing these actions. The participant went on to share an example of how this looks like to them:

> *...Here's this patient that you have. These are the computer systems that you're using. What are the checks that you need to do on the computer system? What do you need to look at? What information do you need to pull out of the system? How would you chart for this specific patient? "How do you interact*

XSL•FO
RenderX

> *with other disciplines within the specific computer system? ... Like that would have been very helpful.* [P3, WC]

Recognizing the changes taking place in health care with respect to technology integration, participants stressed the importance of having both dedicated and focused theoretical and clinical education on core concepts relevant to DH in undergraduate nursing education. Examples of the topics that participants discussed included the "how" and "why" of DH technology, ethical issues involving the use of DH technology, privacy, confidentiality, information security, legal implications, medical devices, and the different DHTs currently in use in hospitals and community settings, as well as newer innovations, technologies, and services that will be introduced in the future. According to one participant:

> *The reality is that tech is really taking over, and it is the future. I'm hoping that manually writing in charts is going to be a thing of the past. And so, to integrate that as early on as possible, the better, it's a reality, it's there. There's no need to leave it on the sidebar until later.* [P2, EC]

Another participant mentioned:

> *I think that having a whole course on digital health would be very interesting, because I feel the topic of artificial intelligence is a big topic, especially recently, and there's a lot of innovation happening with that ... I think it would need to be already applied in the hospitals though for them to fully absorb the information and see the practical application of what they're learning.* [P1, EC]

Participants also emphasized the importance of exposing nursing students to DHT used in clinical practice; creating opportunities for hands-on learning while in the nursing program as opposed to learning about them when opportunities arise in clinical training sites. These experiential learning experiences were viewed as essential to enhance confidence and competence. One participant shared:

> *Definitely, built-in training in school for everyone, not just if you're going to a hospital that uses an EMR. I think everybody needs it.* [P4, WC]

Another participant elaborated:

> *I think having maybe a seminar on how to use particular systems, such as MEDITECH, VAX, if these are systems that are widely used in hospitals, and then in your clinical settings, it'd be nice to kind of have that training beforehand, even just like an introduction to these types of things. Because at school, the training was mostly on the unit, for that specific clinical course.* [P3, WC]

Regarding workplace strategies, participants emphasized the need for creating opportunities for nurses' engagement, particularly younger ones, in technology-related decisions and the role of nursing leadership in this process as one participant explained:

> *It seems like a lot of time, there's not a younger generation involved in the process [technology] ... they're much led by people with experience.* [P7, WC]

Noting the expansive digitalization taking place in health care, another participant added:

> *I definitely think that there should be safety checks if you're really going to be pushing everything into Digital Health, everything needs to be looked after and evaluated or tested properly before it's implemented ... I think you can evaluate it from different ways, like patient outcomes, you could look at the nursing team, the people who are using it day-to-day, you can evaluate them, you can see where they're at with the use.* [P8, EC]

Participants also emphasized the importance of creating opportunities for professional development and ongoing learning as the field of technology continues to evolve:

> *Continuous learning ... that's something that will always be necessary when it comes to technology, based on the advancements that are made every day and hopefully will continue to be made within these units.* [P5, EC]

These strategies were viewed by participants as an indication of support and actions that convey commitment from the organizations and leaders toward enhancing new graduates' abilities to effectively adapt to the current technology-rich work environment and make full use of DHT to improve practice and patient outcomes. A participant described:

> *...that support looks like booking you in for paid training, which the managers do at the site that I'm in. Being advocates for change, and being advocates for a new system, answering concerns that people may have, and having an in-depth knowledge of the system that's coming, and not just, "Oh, this new system's rolling out, and other people will know how to use it.* [P3, WC]

Discussion

Principal Findings

This study explored newly qualified nurses' experiences with digital health in the workplace. Results identified that participants' experiences in this study were influenced by inadequate education about digital health while in their nursing school in addition to challenges they faced upon joining the workplace including variable training and support, as well as technology infrastructure. Nonetheless, participants also identified several factors that facilitated their transition including having access to a wide range of technologies to support care, being adaptable, and positive workplace culture. Inadequate theoretical and clinical education about nursing informatics and digital health during nursing school creates unnecessary challenges for the already overwhelmed newly qualified nurse [5,23]. These nurses not only have to adjust to the complex practice setting and the team culture but also learn how to use the technology independently while providing care. To some of the participants, their abilities to connect the dots and

XSL•FO
RenderX

understand what digital health means were due to events, such as technology implementation or the COVID-19 pandemic, which also made them further question their readiness for digital health.

Educating nurses about digital health and nursing informatics is an area that has received much attention in the literature. Prior research has shown that educators and nursing programs are often challenged and confused about what informatics means, how it relates to nursing practice, and the application of digital health in care delivery [31,32]. In part, this may be attributed to the availability of multiple informatics competency lists; each presenting different perspectives on what to teach students and these are also getting dated with no focus on emerging technologies such as AI [33,34]. Other research has identified an urgent need for explicit and consistent teaching about digital health (existing and emerging technologies) concepts and their integration into clinical care [18,35,36].

Although there are efforts to bridge the theory-practice gap in relation to digital health education, clinical application experiences remain inconsistent and largely influenced by the availability of technology infrastructure in the schools of nursing and the clinical placement settings [11,13,18,20,21,23,34]. Additionally, limited or lacking professional and regulatory standards to inform nursing practice in digital health is also likely contributing to nurses' limited engagement with digital health beyond what is required in a specific work setting. These patterns are further reinforced by a persistent model of basic and condensed computer proficiency training programs in the workplace, limited on-site support and opportunities for continued learning, and vague organizational policies; collectively promoting a task-based approach to working with DHT. While the pandemic has served as a catalyst for more engagement in digital health overall, it also exposed several challenges as a result of abruptly transitioning to virtual care delivery with little support and training for care providers on the technologies used and the changing context of practice [5,37]. These factors raise concerns regarding nurses' capacity to actively participate in the evolving digital health ecosystem continuously being revolutionized by more sophisticated technologies such as AI [20,36-39]; further underscoring the importance of integrated and systematic strategies to address digital health readiness in nursing.

Despite these challenges, in this study, new graduates appreciated the benefits that DHT could provide to improve their practice and patient care. Furthermore, the positive role models and the supportive unit culture enabled a smoother transition, especially for those who experienced a system-wide technology implementation. In that case, senior nurses and newly qualified nurses found themselves in the same boat, and they had to uplift each other into the process. In exchange, new graduates offered to share their digital skills to support more senior nurses who are not as comfortable with technology; thereby, creating reciprocal benefits. Although these strategies proved useful for participants in this study and these behaviors should be encouraged, organizational support should be in place to facilitate nurses' work with DHT [8,24,40].

Newly qualified nurses in this study clearly recognized that technology integration in health care is on the rise and will likely continue to shape their practice and the practice of future nurses joining the profession in the years to come. Therefore, they emphasized the importance of continuing education, mentorship by nurse leaders, investment in quality CIS, ongoing support from health care organizations, and opportunities for nurses to be more engaged in the design, implementation, and evaluation of current and future DHT as paramount strategies to ensure that nurses are better prepared to effectively lead the digital health transformation. They also emphasized the importance of formal education about DH with meaningful linkage to clinical practice applications.

Academic and clinical learning experiences about digital health are intertwined and contribute to shaping nurses' perspectives about their practice in digital health. In educating the next generation of nurses, prelicensure nursing programs should provide early, comprehensive, and focused education about DH to enable nursing students to understand the full spectrum of digital health so they are work-ready upon joining the practice setting as independent practitioners [11,17,18,23,39,41,42]. Such education should also include core concepts of digital health, technological advancements and innovation, the benefits and potential impact of these technologies on the quality and safety of patient care, and professional nursing roles [36,37,39] with clear linkages to informatics competency requirements. This could be achieved by providing relevant theoretical content in the classroom and enriching that with experiential learning experiences in the laboratory, simulation, and clinical practicum throughout the program, as these experiences have been found to be beneficial for bridging the theory-practice gap, specifically in the context of learning about main DH technologies such as electronic records [12,22,39].

To improve nurses' practice in digital health, nursing organizations must also develop clearer and more consistent policies regarding existing and emerging digital health technologies and how these fit within nurses' scope of practice according to their professional roles [8,39]. Consequently, this translates to clearer digital health expectations for schools of nursing, nursing students, and employers [36,37,39,41]. In the workplace, as opposed to proficiency-based training upon onboarding and orientation to clinical practice settings, newly qualified nurses would benefit from structured educational courses to enhance their technology skill acquisition, as well as enable them to bridge their theoretical knowledge with workplace policies and expectations [17,39,43,44]. This also should include exposure to and education about the wide range of medical devices [45] used in the clinical setting (eg, infusion pumps, monitors, and parenteral nutrition devices), having access to consistent DHT support resources, particularly during night shifts and after hours, having opportunities for continuing education in digital health, and providing support for CNEs and nurses serving in clinical preceptor roles, and to all nurses according to their educational needs [4,11,17,39,40]. Collectively, this would improve nurses' inclination, uptake, and competency in digital health technologies, as well as facilitate a healthy organizational culture that values innovation,

XSL•FO
RenderX

change, and a growth mindset; hence, creating a win-win situation.

Furthermore, policies at the professional, organizational, and health system levels directly influence nursing practice in digital health and have implications for the quality of care, patient outcomes, health care service delivery, and quality improvement [1,8,9,38,39]. Currently, there is no digital health strategy for Canada and the eHealth strategy for nurses is outdated. Creating or updating these policies would help outline the digital health vision at the health system level, clarify the professional expectations for nurses according to their scope of practice, as well as other members of the health care team, facilitate the development of supporting and comprehensive digital health infrastructures including adequate funding, education, resource allocation, and strategies that promote innovation, creativity, continuous learning, and improvement, which ultimately facilitates optimal practice with DHT to improve patient and health care system outcomes [38-40]. More research is needed to explore how digital health knowledge gaps experienced by new graduates affect their entry to practice and transition to the workplace and their overall career trajectory. Interventional studies are also needed to evaluate new graduates' actual use of digital health technologies in the workplace and the impact on patient and organizational outcomes. Future studies are also needed to examine the role of clinical preceptors and clinical educators in developing competency in digital health. Although prior research has explored the status of digital health integration across Canadian schools of nursing [20,21], more research is needed that compares curricula between nursing programs and possibly other health professions.

Strengths and Limitations

To our knowledge, this is the first study involving newly graduated nurses and their practice with DHT in the workplace within the Canadian context involving participants from 2 provinces. In addition, the use of qualitative interviews enabled the collection of rich data about participants' experiences. Given that the data collected in this study is unique to the settings included in the research, this should be considered in the interpretation of the results and the transferability of findings to other contexts.

Conclusions

Although the study included participants from EC and WC, findings revealed more similarities rather than differences between participants with respect to digital health education, technology-related challenges, and their influence on nursing practice. Key strengths that newly qualified nurses in this study had included digital capabilities, awareness, and adaptability. However, inconsistent, or inadequate education about digital health during prelicensure nursing education, a proficiency-based approach to working with technology in the workplace, and variable technological infrastructure and support resources available created significant practice-related challenges that can detract nurses from patient care and limit optimal use of DHT. Digital health is the foundation of contemporary health care; therefore, comprehensive education about digital health during nursing school and throughout professional nursing practice, as well as organizational support and policy, are critical pillars. Health systems investing in DHT must also strive to create supportive work environments for nurses to thrive in technologically rich practice settings and further develop their capacity to deliver the digital health future.

Acknowledgments

MK and AA received funding through a Social Sciences and Humanities Research Council Insight Development Grant, a portion of this funding was used to pay for research assistantship costs. Creative artificial intelligence was not used in any part of the writing or analysis of this paper.

Authors' Contributions

MK, AA, and LMN conceptualized the project idea, developed the initial draft, and contributed to the discussion and interpretation of findings. RS and DdC contributed to data collection, analysis, reviewing, and editing of the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Codes, categories and themes of interviews with new graduate nurses employed in the workplace.

[DOCX File , 18 KB - mededu_v10i1e53258_app1.docx]

References

1. Spatharou A, Hieronimus S, Jenkins J. Transforming healthcare with AI: the impact on the workforce and organizations: executive briefing. 2020. URL: https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai [accessed 2024-08-02]
2. Snowdon A. Digital health: a framework for healthcare transformation white paper. HIMSS. URL: https://www.himss.org/news/himss-defines-digital-health-global-healthcare-industry [accessed 2024-02-01]

XSL•FO
RenderX

3. Brown TMH, Bewick M. Digital health education: the need for a digitally ready workforce. Arch Dis Child Educ Pract Ed 2023;108(3):214-217 [FREE Full text] [doi: 10.1136/archdischild-2021-322022] [Medline: 35697475]
4. Wong BLH, Khurana MP, Smith RD, El-Omrani O, Pold A, Lotfi A, et al. Harnessing the digital potential of the next generation of health professionals. Hum Resour Health 2021;19(1):50 [FREE Full text] [doi: 10.1186/s12960-021-00591-2] [Medline: 33853625]
5. McMillan K, Akoo C, Catigbe-Cates A. New graduate nurses navigating entry to practice in the COVID-19 pandemic. Can J Nurs Res 2023;55(1):78-90 [FREE Full text] [doi: 10.1177/08445621221150946] [Medline: 36635915]
6. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
7. Global strategy on digital health 2020-2025.: World Health Organization; 2021. URL: https://cdn.who.int/media/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf?sfvrsn=f112ede5_75 [accessed 2024-04-09]
8. Digital health transformation and nursing practice. International Council of Nurses (ICN). 2023. URL: https://www.icn.ch/what-we-do/position-statements [accessed 2024-02-01]
9. Government of Canada. Nursing retention toolkit: improving the working lives of nurses in Canada. 2024. URL: https://www.canada.ca/en/health-canada/services/health-care-system/health-human-resources/nursing-retention-toolkit-improving-working-lives-nurses.html [accessed 2024-08-02]
10. Canadian Association of Schools of Nursing. Nursing informatics entry-to-practice competencies for registered nurses. 2012. URL: https://www.casn.ca/2014/12/casn-entry-practice-nursing-informatics-competencies/ [accessed 2024-02-01]
11. Kleib M, Nagle L, Furlong K, Paul P, Wisnesky UD, Ali S. Are future nurses ready for digital health?: Informatics competency baseline assessment. Nurse Educ 2022;47(5):E98-E104 [FREE Full text] [doi: 10.1097/NNE.0000000000001199] [Medline: 35324499]
12. Kleib M, Jackman D, Wisnesky UD, Ali S. Academic electronic health records in undergraduate nursing education: mixed methods pilot study. JMIR Nurs 2021;4(2):e26944 [FREE Full text] [doi: 10.2196/26944] [Medline: 34345797]
13. Ewertsson M, Bagga-Gupta S, Allvin R, Blomberg K. Tensions in learning professional identities—nursing students' narratives and participation in practical skills during their clinical practice: an ethnographic study. BMC Nurs 2017;16:48 [FREE Full text] [doi: 10.1186/s12912-017-0238-y] [Medline: 28824335]
14. Kleib M, Nagle L. Development of the Canadian nurse informatics competency assessment scale and evaluation of Alberta's registered nurses' self-perceived informatics competencies. Comput Inform Nurs 2018;36(7):350-358. [doi: 10.1097/CIN.0000000000000435] [Medline: 29668498]
15. Kleib M, Nagle L. Factors associated with Canadian nurses' informatics competency. Comput Inform Nurs 2018;36(8):406-415. [doi: 10.1097/CIN.0000000000000434] [Medline: 29596068]
16. van Houwelingen CTM, Ettema RGA, Kort HSM, Cate OT. Internet-generation nursing students' view of technology-based health care. J Nurs Educ 2017;56(12):717-724 [FREE Full text] [doi: 10.3928/01484834-20171120-03] [Medline: 29206261]
17. Brown J, Pope N, Bosco AM, Mason J, Morgan A. Issues affecting nurses' capability to use digital technology at work: an integrative review. J Clin Nurs 2020;29(15-16):2801-2819. [doi: 10.1111/jocn.15321] [Medline: 32416029]
18. Raghunathan K, McKenna L, Peddle M. Baseline evaluation of nursing students' informatics competency for digital health practice: a descriptive exploratory study. Digit Health 2023;9:20552076231179051 [FREE Full text] [doi: 10.1177/20552076231179051] [Medline: 37274371]
19. Warshawski S. Israeli nursing students' acceptance of information and communication technologies in clinical placements. J Prof Nurs 2020;36(6):543-550. [doi: 10.1016/j.profnurs.2020.08.005] [Medline: 33308554]
20. Nagle L, Kleib M, Furlong K. Digital health in Canadian schools of nursing part aducators' perspectives. Qual Adv Nurs Educ 2020;6(1):1-19 [FREE Full text] [doi: 10.17483/2368-6669.1229]
21. Nagle L, Kleib M, Furlong K. Digital health in Canadian schools of nursing part B: academic administrators' perspectives. Qual Adv Nurs Educ 2020;6(3):1-30 [FREE Full text] [doi: 10.17483/2368-6669.1256]
22. Mollart L, Newell R, Noble D, Geale SK, Norton C, O'Brien AP. Nursing undergraduates' perception of preparedness using patient electronic medical records in clinical practice. AJAN 2021;38(2):44-51. [doi: 10.37464/2020.382.282]
23. Shin EH, Cummings E, Ford K. A qualitative study of new graduates' readiness to use nursing informatics in acute care settings: clinical nurse educators' perspectives. Contemp Nurse 2018;54(1):64-76. [doi: 10.1080/10376178.2017.1393317] [Medline: 29037119]
24. Caton E, Philippou J, Baker E, Lee G. Exploring perceptions of digital technology and digital skills among newly registered nurses and clinical managers. Nurs Manag (Harrow) 2024;31(1):27-33. [doi: 10.7748/nm.2023.e2101] [Medline: 37752873]
25. Doyle L, McCabe C, Keogh B, Brady A, McCann M. An overview of the qualitative descriptive design within nursing research. J Res Nurs 2020;25(5):443-455 [FREE Full text] [doi: 10.1177/1744987119880234] [Medline: 34394658]
26. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. Fam Med Community Health 2019;7(2):e000057 [FREE Full text] [doi: 10.1136/fmch-2018-000057] [Medline: 32148704]
27. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
28. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14 [FREE Full text] [doi: 10.1016/j.npls.2016.01.001]

XSL•FO
RenderX

29. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
30. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
31. Bove LA, Sauer P. Nursing faculty informatics competencies. Comput Inform Nurs 2023;41(1):18-23. [doi: 10.1097/CIN.0000000000000894] [Medline: 36634233]
32. Chauvette A, Kleib M, Paul P. Developing nursing students' informatics competencies—a Canadian faculty perspective. Int J Nurs Educ Scholarsh 2022;19(1):20210165. [doi: 10.1515/ijnes-2021-0165] [Medline: 35697520]
33. Kleib M, Chauvette A, Furlong K, Nagle L, Slater L, McCloskey R. Approaches for defining and assessing nursing informatics competencies: a scoping review. JBI Evid Synth 2021;19(4):794-841. [doi: 10.11124/JBIES-20-00100] [Medline: 33625068]
34. Forman TM, Armor DA, Miller AS. A review of clinical informatics competencies in nursing to inform best practices in education and nurse faculty development. Nurs Educ Perspect 2020;41(1):E3-E7. [doi: 10.1097/01.NEP.0000000000000588] [Medline: 31860501]
35. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020;22(11):e22706 [FREE Full text] [doi: 10.2196/22706] [Medline: 33151152]
36. Booth RG, Strudwick G, McBride S, O'Connor S, López ALS. How the nursing profession should adapt for a digital future. BMJ 2021;373:n1190 [FREE Full text] [doi: 10.1136/bmj.n1190]
37. Booth RG, Strudwick G. Preparing nursing for the virtual care realities of a post-pandemic future. Nurs Leadersh (Tor Ont) 2021;34(4):86-96. [doi: 10.12927/cjnl.2021.26685] [Medline: 35039123]
38. Troncoso EL, Breads J. Best of both worlds: digital health and nursing together for healthier communities. Int Nurs Rev 2021;68(4):504-511. [doi: 10.1111/inr.12685] [Medline: 34133028]
39. Clinical practice in a digital health environment. RNAO. 2024. URL: https://rnao.ca/bpg/guidelines/clinical-practice-digital-health-environment [accessed 2024-04-04]
40. Subramanian S, Kleib M. Leveraging clinical preceptorship to enhance nursing students' readiness in digital health. Qual Adv Nurs Educ 2023;9(3):1-13 [FREE Full text] [doi: 10.17483/2368-6669.1412]
41. Wilson CB, Slade C, Wong WYA, Peacock A. Health care students experience of using digital technology in patient care: a scoping review of the literature. Nurse Educ Today 2020;95:104580. [doi: 10.1016/j.nedt.2020.104580] [Medline: 33065526]
42. Edirippulige S, Samanta M, Armfield NR. Assessment of self-perceived knowledge in e-Health among undergraduate students. Telemed J E Health 2018;24(2):139-144. [doi: 10.1089/tmj.2017.0056] [Medline: 28708457]
43. Lindfors K, Kaunonen M, Huhtala H, Paavilainen E. Newly graduated nurses' evaluation of the received orientation and their perceptions of the clinical environment: an intervention study. Scand J Caring Sci 2022;36(1):59-70. [doi: 10.1111/scs.12963] [Medline: 33522636]
44. McGarity T, Monahan L, Acker K, Pollock W. Nursing graduates' preparedness for practice: substantiating the call for competency-evaluated nursing education. Behav Sci (Basel) 2023;13(7):553 [FREE Full text] [doi: 10.3390/bs13070553] [Medline: 37504000]
45. Ewertsson M, Gustafsson M, Blomberg K, Holmström IK, Allvin R. Use of technical skills and medical devices among new registered nurses: a questionnaire study. Nurse Educ Today 2015;35(12):1169-1174. [doi: 10.1016/j.nedt.2015.05.006] [Medline: 26059922]

Abbreviations

AI: artificial intelligence
CIS: clinical information system
CNE: clinical nurse educator
DHT: digital health technology
EC: Eastern Canada
RA: research assistant
RN: registered nurse
WC: Western Canada

XSL•FO
RenderX

Edited by B Lesselroth, F Pietrantonio, M Montagna, J López Castro, I Said-Criado; submitted 01.10.23; peer-reviewed by N Pope, F Al Dhabbari, A McDaniel; comments to author 22.03.24; revised version received 01.05.24; accepted 14.07.24; published 19.08.24.

Please cite as:
Kleib M, Arnaert A, Nagle LM, Sugars R, da Costa D
Newly Qualified Canadian Nurses' Experiences With Digital Health in the Workplace: Comparative Qualitative Analysis
JMIR Med Educ 2024;10:e53258
URL: https://mededu.jmir.org/2024/1/e53258
doi:10.2196/53258
PMID:

©Manal Kleib, Antonia Arnaert, Lynn M Nagle, Rebecca Sugars, Daniel da Costa. Originally published in JMIR Medical Education (https://mededu.jmir.org), 19.08.2024. 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.

XSL•FO
RenderX

Original Paper

Design, Implementation, and Analysis of an Assessment and Accreditation Model to Evaluate a Digital Competence Framework for Health Professionals: Mixed Methods Study

Francesc Saigí-Rubió[1*], PhD; Teresa Romeu[2*], PhD; Eulàlia Hernández Encuentra[2*], PhD; Montse Guitert[2*], PhD; Erik Andrés[3*], MSc; Elisenda Reixach[3*], PhD

[1]Faculty of Health Sciences, Universitat Oberta de Catalunya, Barcelona, Spain

[2]Faculty of Psychology and Education Sciences, Universitat Oberta de Catalunya, Barcelona, Spain

[3]Fundació TIC Salut i Social, Generalitat de Catalunya, Barcelona, Spain

[*]all authors contributed equally

Corresponding Author:
Francesc Saigí-Rubió, PhD
Faculty of Health Sciences
Universitat Oberta de Catalunya
Rambla del Poblenou, 156
Barcelona, 08018
Spain
Phone: 34 933 263 622
Email: fsaigi@uoc.edu

Abstract

Background: Although digital health is essential for improving health care, its adoption remains slow due to the lack of literacy in this area. Therefore, it is crucial for health professionals to acquire digital skills and for a digital competence assessment and accreditation model to be implemented to make advances in this field.

Objective: This study had two objectives: (1) to create a specific map of digital competences for health professionals and (2) to define and test a digital competence assessment and accreditation model for health professionals.

Methods: We took an iterative mixed methods approach, which included a review of the gray literature and consultation with local experts. We used the arithmetic mean and SD in descriptive statistics, *P* values in hypothesis testing and subgroup comparisons, the greatest lower bound in test diagnosis, and the discrimination index in study instrument analysis.

Results: The assessment model designed in accordance with the competence content defined in the map of digital competences and based on scenarios had excellent internal consistency overall (greatest lower bound=0.91). Although most study participants (110/122, 90.2%) reported an intermediate self-perceived digital competence level, we found that the vast majority would not attain a level-2 Accreditation of Competence in Information and Communication Technologies.

Conclusions: Knowing the digital competence level of health professionals based on a defined competence framework should enable such professionals to be trained and updated to meet real needs in their specific professional contexts and, consequently, take full advantage of the potential of digital technologies. These results have informed the *Health Plan for Catalonia 2021-2025*, thus laying the foundations for creating and offering specific training to assess and certify the digital competence of such professionals.

(JMIR Med Educ 2024;10:e53462) doi:10.2196/53462

KEYWORDS

eHealth literacy; eHealth competencies; digital health; competencies; eHealth; health literacy; digital technology; health care professionals; health care workers

XSL•FO
RenderX

Introduction

Background

The recent COVID-19 pandemic has highlighted the importance and potential of digital health in optimizing the quality, efficiency, and safety of health care [1-4]. Despite this, the adoption of digital tools and technologies in this field has been slow [5,6], and their full implementation in clinical practice has yet to occur [7]. Research has pointed to several factors as potential barriers, including technology, infrastructure, and financial resources [8,9]. However, it is the lack of digital health literacy that most commonly obstructs the implementation of digital health services [6]. Health professionals have been identified as a key factor in the digital transformation of health care [10]. Accordingly, they must be equipped with digital health competences, ranging from basic skills (eg, computer literacy) to more complex ones such as the ability to teach patients how to use technology and digital data sources safely and appropriately. Beyond informing patients about the availability and potential benefits of these technologies, physicians guide them on integrating these tools into their health care routines, playing a pivotal role in this process. For instance, they can guide patients on using portals for test results or mobile apps for medication adherence. However, for in-depth training on specific technologies, other professionals such as nurses or technical support staff may be better suited [2,6,9,11].

In 2016, a survey of 200 health professionals conducted by the European Health Parliament's Digital Skills for Health Professionals (COMPDIG-Salut) Committee found that, in most cases, health professionals felt that they lacked the appropriate skills to cope with the digital revolution in their professional practice [12,13]. Today, there is still a need for accessible, structured, and comprehensive education that will enable future health professionals to make the best use of technology and harness its full potential in terms of quality of care [5,14,15].

Health professionals need to develop digital health competences to keep up with new technologies and ensure that they can provide high-quality patient care [2,5,7,16,17]. To this end, we must first map the specific digital competences needed in health care (to provide the right kind of digital education) [18] and then create a model for assessing and accrediting such competences. While there is a growing number of individual digital health competence frameworks and reviews that focus on specific health care professions or settings [2], there is a lack of standardization in the definition of digital health competences themselves, including discrepancies and overlap among available frameworks and their approach to categorization [7]. This implies a need to continually update the competences required in this field as well as the methods used to assess them [7].

The Professional Dialogue Forum of Catalonia (northeastern Spain), one of the most advanced regions in Europe in the use of digital health technologies [19-21], highlighted "the need to improve information and communication technology (ICT) competences to advance in the use of ICT and the design of telehealth services" as one of the 17 primary current and future challenges facing health professionals [22]. The COMPDIG-Salut project was launched in response to the identified needs, focusing on three aims: (1) defining a specific digital competence framework for health professionals, (2) creating a specific assessment and accreditation model for health professionals, and (3) designing actions to train and qualify health professionals in digital competences. Having determined the current digital competence level among Catalan health professionals [23], it is time to work toward achieving the aims of the COMPDIG-Salut project [12,24].

Objectives

This study had two objectives: (1) to create a specific map of digital competences for health professionals and (2) to define and test a digital competence assessment and accreditation model for health professionals.

Methods

Study Design

The research presented in this paper is the result of collaboration between the TIC Salut Social Foundation (Information and Communication Technologies in Health and Social Care Foundation) and the Universitat Oberta de Catalunya in Spain. As this was an observational exploratory study focusing on the analysis of digital competences in health care, a mixed qualitative and quantitative methodology was used following an iterative approach, and questionnaires were designed for data collection.

Specific Map of Digital Competences for Health Professionals

Narrative Review

A narrative review [25,26] was conducted to explore a broad range of existing digital competence frameworks in the field of health care and identify commonalities and strengths among the specific digital competences they included. We used the Google Scholar database to perform an iterative search for relevant frameworks based on a combination of search terms or keywords and appropriate Boolean operators.

Specifically, we searched for "digital competence framework" OR "digital capabilities framework" and "health professionals" within a search period spanning January 1, 2017, to October 13, 2021. The inclusion of potential frameworks of interest was based on the research team's knowledge and expertise on the topic. Only publications in English and Spanish were considered.

Inclusion and Exclusion Criteria

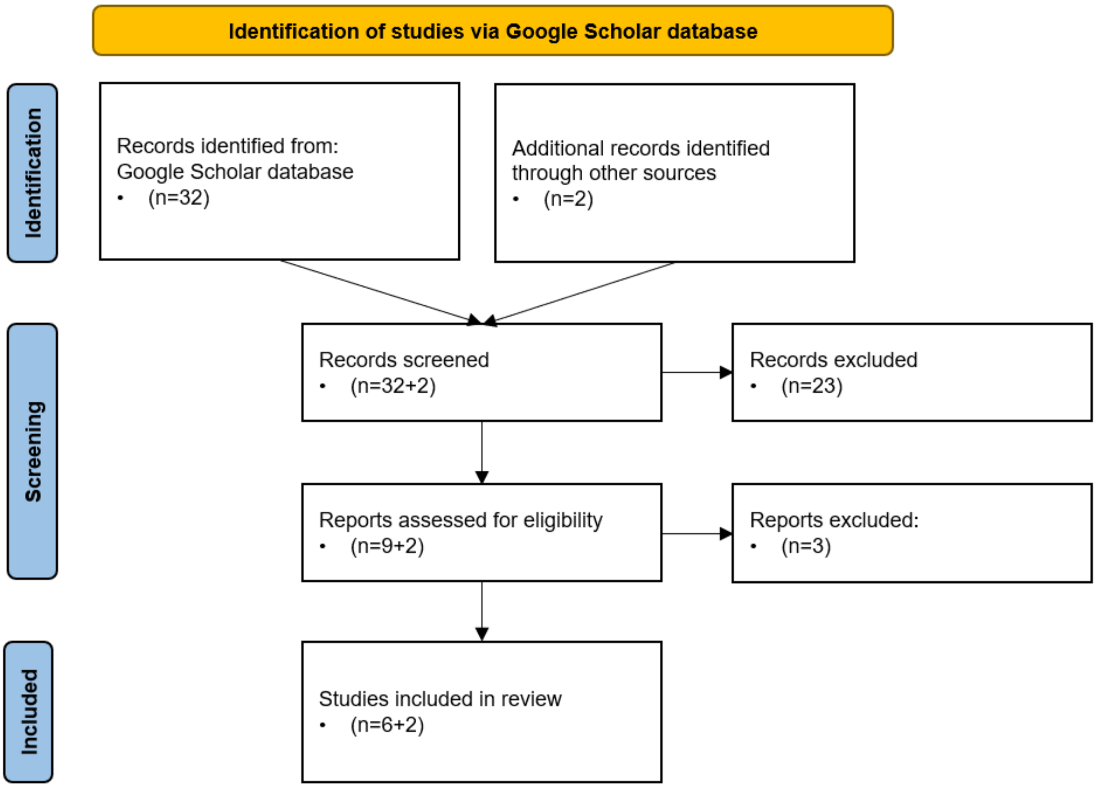
Deciding which frameworks to include in our review required careful and deliberate consideration to avoid bias and ensure valid results. To this end, we established explicit inclusion and exclusion criteria to select complete frameworks (eg, they needed to comprise digital competences specific to health care). Expert synthesis, discussion, and agreement among ≥2 reviewers were required to select frameworks for inclusion in the narrative review and ensure a consistent selection process.

To reduce selection bias and facilitate comparisons between frameworks, we ascertained the specific actions in thematic areas, the number of defined competences, the levels of achievement contained in them, and whether they distinguished

XSL•FO
RenderX

between health professions. We also looked for similarities between them and the European Digital Competence Framework for Citizens (DigComp) [27]. From the 32 results of the initial search, 9 (28%) frameworks were selected based on our inclusion and exclusion criteria. Of these 9 frameworks, 6 (67%) were selected for full-text analysis [28-33]. In total, 2 additional reference frameworks were included because of their relevance to mapping the digital competences of health professionals in Catalonia. These were the framework of digital competences for health professionals developed by the working group of challenge 4 of the Professional Dialogue Forum [22] and the Accreditation of Competence in Information and Communication Technologies (ACTIC) [34], the government of Catalonia's framework for digital competence accreditation for citizens, which is currently in line with DigComp [27]. The latter was included because our proposed framework is closely related to it (Multimedia Appendix 1). The overview flowchart is shown in Figure 1.

Figure 1. Flowchart of the literature search.



Development of the Digital Competence Framework for Health Professionals and Area Specification

To develop the Digital Competence Framework for Health Professionals, we needed to specify the digital competences required of health professionals to manage digital health effectively, critically, and responsibly. Among the various initiatives led by the European Commission to improve people's digital literacy is DigComp, which is "an umbrella or meta-framework for current frameworks, initiatives, curricula, and certifications" [28]. Using DigComp as a reference, we mapped the 6 selected frameworks, projects, and studies to establish relationships and connections between the identified keywords and the most relevant competences. We ordered the competences by similarity to reveal thematic areas and common content (Multimedia Appendix 2).

Axial coding was then applied to this content distribution to define areas, competences, and indicators [35]. Coding was performed using the ATLAS.ti software (version 22; ATLAS.ti Scientific Software Development GmbH; Multimedia Appendix 3).

These areas and competences were validated by a panel of 12 experts using a web-based questionnaire developed by the researchers. Considering the validation criteria (wording, consistency, applicability, and relevance), the experts assessed the clarity and precision of the labeling and the description of the competence areas, validated them or suggested changes, and answered open-ended questions on each item. The experts' feedback was used to refine some of the definitions (Multimedia Appendix 4).

Indicators were then defined for each of the competences to determine which aspects should be assessed for all health professionals. Indicators are characteristics that can be observed through specific tests by either predefined measures or other qualitative information. Finally, 21 professionals of various profiles from the working group of challenge 4 of the Professional Dialogue Forum validated the framework of competence areas and indicators through a web-based questionnaire (Multimedia Appendix 5).

XSL•FO
RenderX

Digital Competence Assessment and Accreditation Model for Health Professionals

Test Creation and Administration

Having developed the Digital Competence Framework for Health Professionals, the next step was to create an assessment and accreditation model. Given the variation in health professionals' roles, we could not reduce this process to a one-size-fits-all test. So, to ensure that it would be relevant to each health professional's activities and duties, 4 professional profiles were defined to account for most cases (Table 1).

Table 1. Professional profiles and their descriptions.

Profile code	Profile	Description
P1	Direct patient care	Professionals who spend >70% of their workday providing direct patient care or services (eg, physicians, nurses, occupational therapists, speech therapists, optometrists, opticians, dental hygienists, and pharmacists)
P2	Indirect patient care	Professionals who spend >70% of their workday providing health care support services (eg, physicians working in biological diagnostic and pathological anatomy services, specialist biologists, specialist physicists, specialist chemists, pharmacists, and dental prosthetists)
P3	Innovation, research, and teaching	Professionals who spend >70% of their workday providing innovation, research, or teaching services (eg, researchers and innovation specialists)
P4	Management	Professionals who spend >70% of their workday managing centers, organizations, departments, services, or teams (eg, executives and middle managers)

The test questions in our assessment and accreditation model had to be linked to definitions of observable behaviors that could be put into practice in different professional settings within the Catalan health care environment. Observable behaviors are understood as practices or actions performed by health professionals as part of their work (eg, finding clinical information in databases, communicating and collaborating remotely with teams or patients, using information management tools, and creating content). As the point of the assessment was to determine respondents' level of digital competence, we considered it appropriate to use assessment scenarios that would present them with challenges that they would need to overcome. Their attempts to deal with situations similar to those in the real world and provide the best possible digital response to the proposed challenges would provide greater insights into their competence level for each indicator. It would also allow respondents to put into practice other skills, such as problem-solving, critical thinking, and the analysis and responsible use of information and communications technology [23].

The test for the 4 professional profiles included a set of 2 cases (scenarios) with a total of 28 questions to be answered within 60 minutes. The test was contextualized for each of the 4 profiles: P1, P2, P3, and P4. The maximum score was 30 points (26 questions counted for a maximum of 1 point each, and 2 questions counted for a maximum of 2 points each). Wrong answers on the multiple-choice questions were scored negatively (−0.2 points). A minimum score of 70% (21/30) was required to pass the test. Participants who scored <21 points were categorized as "suggested level not achieved," and those who scored between 21 and 30 points were categorized as "suggested level achieved."

Implementation and Analysis of the Assessment and Accreditation Model

To analyze the proposed assessment and accreditation model for our competence framework and identify areas for improvement (eg, time allotted, number of test questions, types of test questions, professional profiles, real-life situational approach by profile, assessment scenarios, and suitability of the cases and challenges presented) and validate the proposed level test (ie, to determine whether the proposed level was appropriate), we conducted a pilot study involving a web-based test with legally recognized health professionals [36] and health care social workers employed in Catalonia who reported an intermediate or advanced self-perceived digital competence level.

The web-based test consisted of 3 activities: a level test based on the exam required to obtain the ACTIC 2—intermediate level certificate [23] (activity 1); the test developed for our proposed assessment and accreditation model, as described in the previous section (activity 2); and a feedback questionnaire to understand participants' opinions on the proposed assessment and accreditation model and other aspects of the pilot test (activity 3). The estimated time to complete the 3 activities was 90 minutes. The activities were to be done consecutively and in the specified order.

The ACTIC 2—intermediate level test (activity 1) was used to determine the participants' baseline digital competence level and compare their scores with the results of the proposed assessment and accreditation test. Participants were categorized into 1 of the following 3 groups based on their scores: beginner (0-9.9), basic (10-24.9), and intermediate (25-35).

In relation to the test developed for our proposed assessment and accreditation model (activity 2), Table 2 shows all the variables involved in the pilot study.

XSL•FO
RenderX

Table 2. Variables for which data were collected from participants during the study.

Descriptor	Collection method
Name and surname	Forms (Microsoft Corp) questionnaire
Email address	Microsoft Forms questionnaire
Health profession	Microsoft Forms questionnaire
Self-perceived digital competence level	Microsoft Forms questionnaire
Experience related to digital competence training	Microsoft Forms questionnaire
Official certification in digital competences	Microsoft Forms questionnaire
Professional profile (P1, P2, P3, or P4)	Microsoft Forms questionnaire
Score achieved in activity 1	Moodle (Moodle HQ) questionnaire
Score achieved in activity 2	Moodle (Moodle HQ) questionnaire
Feedback questionnaire	Moodle (Moodle HQ) questionnaire

The feedback questionnaire (activity 3) consisted of 9 questions, 3 (33%) of which were open-ended and 1 (11%) of which asked for the participants' self-perceived level in each of the digital competences defined for health professionals.

Data Collection

To focus the scope of the study, the Catalan Ministry of Health asked relevant professional associations to invite members whom they felt could meet the study's inclusion criteria to volunteer as participants. Volunteer participants were recruited using a Microsoft Forms (Microsoft Corp) questionnaire (Multimedia Appendix 6). After informing them of the purpose of the study, their personal and professional details were collected. If they met the inclusion criteria, they were enrolled in the study and given credentials to access Moodle (Moodle HQ) for the web-based test.

Although the study was scheduled to remain open for 30 days, from March 3 to 31, 2022, it was ultimately extended to April 14, 2022, to increase the response rate. During this period, 2 emails were sent to all candidates to remind them of the study's end date (or to inform them of the extension) and the remaining activities to be completed.

After this period, the test results were reported in accordance with "Good practice in the conduct and reporting of survey research" [31] and the General Data Protection Regulation, where applicable.

Statistical Analysis

When designing the study, we calculated the minimum sample size to ensure significant results with a 10% error rate, the maximum allowed in research studies [37]. On the basis of the results of an exploratory study [23] and the latest available report on the population of health professionals in Catalonia [22], the minimum sample size for the study was set at 168.

Descriptive statistics were performed for professionals who had correctly completed activities 1 and 2, with results presented as absolute and relative frequencies. Arithmetic means and SDs were used for comparative analysis of subgroups according to sociodemographic and professional characteristics, and *P* values were used for hypothesis testing.

The reliability of activity 2 was analyzed by measuring the consistency of its items. In addition, the level of discrimination of the items in relation to the advanced level was evaluated. The arithmetic mean and SD were used in descriptive statistics; *P* values, Bonferroni-adjusted *P* values, and Holm-adjusted *P* values were used in hypothesis testing and subgroup comparisons; and the greatest lower bound (GLB) was chosen to diagnose the test given the lack of homogeneity of the scoring scale, as was done in the exploratory study [23].

The study instrument was analyzed using the participants' scores and the discrimination index (DI) [38]. The DI measures how well an item could discriminate between high-scoring participants (ie, those with strong digital competences) and low-scoring participants in activity 2. DI values between 0 and 0.2 indicate that the item is not discriminating, and negative values imply an inverse relationship between the score on that item and the total score.

Finally, for activity 3, the numerical variables were presented as arithmetic means and SDs, whereas the categorical variables were presented as absolute and relative frequencies.

All responses were analyzed using the R statistical software (version 4.2.0; R Foundation for Statistical Computing). Responses to the open-ended questions (questions 5 and 7) in the feedback questionnaire (activity 3) were analyzed using quantitative content analysis to group them into limited categories [39].

Ethical Considerations

No ethics approval was required due to the type and nature of the study as the Catalan Department of Health is responsible for formulating the general criteria for health planning, setting the objectives, and the levels to be achieved in the topics that are included in the *Health Plan for Catalonia* [40]. All participants were informed about the study's purposes and that their participation was voluntary. Data protection treatment was informed to the participant and before accessing the survey, participants had to provide acceptance.

XSL•FO
RenderX

Results

Specific Map of Digital Competences for Health Professionals

Multimedia Appendix 7 shows the specific map of digital competences for health professionals validated by the 21 individuals (of different profiles) in the working group of challenge 4 of the Professional Dialogue Forum.

Implementation and Analysis of the Digital Competence Assessment and Accreditation Model for Health Professionals

Recruitment

During the study period, we recorded a total of 398 visits to the Microsoft Forms recruitment questionnaire. Of the 398 visitors, 377 (94.7%) agreed to participate and gave their consent for the TIC Salut Social Foundation to process their personal data. After excluding participants who did not meet the inclusion criteria and removing repeated registrations, the total recruited sample comprised 372 participants, who were given access to Moodle to start the 3 programmed activities. A total of 49.5% (184/372) of the participants logged into Moodle and started the designed activities. In the end, 176 professionals completed activity 1 correctly, of whom 122 (69.3%) also completed activity 2 correctly.

For the purposes of this study, we considered a result as valid when the test was completed, allowing for a maximum of 3 consecutive questions with no blank answers.

Sample Description

The sample consisted mainly of health professionals with a direct patient care profile (P1; 95/122, 77.9%), followed by those with a management profile (P4; 13/122, 10.7%); an innovation, research, and teaching profile (P3; 8/122, 6.6%); and an indirect patient care profile (P2; 6/122, 4.9%; Table 3).

XSL•FO
RenderX

Table 3. Descriptive statistics for participants who correctly completed activities 1 and 2 (N=122).

Descriptor	Participants, n (%)
Profile	
P1—direct patient care	95 (77.9)
P2—indirect patient care	6 (4.9)
P3—innovation, research, and teaching	8 (6.6)
P4—management	13 (10.7)
Health profession	
Specialist biologist	1 (0.8)
Dietician or nutritionist	1 (0.8)
Pharmacist	21 (17.2)
Specialist physical chemist	0 (0)
Physical therapist	17 (13.9)
Dental hygienist	7 (5.7)
Nurse	37 (30.3)
Speech therapist	7 (5.7)
Physician	4 (3.3)
Dentist	0 (0)
Optometrist/optician	6 (4.9)
Podiatrist	5 (4.1)
Dental prosthetist	1 (0.8)
Clinical or general psychologist	4 (3.3)
Occupational therapist	9 (7.4)
Health care social worker	2 (1.6)
Self-perceived digital competence level	
Intermediate	110 (90.2)
Advanced	12 (9.8)
ACTIC[a] certification	
No certification	101 (82.8)
ACTIC 1	1 (0.8)
ACTIC 2	14 (11.5)
ACTIC 3	6 (4.9)
Experience at digital health events	
No experience	107 (87.7)
Speaker or trainer	11 (9)
Organizer	2 (1.6)
Both	2 (1.6)

[a]ACTIC: Accreditation of Competence in Information and Communication Technologies.

Responses were received from professionals in all health professions, with the exception of specialist physical chemists and dentists. Nurses represented the largest proportion of the sample (37/122, 30.3%), followed by pharmacists (21/122, 17.2%) and physical therapists (17/122, 13.9%). According to the distribution of health care professions in Catalonia, a representative sample of nurses was obtained; however, this was not the case for physicians [41].

The vast majority of the sample (110/122, 90.2%) reported an intermediate self-perceived digital competence level, with only 9.8% (12/122) of participants considering themselves advanced. Overall, 82.8% (101/122) had no ACTIC certification, 11.5% (14/122) had an ACTIC 2 (intermediate) certification, 4.9%

XSL•FO
RenderX

(6/122) had an ACTIC 3 (advanced) certification, and 0.8% (1/122) had an ACTIC 1 (basic) certification.

Activity 1: ACTIC 2—Intermediate Level Test (Baseline Level)

The mean total score for activity 1 was 24.3 (SD 4.1) points, which was significantly lower (*P*=.03) than the cutoff score for "Intermediate" (25 points). Of the 122 participants, 59 (48.4%) scored between 10 and 24.9 points (basic), and 63 (51.6%) scored ≥25 points (intermediate). On average, the sample took 12.2 (SD 3.8) minutes to complete the baseline level test. There were no outliers in the times recorded (Table 4).

Table 4. Summary of activity 1 results (N=122).

Descriptor	Values
Score, mean (SD)	24.3[a] (4.1)
Score range (points), n (%)	
<10 (beginner)	0 (0)
10-24.9 (basic)	59 (48.4)
≥25 (intermediate)	63 (51.6)
Minutes taken to complete the activity, mean (SD)	12.2 (3.8)

[a]Scores significantly below 25 points (*P*=.03).

The scores followed a normal distribution, with a mode of 25 points. Only 0.8% (1/122) of the participants obtained the maximum possible score on the level test (Multimedia Appendix 8).

Subgroups were compared to determine whether there were any significant differences in the overall scores. It was found that participants who had experience at digital health events scored significantly higher than those who did not (*P*=.03).

The scores for each subgroup were also compared to determine which subgroups scored significantly below 25 points in activity 1 (ie, did not reach intermediate level) and which scored significantly above 25 points (ie, did reach intermediate level). Several subgroups were found to have scored significantly below 25 points, including P1 (direct patient care; *P*=.04), Dental hygienist (*P*=.01), Intermediate level self-perception (*P*=.03), no ACTIC certification or ACTIC 1 (*P*=.02), no experience in digital health events (*P*=.02). Some subgroups scored significantly above 25 points, particularly the group of participants who reported ACTIC 3 certification (*P*=.40). P2 (indirect patient care) professionals and professionals with experience at digital health events scored significantly above 25 points, but this difference did not remain significant after correcting for multiple testing using the Bonferroni and Holm methods (Table 5).

XSL•FO
RenderX

Table 5. Overall scores and subgroup comparisons according to participant characteristics for activity 1.

Subgroup	Overall score, mean (SD)	Group comparison[a]			Intermediate level check					
		Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value	<25 points[b]			≥25 points[c]		
					Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value	Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value
Profile		.23	>.99	.52						
P1—direct patient care (n=95)	24.1 (4.0)				*.01* [d]	*.04*	*.04*	.99	>.99	>.99
P2—indirect patient care (n=6)	27.6 (2.3)				.98	>.99	>.99	*.02*	.08	.08
P3—innovation, research, and teaching (n=8)	24.7 (4.3)				.42	>.99	>.99	.58	>.99	>.99
P4—management (n=13)	24.5 (5.2)				.37	>.99	>.99	.63	>.99	>.99
Health profession[e]		.60	>.99	.60						
Nurse (n=37)	24.39 (4.17)				.19	>.99	>.99	.81	>.99	>.99
Pharmacist (n=21)	25.31 (3.31)				.66	>.99	>.99	.33	>.99	>.99
Physical therapist (n=17)	24.24 (5.28)				.28	>.99	>.99	.72	>.99	>.99
Other (n=44)[f]	23.83 (3.93)				*.02* [d]	.25	.23	.98	>.99	>.99
Specialist biologist (n=1)	—[g]				—	—	—	—	—	—
Dietician or nutritionist (n=1)	—				—	—	—	—	—	—
Dental hygienist (n=7)	20.14 (2.56)				*.001* [h]	*.01*	*.01*	>99	>.99	>*.99*
Speech therapist (n=7)	23.43 (4.75)				.21	>.99	>.99	.79	>.99	>.99
Physician (n=4)	22.75 (5.74)				.25	>.99	>.99	.76	>.99	>.99
Optometrist/optician (n=6)	24.00 (4.12)				.29	>.99	>.99	.71	>.99	>.99
Podiatrist (n=5)	24.10 (2.38)				.22	>.99	>.99	.78	>.99	>.99
Dental prosthetist (n=1)	—				—	—	—	—	—	—
General health or clinical psychologist (n=4)	23.89 (4.23)				.32	>.99	>.99	.68	>.99	>.99
Occupational therapist (n=9)	25.28 (3.16)				.60	>.99	>.99	.40	>.99	>.99
Health care social worker (n=2)	29.00 (0.71)				—	—	—	—	—	—
Self-perceived digital competence level		.05	.26	.20						
Intermediate (n=110)	24.12 (4.18)				*.02* [d]	*.03* [d]	*.03* [d]	.99	>.99	.99
Advanced (n=12)	26.04 (2.85)				.88	>.99	.88	.12	.23	.23
ACTIC[i] certification[e]		.17	.87	.52						

XSL•FO
RenderX

Subgroup	Overall score, mean (SD)	Group comparison[a]			Intermediate level check					
		Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value	<25 points[b]			≥25 points[c]		
					Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value	Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value
No certification/ACTIC 1 (n=102)	24.00 (4.14)				*.008* [h]	*.02* [d]	*.02* [d]	.99	>.99	.99
ACTIC 2 (n=14)	25.04 (3.74)				.51	>.99	>.99	.49	>.99	.97
ACTIC 3 (n=6)	27.92 (2.27)				.99	>.99	>.99	*.01* [d]	*.04* [d]	*.04* [d]
Experience at digital health events[e]		*.006* [j]	*.03* [f]	*.03* [f]						
No experience (n=107)	24.03 (4.21)				*.01* [d]	*.02* [d]	*.02* [d]	.99	>.99	.99
Some experience[k] (n=15)	26.30 (2.50)				.97	>.99	0.97	*.03* [d]	.06	.06

[a]2-tailed *t* test for comparisons between 2 samples and 2-tailed ANOVA for comparisons among >2 samples.

[b]1-tailed *t* test with a defined overall score threshold of <25 points.

[c]1-tailed *t* test with a defined overall score threshold of ≥25 points.

[d]Significantly less or significantly more than 25 points at *P*<.05.

[e]Subgroups with too small a subsample size (n<6) were excluded from the analysis.

[f]Significant differences at *P*<.05.

[g]Subgroups with a subsample size of n=1 were excluded from the analysis.

[h]Significantly less or significantly more than 25 points at *P*<.01.

[i]ACTIC: Accreditation of Competence in Information and Communication Technologies.

[j]Significant differences at *P*<.01.

[k]Includes experience as a speaker, trainer, or organizer.

Activity 2: Proposed Assessment and Accreditation Test for the Map of Digital Competences for Health Professionals

Overview

The mean score for activity 2 was 18.5 (SD 3.7) points, which was very significantly lower (*P*<.001) than the cutoff score for passing the test (21 points). Of the 122 participants, 89 (73%) scored <21 points (did not pass), and 33 (27%) scored ≥21 points (passed). On average, the sample took 34.4 (SD 11.4) minutes to complete the test. There were no outliers in the times recorded (Table 6).

Table 6. Summary of activity 2 results (N=122).

Descriptor	Values
Score, mean (SD)	18.5[a] (3.7)
Score range (points), n (%)	
<21 (did not pass)	89 (73)
≥21 (passed)	33 (27)
Minutes taken to complete the activity, mean (SD)	34.4 (11.4)

[a]Scores significantly below 21 points (*P*<.001).

An internal consistency analysis of activity 2 for P1 showed excellent internal consistency overall (GLB=0.91). The internal consistency of activity 2 could not be determined for the remaining profiles because there were not enough participants in them (Multimedia Appendix 9).

For activity 2, subgroup comparisons revealed significant differences between health professions and between professionals with intermediate and advanced self-perceived competence levels. However, the comparisons did not remain

XSL•FO
RenderX

significant after applying the Bonferroni and Holm adjustments for multiple comparisons.

Some subgroups did not score significantly below 21 points, including P2 professionals (*P*=.83), nurses (*P*=.01), physical therapists (*P*=.22), speech therapists (*P*=.18), physicians (*P*>.99), podiatrists (*P*=.37), occupational therapists (*P*=.78), professionals with an advanced self-perceived competence level (*P*=.16), professionals with ACTIC 3 certification (*P*>.99), and professionals with experience at digital health events (*P*=.23). In some cases, the lack of significant level determination was likely due to the small size of the subgroup (Table 7).

XSL•FO
RenderX

Table 7. Overall scores and subgroup comparisons according to participant characteristics for activity 2.

Subgroup	Overall score, mean (SD)	Group comparison[a]			Intermediate level check					
		Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value	<21 points[b]			≥21 points[c]		
					Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value	Unadjusted *P* value	Bonferroni-adjusted *P* value	Holm-adjusted *P* value
Profile		.24	>.99	.35						
P1—direct patient care (n=95)	18.60 (3.71)				*<.001*[d]	*<.001*[d]	*<.001*[d]	>.99	>.99	>.99
P2—indirect patient care (n=6)	20.14 (2.37)				.21	.83	.21	.79	>.99	>.99
P3—innovation, research, and teaching (n=8)	18.51 (1.77)				*.003*[d]	*.01*[e]	*.006*[d]	>.99	>.99	>.99
P4—management (n=13)	16.74 (4.43)				*.002*[d]	*.008*[d]	*.006*[d]	>.99	>.99	>.99
Health profession[f]		*.03*[g]	.14	.11						
Nurse (n=37)	19.59 (3.59)				*.01*[e]	.11	.07	.99	>.99	>.99
Pharmacist (n=21)	18.59 (3.07)				*<.001*[d]	*.01*[e]	*.007*[d]	>.99	>.99	>.99
Physical therapist (n=17)	19.16 (3.47)				*.02*[e]	.22	.10	.98	>.99	>.99
Other (n=44)[g]	17.28 (3.87)				*<.001*[d]	*<.001*[d]	*<.001*[d]	>.99	>.99	>.99
Specialist biologist (n=1)	—[h]				—	—	—	—	—	—
Dietician or nutritionist (n=1)	—				—	—	—	—	—	—
Dental hygienist (n=7)	12.47 (1.70)				*<.001*[d]	*<.001*[d]	*<.001*[d]	>.99	>.99	>.99
Speech therapist (n=7)	16.87 (3.92)				*.02*[e]	.18	.10	.98	>.99	>.99
Physician (n=4)	19.52 (3.77)				.24	>.99	.32	.76	>.99	>.99
Optometrist/optician (n=6)	15.76 (1.80)				*<.001*[d]	*.005*[d]	*.004*[d]	>.99	>.99	>.99
Podiatrist (n=5)	15.99 (4.49)				*.03*[f]	.37	.14	.97	>.99	>.99
Dental prosthetist (n=1)	—				—	—	—	—	—	—
General health or clinical psychologist (n=4)	19.13 (3.11)				.16	>.99	.32	.84	>.99	>.99
Occupational therapist (n=9)	19.51 (2.73)				.07	.78	.21	.93	>.99	>.99

XSL•FO
RenderX

Subgroup	Overall score, mean (SD)	Group comparison[a]			Intermediate level check					
		Unadjusted P value	Bonferroni-adjusted P value	Holm-adjusted P value	<21 points[b]			≥21 points[c]		
					Unadjusted P value	Bonferroni-adjusted P value	Holm-adjusted P value	Unadjusted P value	Bonferroni-adjusted P value	Holm-adjusted P value
Health care social worker (n=2)	21.67 (0.23)				—	—	—	—	—	—
Self-perceived digital competence level		*.02*[g]	.09	.09						
Intermediate (n=110)	18.29 (3.78)				*<.001*[d]	*<.001*[d]	*<.001*[d]	>.99	>.99	>.99
Advanced (n=12)	20.1 (2.06)				.08	.16	.08	.92	>.99	>.99
ACTIC[i] certification[f]		.12	.59	.35						
No certification/ACTIC 1 (n=102)	18.17 (3.82)				*<.001*[d]	*<.001*[d]	*<.001*[d]	>.99	>.99	>.99
ACTIC 2 (n=14)	19.31 (2.35)				*.008*[d]	*.02*[e]	*.02*[e]	.99	>.99	>.99
ACTIC 3 (n=6)	21.55 (1.70)				.77	>.99	.77	.23	.70	.70
Experience at digital health events[f]		.13	.66	.35						
No experience (n=107)	18.28 (3.67)				*<.001*[d]	*<.001*[d]	*<.001*[d]	>.99	>.99	>.99
Some experience[j] (n=15)	19.83 (3.56)				.11	.23	.11	.89	>.99	>.99

[a]2-tailed *t* test for comparisons between 2 samples and 2-tailed ANOVA for comparisons among >2 samples.

[b]1-tailed *t* test with a defined overall score threshold of <21 points.

[c]1-tailed *t* test with a defined overall score threshold of ≥21 points.

[d]Very significantly less than 21 points at *P*<.01.

[e]Significantly less than 21 points at *P*<.05.

[f]Subgroups with too small a subsample size (n<6) were excluded from the analysis.

[g]Significant differences at *P*<.05.

[h]Subgroups with a subsample size of n=1 were excluded from the analysis.

[i]ACTIC: Accreditation of Competence in Information and Communication Technologies.

[j]Includes experience as a speaker, trainer, or organizer.

Item Dimension (Instrument Diagnosis)

It should be noted that we performed this analysis only for activity 2 and P1 (direct patient care) as this subgroup was large enough for this type of analysis (n>40; Multimedia Appendix 10).

Correlation Between Participants' Performance in Activity 1 and Activity 2

Regarding the correlation between participants' performance in activity 1 and activity 2, there were 2 main groups: those who did not pass either activity 1 or activity 2 (49/122, 40.2%) and those who passed activity 1 but did not pass activity 2 (40/122, 32.8%) (Table 8).

XSL•FO
RenderX

Table 8. Participant classification according to performance in activity 1 and activity 2 (N=122).

Descriptor	Participants, n (%)
Passed activity 1 and passed activity 2	23 (18.9)
Passed activity 1 and did not pass activity 2	40 (32.8)
Did not pass activity 1 and passed activity 2	10 (8.2)
Did not pass activity 1 and did not pass activity 2	49 (40.2)

Of the 33 participants who passed activity 2, a total of 10 (30%) did not pass activity 1. Although these are isolated cases, they call into question the validity of the advanced level of activity 2 and need to be investigated further. No significant differences were found between this subgroup and the remaining professionals in the time taken to complete or in the number of blank answers in activity 1.

Activity 3: Feedback Questionnaire

A total of 94.3% (115/122) of the participants completed activity 3, although not all of them answered every question.

There were no significant differences in the ratings by profile (P1, P2, P3, and P4) or by health profession that would allow for robust conclusions to be drawn (Table 9).

Table 9. Results of the feedback questionnaire (N=115).

Question number and topic	Result
Question 1—level of difficulty of activity 2, mean (SD)	3.6 (0.7)
Question 2—self-perceived level in activity 2, n (%)	
Basic	7 (6.1)
Intermediate	49 (42.6)
Advanced	59 (51.3)
Question 3—wording of the questions, mean (SD)	3.8 (0.9)
Question 4—appropriateness of the challenges, mean (SD)	3.2 (1.2)
Question 5—feedback on the challenges and scenarios	Open-ended question
Question 6—appropriateness of the profiles, mean (SD)	4.2 (0.8)
Question 7—feedback on the profiles	Open-ended question
Question 8—self-perceived level in each of the digital competences defined for health professionals	Multiple-choice question
Question 9—general feedback on the pilot study and the initiative	Open-ended question

When asked whether they would change any of the challenges or scenarios presented in activity 2 as part of our assessment and accreditation model (question 5), a considerable number of professionals (45/94, 48%) indicated that they did not fully identify with them (Multimedia Appendix 11).

When the responses were analyzed by health profession, we found that nurses and physicians generally identified with the challenges more than their counterparts in other health professions did (Multimedia Appendix 12). While 48% (11/23) of nurses and physicians gave positive feedback in this regard, only 31% (15/48) of respondents from other professions said that they adequately identified with the challenges and scenarios.

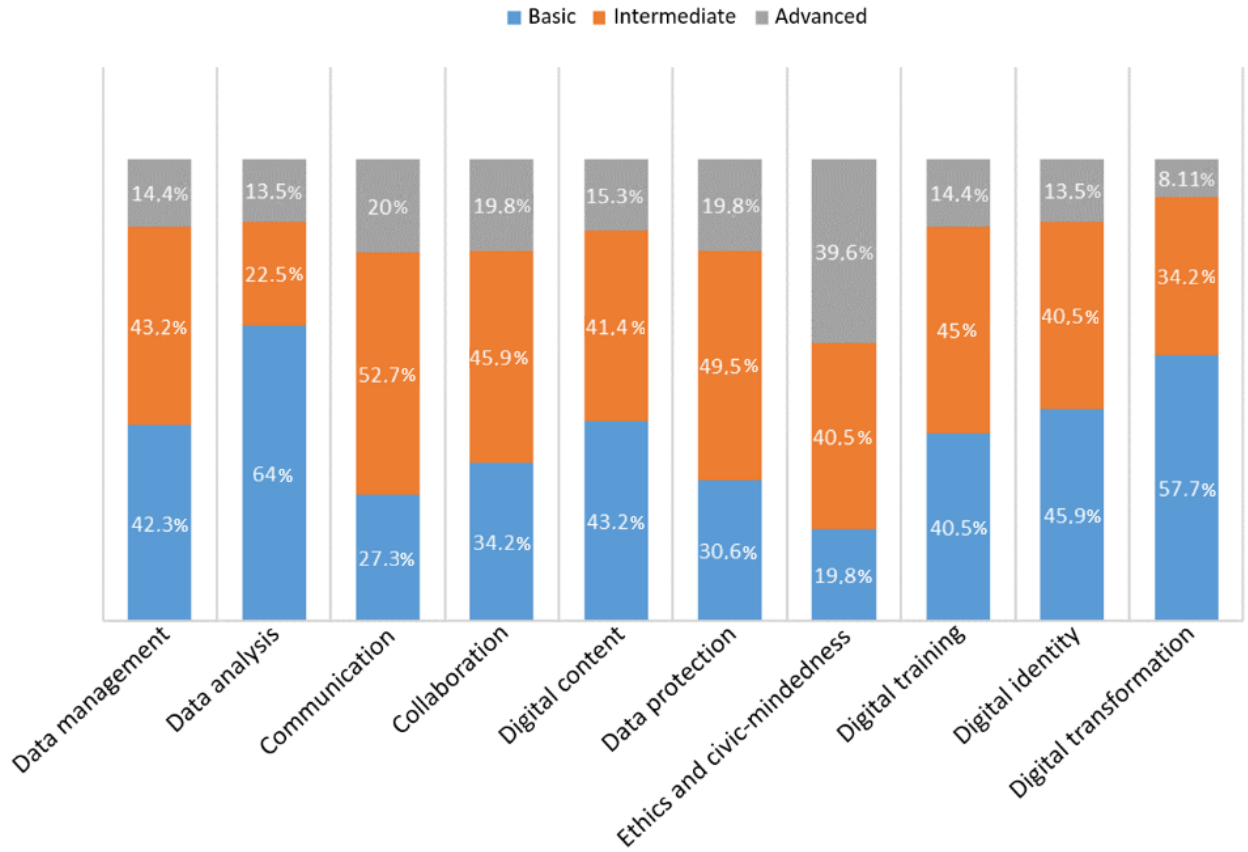
When asked whether they would change any of the profiles in our model (question 7), most participants (49/79, 62%) agreed that the profiles were accurately defined. Of the participants who did not fully agree, some questioned the workday percentage defined for categorization (eg, 70% direct patient care for P1). Others felt that there were hybrid profiles halfway between P1 and P4 or other profiles altogether or indeed that some profiles should be included in others (Multimedia Appendix 13).

It should be noted that the participants who indicated that they did not fully agree with the profiles were either from the private sector or working in pharmaceutical companies.

Regarding the participants' self-perceived level in each of the 10 digital competences defined for health professionals in our framework (question 8), most reported having a basic competence level of data analysis and digital transformation (71/111, 64% and 64/111, 58%, respectively). The competence for which most professionals reported an advanced level was ethics and civic-mindedness (44/111, 40%; Figure 2).

XSL•FO
RenderX

Figure 2. Participants' self-perceived level in each of the digital competences defined for health professionals.



Discussion

Frameworks and Competence Assessment Challenges

Digital health offers a valuable opportunity to make strides toward attaining the United Nations' Sustainable Development Goal of universal health coverage by 2030 [42]. To that end, and given the gradual digitization of the health care sector, health professionals must have the appropriate competences regardless of their specific disciplines [2,5,16,17,43].

Furthermore, the constant evolution of new technologies requires the continuous updating of digital health competence. This entails reviewing both the relevant competences and the methods for properly assessing them [2]. Some discrepancies and overlap still exist among available frameworks, the methods used to conceptualize such frameworks, and the competences they include [2] and also among the health professionals at whom they are aimed [44,45]. The lack of a comprehensive framework applicable to all health professionals warrants the development of frameworks that include different health professionals.

Among the most recent frameworks is the Health Information Technology Competencies. Developed using an iterative method, it is the most complete framework aimed at a broad range of health professionals and medical specialties [30]. In Spain, and more specifically in the Basque Country (northern Spain), the Ikanos project was developed (2020) based on the European DigComp framework as a benchmark for the description of digital competences. Drawing on interviews with experts, the Ikanos framework focuses on profiles based initially on primary care professionals (medical and nursing staff) [29].

The COMPDIG-Salut project was launched for the purposes of designing a specific map of digital competences for health professionals and creating a specific digital competence assessment and accreditation model to enable health professionals to obtain a specific accreditation certificate. Taking an iterative approach, which included a review of the gray literature and consultation with local experts [2], we designed a map of digital competences, consisting of 10 competences under 4 competence areas: data access, management, and analysis; communication and collaboration; digital awareness; and professional development. The second and fourth of these areas have the highest number of competences (3 each). In short, the map includes competences related to essential skills for the digital management and analysis of health and social data and information and includes emerging technological innovation (ie, health apps, artificial intelligence, and autonomous decision support systems) [2,46]. Moreover, and given that digital health entails new methods of communication (teleconsultations and email), it is imperative for health professionals to know how to convey information in a precise, effective, and timely manner to patients, health professionals, and any collaborating parties [47-49]. Effective communication, collaboration, and teamwork are crucial in a health setting, so health professionals must also possess the ability to work effectively as members of interdisciplinary teams [2,50]. The creation and publication of health-related digital content is another basic competence that health professionals should have to ensure high-quality health

XSL•FO
RenderX

care provision and help improve the communities they serve [31,51]. Finally, emphasis is placed on the importance of health professionals complying with the ethical principles and security criteria associated with the appropriate use of digital technologies [52] and also keeping abreast of the regulations on health and social data and information privacy, confidentiality, and protection [2,53,54]. However, we should bear in mind that the evolving nature of digital health and professional practice in the health field—which often outpaces research—requires constant, flexible development of professional competences [55]. Hence, some of the competences in the competence framework defined in the COMPDIG-Salut project may need to be adjusted to incorporate future digital trends, such as addressing challenges in the areas of cybersecurity and the use of artificial intelligence and robotics in professional practice. This implies that the methods for accurately assessing such competences may need to be updated.

Digital Competence Assessment Model

Meanwhile, given the need to assess health professionals' digital competence levels for the purposes of narrowing the digital gap and, by so doing, maintaining health service quality [11,45], a digital competence assessment model for health professionals was developed as part of the COMPDIG-Salut project. Drawing on the competence content defined in the project, the model was designed in accordance with ACTIC's new scenario-based assessment model [23]. In this sense, the assessment model was adapted to the 4 distinct health sector profiles: professionals providing direct patient care; professionals providing indirect patient care; professionals providing innovation, research, or teaching services; and professionals who manage. This not only reflects the impact of the professional role on digital competence [56] but also, with the differentiation of the 4 profiles, facilitates the design of a specific training pathway for each role (Multimedia Appendix 14).

While the focus was on the type of health professional profile (P1, P2, P3, or P4), the most frequent professional categories into which participants fell were nurses, pharmacists, and physiotherapists. The internal consistency of the digital competence assessment model for health professionals in P1 was excellent overall, bearing in mind that it was a proposed assessment model (GLB=0.91). We found that 54% (15/28) of the questions discriminated between health professionals with an advanced level of digital competence and those with a lower level (DI≥0.20) and that 11% (3/28) of the questions were highly discriminating (DI>0.4). However, 7% (2/28) of the items had a negative DI, so these would need to be reviewed because a negative DI implies an inverse relationship between the level and the correct answer, meaning that the wording of the question may not have been sufficiently precise. However, as these values were close to 0, they could be regrouped with the nondifferentiating questions (DI=0-0.2). As the assessment model for the remaining profiles assessed the same competences as those in the P1 test, we expected the internal consistency and the degree of discrimination of the questions to be similar among the profiles. However, it was not possible to directly extrapolate the P1 results.

Most of the questions in the feedback questionnaire (activity 3) could not be robustly interpreted because they did not constitute a standard instrument that had been validated by another study. The values obtained in the Likert-type questions could not be robustly interpreted either because the values were not near the ends of the scale. Many health professionals felt that their roles and challenges were not adequately represented in the assessment model despite the professional profiles being well defined. This was especially the case among health professionals who made up the smallest groups (mostly nonmedical and nonnursing staff and those working in the private or pharmaceutical sectors). This result underscores the necessity of refining the assessment model to meet the diverse digital competence needs of health care professionals, emphasizing the potential requirement for more targeted measures tailored to specific groups, and acknowledging their potential cultural dependencies. Therefore, it would be necessary to either review the profile proposal for the competence assessment model or add scenarios and challenges to it for health professionals who were not reflected in the proposed situations. The sample size of some subgroups (physicians, health professionals with advanced digital competence, P2, P3, and P4) may have been too small to represent the population.

Self-Perceived Digital Competence and Certification

Although most study participants (110/122, 90.2%) reported an intermediate self-perceived digital competence level and only 9.8% (12/122) of participants considered themselves advanced, we found that the vast majority would not attain ACTIC 2 (intermediate) certification [34]. This situation was similar to the one found in an exploratory study conducted in 2021 [23]. Consistent with the low score for the ACTIC 2 certification (activity 1), the total mean score for activity 2, which assessed the defined competence framework, was significantly lower than the defined threshold. In fact, 73% (89/122) of the sample did not pass the tests. Participants with ACTIC 3 certification significantly passed activity 1 but not activity 2. In fact, none of the subgroups studied showed a significant increase in activity 2. This shows that the assessment model is a useful tool as an approach to assessing the competence framework for health professionals, and the results strengthen those obtained in the exploratory study [23]. This also shows that health professionals still need training, that such training goes beyond that required for the main tools currently used, and that merely being exposed to digital media as consumers is insufficient in terms of acquiring the necessary digital skills. More and more faculties of medicine are incorporating digital health training into their programs of study [57,58], and there is an ever-increasing number of initiatives to meet that need [2], such as the Digital Capability Framework developed by Health Education England [59] or the free introduction to eHealth web-based course offered by the EU*US eHealth Work project [60]. The designed map of digital competences could serve as the basis of the knowledge and skills for digital literacy specific to health professionals, to which other competences or areas could be added in accordance with the particularities of the various health professions and medical specialties. This map of digital competences could also inform and contribute to the development of future digital health

XSL•FO
RenderX

training programs for health workers at different stages of their professional careers.

Strengths and Limitations

Our study provides a competence framework and a tool for assessing the digital competences of health professionals that is consistent with the certification available to citizens (ACTIC). It also yields results that strengthen those obtained in the previous exploratory study [23]. However, this study has several limitations. First, several weaknesses in the interpretation of findings from the review should be considered. While narrative reviews are often used in social science research for educational purposes, they may be biased and lack objectivity. That said, they can provide a unique perspective and identify knowledge gaps in the literature [61]. Although a considerable number of reference frameworks from gray literature were included, some may have been overlooked. Moreover, while performing thematic analysis, we found that some frameworks had either vague competence categories or overlaps among categories, which engendered differences of opinion during the classification process. However, the 4 reviewers drew on their experience to develop and clearly define the competence areas and assign the corresponding competences to them, first independently and then jointly through discussion-based agreement to reduce bias and classify the information in the most appropriate way possible. Although the reviewers tried to make the classification process as transparent and replicable as possible, it should be borne in mind that there could be other ways of interpreting and classifying the information and that the categorization might differ in the future with new advances in digital health. Second, while valid results were obtained for 122 participants, the minimum sample size set for the study was not reached (N=168). The platform used for the test (Moodle), the number of activities (3), and the time allotted (estimated at 1.5 hours) were almost certainly the reasons for poor participation in the study, especially given the high workload of the professional group in question. A study assessing both perceived and demonstrated eHealth literacy through a computer simulation of health-related internet tasks also revealed that evaluating demonstrated eHealth literacy via simulations is a challenging endeavor in terms of time [62]. When comparing the results of this study (using Moodle for the test) to those of the previous exploratory study (using Microsoft Forms for the test), it could be said that Moodle might have had an influence on a health professional's decision not to take part in the study [23]. Third, the most highly represented health professionals in this study were those providing direct patient care (P1). Thus, many of the conclusions drawn from the study are limited to that profile. The poor participation of physicians (4/122, 3.3%) meant that they were underrepresented if compared to the 2017 professional population data for Catalonia [22]. As it fell to professional associations to inform their members about this study, the dissemination mechanisms they used are not precisely known. However, their impact was seemingly greater in the associations of nurses, pharmacists, and physiotherapists than in the associations of other health professions. Fourth, the absence of a detailed demographic analysis of the participant population limits our ability to provide a comprehensive understanding of sample characteristics. While demographic characterization could aid in contextualizing the results and their relationship to the simulation performance, such data were not collected in this study in accordance with the General Data Protection Regulation principle of "data minimization" [63]. Consequently, the lack of this analysis may restrict our ability to generalize findings to broader populations and fully comprehend the influence of demographic variables on study outcomes. Fifth, our findings enable us to estimate the prevalence of P1 as the most common profile; however, they do not provide a basis for drawing more specific conclusions. The 2022 report on health care professionals in Catalonia lacks profession categorization [22], and acquiring accurate and current data poses challenges attributable to factors such as professional mobility (eg, change of workplace, the coexistence of public and private sectors, and mobility between various institutions) and liberal professions. Subsequent research endeavors should prioritize the inclusion of all 4 profiles for a more comprehensive understanding. Sixth, the overrepresentation of health professionals with an intermediate self-perceived competence level compared to those considering themselves advanced also limited some of the study conclusions. We will probably need to review our definitions of the different self-perception levels because, perhaps due to the wording used, there was a greater perception of exigency than there really was (eg, advanced user: "I have attained the most advanced digital competences for transforming and innovating in today's digital society."). Moreover, the evaluation of digital health competences should take into account that, despite being notable, there is a moderate correlation between perceived and actual digital competences, as other studies looking at simulation scenarios in health care have shown [62,64]. Despite the considerable similarity between the latter simulation scenarios and those used in this study, previous research has focused on distinct demographic groups within the general population, such as individuals aged ≥50 years [62], or specific cohorts, such as patients with chronic illnesses [64], rather than health care professionals.

Conclusions

Assessing the digital competence level of health professionals based on a defined competence framework should enable such professionals to be trained and updated to meet real needs in their specific professional contexts and, consequently, take full advantage of the potential of digital technologies [65]. The information and data gathered, together with the results of an exploratory study on competence levels conducted in 2021 [23], should be taken as the starting point for promoting relevant strategic policies and actions to ensure that the right resources and conditions are in place for good professional performance [15,66-69]. Faced with the need to improve the digital competence of health professionals working in Catalonia, these results have informed the *Health Plan for Catalonia 2021-2025* [40] and lay the foundations for designing and delivering specific training to first assess and then certify the digital competence of such professionals. Thus, the assessment model presented in this paper—designed in accordance with the competence content defined in the map of digital competences and based on scenarios—has the potential to be applied in diverse countries and languages with appropriate modifications

XSL•FO
RenderX

to meet the specific needs and contexts of health care professionals. This might provide a preliminary perspective for assessing the competence levels and needs of health professionals in different health systems, although further evidence is needed to fully support this claim.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Reference frameworks selected for the analysis and comparative study.
[DOCX File , 18 KB - mededu_v10i1e53462_app1.docx]

Multimedia Appendix 2
Ordering of competences by thematic area for each framework.
[DOCX File , 26 KB - mededu_v10i1e53462_app2.docx]

Multimedia Appendix 3
Axial coding of content to define areas, competences, and indicators.
[PNG File , 252 KB - mededu_v10i1e53462_app3.png]

Multimedia Appendix 4
Web-based questionnaire 1.
[PDF File (Adobe PDF File), 263 KB - mededu_v10i1e53462_app4.pdf]

Multimedia Appendix 5
Web-based questionnaire 2.
[PDF File (Adobe PDF File), 306 KB - mededu_v10i1e53462_app5.pdf]

Multimedia Appendix 6
Information sheet and recruitment questionnaire.
[PDF File (Adobe PDF File), 157 KB - mededu_v10i1e53462_app6.pdf]

Multimedia Appendix 7
Specific map of digital competences for health professionals.
[DOCX File , 21 KB - mededu_v10i1e53462_app7.docx]

Multimedia Appendix 8
Overall distribution of scores on Activity 1.
[DOCX File , 22 KB - mededu_v10i1e53462_app8.docx]

Multimedia Appendix 9
Overall distribution of scores in the proposed assessment and accreditation test for health professionals.
[DOCX File , 21 KB - mededu_v10i1e53462_app9.docx]

Multimedia Appendix 10
Discrimination indexes of the instrument items for profile P1.
[DOCX File , 17 KB - mededu_v10i1e53462_app10.docx]

Multimedia Appendix 11
Question 5: “Feedback on the profile–specific challenges and scenarios.”.
[DOCX File , 13 KB - mededu_v10i1e53462_app11.docx]

Multimedia Appendix 12
Question 5: “Feedback on the profile–specific challenges and scenarios by profession.”.
[DOCX File , 14 KB - mededu_v10i1e53462_app12.docx]

XSL•FO
RenderX

Multimedia Appendix 13
Question 7: “Feedback on the profiles.”.
[DOCX File , 13 KB - mededu_v10i1e53462_app13.docx]

Multimedia Appendix 14
Example scenario 1.
[PDF File (Adobe PDF File), 397 KB - mededu_v10i1e53462_app14.pdf]

References

1. Armaignac DL, Saxena A, Rubens M, Valle CA, Williams LM, Veledar E, et al. Impact of telemedicine on mortality, length of stay, and cost among patients in progressive care units: experience from a large healthcare system. Crit Care Med 2018 May;46(5):728-735 [FREE Full text] [doi: 10.1097/CCM.0000000000002994] [Medline: 29384782]
2. Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. J Med Internet Res 2020 Nov 05;22(11):e22706 [FREE Full text] [doi: 10.2196/22706] [Medline: 33151152]
3. Koehler F, Koehler K, Deckwart O, Prescher S, Wegscheider K, Kirwan B, et al. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. Lancet 2018 Sep 22;392(10152):1047-1057. [doi: 10.1016/S0140-6736(18)31880-4] [Medline: 30153985]
4. Mian A, Khan S. Medical education during pandemics: a UK perspective. BMC Med 2020 Apr 09;18(1):100 [FREE Full text] [doi: 10.1186/s12916-020-01577-y] [Medline: 32268900]
5. Jimenez G, Spinazze P, Matchar D, Koh Choon Huat G, van der Kleij RM, Chavannes NH, et al. Digital health competencies for primary healthcare professionals: a scoping review. Int J Med Inform 2020 Nov;143:104260. [doi: 10.1016/j.ijmedinf.2020.104260] [Medline: 32919345]
6. Schreiweis B, Pobiruchin M, Strotbaum V, Suleder J, Wiesner M, Bergh B. Barriers and facilitators to the implementation of eHealth services: systematic literature analysis. J Med Internet Res 2019 Nov 22;21(11):e14197 [FREE Full text] [doi: 10.2196/14197] [Medline: 31755869]
7. Longhini J, Rossettini G, Palese A. Correction: digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Nov 29;24(11):e43721 [FREE Full text] [doi: 10.2196/43721] [Medline: 36446087]
8. Saigí-Rubió F, Borges do Nascimento IJ, Robles N, Ivanovska K, Katz C, Azzopardi-Muscat N, et al. The current status of telemedicine technology use across the World Health Organization European region: an overview of systematic reviews. J Med Internet Res 2022 Oct 27;24(10):e40877 [FREE Full text] [doi: 10.2196/40877] [Medline: 36301602]
9. Brown J, Pope N, Bosco AM, Mason J, Morgan A. Issues affecting nurses' capability to use digital technology at work: an integrative review. J Clin Nurs 2020 Aug;29(15-16):2801-2819. [doi: 10.1111/jocn.15321] [Medline: 32416029]
10. Burmann A, Tischler M, Faßbach M, Schneitler S, Meister S. The role of physicians in digitalizing health care provision: web-based survey study. JMIR Med Inform 2021 Nov 11;9(11):e31527 [FREE Full text] [doi: 10.2196/31527] [Medline: 34545813]
11. Shiferaw KB, Tilahun BC, Endehabtu BF. Healthcare providers' digital competency: a cross-sectional survey in a low-income country setting. BMC Health Serv Res 2020 Nov 09;20(1):1021 [FREE Full text] [doi: 10.1186/s12913-020-05848-5] [Medline: 33168002]
12. Digital skills for health professionals. European Health Parliament. URL: https://www.healthparliament.eu/digital-skills-health-professionals/ [accessed 2022-07-22]
13. European year of skills 2023. European Commission. URL: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-year-skills-2023_en [accessed 2023-07-15]
14. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://apps.who.int/iris/handle/10665/344249 [accessed 2024-04-29]
15. Brunner M, McGregor D, Keep M, Janssen A, Spallek H, Quinn D, et al. An eHealth capabilities framework for graduates and health professionals: mixed-methods study. J Med Internet Res 2018 May 15;20(5):e10229 [FREE Full text] [doi: 10.2196/10229] [Medline: 29764794]
16. Improving digital literacy. The Royal College of Nursing & Health Education England. 2017. URL: https://www.rcn.org.uk/Professional-Development/publications/pub-006129 [accessed 2023-05-12]
17. Heath S. Digital health literacy: why it’s important and how to improve it. TechTarget. URL: https://www.techtarget.com/patientengagement/feature/Digital-Health-Literacy-Why-Its-Important-and-How-to-Improve-It [accessed 2023-05-25]
18. Milner RJ, Gusic ME, Thorndyke LE. Perspective: toward a competency framework for faculty. Acad Med 2011;86(10):1204-1210. [doi: 10.1097/acm.0b013e31822bd524]
19. Saigí-Rubió F, Vidal-Alaball J, Torrent-Sellens J, Jiménez-Zarco A, López Segui F, Carrasco Hernandez M, et al. Determinants of Catalan public primary care professionals' intention to use digital clinical consultations (eConsulta) in the post-COVID-19 context: optical illusion or permanent transformation? J Med Internet Res 2021 May 31;23(6):e28944 [FREE Full text] [doi: 10.2196/28944] [Medline: 34097638]

XSL•FO
RenderX

20. Pérez Sust P, Solans O, Fajardo JC, Medina Peralta M, Rodenas P, Gabaldà J, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR Public Health Surveill 2020 May 04;6(2):e19106 [FREE Full text] [doi: 10.2196/19106] [Medline: 32339998]
21. Temprana-Salvador J, López-García P, Castellví Vives J, de Haro L, Ballesta E, Rojas Abusleme M, et al. DigiPatICS: digital pathology transformation of the Catalan Health Institute Network of 8 hospitals-planification, implementation, and preliminary results. Diagnostics (Basel) 2022 Mar 30;12(4):852 [FREE Full text] [doi: 10.3390/diagnostics12040852] [Medline: 35453900]
22. Reunió plenària I Fòrum de diàleg professional. Departament de Salut. URL: http://salutweb.gencat.cat/ca/el_departament/eixos-xiv-legislatura/forum-dialeg-professional/etapes/treballs-fase-anterior/i-forum-diagnostic/ [accessed 2022-07-27]
23. Reixach E, Andrés E, Sallent Ribes J, Gea-Sánchez M, Àvila López A, Cruañas B, et al. Measuring the digital skills of Catalan Health Care professionals as a key step toward a strategic training plan: digital competence test validation study. J Med Internet Res 2022 Nov 30;24(11):e38347 [FREE Full text] [doi: 10.2196/38347] [Medline: 36449330]
24. Karnoe A, Furstrand D, Christensen KB, Norgaard O, Kayser L. Assessing competencies needed to engage with digital health services: development of the eHealth literacy assessment toolkit. J Med Internet Res 2018 May 10;20(5):e178 [FREE Full text] [doi: 10.2196/jmir.8347] [Medline: 29748163]
25. Ferrari R. Writing narrative style literature reviews. Med Write 2015 Dec 23;24(4):230-235. [doi: 10.1179/2047480615z.000000000329]
26. Greenhalgh T, Thorne S, Malterud K. Time to challenge the spurious hierarchy of systematic over narrative reviews? Eur J Clin Invest 2018 Jun 16;48(6):e12931 [FREE Full text] [doi: 10.1111/eci.12931] [Medline: 29578574]
27. The digital competence framework. European Commission. URL: https://joint-research-centre.ec.europa.eu/digcomp/digital-competence-framework_en [accessed 2022-07-26]
28. Montero Delgado JA, Merino Alonso FJ, Monte Boquet E, Ávila de Tomás JF, Cepeda Díez JM. Competencias digitales clave de los profesionales sanitarios. Educación Médica 2020 Sep;21(5):338-344. [doi: 10.1016/j.edumed.2019.02.010]
29. Test Ikanos de competencias digitales. Ikanos. URL: https://test.ikanos.eus/ [accessed 2022-07-26]
30. Good practices on DC frameworks. TicSalut Social Foundation. URL: https://ticsalutsocial.cat/wp-content/uploads/2024/07/D1.-Good-practices-on-DC-frameworks.pdf [accessed 2023-05-26]
31. Jahnel T, Pan CC, Pedros Barnils N, Muellmann S, Freye M, Dassow HH, et al. Developing and evaluating digital public health interventions using the digital public health framework DigiPHrame: a framework development study. J Med Internet Res 2024 Sep 12;26:e54269 [FREE Full text] [doi: 10.2196/54269] [Medline: 39264696]
32. A health and care digital capabilities framework. National Health Service. 2018. URL: https://www.hee.nhs.uk/sites/default/files/documents/Digital%20Literacy%20Capability%20Framework%202018.pdf [accessed 2024-04-29]
33. The Topol review. NHS Health Education England. URL: https://topol.hee.nhs.uk/ [accessed 2023-09-18]
34. What is ACTIC. Accreditation of Competence in Information and Communication Technologies. URL: http://actic.gencat.cat/en/actic_informacio/actic_que_es_l_actic_/index.html [accessed 2022-07-26]
35. Charmaz K. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis (Introducing Qualitative Methods series). Thousand Oaks, CA: Sage Publications; 2006.
36. BOE-A-2003-21340 Ley 44/2003, de 21 de noviembre, de ordenación de las profesiones sanitarias. BOE.es. URL: https://www.boe.es/buscar/act.php?id=BOE-A-2003-21340 [accessed 2022-07-26]
37. Rodríguez Del Águila MM, González-Ramírez AR. Sample size calculation. Allergol Immunopathol (Madr) 2014 Sep;42(5):485-492. [doi: 10.1016/j.aller.2013.03.008] [Medline: 24280317]
38. Taib F, Yusoff MS. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. J Taibah Univ Med Sci 2014 Jun;9(2):110-114. [doi: 10.1016/j.jtumed.2013.12.002]
39. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
40. Pla de salut 2021-2025. Departament de Salut, Generalitat de Catalunya. 2021. URL: https://salutweb.gencat.cat/web/.content/_departament/pla-de-salut/pla-de-salut-2021-2025/pla-salut-catalunya-2021-2025.pdf [accessed 2021-04-13]
41. Situació actual i diagnòstic de necessitats de professionals de la salut. Departament de Salut. 2018. URL: https://salutweb.gencat.cat/web/.content/_departament/eixos-xiv-legislatura/1r_forum_dialeg_professional/documents/diagnostic.pdf [accessed 2024-04-29]
42. Sustainable development goals. United Nations Sustainable Development. URL: https://www.un.org/sustainabledevelopment/ [accessed 2023-09-28]
43. Classification of digital health interventions v 1. World Health Organization. 2018. URL: https://apps.who.int/iris/bitstream/handle/10665/260480/WHO-RHR-18.06-eng.pdf [accessed 2024-04-29]
44. Hübner U, Thye J, Shaw T, Elias B, Egbert N, Saranto K, et al. Towards the TIGER international framework for recommendations of core competencies in health informatics 2.0: extending the scope and the roles. Stud Health Technol Inform 2019 Aug 21;264:1218-1222. [doi: 10.3233/SHTI190420] [Medline: 31438119]
45. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]

XSL•FO
RenderX

46. The Topol review: preparing the healthcare workforce to deliver the digital future. The NHS Constitution. URL: https://topol.hee.nhs.uk/ [accessed 2023-07-24]
47. Griffiths FE, Armoiry X, Atherton H, Bryce C, Buckle A, Cave JA, et al. The Role of Digital Communication in Patient–Clinician Communication for NHS Providers of Specialist Clinical Services for Young People [The Long-Term Conditions Young People Networked Communication (LYNC) Study]: A Mixed-Methods Study. Southampton, UK: NIHR Journals Library; 2018.
48. Norgaard O, Furstrand D, Klokker L, Karnoe A, Batterham R, Kayser L. The e-health literacy framework: a conceptual framework for characterizing e-health users and their interaction with e-health systems. Knowl Manag E-Learn 2015;7:540. [doi: 10.34105/j.kmel.2015.07.035]
49. Duffy FF, Fochtmann LJ, Clarke DE, Barber K, Hong S, Yager J, et al. Psychiatrists' comfort using computers and other electronic devices in clinical practice. Psychiatr Q 2016 Sep;87(3):571-584 [FREE Full text] [doi: 10.1007/s11126-015-9410-2] [Medline: 26667248]
50. Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth. World Health Organization. 2016. URL: https://apps.who.int/iris/handle/10665/252529 [accessed 2024-04-29]
51. Tsirintani M. Fake news and disinformation in health care- challenges and technology tools. Stud Health Technol Inform 2021 May 27;281:318-321. [doi: 10.3233/SHTI210172] [Medline: 34042757]
52. Vayena E, Haeusermann T, Adjekum A, Blasimme A. Digital health: meeting the ethical and policy challenges. Swiss Med Wkly 2018 Dec 29;148:w14571 [FREE Full text] [doi: 10.4414/smw.2018.14571] [Medline: 29376547]
53. van Houwelingen CT, Ettema RG, Kort HS, Ten Cate O. Hospital nurses' self-reported confidence in their telehealth competencies. J Contin Educ Nurs 2019 Jan 01;50(1):26-34 [FREE Full text] [doi: 10.3928/00220124-20190102-07] [Medline: 30645656]
54. Kirchberg J, Fritzmann J, Weitz J, Bork U. eHealth literacy of German physicians in the pre-COVID-19 era: questionnaire study. JMIR Mhealth Uhealth 2020 Oct 16;8(10):e20099 [FREE Full text] [doi: 10.2196/20099] [Medline: 33064102]
55. Potts HW. Is e-health progressing faster than e-health researchers? J Med Internet Res 2006 Sep 29;8(3):e24 [FREE Full text] [doi: 10.2196/jmir.8.3.e24] [Medline: 17032640]
56. Konttila J, Siira H, Kyngäs H, Lahtinen M, Elo S, Kääriäinen M, et al. Healthcare professionals' competence in digitalisation: a systematic review. J Clin Nurs 2019 Mar;28(5-6):745-761. [doi: 10.1111/jocn.14710] [Medline: 30376199]
57. Aungst TD, Patel R. Integrating digital health into the curriculum-considerations on the current landscape and future developments. J Med Educ Curric Dev 2020 Jan 20;7:2382120519901275 [FREE Full text] [doi: 10.1177/2382120519901275] [Medline: 32010795]
58. Pathipati AS, Azad TD, Jethwani K. Telemedical education: training digital natives in telemedicine. J Med Internet Res 2016 Dec 12;18(7):e193 [FREE Full text] [doi: 10.2196/jmir.5534] [Medline: 27405323]
59. Workforce, training and education: improving the digital literacy of the workforce. National Health Service, England. 2018. URL: https://www.hee.nhs.uk/our-work/digital-literacy [accessed 2022-07-26]
60. EU*US eHealth work. EHTEL. URL: https://ehtel.eu/activities/eu-funded-projects/euus-ehealth-work.html#:~:text=The%20EU*US%20eHealth%20Work,healthcare%20workforce%20across%20the%20globe [accessed 2022-07-26]
61. Narrative review - an overview. ScienceDirect. URL: https://www.sciencedirect.com/topics/psychology/narrative-review [accessed 2023-07-25]
62. Neter E, Brainin E. Perceived and performed eHealth literacy: survey and simulated performance test. JMIR Hum Factors 2017 Jan 17;4(1):e2 [FREE Full text] [doi: 10.2196/humanfactors.6523] [Medline: 28096068]
63. Article 5: regulation (EU) 2016/679 of the European Parliament and of the Council. European Union. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2024-05-13]
64. van der Vaart R, Drossaert CH, de Heus M, Taal E, van de Laar MA. Measuring actual eHealth literacy among patients with rheumatic diseases: a qualitative analysis of problems encountered using Health 1.0 and Health 2.0 applications. J Med Internet Res 2013 Feb;15(2):e27 [FREE Full text] [doi: 10.2196/jmir.2428] [Medline: 23399720]
65. Jarva E, Oikarinen A, Andersson J, Tomietto M, Kääriäinen M, Mikkonen K. Healthcare professionals' digital health competence and its core factors; development and psychometric testing of two instruments. Int J Med Inform 2023 Mar;171:104995 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.104995] [Medline: 36689840]
66. Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, et al. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020 Aug 14;22(8):e19827 [FREE Full text] [doi: 10.2196/19827] [Medline: 32667899]
67. Socha-Dietrich K. Empowering the health workforce to make the most of the digital revolution. Organization for Economic Cooperation and Development. 2021. URL: https://www.oecd-ilibrary.org/content/paper/37ff0eaa-en [accessed 2024-04-29]
68. Keep M, Janssen A, McGregor D, Brunner M, Baysari MT, Quinn D, et al. Mapping eHealth education: review of eHealth content in health and medical degrees at a metropolitan tertiary institute in Australia. JMIR Med Educ 2021 Aug 19;7(3):e16440 [FREE Full text] [doi: 10.2196/16440] [Medline: 34420920]
69. Echelard JF, Méthot F, Nguyen HA, Pomey MP. Medical student training in eHealth: scoping review. JMIR Med Educ 2020 Sep 11;6(2):e20027 [FREE Full text] [doi: 10.2196/20027] [Medline: 32915154]

XSL•FO
RenderX

Abbreviations

ACTIC: Accreditation of Competence in Information and Communication Technologies
COMPDIG-Salut: Digital Skills for Health Professionals
DI: discrimination index
DigComp: Digital Competence Framework for Citizens
GLB: greatest lower bound

Edited by F Pietrantonio, M Montagna, J López Castro, I Said-Criado; submitted 07.10.23; peer-reviewed by V Grieco, E Brainin; comments to author 21.12.23; revised version received 12.02.24; accepted 17.06.24; published 17.10.24.

Please cite as:
Saigí-Rubió F, Romeu T, Hernández Encuentra E, Guitert M, Andrés E, Reixach E
Design, Implementation, and Analysis of an Assessment and Accreditation Model to Evaluate a Digital Competence Framework for Health Professionals: Mixed Methods Study
JMIR Med Educ 2024;10:e53462
URL: https://mededu.jmir.org/2024/1/e53462
doi:10.2196/53462
PMID:39418092

©Francesc Saigí-Rubió, Teresa Romeu, Eulàlia Hernández Encuentra, Montse Guitert, Erik Andrés, Elisenda Reixach. Originally published in JMIR Medical Education (https://mededu.jmir.org), 17.10.2024. 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.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX