Cureus
Part of SPRINGER NATURE

Open Access Review Article

Applications of Artificial Intelligence in Medical Education: A Systematic Review

Eric Hallquist [1], Ishank Gupta [2], Michael Montalbano [2], Marios Loukas [2, 3]

1. Department of Family Medicine, Prevea Shawano Avenue Health Center, Green Bay, USA 2. Department of Anatomical Sciences, St. George's University School of Medicine, St. George, GRD 3. Department of Clinical Anatomy, Mayo Clinic, Rochester, USA

Corresponding author: Marios Loukas, mloukas@sgu.edu

Review began 01/18/2025
Review ended 02/26/2025
Published 03/01/2025

DOI: 10.7759/cureus.79878

Abstract

Artificial intelligence (AI) models, like Chat Generative Pre-Trained Transformer (OpenAI, San Francisco, CA), have recently gained significant popularity due to their ability to make autonomous decisions and engage in complex interactions. To fully harness the potential of these learning machines, users must understand their strengths and limitations. As AI tools become increasingly prevalent in our daily lives, it is essential to explore how this technology has been used so far in healthcare and medical education, as well as the areas of medicine where it can be applied. This paper systematically reviews the published literature on the PubMed database from its inception up to June 6, 2024, focusing on studies that used AI at some level in medical education, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. Several papers identified where AI was used to generate medical exam questions, produce clinical scripts for diseases, improve the diagnostic and clinical skills of students and clinicians, serve as a learning aid, and automate analysis tasks such as screening residency applications. AI shows promise at various levels and in different areas of medical education, and our paper highlights some of these areas. This review also emphasizes the importance of educators and students understanding AI's principles, capabilities, and limitations before integration. In conclusion, AI has potential in medical education, but more research needs to be done to fully explore additional areas of applications, address the current gaps in knowledge, and its future potential in training healthcare professionals.

Categories: Other, Medical Education
Keywords: machine learning algorithms, medical education assessment, medical examination, natural language processing models, teaching undergraduate and postgraduate

Introduction And Background

Artificial intelligence (AI) refers to technologies designed to mimic human intelligence, and its applications have evolved significantly over the years. Initially, AI was used in narrow fields like digital searching and facial recognition and for tasks such as analyzing cardiac rhythms [1,2]. However, recent advancements in natural language processing (NLP) have greatly expanded AI's capabilities, enabling systems to engage in more complex tasks, such as having conversations in human language [3,4]. When combined with machine learning (ML) algorithms, which allow AI to process large amounts of data and make decisions based on patterns, these systems have the potential to perform tasks autonomously with minimal human input [5]. In medical education, these developments have opened up new opportunities for AI to support learning in various ways. Generative AI models, like Chat Generative Pre-Trained Transformer (ChatGPT; OpenAI, San Francisco, CA), are increasingly being used to answer student queries, create personalized lessons, provide tailored feedback, and even facilitate simulations of virtual patients based on aggregated data [6,7]. In clinical practice, AI is already making an impact, such as using automated systems that analyze heart rhythms with defibrillators [8]. As AI tools gain popularity and adoption in healthcare and education, it becomes crucial for medical educators to comprehend how these tools integrate into their curricula [9-11]. However, to fully capitalize on these technologies, it is essential that medical educators and students understand the strengths and limitations of AI tools. This paper aims to provide a systematic review of how AI is currently being applied in medical education. It focuses on studies that assess AI's role across different educational stages, from medical school to postgraduate training and continuing medical education (CME). The goal of this study is not only to highlight the positive contributions of AI in medical education so far but also to identify gaps in knowledge and encourage further research to explore the full potential of AI in training healthcare professionals.

Materials and methods

Search Strategy

In this systematic review, the authors adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. Two independent researchers (E.H. and I.G.) conducted a literature search on the PubMed database using the keywords "artificial intelligence" AND "medical education" from inception up to June 6, 2024, with a third researcher (M.M.) involved to reach a consensus when needed. The practice of backward citation was also used to locate additional articles. This was done by reviewing the references or citations within an article from the search synthesis to find studies relevant to our research topic. Only PubMed was used because it is a repository for clinical medicine and medical education; other databases or search tools, such as Scopus, Web of Science, or Google Scholar, were not used, as their literature is not limited to biomedical, clinical, or medical sciences, which was our main focus. The search strategy registration was not required by our institution, as our study did not involve human subjects.

Selection Criteria

Articles were selected based on the following inclusion criteria: 1) article discussion of medical education at the level of medical school, postgraduate training, or CME; 2) discussion of AI, including any subset of AI such as ML models, generative AI large language models (LLMs), or NLP models; 3) reporting of numerical data relevant to AI in medical education; and 4) text available in the English language. The exclusion criteria for articles were as follows: 1) discussion of medical education without mentioning AI; 2) discussion of AI without mentioning medical education; 3) discussion of neither AI nor medical education; 4) letters, editorials, expert opinions, and other descriptive reports; and 4) unpublished studies or preprints.

Review

Results

The initial search of the PubMed database yielded 781 articles, with another 286 identified through backward citation searching. After applying inclusion and exclusion criteria to the 1,067 articles, 195 articles were sought for retrieval, and all 195 articles were retrieved. After retrieval, 113 reports were further excluded after a review of the entire text, thereby leaving 82 articles for final inclusion in this review (Figure 1).

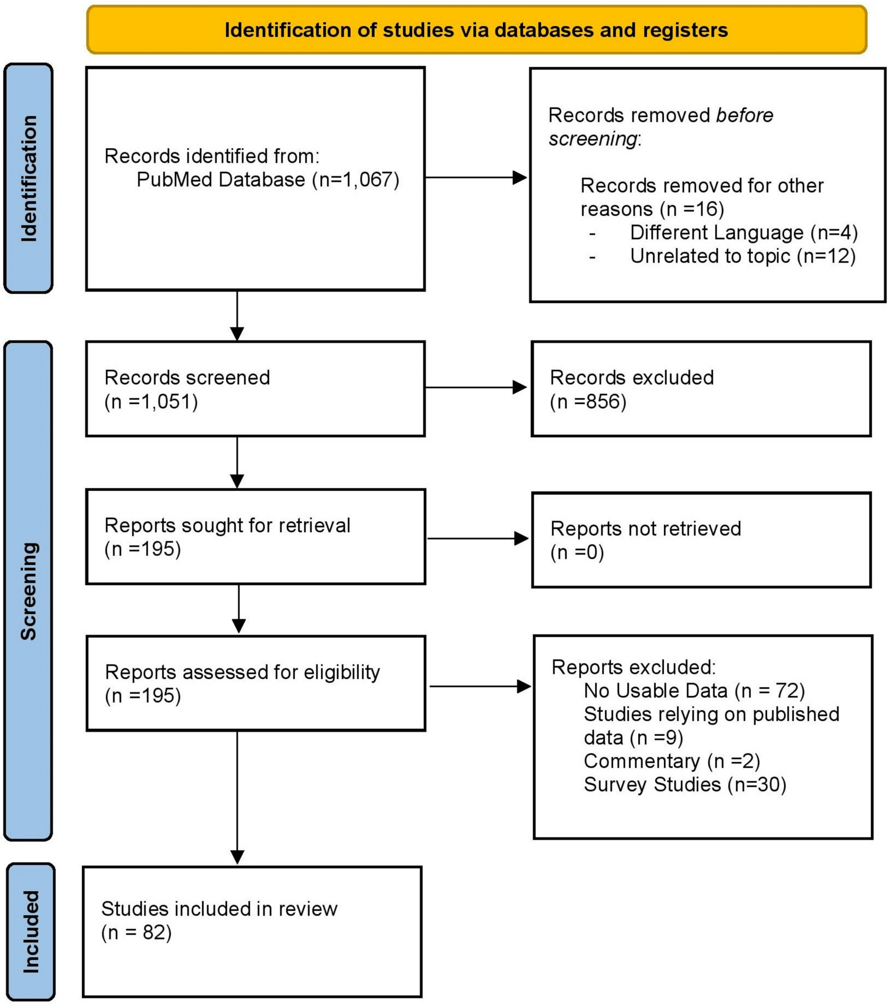

FIGURE 1: The PRISMA flowchart

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Image credit: This is an original image created by the author Ishank Gupta

Assessing AI's Clinical Knowledge and Decision-Making Through Standardized Assessments

Fourteen studies employed the LLM ChatGPT as their AI tool of choice [12-25]. Among these, six studies directly compared the performance of GPT-3.5 and GPT-4.0 [12,18,20,22-24], with GPT-4 consistently demonstrating superior performance over GPT-3.5. The model was evaluated across various disciplines of medicine: surgery [12], parasitology [13], microbiology [14], neurology [18], and bioethics [21]. Three studies compared ChatGPT's performance with human test-takers [13,16,24]. Eight studies assessed ChatGPT's performance on national medical board exams [15,17,19,20,22-25], including the German medical licensing examination [24], the European Exam in Core Cardiology [15], the Family Medicine Board Exam in Taiwan [17], and the United States Medical Licensing Examination (USMLE) Step 3 study question bank [23]. Some studies did not specify the ChatGPT version used [17], and in four studies, questions in languages other than English were utilized [17,20,24,25]. A common limitation across these studies was ChatGPT's inability to interpret questions involving tables, graphs, or images, restricting evaluations to text-based questions only.

Using AI Tools to Craft Medical Illness Scripts and Exam Questions

Several studies utilized ChatGPT to generate medical examination questions and illness scripts. Yanagita et al. assessed whether ChatGPT-4 could create illness scripts for diseases according to Japanese medical education standards [26]. Another study aimed to compare script concordance tests (SCTs) generated by ChatGPT with those created by clinical experts [27]. Hudon et al. used a mixed-method approach, evaluating three SCTs generated by ChatGPT alongside three SCTs crafted by clinical experts and gathering feedback through a web-based survey from clinician-educators and resident doctors specializing in psychiatry in

Quebec, Canada [27]. Klang et al. generated 210 multiple-choice questions (MCQs) using ChatGPT-4, and Cheung et al. used ChatGPT to generate 50 MCQs for medical examinations [28,29]. These studies are listed in Table *1*.

Study	Study purpose	AI tool used	Number of questions/illness scripts	Study features	Results	Areas of errors in AI-generated content	Conclusions
Yanagita et al. [26]	Aims to investigate whether AI can generate illness scripts	GPT-4	184 disease illness scripts	Three physicians assessed each illness script using a three-tier grading scale	The illness scripts received “A,” “B,” and “C” ratings of 56.0% (103/184), 28.3% (52/184), and 15.8% (29/184), respectively	Cardiovascular and psychiatric systems had the highest number of “C” ratings	GPT-4 generates illness scripts rapidly and with high quality
Hudon et al. [27]	Comparing SCTs generated by ChatGPT with those produced by clinical experts	ChatGPT	Three ChatGPT-generated SCTs vs. three expert-created SCTs	102 educators evaluated the six SCTs based on three preset criteria	No significant distinctions between the two types of SCT scenario (p = 0.84)	-	ChatGPT is useful for developing educational materials
Klang et al. [28]	Writing multiple choice examination questions for medical students	GPT-4	210 MCQs generated by AI	AI-generated questions reviewed by specialist physicians. Physicians were blinded to the source of the questions	One question (0.5%) was defined as false; 15% of questions required revisions	Inaccurate medical terminology and inaccuracies in age-sensitivity, gender-sensitivity, and geographic sensitivity	GPT-4 can assist in creating multiple-choice medical exam questions
Cheung et al. [29]	Comparison of graduate medical exam MCQs: University Professors vs. ChatGPT-generated	ChatGPT	50 MCQs generated by ChatGPT and 50 MCQs drafted by two university professors	Evaluated by five independent assessors. Scoring is based on a standardized system	No significant difference in question quality among the groups	In the relevance domain, the AI was inferior to humans	ChatGPT can generate MCQs for medical graduate exams

TABLE 1: Overview and description of studies where AI tools were used to generate multiple choice medical examination questions and illness scripts

AI: artificial intelligence; GPT: Generative Pre-Trained Transformer; SCTs: script concordance tests; MCQs: multiple-choice questions

AI as a Learning Aid in Medical School, Residency Training, and CME

Thirty publications examined the role of AI in enhancing medical education, encompassing a range of applications. Fifteen studies focused on using AI as a teaching tool for medical students [30-44]. Nine explored AI's applications in residency training [35,36,45-51], and six investigated AI's potential role in CME [31,35,36,52-54]. Price et al. proposed an ML model to enhance outcomes in medical specialty board recertification among physicians [54]. Six studies suggested that AI can improve training in diagnosis, communication, and radiograph interpretation for medical professionals [38,43-45,48,52]. These studies are listed in Table *2*.

Study	Study group	Control group present	Study summary
Monlezun et al. [30]	Medical students	No	Machine learning-enhanced causal inference analysis of a multisite cohort to boost medical trainees' skills in counseling patients about nutrition and also improve their own dietary habits
Ruberto et	Medical students and postresidency	No	A simulation platform that adjusts to a participant's cognitive load in real-time, offering

al. [31]	or CME		potential for advancing expertise in resuscitation medicine
Yang and Shulruf [32]	Medical students	Yes	The addition of expert-led + AI-assisted tutoring to the surgical curriculum showed potential benefits, with the expert-led + AI group outperforming and improving more in the end-of-surgical block OSCE compared to the expert-led group
Nakawala et al. [33]	Medical students	Yes	A context-aware software framework for thoracocentesis training in surgical workflows yielded results comparable to traditional mentor-based training
Fazlollahi et al. [34]	Medical students	No	A randomized clinical trial showed that VOA feedback led to better performance and skill transfer than remote expert instruction in surgical training
Allen et al. [35]	Medical students, residents, and postresidency or CME	No	SVMs offer more accurate predictions of competency in laparoscopic training tasks than conventional methods of evaluation
Mirchi et al. [36]	Medical students, residents, and postresidency or CME	No	The virtual operative assistant, an AI tool, accurately classified skilled and novice participants, highlighting the potential of integrating AI and virtual reality simulation into surgical education
Del Blanco et al. [37]	Medical students	Yes	A game-like simulation to improve novices' perceptions and performance during their first operating theater experience. The simulation was effective in reducing fears and errors, and enhancing perceived knowledge and collaboration among the students
Cheng et al. [38]	Medical students	Yes	An AI-based medical image learning system significantly enhanced medical students' ability to identify hip fractures on pelvic X-rays
Wang et al. [39]	Medical students	No	AIteach (AITEACH Limited, Cambridge, UK), a virtual case system using NLP and hospital records, generated clinical cases for medical students, improving their clinical thinking skills
Denny et al. [40]	Medical students	No	A novel electronic advisor system using NLP to identify geriatric medicine competencies by analyzing medical students' clinical notes. Such models can be used to assess specific competencies among trainees
Maicher et al. [41]	Medical students	No	A virtual standardized patient system that assesses students' history-taking skills by understanding, responding, and providing immediate feedback on their performance
Hamdy et al. [42]	Medical students	No	The online VICEE effectively assessed medical students' non-psychomotor clinical competencies
Suebnukarn and Haddawy [43]	Medical students	No	COMET, an intelligent tutoring system, assessed clinical reasoning in problem-based learning among students. Such clinical reasoning models can be combined with traditional tutoring strategies to effectively emulate human tutor hints
Woo et al. [44]	Medical students	No	An intelligent tutoring system capable of engaging in a natural language dialogue with a student
Merritt et al. [45]	Residents	No	An AI-driven simulation platform to evaluate and provide real-time feedback to residents on a standardized, simulated conversation
Bissonnette et al. [46]	Residents	No	Machine learning algorithms were used to assess surgical performance among trainees executing a virtual reality hemilaminectomy
Lin et al. [47]	Residents	No	A novel hybrid prediction algorithm, CBCF, predicts the difficulty level of each case to create personalized training programs for radiology trainees
Hershberger et al. [48]	Residents	No	The NLP model, ReadMI, was used to enhance motivational interviewing skills among residents to encourage patients to change high-risk lifestyle behaviors

TABLE 2: Studies showing AI as a learning aid

CME: continuing medical education; AI: artificial intelligence; OSCE: objective structured clinical examination; VOA: virtual operative assistant; SVMs: support vector machines; COMET: collaborative intelligent tutoring system; NLP: natural language processing; VICEE: Virtual Clinical Encounter Examination; CBCF: content-boosted collaborative filtering

NLP, Motion Analysis, Clinical Skills, and Building Competence in Stakeholders

Thirty-four studies examined ML's capabilities in automating analysis tasks rather than direct applications to medical education [55-88].

Studies using NLP: Two studies evaluated the clinical competency of residents using NLP in clinical assessments [57,58]. Seven investigated the use of AI to assess medical school or residency applications [63-69]. One paper explored the potential of AI in assisting with writing letters of recommendation [76]. One study used AI in a virtual Q&A session for fellowship applicants, finding that many applicants considered the Chatbot helpful [56]. One used NLP to evaluate medical students' clinical notes to assess their exposure to the American Association of Medical College geriatric competencies [72]. Two reported on AI's role in analyzing feedback from clinical supervisors and its correlation with clinical entrustment decisions [74,75].

AI in motion analysis and clinical skills: Four studies used AI models to assess and evaluate clinical skill expertise via operational exposure and surgical hand movements [59-62]. Another focus is predicting correct answers based on students' eye movements while viewing whole slide images [71]. One study used ChatGPT-assisted training to enhance clinical skills among pediatric trainees [73], and three focused on AI in medical imaging [85-87]. Additionally, one study examined AI's ability to detect racial stereotypes in medical scenarios [70].

Building competency in AI among stakeholders: Twelve papers addressed building competence in stakeholders regarding AI. Of these, eight focused on teaching AI fundamentals at the medical school level and postgraduate training [77-84] and one on the current state of AI education in German universities [88]. Abid et al. proposed a one-month elective on AI and ML for fourth-year medical students [81]. One study used surveys to measure the use of ChatGPT among faculty members [55].

Discussion

Advancements of AI in Medical Education and Assessments

GPT-3.5 was launched in November of 2022, and GPT-4 was launched in March of 2023. GPT-4 has demonstrated significant advancements in accuracy and clinical reasoning compared to its predecessor, GPT-3.5, in medical assessments. Notably, while ChatGPT-3.5 failed the UK Radiology Fellowship Exam, ChatGPT-4 passed with a score of 75.5% [22]. Meyer et al. observed that while GPT-4 successfully passed all three German medical licensing examinations tested, GPT-3.5 only passed one out of three [24]. These improvements position GPT-4 as a more reliable tool for delivering precise responses to complex medical queries [12,18,24]. In other studies, ChatGPT achieved an accuracy rate of 76.4% in surgical subspecialties [12], 60.8% in parasitology [13], 80% in microbiology [14], 64% in neurology [18], and an accuracy rate of 59.6% in bioethics [21]. Chen et al. found that ChatGPT excelled in bioethical questions related to death and physician-patient relationships but faced challenges with topics like abuse and informed consent [21]. Therefore, current AI capabilities may not be consistent enough to handle complex ethical decisions in medicine autonomously. This necessitates and emphasizes the need for human oversight, especially in critical scenarios [21].

In several studies, AI tackled difficult medical licensing or subspecialty board exams, which consist of complex questions requiring years of preparation, deductive reasoning, and extensive knowledge. In one study by Friederichs et al., ChatGPT answered 65.5% of progress test questions correctly, surpassing German medical students in their first three years of study [16]. ChatGPT also achieved a passing score on the European Exam in Core Cardiology with an accuracy rate of approximately 60% [15] but failed the 2022 Family Medicine Board Exam in Taiwan with only 41.6% accuracy [17]. Additionally, ChatGPT performed well on the USMLE Step 3 study question bank, achieving an overall accuracy of 84.7% [23]. In two studies, ChatGPT outperformed human test-takers [16,24], whereas in one study, it performed less effectively [13]. This shows the evolving capacity of AI to comprehend and process intricate medical information in a very short period. However, there are varying performance outcomes across studies when handling questions in languages other than English, where the model's linguistic database is less robust [17,24,25]. Despite the challenges, ChatGPT, even in its beta version, achieved satisfactory or high percentile scores, highlighting its growing capabilities in medical knowledge [15,19,20,22-25]. However, there is a need for dedicated research to better understand the factors influencing AI's test performance, including language disparities, and to determine whether the assessed knowledge is universally applicable or specific to a particular country [89]. Close collaboration between AI developers and medical professionals will be essential to ensure the effectiveness of these tools [18]. This partnership becomes crucial, especially when considering the ethical implications of AI applications in healthcare.

Using AI Tools to Craft Medical Illness Scripts and Exam Questions

ChatGPT is a deep learning-based language model trained on vast amounts of text data to generate human-like responses. While it does not store information between interactions, it can generate medical examination questions and illness scripts based on patterns learned during its training. Several studies have demonstrated its ability to produce such content. In the study by Yanagita et al., three physicians evaluated the AI-generated illness scripts for 184 diseases using a grading scale: "A" for sufficient, "B" for partially lacking but acceptable, and "C" for deficient. The results showed that GPT-4 successfully generated complete illness scripts for all diseases, with 56.0% rated "A," 28.3% rated "B," and 15.8% rated "C." This study concluded that GPT-4 could promptly create useful illness scripts [26]. In the study by Hudon et al., SCTs generated by ChatGPT were compared with those created by clinical experts and found that respondents found it challenging to distinguish between the AI-generated scripts and those generated by clinicians. Although ChatGPT demonstrated its capability to produce SCTs aligned with the authors ' predetermined clinical criteria, some outputs were criticized for oversimplifying or caricaturing scenarios [27]. Hudon et al. underscored ChatGPT's potential as a tool for writing SCT but highlighted the need for further refinement to ensure clinical accuracy and educational effectiveness [27].

Klang et al. generated 210 MCQs using GPT-4, which were then reviewed by physicians unaware of their origin. Of these, only one question was classified as "false," and 15% required some revisions. The study concluded that generative AI is a valuable supplementary tool for developing medical exam questions, provided specialist oversight exists [28]. Similarly, Cheung et al. used ChatGPT to generate 50 MCQs for medical examinations, finding their quality comparable to those written by university professors [29]. Therefore, generative AI tools like ChatGPT are highly efficient for producing high-quality medical scripts and clinical vignettes more quickly than their human counterparts. However, despite their advantages, these tools are prone to errors and should be used as adjunctive resources, with thorough review by experienced medical professionals.

Using AI to Enhance Learning, Assessment, and Diagnostic Capabilities

AI holds significant promise as a learning aid, particularly in training simulations. Wang et al. demonstrated this by using a virtual patient simulator that allowed students to interact using natural language. Students who engaged with this system showed improvements in clinical skills, as assessed automatically by the simulator based on their performance [39]. Similarly, Merritt et al. found that residents using AI-driven simulated case presentations, with AI acting as a primary care physician, reported enhanced communication skills and confidence in future interactions [45]. Hershberger et al. further supported this, showing that AI analysis of resident interview transcripts had moderate agreement with human raters (Kappa = 0.52) and a narrower range of agreement (Kappa = 0.313-0.658) when evaluating reflective statements, open-ended questions, closed-ended questions, and readiness-to-change assessments [48]. These findings underscore AI's potential to augment learning outcomes and assessment capabilities in educational settings, particularly in healthcare training. Overall, the results were promising, with AI tutors showing partial agreement with human decision-making and significant improvements across various learner groups in medicine: medical students, residents, and those in CME.

AI’s utility can extend beyond language skills into diagnostics. Cheng et al. applied AI to generate fracture probability maps on X-rays, significantly enhancing students' accuracy in identifying hip fractures, even without AI assistance [38]. Additionally, McFadden and Crim reported that primary care physicians at a CME conference who used AI-driven simulations showed marked improvement in diagnosing case vignettes, indicating enhanced diagnostic capabilities [52]. Suebnukarn and Haddawy developed the collaborative intelligent tutoring system, an automated tutor for small groups, which made choices consistent with human tutors 62%-83% of the time [43]. Another AI tutor, CIRCSIM-Tutor (Illinois Institute of Technology, Chicago, IL, and Rush College of Medicine, Chicago, IL), demonstrated significant learning gains in students, particularly in problem-solving questions, as evidenced by pretest and posttest comparisons [44]. While these studies report that AI improved diagnostic skills, communication skills, and radiograph interpretation for medical professionals, many studies were limited by the absence of control groups and relied on self-assessment for measuring outcomes.

Need for Structured AI Education in Medicine

Understanding the basic principles of AI is essential for its effective use, yet this knowledge may be limited, given the technology's relatively recent development. One study revealed that 66% (n = 29) of the surveyed faculty had utilized ChatGPT [55]. Understanding the normal operating parameters and limitations of AI is essential to ensure its effective application and to prevent undesired outcomes. Brief introductions to AI have proven beneficial in various contexts and cultures. For instance, a 2022 study of students in Lebanon found that those who received AI education from their medical school were more knowledgeable about AI than those who did not [90]. Another 2022 study in Germany demonstrated that an online class improved students' self-perceived AI readiness [86]. Similarly, a 10-hour pilot class for medical students at the University of British Columbia helped students better understand ML concepts [83,86]. Most learners reported increased confidence and satisfaction with AI-related courses [77-88]. These studies underscore the importance of AI education and indicate a positive trend toward embracing AI in the medical field to enhance both learning and clinical practices. Most students who completed the elective proposed by Abid et al. reported high satisfaction and increased confidence and understanding of AI [81]. The authors recommended incorporating such initiatives into medical school curricula to enhance AI literacy among future physicians [81]. However, caution is advised when relying on self-reported competence, as definitions can vary, and self-assessed competence may not always align with objective measures [91].

Limitations

The study on AI in medical education reveals several key limitations. Many studies included were qualitative, making it difficult to generalize the results or draw definitive conclusions. Additionally, confidence intervals or p values were often unavailable, and while the authors tried to quantify the findings, most studies remained qualitative by nature. This is likely due to the relatively recent introduction of AI models like ChatGPT to the public, meaning the existing literature consists of smaller, less structured studies. Many studies also lacked control groups or relied on self-assessment, which could introduce bias. The review was also limited to studies from PubMed, leaving out potentially relevant research from other sources. Furthermore, only studies in the English language were included, and unpublished data or preprints, which may provide more robust or cutting-edge conclusions, were not considered. ChatGPT's inability to interpret questions involving tables, graphs, or images restricted evaluations to text-based questions only. While AI showed promise in generating medical content, human oversight was often necessary to correct errors in the outputs. Additionally, AI faced challenges in ethical decision-making, emphasizing the need for human oversight.

Conclusions

AI holds great promise as a learning aid and teaching tool in medical education. While research on its application in this field is encouraging, further studies with objective assessments and control groups are necessary to draw more definitive comparisons. Meanwhile, techniques like fine-tuning can enhance AI accuracy for specific tasks. Before integrating AI into teaching or personal study, educators and students should understand the principles, capabilities, and limitations of ML. Although comfort levels with AI vary among different groups, brief introductory sessions have been shown to effectively reduce unease when encountering this technology.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Marios Loukas, Eric Hallquist, Ishank Gupta, Michael Montalbano

Acquisition, analysis, or interpretation of data: Marios Loukas, Eric Hallquist, Ishank Gupta, Michael Montalbano

Critical review of the manuscript for important intellectual content: Marios Loukas, Ishank Gupta, Michael Montalbano

Supervision: Marios Loukas, Michael Montalbano

Drafting of the manuscript: Eric Hallquist, Ishank Gupta

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

Eric Hallquist and Ishank Gupta contributed equally to the work and should be considered co-first authors.

References

1. Haenlein M, Kaplan A: A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. Calif Manag Rev. 2019, 61:5-14. 10.1177/0008125619864925
2. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al.: An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019, 394:861-7. 10.1016/S0140-6736(19)31721-0
3. Hirschberg J, Manning CD: Advances in natural language processing . Science. 2015, 349:261-6. 10.1126/science.aaa8685
4. Aljanabi M: ChatGPT: future directions and open possibilities . Mesopotamian J Cybersecur. 2023, 2023:16-7. 10.58496/MJCS/2023/003
5. Dike HU, Zhou Y, Deveerasetty KK, Wu Q: Unsupervised learning based on artificial neural network: a review. 2018 IEEE Int Conf Cyborg Bionic Syst. 2018, 322-7. 10.1109/CBS.2018.8612259

6. Firat M: What ChatGPT means for universities: perceptions of scholars and students . J Appl Learn Teach. 2023, 6:1. 10.37074/jalt.2023.6.1.22
7. Furini M, Gaggi O, Mirri S, Montangero M, Pelle E, Poggi F, Prandi C: Digital twins and artificial intelligence: as pillars of personalized learning models. Commun ACM. 2022, 65:98-104. 10.1145/3478281
8. Figuera C, Irusta U, Morgado E, et al.: Machine learning techniques for the detection of shockable rhythms in automated external defibrillators. PLoS One. 2016, 11:e0159654. 10.1371/journal.pone.0159654
9. Meskó B, Topol EJ: The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023, 6:120. 10.1038/s41746-023-00873-0
10. Chen L, Chen P, Lin Z: Artificial intelligence in education: a review . IEEE Access. 2020, 8:75264-78. 10.1109/ACCESS.2020.2988510
11. Popenici SA, Kerr S: Exploring the impact of artificial intelligence on teaching and learning in higher education. Res Pract Technol Enhanc Learn. 2017, 12:22. 10.1186/s41039-017-0062-8
12. 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, 104:269-73. 10.4174/astr.2023.104.5.269
13. 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. 10.3352/jeehp.2023.20.1
14. Das D, Kumar N, Longjam LA, Sinha R, Deb Roy A, Mondal H, Gupta P: Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus. 2023, 15:e36034. 10.7759/cureus.36034
15. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S: ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?. Eur Heart J Digit Health. 2023, 4:279-81. 10.1093/ehjdh/ztad029
16. Friederichs H, Friederichs WJ, März M: ChatGPT in medical school: how successful is AI in progress testing? . Med Educ Online. 2023, 28:2220920. 10.1080/10872981.2023.2220920
17. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ: ChatGPT failed Taiwan's Family Medicine Board Exam . J Chin Med Assoc. 2023, 86:762-6. 10.1097/JCMA.0000000000000946
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:e000451. 10.1136/bmjno-2023-000451
19. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D: 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. 10.2196/45312
20. 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. 10.2196/48002
21. Chen J, Cadiente A, Kasselman LJ, Pilkington B: Assessing the performance of ChatGPT in bioethics: a large language model's moral compass in medicine. J Med Ethics. 2024, 50:97-101. 10.1136/jme-2023-109366
22. Ariyaratne S, Jenko N, Mark Davies A, Iyengar KP, Botchu R: Could ChatGPT pass the UK radiology fellowship examinations?. Acad Radiol. 2024, 31:2178-82. 10.1016/j.acra.2023.11.026
23. Knoedler L, Alfertshofer M, Knoedler S, et al.: 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. 10.2196/51148
24. 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. 10.2196/50965
25. 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, 24:143. 10.1186/s12909-024-05125-7
26. 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, 24:536. 10.1186/s12909-024-05534-8
27. 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. 10.2196/54067
28. Klang E, Portugez S, Gross R, et al.: Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ. 2023, 23:772. 10.1186/s12909-023-04752-w
29. Cheung BH, Lau GK, Wong GT, et al.: ChatGPT versus human in generating medical graduate exam multiple choice questions-a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023, 18:e0290691. 10.1371/journal.pone.0290691
30. Monlezun DJ, Dart L, Vanbeber A, et al.: Machine learning-augmented propensity score-adjusted multilevel mixed effects panel analysis of hands-on cooking and nutrition education versus traditional curriculum for medical students as preventive cardiology: multisite cohort study of 3,248 trainees over 5 years. Biomed Res Int. 2018, 2018:5051289. 10.1155/2018/5051289
31. Ruberto AJ, Rodenburg D, Ross K, et al.: The future of simulation-based medical education: adaptive simulation utilizing a deep multitask neural network. AEM Educ Train. 2021, 5:e10605. 10.1002/aet2.10605
32. Yang YY, Shulruf B: An expert-led and artificial intelligence system-assisted tutoring course to improve the confidence of Chinese medical interns in suturing and ligature skills: a prospective pilot study. J Educ Eval Health Prof. 2019, 16:7. 10.3352/jeehp.2019.16.7
33. Nakawala H, Ferrigno G, De Momi E: Development of an intelligent surgical training system for thoracentesis. Artif Intell Med. 2018, 84:50-63. 10.1016/j.artmed.2017.10.004
34. Fazlollahi AM, Bakhaidar M, Alsayegh A, et al.: Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. JAMA Netw Open. 2022, 5:e2149008. 10.1001/jamanetworkopen.2021.49008

35. Allen B, Nistor V, Dutson E, Carman G, Lewis C, Faloutsos P: Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks. Surg Endosc. 2010, 24:170-8. 10.1007/s00464-009-0556-6
36. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF: The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. PLoS One. 2020, 15:e0229596. 10.1371/journal.pone.0229596
37. Del Blanco Á, Torrente J, Fernández-Manjón B, Ruiz P, Giner M: Using a videogame to facilitate nursing and medical students' first visit to the operating theatre. A randomized controlled trial. Nurse Educ Today. 2017, 55:45-53. 10.1016/j.nedt.2017.04.026
38. Cheng CT, Chen CC, Fu CY, et al.: Artificial intelligence-based education assists medical students' interpretation of hip fracture. Insights Imaging. 2020, 11:119. 10.1186/s13244-020-00932-0
39. Wang M, Sun Z, Jia M, et al.: Intelligent virtual case learning system based on real medical records and natural language processing. BMC Med Inform Decis Mak. 2022, 22:60. 10.1186/s12911-022-01797-7
40. Denny JC, Spickard A 3rd, 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, 56:292-9. 10.1016/j.jbi.2015.06.004
41. Maicher KR, Zimmerman L, Wilcox B, et al.: Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach. 2019, 41:1053-9. 10.1080/0142159X.2019.1616683
42. Hamdy H, Sreedharan J, Rotgans JI, et al.: Virtual Clinical Encounter Examination (VICEE): a novel approach for assessing medical students' non-psychomotor clinical competency. Med Teach. 2021, 43:1203-9. 10.1080/0142159X.2021.1935828
43. Suebnukarn S, Haddawy P: A Bayesian approach to generating tutorial hints in a collaborative medical problem-based learning system. Artif Intell Med. 2006, 38:5-24. 10.1016/j.artmed.2005.04.003
44. Woo CW, Evens MW, Freedman R, et al.: An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning. Artif Intell Med. 2006, 38:25-46. 10.1016/j.artmed.2005.10.004
45. Merritt C, Glisson M, Dewan M, Klein M, Zackoff M: Implementation and evaluation of an artificial intelligence driven simulation to improve resident communication with primary care providers. Acad Pediatr. 2022, 22:503-5. 10.1016/j.acap.2021.12.013
46. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF: Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task. J Bone Joint Surg Am. 2019, 101:e127. 10.2106/JBJS.18.01197
47. Lin H, Yang X, Wang W: A content-boosted collaborative filtering algorithm for personalized training in interpretation of radiological imaging. J Digit Imaging. 2014, 27:449-56. 10.1007/s10278-014-9678-z
48. Hershberger PJ, Pei Y, Bricker DA, et al.: Advancing motivational interviewing training with artificial intelligence: ReadMI. Adv Med Educ Pract. 2021, 12:613-8. 10.2147/AMEP.S312373
49. Muntean GA, Groza A, Marginean A, Slavescu RR, Steiu MG, Muntean V, Nicoara SD: Artificial intelligence for personalised ophthalmology residency training. J Clin Med. 2023, 12:1825. 10.3390/jcm12051825
50. Thanawala R, Jesneck J, Shelton J, Rhee R, Seymour NE: Overcoming systems factors in case logging with artificial intelligence tools. J Surg Educ. 2022, 79:1024-30. 10.1016/j.jsurg.2022.01.013
51. Booth GJ, Ross B, Cronin WA, et al.: Competency-based assessments: leveraging artificial intelligence to predict subcompetency content. Acad Med. 2023, 98:497-504. 10.1097/ACM.0000000000005115
52. McFadden P, Crim A: Comparison of the effectiveness of interactive didactic lecture versus online simulation-based CME programs directed at improving the diagnostic capabilities of primary care practitioners. J Contin Educ Health Prof. 2016, 36:32-7. 10.1097/CEH.0000000000000061
53. Kobayashi M, Katayama M, Hayashi T, Hashiyama T, Iyanagi T, Une S, Honda M: Effect of multimodal comprehensive communication skills training with video analysis by artificial intelligence for physicians on acute geriatric care: a mixed-methods study. BMJ Open. 2023, 13:e065477. 10.1136/bmjopen-2022-065477
54. Price DW, Wingrove P, Bazemore A: Machine learning to identify clusters in family medicine diplomate motivations and their relationship to continuing certification exam outcomes: findings and potential future implications. J Am Board Fam Med. 2024, 37:279-89. 10.3122/jabfm.2023.230369R1
55. Cross J, Robinson R, Devaraju S, et al.: Transforming medical education: assessing the integration of ChatGPT into faculty workflows at a Caribbean medical school. Cureus. 2023, 15:e41399. 10.7759/cureus.41399
56. Yi PK, Ray ND, Segall N: A novel use of an artificially intelligent Chatbot and a live, synchronous virtual question-and answer session for fellowship recruitment. BMC Med Educ. 2023, 23:152. 10.1186/s12909-022-03872-z
57. Abbott KL, George BC, Sandhu G, et al.: Natural language processing to estimate clinical competency committee ratings. J Surg Educ. 2021, 78:2046-51. 10.1016/j.jsurg.2021.06.013
58. Latifi S, Gierl MJ, Boulais AP, De Champlain AF: Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination. Eval Health Prof. 2016, 39:100-13. 10.1177/0163278715605358
59. Uemura M, Tomikawa M, Miao T, et al.: Feasibility of an AI-based measure of the hand motions of expert and novice surgeons. Comput Math Methods Med. 2018, 2018:9873273. 10.1155/2018/9873273
60. Baloul MS, Yeh VJ, Mukhtar F, et al.: Video commentary & machine learning: tell me what you see, I tell you who you are. J Surg Educ. 2022, 79:e263-72. 10.1016/j.jsurg.2020.09.022
61. Johnsson V, Søndergaard MB, Kulasegaram K, et al.: Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. Med Educ. 2024, 58:105-17. 10.1111/medu.15190
62. Ryder CY, Mott NM, Gross CL, et al.: Using artificial intelligence to gauge competency on a novel laparoscopic training system. J Surg Educ. 2024, 81:267-74. 10.1016/j.jsurg.2023.10.007
63. Mahtani AU, Reinstein I, Marin M, Burk-Rafel J: A new tool for holistic residency application review: using natural language processing of applicant experiences to predict interview invitation. Acad Med. 2023, 98:1018-21. 10.1097/ACM.0000000000005210

64. Triola MM, Reinstein I, Marin M, Gillespie C, Abramson S, Grossman RI, Rivera R Jr: Artificial intelligence screening of medical school applications: development and validation of a machine-learning algorithm. Acad Med. 2023, 98:1036-43. 10.1097/ACM.0000000000005202
65. Burk-Rafel J, Reinstein I, Feng J, et al.: Development and validation of a machine learning-based decision support tool for residency applicant screening and review. Acad Med. 2021, 96:S54-61. 10.1097/ACM.0000000000004317
66. Keir G, Hu W, Filippi CG, Ellenbogen L, Woldenberg R: Using artificial intelligence in medical school admissions screening to decrease inter- and intra-observer variability. JAMIA Open. 2023, 6:ooad011. 10.1093/jamiaopen/ooad011
67. Kibble J, Plochocki J: Comparing machine learning models and human raters when ranking medical student performance evaluations. J Grad Med Educ. 2023, 15:488-93. 10.4300/JGME-D-22-00678.1
68. Koleilat I, Bongu A, Chang S, Nieman D, Priolo S, Patel NM: Residency application selection committee discriminatory ability in identifying artificial intelligence-generated personal statements. J Surg Educ. 2024, 81:780-5. 10.1016/j.jsurg.2024.02.009
69. Drum B, Shi J, Peterson B, Lamb S, Hurdle JF, Gradick C: Using natural language processing and machine learning to identify internal medicine-pediatrics residency values in applications. Acad Med. 2023, 98:1278-82. 10.1097/ACM.0000000000005352
70. Lin CC, Akuhata-Huntington Z, Hsu CW: Comparing ChatGPT's ability to rate the degree of stereotypes and the consistency of stereotype attribution with those of medical students in New Zealand in developing a similarity rating test: a methodological study. J Educ Eval Health Prof. 2023, 20:17. 10.3352/jeehp.2023.20.17
71. Walkowski S, Lundin M, Szymas J, Lundin J: Exploring viewing behavior data from whole slide images to predict correctness of students' answers during practical exams in oral pathology. J Pathol Inform. 2015, 6:28. 10.4103/2153-3539.158057
72. Chen Y, Wrenn J, Xu H, Spickard A II, Habermann R, Powers J, Denny JC: Automated assessment of medical students' clinical exposures according to AAMC geriatric competencies. AMIA Annu Symp Proc. 2014, 2014:375-84.
73. Ba H, Zhang L, Yi Z: Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods. BMC Med Educ. 2024, 24:558. 10.1186/s12909-024-05565-1
74. Spadafore M, Yilmaz Y, Rally V, et al.: Using natural language processing to evaluate the quality of supervisor narrative comments in competency-based medical education. Acad Med. 2024, 99:534-40. 10.1097/ACM.0000000000005634
75. Gin BC, Ten Cate O, O'Sullivan PS, Hauer KE, Boscardin C: Exploring how feedback reflects entrustment decisions using artificial intelligence. Med Educ. 2022, 56:303-11. 10.1111/medu.14696
76. Mansour J, Burman M, Bernstein M, Sandman E, Yammine K, Daher M, Martineau PA: Should my recommendation letter be written by artificial intelligence?. Can J Surg. 2024, 67:E243-6. 10.1503/cjs.009623
77. Lindqwister AL, Hassanpour S, Lewis PJ, Sin JM: AI-RADS: an artificial intelligence curriculum for residents . Acad Radiol. 2021, 28:1810-6. 10.1016/j.acra.2020.09.017
78. Lindqwister AL, Hassanpour S, Levy J, Sin JM: AI-RADS: successes and challenges of a novel artificial intelligence curriculum for radiologists across different delivery formats. Front Med Technol. 2022, 4:1007708. 10.3389/fmedt.2022.1007708
79. Hu R, Rizwan A, Hu Z, Li T, Chung AD, Kwan BY: An artificial intelligence training workshop for diagnostic radiology residents. Radiol Artif Intell. 2023, 5:e220170. 10.1148/ryai.220170
80. Mosch L, Agha-Mir-Salim L, Sarica MM, Balzer F, Poncette AS: Artificial intelligence in undergraduate medical education. Challenges of Trustable AI and Added-Value on Health. IOS Press, Amsterdam, The Netherlands; 2022. 294:821-2. 10.3233/SHTI220597
81. 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. 10.2196/46500
82. Tsopra R, Peiffer-Smadja N, Charlier C, et al.: Putting undergraduate medical students in AI-CDSS designers' shoes: an innovative teaching method to develop digital health critical thinking. Int J Med Inform. 2023, 171:104980. 10.1016/j.ijmedinf.2022.104980
83. Fan KY, Hu R, Singla R: Introductory machine learning for medical students: a pilot . Med Educ. 2020, 54:1042-3. 10.1111/medu.14318
84. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R: Grounded in reality: artificial intelligence in medical education. JAMIA Open. 2023, 6:ooad037. 10.1093/jamiaopen/ooad037
85. Hedderich DM, Keicher M, Wiestler B, et al.: AI for doctors-a course to educate medical professionals in artificial intelligence for medical imaging. Healthcare (Basel). 2021, 9:1278. 10.3390/healthcare9101278
86. Laupichler MC, Hadizadeh DR, Wintergerst MW, von der Emde L, Paech D, Dick EA, Raupach T: Effect of a flipped classroom course to foster medical students' AI literacy with a focus on medical imaging: a single group pre-and post-test study. BMC Med Educ. 2022, 22:803. 10.1186/s12909-022-03866-x
87. Wang D, Huai B, Ma X, et al.: Application of artificial intelligence-assisted image diagnosis software based on volume data reconstruction technique in medical imaging practice teaching. BMC Med Educ. 2024, 24:405. 10.1186/s12909-024-05382-6
88. Agha-Mir-Salim L, Mosch L, Klopfenstein SA, Wunderlich MM, Frey N, Poncette AS, Balzer F: Artificial intelligence competencies in postgraduate medical training in Germany. Challenges of Trustable AI and Added-Value on Health. IOS Press, Amsterdam, The Netherlands; 2022. 294:805-6. 10.3233/SHTI220589
89. Hegghammer T: OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. J Comput Soc Sci. 2022, 5:861-82. 10.1007/s42001-021-00149-1
90. Doumat G, Daher D, Ghanem NN, Khater B: Knowledge and attitudes of medical students in Lebanon toward artificial intelligence: a national survey study. Front Artif Intell. 2022, 5:1015418. 10.3389/frai.2022.1015418
91. Gabbard T, Romanelli F: The accuracy of health professions students' self-assessments compared to objective measures of competence. Am J Pharm Educ. 2021, 85:8405. 10.5688/ajpe8405