QMJ
A peer reviewed journal
QATAR MEDICAL JOURNAL

RESEARCH PAPER

Performance of ChatGPT in emergency medicine residency exams in Qatar: A comparative analysis with resident physicians

Haris Iftikhar[1]*, Shahzad Anjum[1], Zain A. Bhutta[1], Mavia Najam[2], Khalid Bashir[1]

Address for Correspondence:
Haris Iftikhar[1]*

[1]Emergency Medicine, Hamad General Hospital, Doha, Qatar

[2]Department of Medical Education, Hamad Medical Corporation, Doha, Qatar

*Email: haris.ifti@gmail.com

https://doi.org/10.5339/qmj.2024.61

Submitted: 28 November 2023
Accepted: 09 September 2024
Published: 11 November 2024

Cite this article as: Iftikhar H, Anjum S, Bhutta ZA, Najam M, Bashir K. Performance of ChatGPT in emergency medicine residency exams in Qatar: A comparative analysis with resident physicians, Qatar Medical Journal 2024(4):61 https://doi.org/10.5339/qmj.2024.61

QSCIENCE
HAMAD BIN KHALIFA UNIVERSITY PRESS

ABSTRACT

Introduction: The inclusion of artificial intelligence (AI) in the healthcare sector has transformed medical practices by introducing innovative techniques for medical education, diagnosis, and treatment strategies. In medical education, the potential of AI to enhance learning and assessment methods is being increasingly recognized. This study aims to evaluate the performance of OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT) in emergency medicine (EM) residency examinations in Qatar and compare it with the performance of resident physicians.

Methods: A retrospective descriptive study with a mixed-methods design was conducted in August 2023. EM residents' examination scores were collected and compared with the performance of ChatGPT on the same examinations. The examinations consisted of multiple-choice questions (MCQs) from the same faculty responsible for Qatari Board EM examinations. ChatGPT's performance on these examinations was analyzed and compared with residents across various postgraduate years (PGY).

Results: The study included 238 emergency department residents from PGY1 to PGY4 and compared their performances with ChatGPT. ChatGPT scored consistently higher than resident groups in all examination categories. However, a notable decline in passing rates was observed among senior residents, indicating a potential misalignment between examination performance and practical competencies. Another likely reason can be the impact of the COVID-19 pandemic on

their learning experience, knowledge acquisition, and consolidation.

Conclusion: ChatGPT demonstrated significant proficiency in the theoretical knowledge of EM, outperforming resident physicians in examination settings. This finding suggests the potential of AI as a supplementary tool in medical education.

1. INTRODUCTION

The integration of artificial intelligence (AI) in healthcare has revolutionized medical practices, introducing novel approaches to medical education, diagnosis, and treatment strategies. Among the groundbreaking advancements, OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT) stands out for its remarkable text generation and interpretation capabilities. This study focuses on evaluating the performance of ChatGPT in emergency medicine (EM) residency examinations in Qatar, juxtaposed against resident physicians' performance.

In the realm of medical education, AI's potential to enhance learning and assessment methods has been increasingly acknowledged. Studies have shown that AI, especially in the form of large language models like ChatGPT, holds promise in various medical educational contexts.[1,2] The ability of AI to process and analyze vast amounts of data offers an unparalleled advantage in creating a more dynamic and interactive learning environment. This is particularly relevant in EM, a field that demands quick decision-making and a thorough understanding of diverse medical scenarios.[3]

AI's role in educational assessment, particularly in the format of multiple-choice questions (MCQs), is a burgeoning field of interest. MCQs are a widely used tool in medical examinations due to their objectivity and ability to cover a broad range of topics efficiently.[4] The effectiveness of AI in generating and answering MCQs can provide insights into its potential as an educational tool in medicine. Furthermore, the comparison of AI's performance with that of human trainees offers valuable perspectives on the strengths and limitations of both, and how they can complement each other in medical education.[5,6]

EM, characterized by its fast-paced and high-pressure environment, requires residents to rapidly assimilate a wide range of medical knowledge and apply it effectively. The training programs in Qatar have been designed to equip residents with the necessary skills and knowledge to manage critical cases efficiently.[3] However, there is a paucity of research on the application of AI in this specific medical field, especially in the Middle Eastern context. This study aims to bridge this gap by analyzing and comparing the performances of ChatGPT and EM residents in residency examinations.

2. METHODS

This retrospective descriptive study, employing a mixed-methods approach, was initiated following the receipt of ethical approval on August 23, 2023 (MRC 01-23-502) from Hamad Medical Corporation (HMC) Medical Research Centre. We adopted a convenience sampling strategy for our research. The inclusion criteria for the study were straightforward: we incorporated the scores of all EM residents enrolled in the Department of Emergency Medicine Residency Training Program. Conversely, we excluded residents whose scores were unavailable in the program database due to absence from the examination for reasons like annual, casual, or sick leave, ensuring a comprehensive and representative sample of current trainees.

In terms of data collection, we collated anonymized performance data of EM trainees. This data encompassed their examination scores between October 2021 and September 2022, consisting of a total of five examinations (Oct-21, Dec-21, Feb-22, Jun-22, and Sep-22). The examinations in question consisted of single-best-answer MCQs devised by faculty members who were also involved in constructing the Qatari Board EM examinations. Each examination consists of 40 MCQ questions, and every question has four options to choose from. The examinations were conducted using paper and pencil in a designated examination hall, where residents had 60 minutes to complete the examination. All residents from various postgraduate years (PGY) levels took the same examination simultaneously. The topics of each

examination were distinct, and each year, new examinations were designed by updating the question bank with evidence-based information from the core faculty and examination committee. For the validation of correct answers in MCQ examinations, a comprehensive approach was taken. EM program directors and core faculty, who are subject matter experts, initially reviewed and aligned questions with the curriculum, followed by rigorous cross-referencing with authoritative medical sources. These questions underwent a thorough peer review and were pilot tested with all residents yearly to assess clarity and difficulty. The peer-reviewed and pilot-tested MCQs are added to the department's confidential MCQ pool, from which future examinations are made. After the examinations, feedback from participants and educators also facilitated regular updates, ensuring the examination's reliability in reflecting the evolving field of EM.

In a unique aspect of our study, these examinations were administered to ChatGPT 4.0 (paid version) in May 2023 for performance assessment. During this time, ChatGPT 4.0 did not possess image recognition capabilities; therefore, any images in the examinations were converted into text descriptions for evaluation. MCQs were transcribed into ChatGPT in batches of 10, following which the answers provided by the AI-powered assistant were verified, documented, and evaluated. The performance of ChatGPT was then scored and juxtaposed against the scores of EM residents across different PGY, ranging from PGY1 to PGY4. This comparison, sanctioned by the Institutional Review Board (IRB), aimed to critically evaluate ChatGPT's proficiency and familiarity with current EM clinical guidelines, as cited in references.[7,8]

For the statistical analysis of our data, we employed STATA software. The analysis included using linear regression to assess associations between category-level scores and examining performance differences utilizing Chi-square tests, Fisher's exact test, and univariable logistic regression. Throughout this analytical process, we maintained a significance level of $P < 0.05$. It is important to note that our study meticulously adhered to the strengthening the reporting of observational studies in epidemiology (STROBE) guidelines to ensure the robustness and reliability of our research findings.

3. RESULTS

In this study, we evaluated the performances of 238 EM residents from the emergency department, spanning from PGY1 to PGY4, and compared them with the AI language model, ChatGPT, as detailed in Table 1. The participant demographics were diverse, encompassing 58 PGY1 residents (23.8%), 61 PGY2 residents (25.1%), 66 PGY3 residents (27.2%), and 53 PGY4 residents (21.8%), with a gender distribution showing a male-to-female ratio of approximately 2:1. The minimum passing scores for each PGY level were as follows: PGY4: 60%, PGY3: 55%, PGY2: 50%, and PGY1: 45%. Notably, each examination comprised 40 questions, and the maximum score for each was 40.

Focusing on the performance of EM trainees, the results were indicative of a progressive learning curve throughout the residency program. PGY1 residents, at the outset of their training, scored an average of 18 ± 3.5, laying down a foundational knowledge base. This was followed by a slight improvement among PGY2 residents, who achieved an average score of 19.4 ± 3.2. The upward trend continued with PGY3 residents, who demonstrated further progress, scoring an average of 21.1 ± 3.8. The most senior group, the PGY4 residents, attained the highest average score of 21.9 ± 4.2, reflecting their continuous learning and skill development throughout the residency program.

A remarkable aspect of our study was the comparison with ChatGPT. Surprisingly, ChatGPT outperformed all resident groups, scoring an impressive average of 25.8 ± 2.6. Statistical analysis underscored significant differences in performance between the EM trainees and ChatGPT. This difference pointed towards ChatGPT's advanced proficiency in medical knowledge and problem-solving.[9,10]

An intriguing pattern emerged when analyzing the passing rates trend. Contrary to expectations, 64.7% of PGY1 residents passed the examinations,

Table 1. Comparison of marks obtained and percentages for trainees and ChatGPT for examination sessions.

Variable	Category	PGY1 *N* = 58	PGY2 *N* = 61	PGY3 *N* = 66	PGY4 *N* = 53	GPT *N* = 5	Total	*p*-value
Gender, *n* (%)	Male	28 (48.3)	43 (70.5)	55 (83.3)	39 (73.5)		165 (69.3)	
	Female	30 (51.7)	18 (29.5)	11 (16.6)	14 (26.4)		73 (30.6)	
Obtained mean examination score, Mean ± SD	Total	18 ± 3.5	19.4 ± 3.2	21.1 ± 3.8	21.9 ± 4.2	25.8 ± 2.6	20.4 ± 3.9	< 0.05
	Oct-21	14.4 ± 2.9	17.1 ± 2.5	21.3 ± 2.3	22.7 ± 3.7	23	19.1 ± 4.2	
	Dec-21	20.8 ± 1.9	21.3 ± 3.4	21.9 ± 3.1	23.8 ± 2.1	26	21.9 ± 2.7	
	Feb-22	20.8 ± 2.8	22.3 ± 2.8	24.7 ± 4.1	25.9 ± 2.4	25	23.4 ± 3.6	
	Jun-22	17.9 ± 1.8	18.3 ± 2.8	19.2 ± 2.1	18.6 ± 4.2	25	18.6 ± 2.8	
	Sep-22	17.5 ± 2.2	18.3 ± 2.5	17.3 ± 3	18.5 ± 3.4	30	18.2 ± 3.3	
Percentage mean examination score, Mean ± SD	Total	46.4 ± 8.4	48.6 ± 8.1	52.8 ± 9.6	54.7 ± 10.6	65.5 ± 6.5	51.0 ± 9.9	< 0.05
	Oct-21	36.3 ± 7.2	42.7 ± 6.3	53.2 ± 5.7	56.7 ± 9.1	57.5	47.7 ± 10.4	
	Dec-21	52.1 ± 4.8	53.3 ± 5.9	54.8 ± 7.8	59.5 ± 5.2	65	54.9 ± 6.7	
	Feb-22	52.1 ± 7.1	55.6 ± 7.2	61.7 ± 10.1	64.7 ± 5.9	67.5	58.6 ± 9.1	
	Jun-22	44.7 ± 4.5	45.7 ± 7.2	47.9 ± 5.1	46.5 ± 10.4	62.5	46.6 ± 7.1	
	Sep-22	43.7 ± 5.4	45.7 ± 6.2	43.2 ± 7.5	46.2 ± 8.6	75	45.6 ± 8.4	
Status, *n* (%)	Pass	33 (64.7)	28 (47.5)	26 (40.6)	21 (42)		108 (48.2)	
	Fail	18 (35.3)	31 (52.5)	38 (59.4)	29 (58)		116 (51.7)	
	Absent	7 (12.1)	2 (3.3)	2 (3.0)	3 (5.6)			
Pass, *n* (% out of passed candidates)	Oct-21	2 (6.1)	2 (7.1)	6 (23.1)	4 (19.1)		14 (30.4)	
	Dec-21	12 (36.3)	8 (28.6)	8 (30.7)	6 (28.6)		34 (70.8)	
	Feb-22	10 (30.3)	9 (32.1)	10 (38.5)	9 (42.3)		38 (79.2)	
	Jun-22	7 (21.2)	5 (17.8)	2 (7.7)			14 (35)	
	Sep-22	2 (6.1)	4 (14.3)		2 (9.5)		8 (19.1)	

PGY: postgraduate years; *n*: Number; SD: standard deviation.

whereas only 42% of PGY4 residents achieved passing scores.

4. DISCUSSION

The findings of this study offer a compelling insight into the capabilities of AI, particularly ChatGPT, in the context of EM education. ChatGPT's performance, outstripping that of resident physicians across all PGY, marks a significant milestone in the intersection of AI and medical education. This achievement corroborates the growing body of research underscoring the potential of AI as a supplementary tool in various educational settings, including medical training.[1]

The study used a retrospective descriptive methodology with a mixed-methods design, providing a comprehensive analysis of the performances of both ChatGPT and EM residents. By including a wide range of participants across different PGY and comparing their performances on standardized examinations, the study offers a robust evaluation of ChatGPT's capabilities in the context of EM. This approach aligns with the current trends in medical education research, which emphasize the need for innovative methods to enhance learning and assessment processes.[10,11]

The consistent improvement in scores across the EM training years highlights the efficacy of the existing residency programs in Qatar in enhancing medical knowledge and clinical decision-making skills.[3] However, the performance of ChatGPT indicates the added value that AI can bring to medical education, especially in standard theoretical knowledge and problem-solving abilities. This observation is in line with studies demonstrating AI's proficiency in processing vast amounts of data and applying knowledge in diverse scenarios.[1,12]

The decline in passing rates among senior residents raises questions about the alignment between examination formats and the practical competencies developed throughout the residency. This suggests a potential gap in the assessment methods used in EM training, where theoretical knowledge may not entirely capture the complexities of real-life medical scenarios. This finding echoes the sentiments of recent literature advocating for a more holistic approach in medical education, emphasizing the integration of practical skills and humanistic qualities alongside theoretical knowledge.[5,13] Another likely reason can be the impact of the COVID-19 pandemic on their learning experience, knowledge acquisition, and consolidation. To effectively prepare for future pandemics, we must prioritize the safety and security of our teaching environment while ensuring residents' well-being. This can only be achieved through rigorous safeguards, including adequate support in medical education and supervision.[14,15]

Recently, ChatGPT demonstrated remarkable success in medical licensing and other examinations and continues to improve. ChatGPT achieved a passing score for a third-year medical student by surpassing the 60% threshold on the NBME-Free Step-1 dataset.[6] GPT-4 surpassed GPT-3 and successfully passed all six years of the Japanese medical examinations, demonstrating its potential in a different language.[9] ChatGPT performed well on the European Examination in Core Cardiology (EECC) multiple choice questions, consistently scoring above 60%, which was usually the pass mark for the candidates.[16] In the European Board Examination of the Nuclear Medicine Section, ChatGPT was put to the test with fifty MCQs. ChatGPT provided a definitive answer 34% of the time. However, the mean probability of randomly selecting the correct answer was only 0.24, indicating that ChatGPT possesses some knowledge rather than simply guessed.[1,8]

Completing medical registration in Germany involves passing three state examinations. The first examination (M1) covers pre-clinical subjects, while the second examination (M2) is a written test that assesses medical specialties. ChatGPT passed German medical examinations M1 and M2 with 60.1% and 66.7% with a passing grade on each exam.[8]

In another study of 50 questions from EM, human test-takers achieved a mean correct response rate of 83.7%, outperforming GPT-3.5 (61.4%) but not GPT-4 (82.1%). The study found a statistically significant difference in performance between GPT-3.5 and both human beings and GPT-4 (mean difference: 21.5%), while no statistically significant

difference was observed between human performance and GPT-4 (mean difference: 1.6%).[17]

A study on the bar examination found that GPT-4 can pass the uniform bar examination without prior training, challenging the assumption that domain-specific models would struggle. With a deep understanding of legal concepts and excellent reading and writing skills, large language models like GPT-4 can meet the same standard as human lawyers.[10]

The study's results underscore the need for further research into the application of AI in medical education, particularly in specialties like EM that require a unique blend of knowledge, skills, and decision-making abilities. While ChatGPT has shown proficiency in theoretical knowledge, its current limitation in image analysis, a crucial aspect of EM diagnostics, points towards areas needing improvement. Future iterations of AI tools capable of interpreting clinical images could significantly enhance their utility in medical education and practice.[3,18]

Moreover, the study's findings highlight the dynamic nature of AI development and the challenges it poses in establishing a consistent assessment framework. As AI tools rapidly evolve, keeping pace with their capabilities and integrating them effectively into medical education becomes imperative. This aligns with recent discussions on the role of AI in reshaping educational paradigms, advocating for adaptive and flexible teaching and assessment strategies that can accommodate the rapid advancements in technology.[19,20]

ChatGPT achieved an impressive average score of 25.8 ± 2.6 out of 40. However, it is still far from perfect considering its percentage out of the total. Even though advanced, ChatGPT still does not fully grasp the complexities of medical scenarios, where decision-making often involves synthesizing information from diverse sources.[21] There is a concern that AI and ChatGPT may inadvertently reinforce cognitive biases in its training data, leading to potential errors in clinical reasoning and decision-making.[22] It is unclear whether AI models can process real-time patient data or adjust their recommendations dynamically during ongoing care.[23] Additionally, AI might lack the ability to make ethical judgments, which is often required in EM, such as in cases involving triage, end-of-life care, or complex consent issues.[24] There is a risk that AI models might generate recommendations that are not evidence-based or that contradict established medical guidelines, potentially leading to harmful outcomes if relied upon inappropriately.[25] While ChatGPT has the potential for benefits in EM education, it is important to be cautious as it also has limitations in clinical training and practice.

The study not only demonstrates ChatGPT's formidable capabilities in understanding and applying medical knowledge but also suggests a potential transformative role for AI in medical education and healthcare in general. The integration of AI, with its unique strengths and evolving capabilities, alongside traditional training methods, could significantly enhance the learning experience and the overall quality of medical education. However, the art of medicine, with its inherent humanistic qualities and practical skills, remains a domain that AI cannot replace. The complementary use of AI and human expertise could be the key to optimizing medical education and improving healthcare outcomes.[11,17,26]

We want to acknowledge that there are several limitations in our study. The study was conducted in a single facility and program, which may limit the generalizability of the findings. Future research should aim for a more diverse sample to ensure broader applicability and relevance across various healthcare settings. This study only included MCQs, and we cannot compare the performance of ChatGPT on other types of questions, such as short-answer questions (SAQ) or objective structured clinical examinations (OSCEs), which are also commonly used in international medical examinations.[27–29] Furthermore, at the time of the study, ChatGPT did not have image recognition capabilities, so we could not assess its image recognition abilities and knowledge. Lastly, we should also consider that the potential impact of the pandemic may have affected different groups of residents at various stages of their medical education, which may have affected their scores and influenced ChatGPT's performance comparison.

5. CONCLUSION

The study analyzed the performance of EM trainees during their residency program and found that their knowledge and scores improved as they progressed through the program. However, an AI-powered language model, ChatGPT, outperformed all the resident groups in medical knowledge and clinical problem-solving.

In summary, this study not only contributes to the emerging body of literature on AI in medical education but also provides practical insights that could shape future training and assessment strategies in EM. As AI continues to evolve and integrate into various aspects of healthcare, understanding its potential and limitations becomes crucial for optimizing its benefits in medical education and practice.

In light of these findings, future research should focus on exploring the practical applications of AI in various medical specialties, assessing its impact on both theoretical and practical competencies, and developing strategies to integrate AI effectively into medical training programs. The ultimate goal is to harness the full potential of AI in enhancing the quality of medical education and healthcare delivery, thereby benefiting patients and healthcare providers alike.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

NA.

AUTHORSHIP DECLARATION

All authors agree with the content of the manuscript.

MRC APPROVAL

Ethical approval was obtained from the Hamad Medical Corporation's Institutional Review Board (MRC 01-23-502).

AUTHORS' CONTRIBUTIONS

HI, SA, MN, KB: Conceptualization; Writing—original draft preparation; Data curation; Funding acquisition; Methodology; Project administration. ZB: Formal analysis; Validation; Writing—review and editing.

REFERENCES

1. Alberts IL, Mercolli L, Pyka T, Prenosil G, Shi K, Rominger A, et al. Large language models (LLM) and ChatGPT: What will the impact on nuclear medicine be? *Eur J Nucl Med Mol Imaging*. 2023;50(6):1549–52. https://doi.org/10.1007/s00259-023-06172-w
2. Bashir K, Azad AM, Hereiz A, Bashir MT, Masood M, Elmoheen A. Current use, perceived barriers, and learning preference of point of care ultrasound (Pocus) in the emergency medicine in qatar – A mixed design. *Open Access Emerg Med*. 2021;13:177–82. https://doi.org/10.2147/OAEM.S304153
3. Chartier C, Gfrerer L, Knoedler L, Austen WG. Artificial intelligence-enabled evaluation of pain sketches to predict outcomes in headache surgery. *Plast Reconstr Surg*. 2023;151(2):405–11. https://doi.org/10.1097/PRS.0000000000009855
4. Moss E. Multiple choice questions: Their value as an assessment tool. *Curr Opin Anaesthesiol*. 2001;14(6):661–6. https://doi.org/10.1097/00001503-200112000-00011
5. Deng J, Lin Y. The benefits and challenges of ChatGPT: An overview. *Front Comput Intell Syst*. 2023;2(2):81–3. https://doi.org/10.54097/fcis.v2i2.4465
6. 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;9:1. https://doi.org/10.2196/45312
7. Haque MU, Dharmadasa I, Sworna ZT, Rajapakse RN, Ahmad H. I think this is the most disruptive technology: Exploring sentiments of ChatGPT early

adopters using Twitter data. 2022 Dec 12 [cited 2023 Dec 20]. Available from: https://arxiv.org/abs/2212.05856v1
8. 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;120(21–22):373–4. https://doi.org/10.3238/arztebl.m2023.0113
9. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. 2023. Available from: http://arxiv.org/abs/2303.18027
10. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 passes the bar exam. *SSRN Electr J*. 2023;4389233.
11. Zielinski C, Winker M, Aggarwal R, Ferris L, Heinemann M, Lapeña JF, et al. Chatbots, ChatGPT, and Scholarly Manuscripts: WAME recommendations on ChatGPT and Chatbots in relation to scholarly publications. *Natl Med J India*. 2023 Jan 1 [cited 2023 Dec 20];36(1):1–4. https://doi.org/10.25259/NMJI_365_23
12. Wu J, Wu X, Qiu Z, Li M, Zheng Y, Yang J. Qualifying Chinese medical licensing examination with knowledge enhanced generative pre-training model. 2023. Available from: http://arxiv.org/abs/2305.10163
13. Naaz S, Asghar A. Artificial intelligence, nano-technology and genomic medicine: The future of anaesthesia. *J Anaesthesiol Clin Pharmacol*. 2022;38(1):11–7. https://doi.org/10.4103/joacp.JOACP_139_20
14. Ferrel MN, Ryan JJ. The impact of COVID-19 on medical education. *Cureus*. 2020;12:3. https://doi.org/10.7759/cureus.7492
15. Harky A, Karimaghaei D, Katmeh H, Hewage S. The impact of COVID-19 on medical examinations. *Acta Biomed*. 2020;91(4):e2020135. https://doi.org/10.23750/abm.v91i4.10487
16. 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;4(3):279–81. https://doi.org/10.1093/ehjdh/ztad029
17. Jarou ZJ, Dakka A, McGuire D, Bunting L. ChatGPT Versus Human performance on emergency medicine board preparation questions. *Ann Emerg Med*. 2023;83(1):87–88. https://doi.org/10.1016/j.annemergmed.2023.08.010
18. Knoedler L, Odenthal J, Prantl L, others. (2023). Artificial intelligence-enabled simulation of gluteal augmentation: A helpful tool in preoperative outcome simulation? *J Plast Reconstr Aesthet Surg*. 80:94–101. https://doi.org/10.1016/j.bjps.2023.01.039
19. King MR. The future of AI in medicine: A perspective from a Chatbot. *Ann Biomed Eng*. 2023;51(2):291–295. https://doi.org/10.1007/s10439-022-03121-w
20. Zhai X. ChatGPT user experience: Implications for education. *SSRN Electron J*. 2023;4312418.
21. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018 Nov 11 [cited 2024 Aug 15];178(11):1544. https://doi.org/10.1001/jamainternmed.2018.3763
22. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017 Aug 8;318(6):517–8. https://doi.org/10.1001/jama.2017.7797
23. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019 Jan;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7
24. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, Floridi L. The ethics of AI in health care: A mapping review. *Soc Sci Med*. 2020 Sep 1;260:113172. https://doi.org/10.1016/j.socscimed.2020.113172
25. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018 Oct;2(10):719–31. https://doi.org/10.1038/s41551-018-0305-z
26. Stokel-Walker C. ChatGPT listed as author on research papers: Many scientists disapprove. *Nature*. 2023;613(7945):620–1. https://doi.org/10.1038/d41586-023-00107-z
27. Majumder MA, 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 5:387–97. https://doi.org/10.2147/AMEP.S197275
28. 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 Jan 1;24(1):1649959. https://doi.org/10.1080/10872981.2019.1649959
29. Pham H, Trigg M, Wu S, O'Connell A, Harry C, Barnard J, Devitt P. Choosing medical assessments: Does the multiple-choice question make the grade? *Educ Health (Abingdon)*. 2018 May 1;31(2). https://doi.org/10.4103/efh.EfH_229_17