Journal of the American Medical Informatics Association, 2023, **30(9)**, 1558–1560
https://doi.org/10.1093/jamia/ocad104
Advance access publication 19 June 2023
Brief Communications

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD

Brief Communications

ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification?

Yaa Kumah-Crystal[1,*], Scott Mankowitz[2], Peter Embi[3], and Christoph U. Lehmann[4]

[1]Department of Biomedical Informatics, Pediatric Endocrinology, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[2]Department of Clinical Informatics, Overlook Medical Center, Summit, New Jersey, USA
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[4]Clinical Informatics Center, UT Southwestern Medical Center, Dallas, Texas, USA

*Corresponding Author: Yaa Kumah-Crystal, MD, MPH, Department of Biomedical Informatics, Pediatric Endocrinology, Vanderbilt University Medical Center, 3401 West End Ave, Suite 600 Office 6023, Nashville, TN 37203, USA; yaa.kumah@vumc.org

ABSTRACT

We aimed to assess ChatGPT's performance on the Clinical Informatics Board Examination and to discuss the implications of large language models (LLMs) for board certification and maintenance. We tested ChatGPT using 260 multiple-choice questions from Mankowitz's Clinical Informatics Board Review book, omitting 6 image-dependent questions. ChatGPT answered 190 (74%) of 254 eligible questions correctly. While performance varied across the Clinical Informatics Core Content Areas, differences were not statistically significant. ChatGPT's performance raises concerns about the potential misuse in medical certification and the validity of knowledge assessment exams. Since ChatGPT is able to answer multiple-choice questions accurately, permitting candidates to use artificial intelligence (AI) systems for exams will compromise the credibility and validity of at-home assessments and undermine public trust. The advent of AI and LLMs threatens to upend existing processes of board certification and maintenance and necessitates new approaches to the evaluation of proficiency in medical education.

Key words: large language models, ChatGPT, Clinical Informatics Board Examination, medical education, artificial intelligence

BACKGROUND AND SIGNIFICANCE

The use of large language models (LLMs) such as OpenAI's ChatGPT[1] has demonstrated remarkable potential in answering knowledge questions and passing exams. This is exemplified by its recent accomplishments in passing the United States Medical Licensing Examination (USMLE), where GPT-3[2,3] surpassed the 60% passing threshold, and GPT-4 subsequently achieved an impressive accuracy rate of over 80%.[4] While the USMLE Step 2 and 3 exams are designed to assess the application of a broad range of general medical knowledge, the Clinical Informatics Board Exam (CIBE) focuses on health information technology.[5,6] Exam questions can be categorized into recall, application, and rationalizing questions. Recall questions are likely easiest for a LLMs to address; however, a high proportion of application (eg, apply decision support principles on a presented problem) and rationalizing (eg, determine the output from a programming code sample) questions may require application of domain-specific principles and expertise, which could be more difficult for an LLM like ChatGPT to solve especially if the questions present unique and novel challenges not included in the training set.

High-stakes exams, such as the USMLE and the initial certifying exam for clinical informatics boards are typically proctored as they play a vital role in determining certification outcomes and serve as essential gateways to professional advancement and recognition. In contrast, maintenance of certification (MOC) for the CIBE and various other medical certifications have shifted to a self-paced, remote format that fosters continued proficiency.[7] The widespread availability and accessibility of tools like ChatGPT, however, raise concerns about the potential misuse of such resources during nonproctored MOC exams. Such misuse could invalidate the testing apparatus and subsequently challenge credentialing organizations' ability to accurately evaluate an individual's continued expertise and proficiency in their respective fields.[8]

OBJECTIVES

Our study assessed ChatGPT's ability to pass practice exams for the CIBE and its performance in the core competencies of the CIBE. We discuss the potential implications of using LLMs for board certification and maintenance.

MATERIALS AND METHODS

We used a corpus of 260 multiple-choice questions from Mankowitz' Clinical Informatics Board Review book published in 2018.[8] The questions represent the knowledge areas tested on the examination administered by the American Board of Preventive Medicine. Questions were categorized according to the Core Content for the Subspecialty of Clinical Informatics.[9,10] Questions depending on the use of images were omitted. Each question was entered into ChatGPT 3.5 with a brief preamble requesting justification why the answer suggested by ChatGPT was correct. The question was

Received: 23 March 2023. **Revised:** 31 May 2023. **Editorial Decision:** 1 June 2023. **Accepted:** 8 June 2023

considered answered correctly if ChatGPT could identify the answer that correlated to the book's answer key.

RESULTS

Of 260 questions, 6 (3%) were excluded because they relied on visual stimuli to deliver the context, leaving 254 (97%) questions available for analysis. Of the remaining 254 questions, ChatGPT answered 190 (74%) correctly. Categorized based on the Clinical Informatics Core Content Areas schema, ChatGPT performed from best to worst in (1) Fundamental Knowledge and Skills (85%), (2) Leadership and Professionalism (76%), (3) Data Governance and Data Analytics (74%), (4) Enterprise Information Systems (72%), and (5) Improving Care Delivery and Outcomes (71%). Chi-square analysis did not reveal any statistically significant differences across these categories ($X^2(4) = 0.59$, $P = .96$; Table 1).

Table 1. ChatGPT's performance on categories of CIBE questions

Clinical informatics category	Correct/total
Fundamental knowledge and skills	28/33 (85%)
Leadership and professionalism	52/68 (76%)
Data governance and data analytics	17/23 (74%)
Enterprise information systems	28/39 (72%)
Improving care delivery and outcomes	65/91 (71%)
Total	190/254 (75%)

DISCUSSION

Our study demonstrated that ChatGPT has the capability to answer multiple-choice questions with a high degree of accuracy. It is estimated that the average number of correctly answered questions on the CIBE exam is ∼60% thus ChatGPT's 74% performance hypothetically suggests its capacity to meet certification maintenance standards.[11]

In the Clinical Informatics practice set questions, the Fundamental Knowledge and Skills practice area includes more recall-based questions, while the other areas emphasize application and reasoning. ChatGPT performed slightly better in the Fundamental Knowledge and Skills section at 85%, but the difference was not statistically significant. Our observation suggests that ChatGPT may find novel questions requiring knowledge-based reasoning or application of knowledge to a new problem more challenging than recall-type questions. Future studies could use larger question sets to determine whether significance arises with an expanded sample size.

The goal of MOC exams is to reinforce concepts and principles, aiding exam-takers in understanding their skills, knowledge, and comprehension.[12] In the current, self-paced, open-book MOC exam format, primary texts and search engines such as Google are permitted to encourage reflective learning and critical thinking by enabling individuals to identify knowledge gaps and draw connections between concepts while seeking explanations.[13] This also mimics real-world scenarios in which clinicians are expected to confirm their intuition with authoritative sources.

While it has been recognized that artificial intelligence (AI) can outperform humans in computational activities like chess,[14] with some notorious reports of computer chess programs used to aid human cheating,[15] the advent of LLMs that mimic the effective application of concepts and principles in the USMLE and CIBE usher in a new era in psychometric testing. Exam takers would historically approach questions from a bottom-up perspective, where specific details of the question are analyzed, and previous knowledge and experiences are drawn upon to understand and solve the problem. Alternatively, a top-down approach can be used, where concepts are recalled from memory and reasoning is applied to answer the question. ChatGPT has shown that its training allows it to make associations that mimic the effective application of clinical informatics concepts and principles, which enables the model to derive answers requiring neither cognitive approach to be exercised.[16]

A fundamental aspect of the current reflective learning model of MOC exams that permit access to online resources is that individuals must still process and assimilate the information found online to determine the correct answer to the exam questions. However, when using LLMs like ChatGPT, exam takers can simply manually enter or automaticaly scrape the question into the freely available web interface and be given an instantaneous result. This transaction requires no prior knowledge of theory or application and eliminates the need for reflection, reasoning, and understanding but can still result in a passing score. This issue calls into question the validity and credibility of the cognitive assessment and could ultimately undermine the public's trust in the board certification process.

LIMITATIONS

Our study was limited by the fact that the sample of questions used in the study was derived from a single source, Mankowitz's Clinical Informatics Board Review book, which may not represent the full range of question types and content encountered on the actual CIBE. Additionally, questions that contained images could not be evaluated by GPT-3.5. Our assessment centered on GPT-3.5 due to its free and widespread availability and accessibility online, which could make it an appealing resource for exam-takers. Subsequent research should explore the performance of more advanced models, such as GPT-4, on board exams, as these newer iterations have yielded even higher levels of proficiency on other comparable assessments and can also interpret images.

CONCLUSION

The increasing ease of access and growing popularity of user-friendly LLMs like ChatGPT raise significant concerns regarding their use in board certification exams. Using LLMs to provide answers may undermine the validity of knowledge assessment because it requires neither subject matter expertise nor reflective learning to obtain a passing score. It is crucial, therefore, to explore new approaches to evaluating and measuring mastery.

To maintain credibility and promote learning, testing must be adapted to incentivize learning instead of solely focusing on obtaining a passing grade. This may involve more complex and novel question types, as well as introducing images or pictographs that cannot be easily interpreted by today's LLMs. However, as newer GPT models are becoming more proficient at interpreting images and producing rational responses, this may not be a permanent solution.[17] In some situations, there may be a need to consider reverting to proctored, in-person exams. However, a more innovative approach might be to develop dynamic assessment techniques that incorporate interactions with LLMs as part of the exam itself. This could showcase the users' ability to demonstrate proficiency in the evolving technology landscape as life-long learners.

In the interim, while the Board of Examiners determines how to reshape testing, it is essential to share guidelines and expectations for how users should and should not engage with LLMs during testing. Ultimately, we must ensure that the integrity of ongoing board certification exams is upheld to maintain public trust in board-certified professionals.

FUNDING

Dr. Lehmann's efforts were supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award Number UL1 TR003163. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH (National Institutes of Health). No other specific grant from any funding agency in the public, commercial or not-for-profit sectors was used for this publication.

AUTHOR CONTRIBUTIONS

YK-C contributed to conceptualization, design, methodology, and writing. SM contributed to design, methodology, writing, and editing. PE contributed to review and editing. CUL contributed to methodology, writing, review, and editing.

ACKNOWLEDGEMENTS

We express our gratitude to Sydney Roth for her valuable data analysis and editing assistance, which greatly enhanced the article's quality and clarity.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

1. OpenAI. ChatGPT. https://openai.com/blog/chatgpt/. Accessed February 1, 2023.
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; 2 (2): e0000198.
3. 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; 9: e45312.
4. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXIV:2303.13375; March 20, 2023.
5. United States Medical Licensing Examination. https://www.usmle.org/. Accessed February 1, 2023.
6. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ* 2019; 5 (1): e13930.
7. Clinical Informatics. https://www.theabpm.org/become-certified/subspecialties/clinical-informatics/. Accessed February 1, 2023.
8. Mankowitz S. *Clinical Informatics Board Review and Self Assessment*. Cham: Springer International Publishing; 2018.
9. Gardner RM, Overhage JM, Steen EB, *et al.*; AMIA Board of Directors. Core content for the subspecialty of clinical informatics. *J Am Med Inform Assoc* 2009; 16 (2): 153–7.
10. Silverman HD, Steen EB, Carpenito JN, *et al.* Domains, tasks, and knowledge for clinical informatics subspecialty practice: results of a practice analysis. *J Am Med Inform Assoc* 2019; 26 (7): 586–93.
11. Lehmann CU, Silverman HD, Gardner RM, Safran C, Gadd C. *Clinical Informatics Subspecialty Certification and Training. Informatics Education in Healthcare*. Cham: Springer; 2020.
12. Hawkins RE, Lipner RS, Ham HP, *et al.* American Board of Medical Specialties Maintenance of Certification: theory and evidence regarding the current framework. *J Contin Educ Health Prof* 2013; 33 Suppl 1: S7–19.
13. Zagury-Orly I, Durning SJ. Assessing open-book examination in medical education: the time is now. *Med Teach* 2021; 43 (8): 972–3.
14. Gulko B. *Is Chess Finished? Commentary*. New York, NY: American Jewish Committee; 1997: 45
15. Beaton A, Robinson J. *Chess Investigation Finds That U.S. Grandmaster 'Likely Cheated' More Than 100 Times*. October 7, 2022. Accessed May 5, 2022.
16. Liévin V, Hother CE, Winther O. Can large language models reason about medical questions?. arXiv preprint arXiv:2207.08143; July 17, 2022.
17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25 (1): 44–56.