ORIGINAL ARTICLE

DOI: 10.37047/jos.galenos.2025.2024.106532
J Oncol Sci 2025;11(1):36-39

Evaluation of 2024 Turkish Medical Oncology Board Exam with ChatGPT

Emir Gökhan KAHRAMAN, Olçun Ümit ÜNAL

University of Health Sciences Türkiye, İzmir City Hospital, Clinic of Medical Oncology, İzmir, Türkiye

ABSTRACT

Objective: This study aims to assess ChatGPT-4's performance on the Turkish Medical Oncology Board Exam questions, highlighting its potential uses and limitations in medical specialty evaluations.

Material and Methods: ChatGPT-4 was presented with each question from the 2024 Turkish Medical Oncology Proficiency Exam. Answers were determined to be correct or incorrect by comparison with the official answer key.

Results: The overall accuracy of ChatGPT-4.0 in this study was 64% out of 100 questions. For the fact-based questions (45 items), which require knowledge of specific information, such as molecules and side effects, ChatGPT-4o demonstrated an accuracy of 75.5%, with 34 correct responses. However, in the case-based questions (55 items) that require clinical judgment, its accuracy dropped to 54.5% (correct responses of 30). All these results highlight strengths of ChatGPT-4o on fact-driven questions but expose its limitations in scenarios needing nuanced decision-making.

Conclusion: Oncological clinical decision-making necessitates a nuanced approach that extends beyond standardized guidelines, integrating individual patient variables such as medical history, comorbidities, and therapeutic responses. While artificial intelligence (AI) systems demonstrate proficiency in processing guideline-driven data, they exhibit limitations in contextual clinical judgment requiring physician expertise. This study observed ChatGPT-4's superior performance on knowledge-based assessments (75.5% accuracy), attributable to its training on the American Society of Clinical Oncology/the European Society for Medical Oncology frameworks. However, its accuracy declined significantly in case-based evaluations (54.5%), highlighting challenges in personalized care integration. These findings underscore the indispensable role of clinician judgment in navigating complex, individualized treatment landscapes. Enhancing AI's clinical utility requires training on real-world patient data, though ethical constraints-particularly General Data Protection Regulation compliance-limit access to such datasets. Institution-specific AI tools leveraging anonymized records may bridge this gap, pending technological and regulatory advancements.

Keywords: Medical oncology; medical board exams; large language model; clinical reasoning

INTRODUCTION

Recent advancement in NLP, especially the introduction of ChatGPT-4 and subsequent models, has vastly changed medical education and assessment by increasing the capability to address complex board examination questions.[1,2] ChatGPT-4o, updated to include guidelines from major professional bodies such as the American Society of Clinical Oncology and the European Society for Medical Oncology, demonstrated improved clinical reasoning skills and positioned itself as a promising support tool for practicing clinicians and trainees alike.[3,4] ChatGPT-4's ability to pass high-stakes examinations, as demonstrated in the report by Kung et al.[4] on the successful performance of ChatGPT-4 in the United States Medical Licensing Examination, further points to its potential value in medical settings.[5] While earlier versions were particularly good at fact-based questions and could not perform well in case-based scenarios that required subtlety in judgment, the recent improvements have enhanced the capability of ChatGPT in understanding context.[6,7] These newer versions also have their limitations with regard to distinguishing subtle clinical cues, hence a need for continued research to refine their use in medical training.[3] In this context, the

Correspondence: Emir Gökhan KAHRAMAN MD,
University of Health Sciences Türkiye, İzmir City Hospital, Clinic of Medical Oncology, İzmir, Türkiye
E-mail: emirgokhan@gmail.com
ORCID ID: orcid.org/0000-0001-5303-6590

Received: 01.11.2024 **Accepted:** 01.03.2025 **Publication Date:** 29.04.2025

Cite this article as: Kahraman EG, Ünal OÜ. Evaluation of 2024 Turkish Medical Oncology Board Exam with ChatGPT. J Oncol Sci. 2025;11(1):36-39

Available at www.jos.galenos.com.tr

current study has aimed at the performance of ChatGPT-4o at addressing questions from the Turkish Medical Oncology Board Exam, reflecting both the potential benefits and challenges encountered in specialty assessments.

MATERIAL AND METHODS

In this study, ChatGPT-4 was systematically tested using the 2024 Turkish Medical Oncology Board Examination questions. The examination questions were received from the official website of the Turkish Society of Medical Oncology (www.kanser.org) and were presented to ChatGPT-4o without translation and verbatim, to ensure their original context and integrity were preserved. A total of 100 questions were analyzed, consisting of 55 case-based questions (clinical scenarios requiring decision-making) and 45 knowledge-based questions (factual recall of drug mechanisms, side effects, and guideline recommendations).

Each question was entered into ChatGPT-4 line by line and transferred to the ChatGPT-4 without any adaptation or translation as it appeared on the Turkish Society of Medical Oncology platform (www.kanser.org). The responses of the model were noted and then compared with the official answer key published by the Society. Responses were classified as correct or incorrect based on this comparison, thus allowing a direct evaluation of ChatGPT-4's accuracy across question types.

Case-based questions evaluated the model's capability to synthesize clinical facts and suggest patient-specific management strategies, while knowledge-based questions included factoid recall, such as medication mechanisms, or guideline-endorsed protocols. These items were analyzed for accuracy rates of the two categories to find the difference in performance.

This study aims to determine the degree to which ChatGPT-4 can simulate clinical reasoning in oncology and to outline the usefulness and limitations of its application in the assessment of medical oncology competence.

RESULTS

Analysis of ChatGPT-4o's response to the 2024 Turkish Medical Oncology Board Exam demonstrated different performances between clinical and direct knowledge-based questions. Out of 100 questions, 55 were scenario-based clinical questions, while 45 were direct information-focused ones. ChatGPT-4o answered 64% of all questions correctly; looking at it from another point of view, there was an obvious difference between the types of questions. On the direct knowledge questions, which required recalling specific facts such as drug mechanisms or side effects, ChatGPT-4o performed well, with 34 correct answers out of 45 for a 75.5% success rate within the knowledge category (Figure 1).

There is evidence of the model's strength in retaining and recalling guideline-based medical information.

In contrast, ChatGPT-4o's performance on clinical questions, which demand a more interpretative, case-based approach, demonstrated reduced accuracy. The model correctly answered 30 of 55 clinical questions for a success rate of 54.5%. The lower accuracy observed is consistent with what has been seen in artificial intelligence (AI)-driven models whenever there are complex, individualized treatments where human clinical judgment and contextual understanding come into play. These results emphasize that, while the AI has shown proficiency in direct knowledge recall, there are still many challenges for it to overcome in effectively adapting guideline-based information to nuanced clinical contexts.

DISCUSSION

The objective of this study is to evaluate the performance of ChatGPT-4 on the 2024 Turkish Medical Oncology Proficiency Examination, both in terms of recall of facts and clinical judgment. The performance of the model was considerably better on fact-type questions (accuracy 75.5%) while overall accuracy was 64%-particularly for questions related to oncological drugs and side effects. The results were consistent with earlier studies by Barbour and Barbour[6] and Kung et al.[4] showing that artificial intelligence models perform better on knowledge-based, structured questions. Their performance reduced to 54.5% when case-based questions were present, which required critical thinking as well as patient-specific decision-making.

One of the major reasons for this divergence is the complexity of individualized patient management. While the answers produced by ChatGPT-4 are based on documented oncology literature and guidelines, such as those released by the NCCN,

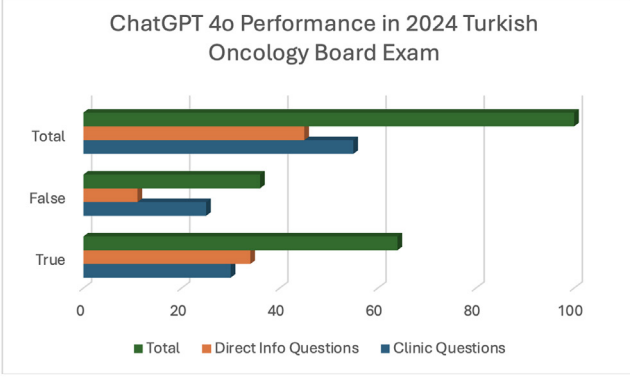

FIGURE 1: ChatGPT performance in 2024 Turkish Oncology Board Exam.

the American Society of Clinical Oncology (ASCO), and the European Society for Medical Oncology (ESMO), real-time medical decision-making goes beyond predetermined protocols. Physicians need to consider a number of variables, including the patient's comorbidities, functional status, socioeconomic status, access to healthcare services, and health insurance plans because each of these variables affects treatment decisions but is not directly addressed by clinical guidelines. Thus, although these guidelines are useful sources, they cannot substitute for physicians' clinical judgment, particularly in complex cases.

These issues have also been found in other branches of medicine. A study in urology[7] shows how AI is not able to deliver standardized responses to intricate, patient-specific situations. In addition, a study concluded that AI was not flexible in clinical decision-making, substantiating its more structured nature.

The findings of the study indicate that, while ChatGPT-4.0 performed well in evidence-based assessments, it faced difficulties with case-based reasoning. According to a study,[7-10] there is still more to be done to enhance the ability of artificial intelligence for balancing theoretical knowledge and the dynamics of real-world clinical situations, particularly in oncology.

CONCLUSION

Clinical decision-making in oncology is not simply following guidelines-it is weighing each patient's unique medical history, comorbidities, and response to treatment. Two patients may share the same cancer type and stage but could need different therapeutic approaches depending on age, genetic factors, or overall health status. AI models like ChatGPT-4o excel at reading medical literature and standard operating procedures, but are behind when faced with complicated, case-by-case decisions that only a physician's experience can supply.

ChatGPT-4 demonstrated remarkable efficacy in answering knowledge-based questions on the Turkish Medical Oncology Proficiency Exam, owing mainly to its reliance on the guidelines provided by ASCO and ESMO.

While ChatGPT-4.0 was good at knowledge-based questions, it struggled with case-based situations involving clinical judgment and individualized patient care. Even with further advancement of artificial intelligence, the complexity of oncology decision-making continues to heavily rely on the experience of doctors to assess individual patient variables and decide on the best treatment regimens.

To perform better in this area, the AI would need to rely on actual patient cases rather than relying solely on medical guidelines and textbooks. This does create ethical and legal problems, particularly with patient privacy laws such as General Data Protection Regulation, limiting access to actual clinical data. Due to such restrictions, general AI models such as ChatGPT-4 could always fall short in patient-specific decision-making.

A better option would be to develop AI models in hospitals or medical institutions, where anonymized patient information could be used subject to privacy legislation. With enhanced technology and declining costs of computing, these expert models might give more accurate clinical guidance without breaching patient confidentiality.

Ethics

Ethics Committee Approval: Ethics committee approval is not required for this study.

Informed Consent: Informed consent approval is not required for this study.

Footnotes

Authorship Contributions

Surgical and Medical Practices: E.G.K., O.Ü.Ü., Concept: E.G.K., O.Ü.Ü., Design: E.G.K., O.Ü.Ü., Data Collection or Processing: E.G.K., O.Ü.Ü., Analysis or Interpretation: E.G.K., O.Ü.Ü., Literature Search: E.G.K., O.Ü.Ü., Writing: E.G.K., O.Ü.Ü.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in neural information processing systems 33(2020):1877-1901.OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774; 2023. [Crossref]
2. Boillat T, Nawaz FA, Rivas H. Readiness to Embrace artificial intelligence among medical doctors and students: questionnaire-based study. JMIR Med Educ. 2022;8(2):e34973. [Crossref] [PubMed] [PMC]
3. Lower K, Seth I, Lim B, Seth N. ChatGPT-4: transforming medical education and addressing clinical exposure challenges in the post-pandemic era. Indian J Orthop. 2023;57(9):1527-1544. [Crossref] [PubMed] [PMC]
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [Crossref] [PubMed] [PMC]
5. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;25:e48568. [Crossref] [PubMed] [PMC]
6. Barbour AB, Barbour TA. A radiation oncology board exam of ChatGPT. Cureus. 2023;15(9):e44541. [Crossref] [PubMed] [PMC]
7. Touma NJ, Caterini J, Liblk K. Is ChatGPT ready for primetime? Performance of artificial intelligence on a simulated Canadian

urology board exam. Can Urol Assoc J. 2024;18(10):329-332. [Crossref] [PubMed] [PMC]

8. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. Cancer Cell. 2021;39(7):916-927. [Crossref] [PubMed] [PMC]
9. Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open. 2023;5(2):e000530. [Crossref] [PubMed] [PMC]
10. Kolla L, Parikh RB. Uses and limitations of artificial intelligence for oncology. Cancer. 2024;130(12):2101-2107. [Crossref] [PubMed] [PMC]