Original Article

DOI: 10.4274/uob.galenos.2025.2025.1.2

Bull Urooncol 2025;24(2):40-46

ChatGPT-3.5 and ChatGPT-4 Performance in Testicular Cancer: A Comparative Study

Ümit Uysal[1], Murat Uçar[2] ,Süleyman Sağır[3]

[1]*Health Sciences University Türkiye, Adana City Training and Research Hospital, Clinic of Urology, Adana, Türkiye*
[2]*Alanya Alaaddin Keykubat University Faculty of Medicine, Department of Urology, Antalya, Türkiye*
[3]*Mardin Training and Research Hospital, Clinic of Urology, Mardin, Türkiye*

Abstract

Objective: The aim of our study is to assess the reliability of Chat Generative Pre-trained Transformer (ChatGPT), compare the performance of ChatGPT-4 to ChatGPT-3.5, and explore its potential roles in healthcare decision-making.
Materials and Methods: Thirty questions related to testicular cancer were prepared, based on the 2023 European Association of Urology guidelines and clinical experience. These questions were systematically posed to ChatGPT-3.5 and ChatGPT-4, and responses were rated by three independent urologists using a six-point Likert scale. The median score from the three specialists was used as the final score.
Results: Both ChatGPT versions provided an incorrect answer to one question, scoring a one. For GPT-3.5 and GPT-4, the percentage of responses considered incorrect by the urologists was 20% and 13.3%, respectively, while correct responses (scoring 3 or higher) accounted for 80% and 86.7%. For general information-diagnosis questions, GPT-3.5 and GPT-4, had average scores of 4.29 and 4.80, with median values of 4.27 and 4.67. For treatment follow-up questions, average scores were 3.60 and 4.16, with median values of 3.60 and 4.20. GPT 4 generally outperformed GPT-3.5, but the difference was not statistically significant ($p>0.05$).
Conclusion: Our study shows that ChatGPT-4 is more reliable and accurate than ChatGPT-3.5 in testicular cancer-related queries. Continued development of its database and clinical capabilities could optimize ChatGPT's utility in healthcare.
Keywords: Artificial intelligence, ChatGPT, natural language processing, testicular cancer

Introduction

To improve the survival rates of cancer patients, rapid diagnosis and optimal treatments are essential. These patients seek various sources of information to address their health concerns but are often exposed to misinformation on platforms such as Google and YouTube (1). In this context, natural language processing (NLP) models have the potential to enhance patients' access to accurate medical information. Large language models (LLMs) should be evaluated for their accuracy in providing medical information. Artificial intelligence (AI) programs have demonstrated diagnostic accuracy comparable to that of medical professionals and have even outperformed physicians in delivering high-quality, empathetic responses to patient inquiries (2,3).

One of the LLMs, the Chat Generative Pre-trained Transformer (ChatGPT), is an NLP tool capable of understanding and generating human-like text (4). Developed by OpenAI, ChatGPT was launched in November 2022 and has been widely used by millions of users for information retrieval and task completion. ChatGPT-4, an advanced version provided by OpenAI, offers improvements over its predecessor, ChatGPT-3.5. This model is reported to have enhanced reasoning capabilities and a significantly larger knowledge base, enabling it to solve complex problems with greater accuracy (5). Trained on extensive datasets, ChatGPT possesses the ability to generate human-like text rapidly. Its rapid adoption highlights its accessibility and ease of use (6). In the medical field, it holds the potential to assist healthcare professionals in various aspects, including patient education, diagnosis, and treatment planning (7).

ChatGPT's success in the United States Medical Licensing Examination (USMLE) suggests that AI has the potential to revolutionize medicine (8). ChatGPT-4 is anticipated to enhance

Cite this article as: Uysal Ü, Uçar M, Sağır S. ChatGPT-3.5 and ChatGPT-4 performance in testicular cancer: a comparative study. Bull Urooncol. 2025;24(2):40-46.

Address for Correspondence: Ümit Uysal MD, Health Sciences University Türkiye, Adana City Training and Research Hospital, Clinic of Urology, Adana, Türkiye
E-mail: uysldr.74@gmail.com **ORCID:** orcid.org/0000-0002-9340-4260
Received: 15.01.2025 **Accepted:** 22.04.2025 **Publication Date:** 25.05.2025

clinical accuracy and reduce error rates. While ChatGPT-3.5 achieved a 60% success rate in the USMLE, ChatGPT-4 significantly improved this performance, reaching 87% (9,10). At the time of our study, ChatGPT-3.5 was available for free, while ChatGPT-4 was accessible through a subscription model, with claims of improved accuracy and speed (11).

In this study, we aimed to compare the reliability of responses provided by ChatGPT-3.5 and ChatGPT-4 to urology related questions based on the strong recommendations and clinical expertise outlined in the 2023 European Association of Urology (EAU) guidelines on testicular cancer. Our objective was to assess the practicality of AI in healthcare, particularly for users with limited resources. This study is expected to provide valuable insights into the benefits and limitations of AI models in clinical education and medical decision-making.

Materials And Methods

In our study, a total of 30 questions at three different levels of difficulty-basic, intermediate, and advanced-were prepared by three expert urologists: Ümit Uysal (ÜU), Süleyman Sağır (SS), Murat Uçar (MU), each with a minimum of four years of clinical experience. The questions were developed using high-grade recommendations from the testicular cancer section of the 2023 EAU guidelines as well as clinical expertise. Two of the urologists ÜU, MU are certified as Fellows of the European Board of Urology. Only questions written in English were included in the study. This rigorous question development and evaluation process was carefully conducted to enhance the reliability of the responses. In the development of the questions, clinical practice-oriented scenarios, up-to-date information from the literature, and expert opinions were taken into account. Furthermore, the questions were reviewed and validated by an expert panel of three urologists ÜU, SS, MU in terms of their relevance to clinical practice, adherence to current guidelines, and overall validity. This structured approach was designed to enhance the reproducibility and reliability of the study. On April 3, 2024, all questions were systematically submitted to both ChatGPT-3.5 and ChatGPT-4. Subsequently, each response was independently evaluated by three urology specialists ÜU, SS, MU based on the 2023 EAU guidelines and their own clinical experience. More specifically, the accuracy of the responses was rated using a six-point Likert scale: 1 indicating completely incorrect; 2 indicating more incorrect than correct; 3 indicating equally incorrect and correct; 4 indicating more correct than incorrect; 5 indicating almost correct; and 6 indicating completely correct (12). To enhance the reliability of the evaluations made by the experts, each response was independently scored, and the final score was determined by calculating the median. Although consensus-based approaches such as the Delphi method were not used in our study, the potential of such methods to improve inter-rater consistency can be investigated in future research. This study did not involve any human subjects or health data; therefore, ethical approval and patient informed consent were not required.

Statistical Analysis

Data analysis was performed using SPSS 24.0 software package (SPSS Inc., Chicago, IL). Descriptive statistics were calculated. Differences in scores between the two ChatGPT models, and the differences for each question group, were evaluated using the Wilcoxon test.

Results

In Table 1, the question "When should cranial imaging be performed in testicular cancer?" had the same average score for ChatGPT-3.5 and ChatGPT-4, both receiving 6.00 points. Both models received equal scores. In contrast, for the question "What is the most appropriate treatment for a patient with germ cell neoplasia in situ in a solitary testis?", both models scored 1.00, indicating that both models, including ChatGPT provided completely incorrect answers.

In our study, 20% of responses from ChatGPT-3.5 were evaluated as incorrect by specialists, while 80% of responses, scoring 3 or higher, were considered correct. This result indicates that the majority of responses from ChatGPT-3.5 were deemed correct. For ChatGPT-4, the percentage of incorrect responses was lower at 13.3%, and the percentage of correct responses was higher at 86.7%. This demonstrates that ChatGPT-4's responses were more accurate and reliable than those of GPT-3.5. While both models exhibited high accuracy, ChatGPT-4 provided fewer incorrect and more accurate responses according to the specialist physicians.

As shown in Table 2, for ChatGPT-4, 20.0% of the responses in the general information-diagnosis category received 5 points, and 13.3% received 6 points. The proportion of responses receiving low scores was quite small, with only 3.3% receiving 2 points. This indicates that ChatGPT-4 provided responses at a higher level of accuracy in this category. In the treatment-follow-up category, 13.3% of the responses received 5 or 6 points, while 6.7% received 2, 3, and 4 points. These results show that ChatGPT-4 also achieved high accuracy in this category, with responses generally receiving higher scores.

ChatGPT-4 provided more accurate and reliable responses than ChatGPT-3.5, with higher scores in both the general information-diagnosis and treatment-follow-up categories. ChatGPT-3.5 received moderately high scores in the general information-diagnosis category compared to its wider distribution of scores in the treatment-follow-up category. This demonstrates that ChatGPT-4 performed better .

When examining the responses of ChatGPT-3.5 and ChatGPT-4, that were evaluated as correct and incorrect by specialist physicians in the general information-diagnosis and treatment-follow-up subcategories, 43.3% of the responses in the general information-diagnosis category for ChatGPT-3.5 were evaluated as correct, while 6.7% were considered incorrect. This shows that ChatGPT-3.5 had a high rate of correct answers in this category, although some responses were evaluated as incorrect. In the treatment-follow-up category, 36.7% of the responses were evaluated as correct, while 13.3% were evaluated as incorrect. Although ChatGPT-3.5 generally tended to provide correct answers in this category, the rate of incorrect answers was higher compared to the general information-diagnosis category.

For ChatGPT-4, the correct response rate in the general information-diagnosis category was quite high at 46.7%, while

Table 1. Descriptive statistics of the evaluation scores of three expert physicians for ChatGPT-3.5 and ChatGPT-4's responses										
	ChatGPT-3.5					ChatGPT-4				
	Min.	Max.	X	SS	Median	Min.	Max.	X	SS	Median
General information-diagnosis questions										
What should a physician do first when a male patient presents to the urology clinic with suspected testicular cancer?	4.00	4.00	4.00	0.00	4.00	4.00	5.00	4.33	0.58	4.00
Which recurring genetic marker is associated with invasive GHNIS*?	4.00	4.00	4.00	0.00	4.00	6.00	6.00	6.00	0.00	6.00
What are the epidemiological risk factors for testicular cancer?	3.00	4.00	3.67	0.58	4.00	5.00	6.00	5.33	0.58	5.00
Which serum tumor marker might increase in a patient with a pathology report of "pure seminoma"?	4.00	5.00	4.33	0.58	4.00	5.00	6.00	5.67	0.58	6.00
In a male patient with "gynecomastia" detected during physical examination, which types of testicular cancer should be considered?	2.00	3.00	2.33	0.58	2.00	3.00	4.00	3.33	0.58	3.00
Is the sensitivity and specificity of micro RNA high in diagnosing and monitoring testicular cancer?	4.00	5.00	4.33	0.58	4.00	5.00	6.00	5.67	0.58	6.00
When should scrotal MRI be performed in a patient suspected of having testicular cancer?	3.00	3.00	3.00	0.00	3.00	3.00	4.00	3.33	0.58	3.00
Is the sensitivity and specificity of CT high in detecting lymph node metastasis in testicular cancer?	4.00	4.00	4.00	0.00	4.00	4.00	5.00	4.67	0.58	5.00
Is there a role for FDG PET-CT in testicular cancer?	4.00	4.00	4.00	0.00	4.00	5.00	6.00	5.33	0.58	5.00
When should cranial imaging be performed in testicular cancer?	6.00	6.00	6.00	0.00	6.00	6.00	6.00	6.00	0.00	6.00
Is there a role for bone scanning in staging testicular cancer?	2.00	3.00	2.33	0.58	2.00	2.00	3.00	2.33	0.58	2.00
What should be done for a male patient with a retroperitoneal mass normal hCG and AFP levels, and no palpable testicular mass?	6.00	6.00	6.00	0.00	6.00	5.00	5.00	5.00	0.00	5.00
Is routine contralateral biopsy performed in testicular cancer?	4.00	5.00	4.67	0.58	5.00	4.00	5.00	4.33	0.58	4.00
What should be done to preserve fertility in a male patient diagnosed with testicular cancer?	6.00	6.00	6.00	0.00	6.00	5.00	6.00	5.33	0.58	5.00
What should be considered if serum tumor markers remain elevated after an orchiectomy performed for suspected testicular cancer?	5.00	6.00	5.67	0.58	6.00	5.00	6.00	5.33	0.58	5.00
Treatment-follow-up questions										
16. Is there a role for testis-sparing surgery in testicular cancer?	2.00	2.00	2.00	0.00	2.00	3.00	3.00	3.00	0.00	3.00
17. Why is scrotal orchiectomy not recommended in the surgical treatment of testicular cancer?	4.00	5.00	4.67	0.58	5.00	6.00	6.00	6.00	0.00	6.00
18. What is the most appropriate treatment for a patient diagnosed with GHNIS in a solitary testis?	**1.00**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.00**	**1.00**

Table 1. Continued

	ChatGPT-3.5					ChatGPT-4				
	Min.	Max.	X	SS	Median	Min.	Max.	X	SS	Median
19. Is adjuvant radiotherapy routinely performed for stage 1 seminomas?	5.00	6.00	5.33	0.58	5.00	5.00	5.00	5.00	0.00	5.00
20. What should be the next treatment plan if tumor size is 5 cm and rete testis invasion is present in a patient with stage 1 seminoma?	2.00	3.00	2.67	0.58	3.00	3.00	4.00	3.33	0.58	3.00
21. What is the treatment option for a high-risk clinical stage 1 non-seminoma patient with vascular invasion?	4.00	4.00	4.00	0.00	4.00	5.00	5.00	5.00	0.00	5.00
22. Should we immediately perform orchiectomy in a life-threatening situation with widespread metastases in a patient with a testicular mass?	6.00	6.00	6.00	0.00	6.00	5.00	5.00	5.00	0.00	5.00
23. If a patient who underwent orchiectomy for suspected testicular cancer is diagnosed with stage 1 seminoma and refuses adjuvant chemotherapy and radiotherapy, what should be recommended?	5.00	6.00	5.67	0.58	6.00	6.00	6.00	6.00	0.00	6.00
24. What is the recommended minimum follow-up schedule for clinical stage I seminoma after active surveillance or adjuvant treatment (chemotherapy or radiotherapy)?	3.00	3.00	3.00	0.00	3.00	4.00	4.00	4.00	0.00	4.00
25. Should a testicular prosthesis be recommended to all patients who undergo orchiectomy for testicular cancer?	2.00	2.00	2.00	0.00	2.00	2.00	2.00	2.00	0.00	2.00
26. What should be the next treatment step if recurrence occurs after nerve-sparing RPLND in clinical stage 1 non-seminoma?	4.00	4.00	4.00	0.00	4.00	5.00	6.00	5.67	0.58	6.00
27. What is the alternative treatment to chemotherapy for a patient with clinical stage 2b seminoma?	2.00	2.00	2.00	0.00	2.00	2.00	2.00	2.00	0.00	2.00
28. What chemotherapy protocol should be applied if bleomycin cannot be administered in a patient with advanced metastatic non-seminomatous testicular cancer?	4.00	4.00	4.00	0.00	4.00	5.00	6.00	5.67	0.58	6.00
29. How should thromboprophylaxis be performed to prevent thromboembolic events in a young male patient receiving chemotherapy for testicular cancer?	3.00	4.00	3.33	0.58	3.00	4.00	4.00	4.00	0.00	4.00
30. What is the minimum duration of contraception recommended after completing treatment for testicular cancer?	4.00	5.00	4.33	0.58	4.00	4.00	5.00	4.67	0.58	5.00

*Germ cell neoplasia in situ, ChatGPT: Chat Generative Pre-trained Transformer, Min.: Minimum, Max.: Maximum, SS: Standard score, RNA: Ribonucleic acid, MRI: Magnetic resonance imaging, FDG: Fluorodeoxyglucose, PET: Positron emission tomography, CT: Computed tomography, hCG: Human chorionic gonadotropin, AFP: Alpha-fetoprotein, RLND: Retroperitoneal lymph node dissection

the incorrect response rate remained low at 3.3%. This indicates that ChatGPT-4 performed very well in this category and largely provided correct responses. In the treatment-follow-up category, the correct response rate was 40%, while the incorrect response rate was 10.0%. This shows that ChatGPT-4 was generally successful in this category as well, although there were a few incorrect responses. ChatGPT-4 had higher accuracy rates than ChatGPT-3.5 in both the general information-diagnosis and treatment-follow-up categories. In the general information-diagnosis category, ChatGPT-4 provided more accurate responses with fewer errors compared to ChatGPT-3.5. Although ChatGPT-4 was more successful in the treatment-follow-up category than ChatGPT-3.5, both systems demonstrated similar accuracy rates. These results indicate that ChatGPT-4 generally provided more reliable and accurate responses, compared to ChatGPT-3.5.

As shown in Table 3, the average score for the ChatGPT-3.5 model in general information-diagnosis questions was 4.29, with a median of 4.27, while the average score for the ChatGPT-4 model was 4.80, with a median of 4.67. According to the results

of the Wilcoxon test, the z-value was -1.633 and the p-value was 0.102, indicating that there was no statistically significant difference between the two models. For the treatment-follow-up questions, the average score for the ChatGPT-3.5 model was 3.60, with a median of 3.60, while the average score for the ChatGPT-4 model was 4.16, with a median of 4.20. According to the Wilcoxon test results, the z-value was -1.633 and the p-value was 0.102, again, showing no statistically significant difference between the two models. The results of the Wilcoxon test indicated that there was no statistically significant difference between the ChatGPT-3.5 and ChatGPT-4 models for both general information and diagnosis and treatment and follow-up questions (p>0.05). However, it was observed that the ChatGPT-4 model had higher average scores in both categories. This suggests that ChatGPT-4 generally performed better than ChatGPT-3.5, although the difference was not statistically significant.

Table 2. Score distribution of ChatGPT-3.5 and ChatGPT-4 responses based on subcategories (general information-diagnosis and treatment-follow-up)

		Score	n	%
ChatGPT-3.5	**General information-diagnosis**	2	2	6.7
		3	1	3.3
		4	7	23.3
		5	1	3.3
		6	4	13.3
	Treatment-follow-up	1	1	3.3
		2	3	10.0
		3	3	10.0
		4	4	13.3
		5	2	6.7
		6	2	6.7
ChatGPT-4	**General information-diagnosis**	2	1	3.3
		3	2	6.7
		4	2	6.7
		5	6	20.0
		6	4	13.3
	Treatment-follow-up	1	1	3.3
		2	2	6.7
		3	2	6.7
		4	2	6.7
		5	4	13.3
		6	4	13.3

ChatGPT: Chat Generative Pretrained Transformer

Discussion

In recent years, advancements in NLP technologies and deep learning hardware have led to significant progress in the field of LLMs. ChatGPT, a state-of-the-art LLM built upon ChatGPT-3.5 and GPT-4, demonstrates exceptional capabilities in general language comprehension and reasoning (13). The AI chatbot ChatGPT has shown promising performance across various domains, including medical science, business, and law. However, its accuracy in handling medical queries requiring domain-specific expertise, particularly in the field of urology, remains uncertain. The purpose of this study was to assess the ability and performance of ChatGPT in responding to 30 questions, prepared based on high-level recommendations from the testicular cancer section of the 2023 EAU guidelines, as well as clinical experience. Furthermore, we aimed to determine whether there is a significant performance difference between ChatGPT-3.5 and ChatGPT-4, with the goal of clarifying their potential roles in healthcare decision-making processes.

Various studies have demonstrated that GPT-4 generally achieves a higher accuracy rate compared to GPT-3.5. In a study comparing the performance of ChatGPT-3.5 and GPT-4 on standard urology multiple-choice questions, a total of 700 questions were presented to both models, and the results were analyzed. GPT-4 exhibited a higher accuracy rate than GPT-3.5 (44.4% vs. 30.9%). Notably, GPT-4 was found to be more successful in areas such as urologic oncology, sexual medicine, and pediatric urology (14). Similarly, in another study comparing the performance of ChatGPT-3.5 and ChatGPT-4 in European Board of Urology examinations, ChatGPT-4 demonstrated significantly better accuracy across all exams compared to ChatGPT-3.5 (15). Tsai et al. (16) demonstrated in their study that ChatGPT-4 outperformed ChatGPT-3.5 in terms of quality,

Table 3. A statistical comparison of the responses provided by ChatGPT-3.5 and ChatGPT-4 models to general information-diagnosis, treatment-follow-up questions

Min.		Evaluation of ChatGPT-3.5					Evaluation of ChatGPT-4					Wilcoxon	
		Max.	X	SD	Median	Min	Max.	X	SD	Median	Z	p-value	
General information-diagnosis		4.07	4.53	4.29	0.23	4.27	4.47	5.27	4.80	0.42	4.67	-1.633	0.102
Treatment-follow-up		3.40	3.80	3.60	0.20	3.60	4.00	4.27	4.16	0.14	4.20	-1.633	0.102
Wilcoxon test	Z	-1.633					-1.633						
	p-value	0.102					0.102						

ChatGPT: Chat Generative Pretrained Transformer, Min.: Minimum, Max.: Maximum, SD: Standard deviation

adherence to clinical guidelines, and alignment with expert opinions when providing cancer treatment recommendations. Another study comparing the diagnostic capabilities of GPT-3.5 and GPT-4.0 in surgery revealed that GPT-4.0 exhibited higher accuracy for both primary and secondary diagnoses, indicating significant diagnostic potential (17). In a study examining the performance of GPT-4 in orthopedic surgery board questions, GPT-4 accurately answered 63.4% of the questions, while GPT-3.5 correctly answered only 46.3%. GPT-4 demonstrated significantly better performance on orthopedic board-style questions (18). Another study assessing the accuracy of ChatGPT references in the disciplines of head and neck surgery and otolaryngology showed that ChatGPT-4.0 performed better in terms of reliability compared to version 3.5 (19).

Other studies in the field of urology have also demonstrated the superior performance of GPT-4. For instance, in a comparative analysis of advanced AI strategies in renal oncology, another study compared GPT-3.5 and GPT-4.0. The average accuracy rates of responses to 30 questions related to renal cell carcinoma, prepared by urology specialists, were 67.08% for ChatGPT-3.5 and 77.50% for ChatGPT-4.0. ChatGPT-4.0 outperformed ChatGPT-3.5 with a significantly higher accuracy rate (20). In another study evaluating the performance of ChatGPT-4 in answering questions related to urolithiasis, it was found that ChatGPT accurately and satisfactorily responded to more than 95% of the urolithiasis-related questions (21). Furthermore, a study investigating ChatGPT's performance in the diagnosis and treatment of urological trauma concluded that ChatGPT demonstrated a highly competent and reliable performance in managing urological trauma cases (22).

On the other hand, a study examining the quality of ChatGPT-4.0's responses to frequently asked popular questions about prostate, bladder, kidney, and testicular cancers, as well as questions selected from the 2023 EAU Oncology guidelines, revealed mixed findings. While ChatGPT demonstrated commendable accuracy rates when answering popular questions related to urologic cancers, its performance in providing responses consistent with EAU guideline-based questions was found to be unsatisfactory (23). Similarly, another study assessing ChatGPT-4's responses to 195 clinical questions related to prostate cancer, prepared with consideration of the EAU 2023 guidelines, demonstrated that ChatGPT exhibited poor accuracy (24). Furthermore, a study evaluating ChatGPT's performance on standard multiple-choice urology examinations also reported suboptimal performance (14). In our study, we observed that the responses provided by ChatGPT-3.5 to the questions related to testicular cancer, mostly received moderate scores (4 points), whereas the responses from ChatGPT-4 received higher scores (5 and 6 points). This finding suggests that GPT-4 provided more accurate or satisfactory answers as evaluated by expert clinicians. The assessment of GPT-4 revealed that the incorrect response rate was 13.3%, which was lower than that of GPT-3.5. Meanwhile, the correct response rate was higher for GPT-4, reaching 86.7%. Overall, both systems demonstrated high accuracy rates; however, GPT-4 provided fewer incorrect answers and more accurate responses compared to GPT-3.5. Furthermore, GPT-4 achieved higher scores than GPT-3.5 in both general knowledge and diagnosis and treatment and follow-up categories. GPT-3.5, on the other hand, predominantly received moderate scores in the general knowledge-diagnosis category and demonstrated a broader distribution of scores in the treatment-follow-up category. In our study, although not statistically significant, ChatGPT-4 demonstrated better performance than ChatGPT-3.5 by providing more comprehensive answers. This suggests that ChatGPT-4 has the potential to be an effective supportive tool in diagnostic, therapeutic, and clinical decision-making processes related to testicular cancer. However, both models exhibited limitations in answering certain questions. This finding underscores the importance of human oversight when employing AI applications, particularly in healthcare-related topics. The study also emphasizes the importance of continuous improvement to ensure the effectiveness and reliability of ChatGPT, a supportive tool used in clinical practice, emphasizing the importance of continuous improvement to ensure its effectiveness and reliability in assisting healthcare professionals with diagnostic and therapeutic decision-making processes. Although ChatGPT-4 demonstrates significant advancements in providing responses to questions related to testicular cancer, the best use cases and ethical considerations have not yet been fully clarified. Further detailed studies are required to determine whether these models can reliably serve as clinical aids in medical practice.

Study Limitations

This study focuses exclusively on testicular cancer, which limits the generalizability of its findings to other oncological or urological conditions. Although the responses were evaluated by experts, variability in assessments may occur across different expert panels. In future studies, we aim to obtain more objective results by including evaluations from independent urologists and employing consensus-based approaches such as the Delphi method. Additionally, the model was tested solely using the 2023 EAU guidelines and clinical experience. Incorporating additional authoritative sources-such as the American Urological Association guidelines, Campbell-Walsh Urology, Smith & Tanagho's General Urology, and other prominent urological guidelines and reference texts-may enhance the model's accuracy and comprehensiveness. The subscription-based structure and limited accessibility of the platform may also pose a barrier for users with constrained resources. Furthermore, it remains unclear whether the EAU guidelines were directly included in ChatGPT's training data, which may limit the alignment of its responses with these guidelines.

Conclusion

Although ChatGPT-4 provides more accurate and satisfactory responses compared to ChatGPT-3.5 in specific urological topics such as testicular cancer, it is not entirely flawless. Its occasional blending of correct and incorrect information may pose risks for healthcare professionals. This highlights the necessity of expert validation and supervised systems in the integration of AI-based models into clinical practice. In this context, it is evident that such technologies should be positioned solely as supportive tools. In the future, the development of customized AI systems trained exclusively on urology-specific data, by leveraging open-source LLMs (e.g., DeepSeek, LLaMA, Mistral), may enable the creation

of more reliable, specialized, and clinically applicable AI solutions. This approach could enhance accuracy and trustworthiness, particularly in niche areas such as testicular cancer. The present study may serve as a foundational step toward the development of urology-specific LLMs. Ultimately, this could contribute to the creation of more tailored solutions that support the safe, ethical, and effective use of AI in healthcare.

Ethics

Ethics Committee Approval: This study did not involve any human subjects or health data; therefore, ethical approval was not required.

Informed Consent: This study did not include any human subjects or health data, patient consent was not required.

Acknowledgements

Publication: The results of the study were not published in full or in part in form of abstracts.

Contribution: There is not any contributors who may not be listed as authors.

Footnotes

Authorship Contributions

Surgical and Medical Practices: Ü.U., Concept: Ü.U., M.U., Design: Ü.U, S.S., Data Collection or Processing: Ü.U, M.U., Analysis or Interpretation: Ü.U., S.S., Literature Search: Ü.U, Writing: Ü.U.

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. Park H, Kang E, Kim Y, et al. Analysis of the spread of misinformation about lung cancer on YouTube: based on source of information. Korean J Fam Pract. 2013;13:152-158.
2. Zhu J, Shen B, Abbasi A, Hoshmand-Kochi M, et al. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. PLoS One. 2020;15:e0236621.
3. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183:589-596.
4. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023;11:1166120.
5. Slowik C, Kaiser F. GPT 3 vs. GPT 4. open AI language models comparison. Neoteric. 2023.
6. Hartmann J, Schwenzow J, Witte M. The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv. 2023.
7. Biswas SS. Role of chat GPT in public health. Ann Biomed Eng. 2023;51:868-869.
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
9. Liévin V, Hother CE, Motzfeldt AG, et al. Can large language models reason about medical questions? Patterns (N Y). 2024;5:100943.
10. Nori H, King N, McKinney SM, et al. Capabilities of GPT-4 on medical challenge problems. arXiv. 2023.
11. Deebel NA, Terlecki R. ChatGPT performance on the American Urological Association self-assessment study program and the potential influence of artificial intelligence in urologic training. Urology. 2023;177:29-33.
12. Smidt N, van der Windt DA, Assendelft WJ, et al. Corticosteroid injections, physiotherapy, or a wait-and-see policy for lateral epicondylitis: a randomised controlled trial. Lancet. 2002;359:657-662.
13. Chen Q, Sun H, Liu H, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. Bioinformatics. 2023;39:btad557.
14. Yudovich MS, Makarova E, Hague CM, Raman JD. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. J Educ Eval Health Prof. 2024;21:17.
15. Schoch J, Schmelz HU, Strauch A, et al. Performance of ChatGPT-3.5 and ChatGPT-4 on the European Board of Urology (EBU) exams: a comparative analysis. World J Urol. 2024;42:445.
16. Tsai CY, Cheng PY, Deng JH, et al. ChatGPT v4 outperforming v3.5 on cancer treatment recommendations in quality, clinical guideline, and expert opinion concordance. Digit Health. 2024;10:20552076241269538.
17. Liu J, Liang X, Fang D, et al. The diagnostic ability of GPT-3.5 and GPT-4.0 in surgery: comparative analysis. J Med Internet Res. 2024;26:e54985.
18. Hofmann HL, Guerra GA, Le JL, et al. The rapid development of artificial intelligence: GPT-4's performance on orthopedic surgery board questions. Orthopedics. 2024;47:e85-e899.
19. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol. 2023;280:5129-5133.
20. Liang R, Zhao A, Peng L, et al. Enhanced artificial intelligence strategies in renal oncology: iterative optimization and comparative analysis of GPT 3.5 versus 4.0. Ann Surg Oncol. 2024;31:3887-3893.
21. Cakir H, Caglar U, Yildiz O, et al. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. Int Urol Nephrol. 2024;56:17-21.
22. Li J, Yi X, Han Z, et al. The theranostic performance of Chat-GPT against urological trauma. Int J Surg. 2024;110:4485-4487.
23. Ozgor F, Caglar U, Halis A, et al. Urological cancers an d ChatGPT: assessing the quality of information and possible risks for patients. Clin Genitourin Cancer. 2024;22:454-457.e4.
24. ombardo R, Gallo G, Stira J, et al. Quality of information and appropriateness of open AI outputs for prostate cancer. Prostate Cancer Prostatic Dis. 2025;28:229-231.