ORIGINAL PAPER

UROLOGICAL ONCOLOGY

Evaluation of DeepSeek-R1 and ChatGPT-4o as educational sources for upper tract urothelial carcinoma

Wojciech Krajewski[1*], Jan Łaszkiewicz[2*], Łukasz Biesiadecki[2*], Wojciech Tomczak[2], Łukasz Nowak[1], Piotr Łaszkiewicz[3], Joanna Chorbińska[1], Francesco Del Giudice[4,5], Benjamin I. Chung[5], Tomasz Szydełko[2]

[1]Department of Minimally Invasive and Robotic Urology, University Center of Excellence in Urology, Wroclaw Medical University, Poland
[2]University Centre of Excellence in Urology, Wroclaw Medical University, Poland
[3]Nova School of Science and Technology, Universidade Nova de Lisboa, Lisbon, Portugal
[4]Department of Maternal Infant and Urologic Sciences, "Sapienza" University of Rome, Policlinico Umberto I Hospital, Rome, Italy
[5]Department of Urology, Stanford University School of Medicine, Stanford, CA, United States

[*]These authors are co-first authors.

Citation: Krajewski W, Łaszkiewicz J, Biesiadecki Ł, et al. Evaluation of DeepSeek-R1 and ChatGPT-4o as educational sources for upper tract urothelial carcinoma. Cent European J Urol 2026; 79: 1-8.

Article history
Submitted: Oct. 1, 2025
Accepted: Nov. 4, 2025
Published online: Jan. 13, 2025

Introduction Upper tract urothelial carcinoma (UTUC) is associated with poor survival outcomes. Therefore, providing reliable information about UTUC is crucial. Recently, chatbots powered by large language models have become a widely used information source. Our aim was to evaluate and compare responses generated by ChatGPT-4o and DeepSeek-R1 to patient-important questions regarding UTUC.

Material and methods A set of 43 questions assigned into four categories (general information, symptoms and diagnosis, treatment, prognosis) was curated. Each question was entered into DeepSeek-R1 and ChatGPT-4o. Answers were rated by two urologists using a scale from 1 (completely incorrect) to 4 (fully correct). The median score was calculated for each question. Median scores ≥3 were considered accurate. The repeatability of responses was evaluated using cosine similarity. The number of words in responses was counted.

Results The median scores for DeepSeek-R1 and ChatGPT-4o were both 3.5. There was no statistically significant difference between the scores assigned to two chatbots for all questions ($p = 0.35$), nor for any particular category.

DeepSeek-R1 and ChatGPT-4o provided satisfactory answers for 93% and 91% of the evaluated questions, respectively. No potentially dangerous information was found. Both models consistently generated responses with moderate-high similarity (cosine similarity >0.5), except in one query. Finally, DeepSeek-R1 provided significantly longer answers than ChatGPT-4o ($p < 0.001$).

Conclusions Both DeepSeek-R1 and ChatGPT-4o predominantly provide satisfactory responses to patient-important questions about UTUC. Artificial intelligence chatbots demonstrate potential as the first-line information sources for patients but struggle with highly specialized inquiries and thus cannot replace expert medical advice.

Corresponding author
Łukasz Biesiadecki
University Center of
Excellence in Urology,
Wrocław Medical
University,
213 Borowska St.,
50-556 Wrocław, Poland
lbiesiadecki2001@gmail.com

Key Words: AI ‹› artificial intelligence ‹› upper tract urothelial carcinoma ‹› ChatGPT ‹› DeepSeek

INTRODUCTION

Urothelial carcinoma (UC) is the second most common urological malignancy [1]. Most UCs derive from the urinary bladder epithelium. However, 5–10% develop in the ureter and pelvicalyceal system, where they are classified as upper tract urothelial carcinoma (UTUC). Although UTUC is relatively rare, its incidence has steadily increased over recent decades [2]. This rise is largely attributed to an aging population and advancements in diagnostic techniques [3].

doi: 10.5173/ceju.2025.0238

Despite improvements in management, UTUC is still associated with poor survival outcomes. **The estimated 5-year cancer-specific survival rate is approximately 50% for patients with pT2/pT3 stage disease [4].** Therefore, ensuring patient compliance with treatment and follow-up protocols is crucial for improving therapeutic outcomes. A key factor in achieving this is to provide accessible and comprehensible information sources about UTUC for the general population.

Unfortunately, trustworthy sources often use complex and technical language, which is difficult for the patients to understand [5]. Nowadays, the internet has become a popular source of health information for patients with oncologic diseases [6]. However, patients may not be able to distinguish reliable data from misinformation.

In recent years we have witnessed remarkable advancements in artificial intelligence (AI), **including in the field of medicine [7]. Patients may use chatbots powered by large language models (LLMs) as an accessible and user-friendly source of medical information.** However, LLMs have the potential to reproduce existing biases and disseminate misinformation without built-in verification mechanisms [8]. A recent study evaluated the responses generated by ChatGPT-3.5 to patient-important questions regarding UTUC, yielding moderate results [9]. It is essential to reassess the capabilities of newer LLMs: ChatGPT-4o (OpenAI, San Francisco, CA, USA) and DeepSeek-R1 (DeepSeek, Hangzhou, Zhejiang, China), to determine whether they are reliable sources of information about UTUC.

This study aims to evaluate and compare responses generated by ChatGPT-4o and DeepSeek-R1 to patient-important questions regarding UTUC.

MATERIAL AND METHODS

In order to identify commonly asked questions about UTUC, we reviewed patient-oriented websites dedicated to UTUC and existing studies evaluating AI responses to patient queries. Additionally, questions asked by patients admitted to our department for UTUC management were recorded. A comprehensive set of 43 relevant questions with varying difficulty was curated by two attending urologists specialising in UC. These questions were assigned into four categories: general information, symptoms and diagnosis, treatment, and prognosis (Table 1).

Each question was entered into two LLMs: DeepSeek-R1 with activated DeepThink function and ChatGPT-4o. Questions were entered individu-

Table 1. *List of questions about upper tract urinary carcinoma (UTUC) and median scores of large language models' (LLMs') responses*

	Median score of ChatGPT-4o	Median score of DeepSeek-R1
General information		
1. What is UTUC?	3.5	3.5
2. How common is UTUC?	3.5	4
3. What are the risk factors of UTUC?	3	4
4. What is the difference between UTUC, bladder cancer and kidney cancer?	3.5	3.5
5. Can UTUC occur in both kidneys (bilateral UTUC)?	4	3.5
6. Are there genetic or hereditary factors linked to UTUC?	3	4
7. Can UTUC spread to other organs?	3.5	4
Symptoms and diagnosis		
8. What are the symptoms of UTUC?	4	3.5
9. What symptoms distinguish UTUC from bladder cancer and kidney cancer?	3.5	3
10. Could UTUC symptoms be caused by other conditions?	3.5	3.5
11. Can UTUC cause pain?	4	3.5
12. Can UTUC affect kidney function?	4	4
13. How is UTUC diagnosed?	3.5	4
14. What are the stages of UTUC?	3.5	3.5
15. What is the difference between low-risk and high-risk UTUC?	2.5	3
16. Are blood and urine tests useful for diagnosing UTUC?	4	4
17. Can an ultrasonography detect UTUC?	4	4
18. What role does cytology play in UTUC diagnosis?	3.5	4
19. Does negative cytology result rule out UTUC?	4	4
20. Which method is better for diagnosing UTUC: CT or MRI?	3.5	4
21. Can UTUC be detected early?	3.5	4
Treatment		
22. How is UTUC treated?	3.5	3
23. Is it possible to cure UTUC without surgery?	2.5	2.5
24. When is surgery necessary in UTUC?	2.5	3.5
25. Are there non-surgical treatment options for UTUC?	3	2.5
26. What is a radical nephroureterectomy?	4	3.5
27. What does kidney-sparing surgery mean in UTUC?	4	3

Table 1. *Continued*

	Median score of ChatGPT-4o	Median score of DeepSeek-R1
28. When is kidney-sparing surgery an option in UTUC?	3.5	4
29. Are chemotherapy and radiation therapy used for treating UTUC?	4	3.5
30. Are there any new advancements in UTUC treatment?	3.5	2.5
31. Which UTUC treatment options have the fewest complications?	4	2.5
32. How is bilateral UTUC treated?	4	4
33. What are the risks and complications of UTUC treatment?	4	4
Prognosis		
34. What are the complications of UTUC?	4	4
35. How long can I live with UTUC?	4	3.5
36. What is the survival rate for UTUC?	4	3.5
37. What is the risk of UTUC recurrence?	3.5	4
38. What is the risk of metastatic disease in UTUC?	4	4
39. How often should I have follow-up visits after UTUC treatment?	3.5	3
40. What factors affect the prognosis of UTUC?	4	3
41. What lifestyle changes should I make to reduce risk of UTUC recurrence?	4	3.5
42. Can I still drink alcohol or smoke if I have UTUC?	4	4
43. What are the chances of developing bladder cancer after UTUC?	4	3

CT – computed tomography; MRI – magnetic resonance imaging; UTUC – upper tract urinary carcinoma

ally into separate chat sessions without any additional context or clarification, in English language, on February 7, 2025. All responses generated by the LLMs were recorded without modification. For DeepSeek-R1, only the final responses were registered, without preceding reasoning. The number of words in responses was counted.

Each question was paired with the corresponding responses from the LLMs and compiled into an assessment questionnaire. Responses were rated using a four-point scale: from 1 (completely incorrect or containing potentially dangerous information) to 4 (fully correct, requiring no further expert clarification). In addition to numerical scoring, any responses containing potentially dangerous information were listed separately. Two experienced urologists specialising in urothelial carcinoma independently assessed the responses using the 2024 EAU Guidelines on Upper Urinary Tract Urothelial Carcinoma as a reference [2]. The median score was calculated for each question from the ratings of two evaluators. Responses with median scores of ≥3 were considered sufficiently accurate by the assessing urologists for preliminary sources of information.

To assess the repeatability of LLMs responses, we obtained a second set of responses on February 22, 2025, following the same procedure. Then, we evaluated the repeatability of responses over time. To do so, cosine similarity was used – a metric for assessing textual consistency between the two responses generated by the same LLM for each question. Cosine similarity scores were calculated using a formula in the Python programming language.

Continuous parametric variables were reported as mean (standard deviation [SD]), while ordinal and nonparametric variables were presented as median (interquartile range [IQR]). For comparative analysis, the Wilcoxon Signed-Rank test was performed for paired nonparametric variables, while the independent t-test was used for independent parametric variables. A p-value of <0.05 was considered statistically significant. All statistical analyses were conducted using Statistica 13 (TIBCO Software Inc., Palo Alto, CA, USA).

Bioethical standards

Due to the nature of the study, the consent of the bioethics committee was not required.

RESULTS

Response quality

In the collective median score distribution analysis, the following results were observed for DeepSeek-R1: a score of 4 was assigned 21 times (48.8%); a score of 3.5 – 16 times (37.2%); a score of 3 – 3 times (7%); and a score of 2.5 – 3 times (7%). For ChatGPT-4o: a score of 4 was assigned 19 (44.2%); a score of 3.5 – 13 times (30.2%); a score of 3 – 7 times (16.3%); and a score of 2.5 – 4 times (9.3%) (Table 1; Figure 1). The lowest median score (2.5) for DeepSeek-R1 was assigned to responses for the three questions: “What is the difference between low-risk and high-risk UTUC?”, “Is it possible to cure UTUC without surgery?”, “When is surgery necessary in UTUC?”. Similarly, ChatGPT-4o responses with the lowest median score (2.5) were observed in answers to: “Is it possible to cure UTUC without surgery?”, “Are there non-surgical treatment options for UTUC?”, “Are there any new advancements in UTUC treatment?”, “Which UTUC treatment options have the fewest complications?” (Table 1).

Considering the ratings of individual evaluators, for DeepSeek-R1, the median score assigned by the first evaluator was 4 (4–4), while the median score assigned by the second evaluator was 4 (3–4). Similarly, for ChatGPT-4o the first evaluator assigned the median score of 4 (4–4), while the second evaluator assigned the median score of 3 (3–4). For both LLMs, the difference between the two evaluators' assessments was statistically significant ($p < 0.001$).

The median scores for DeepSeek-R1 and Chat-GPT-4o were 3.5 (3.5–4), and 3.5 (3.25–4), respectively. There was no statistically significant difference between the scores assigned to two LLMs (p-value = 0.35).

Responses generated by Deep-Seek-R1 achieved the highest median scores in the prognosis category – 4 (4–4), whereas its lowest scores were observed in the general information category 3.5 (3.25–4). In contrast, ChatGPT-4o's responses received the highest scores in the general information and symptoms and diagnosis categories 4 (3.5–4) and the lowest in the treatment category 3.25 (2.5–3.625). There were no statistically significant differences between the median scores of ChatGPT-40 and DeepSeek-R1 in any category (Table 2).

Nonetheless, DeepSeek-R1 and ChatGPT-4o provided satisfactory answers for 93% and 91% of the evaluated questions, respectively.

Importantly, the evaluators did not identify any potentially dangerous information in any responses generated by either LLM.

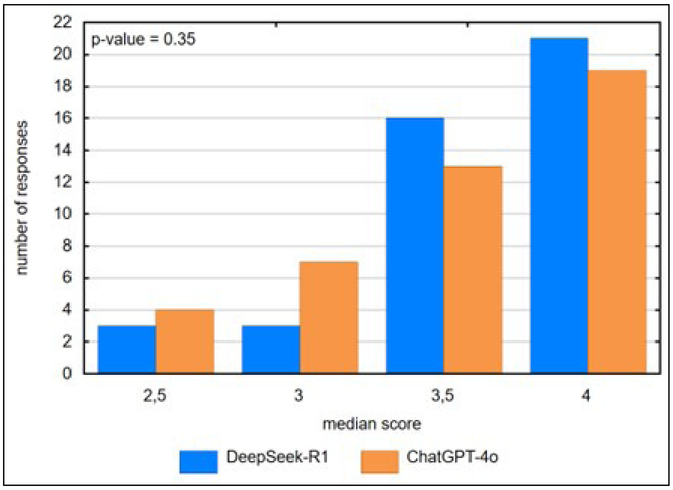

Figure 1. *Median scores of large language models' responses.*

Response repeatability

The mean cosine similarity score for responses generated by DeepSeek-R1 and ChatGPT-4o was 0.719 (0.064) and 0.694 (0.089), respectively (Figure 2). Both LLMs exhibited variability in response repeatability, with DeepSeek-R1 demonstrating slightly greater response consistency. However, the difference was not statistically significant ($p = 0.15$).

Across all evaluated questions, both models consistently generated responses with moderate to high similarity (cosine similarity >0.5), with one exception. ChatGPT-4o answers to the query “How common is UTUC?” yielded the lowest cosine similarity score (0.469), which corresponds to low similarity of two responses (cosine similarity >0.5).

Response lengths

For further analysis, the number of words in responses were counted. DeepSeek-R1 consistently

Table 2. *Median scores of DeepSeek-R1 and ChatGPT-4o stratified into questions' categories*

	Median score (IQR) DeepSeek-R1	Median score (IQR) ChatGPT-4o	p-value
All responses	3.5 (3.5–4)	3.5 (3.25–4)	0.35
General information	3.5 (3.25–3.5)	4 (3.5–4)	0.14
Symptoms and diagnosis	3.5 (3.5–4)	4 (3.5–4)	0.53
Treatment	3.75 (3.375–4)	3.25 (2.5–3.625)	0.14
Prognosis	4 (4–4)	3.5 (3.125–4)	0.13

IQR – interquartile range

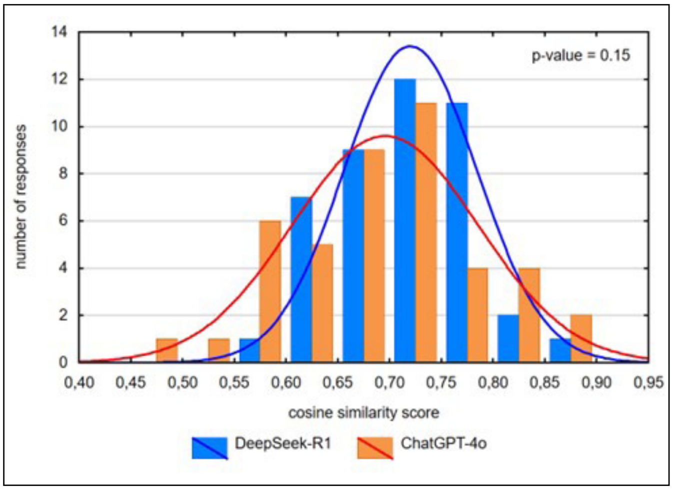

Figure 2. *Cosine similarity score of large language models' (LLMs') responses obtained in two days.*

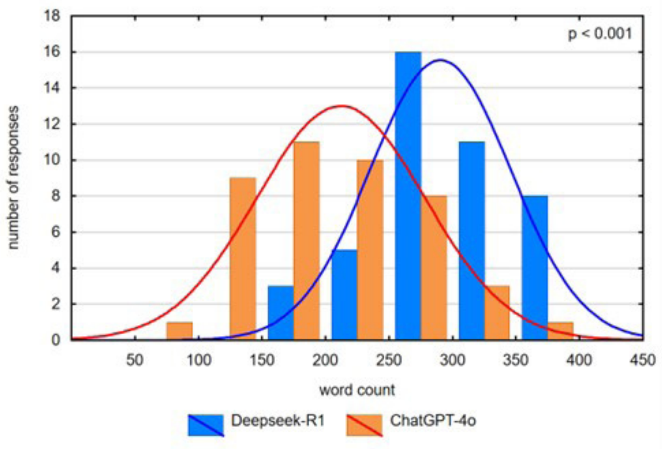

Figure 3. *Word count in large language models' (LLMs') responses.*

provided significantly longer answers than ChatGPT-4o ($p < 0.001$), with a mean word count of 288.93 (55.17) and 211.00 (65.99) words, respectively (Figure 3).

DISCUSSION

In the present study, we conducted a comparative analysis of responses generated by DeepSeek-R1 and ChatgPT-4o to commonly asked questions about UTUC. Two experts specialising in UC assessed LLMs' responses using an ordinal 4-point rating scale. Additionally, we evaluated response repeatability through cosine similarity analysis, as well as the length of the answers.

To our knowledge, this is the first research to evaluate DeepSeek-R1, as a patient information source on UTUC. Furthermore, this study represents the first comparative analysis of the two state-of-the-art LLMs: ChatGPT-4o and DeepSeek-R1 in the field of urology.

This analysis builds upon our previous research [9], which evaluated the performance of ChatGPT-3.5 in providing patient information about UTUC based on 16 patient-centered queries. The aforementioned paper identified limitations in ChatGPT-3.5's ability to generate information about UTUC, particularly in highly specialised aspects. These restrictions may be attributed to the low incidence of UTUC and the rapid advancements in its treatment [2], which contribute to the presence of misleading and contradictory information online. The lack of accessible, reliable, and user-friendly medical resources is particularly concerning for rare and aggressive malignancies, such as UTUC, where patient compliance is crucial in treatment outcomes. Therefore, further evaluation of advanced LLMs is essential to provide a reliable patient information source for UTUC.

ChatGPT-4o demonstrated significant improvements over ChatGPT-3.5 in generating accurate and detailed treatment recommendations for urological cancers, aligning more closely with clinical guidelines and expert opinions [10]. The core knowledge base of ChatGPT-4o was last updated in June 2024, and the model is capable of retrieving information from the web, allowing it to generate more up-to-date responses.

In January 2025, DeepSeek introduced DeepSeek-R1, a model designed to compete directly with ChatGPT-4o, based on an open-source architecture with better cost efficiency than OpenAI's model [11].

Both LLMs achieved a median response score of 3.5, with ChatGPT-4o demonstrating slightly greater variability in answer quality. However, there was no statistically significant difference between the performance of the two chatbots. This indicates that both models provide comparable levels of information quality about UTUC for patients.

The significant variations between the scores assigned by the two evaluators reflects the subjective nature of the rating process. Personal biases and preconceived notions about AI may have further influenced assessment. However, these differing approaches contributed to a balanced final evaluation.

During evaluation, responses with a median score of 3 or higher were considered sufficiently accurate to serve as preliminary information sources. Based on this criterion, both DeepSeek-R1 and ChatGPT-4o provided satisfactory responses for over 90% of the commonly asked patient questions. This suggests that both LLMs can serve as a relatively reliable first-line information source for patients, in most cases. However, LLMs cannot replace specialist medical consultation. We recommend that all AI-generated medical information should include a disclaimer stating that a direct consultation with a specialist remains the most reliable source of information. Notably, some responses already incorporated recommendations for specialist consultations.

Notably, most of the lowest-rated responses belonged to the Treatment category, with one exception from the Symptoms and Diagnosis category. Providing accurate answers to these questions required up-to-date knowledge of UTUC treatment, which can only be provided by experienced urologists. This is particularly important, because patients with oncologic diseases may seek alternative treatment options. Contradictory or unclear information about surgery could discourage them from appropriate medical care.

Importantly, none of the responses were identified as potentially dangerous to patients. Based on the evaluation of 43 queries about UTUC, both DeepSeek-R1 and ChatGPT-4o can be considered as safe tools for preliminary information searching.

DeepSeek-R1 performed best in the Prognosis category, but worst in the General information category. In contrast, ChatGPT-4o performed best in the general information and the symptoms and diagnosis categories, but struggled in the treatment category. There was no statistically significant difference between the performance of the two LLMs across any category. This study presented the same results about the performance of ChatGPT across categories as our previous study [9]. Differences in responses between DeepSeek-R1 and ChatGPT-4o are likely attributable to variations in their training data and underlying algorithms. However, an insight into individual questions leads to the common conclusion for both LLMs – chatbots provide comprehensive responses on the basic aspects of UTUC, but struggle with highly specialised topics across all categories.

Since a single response from a LLM does not allow for general conclusions, we utilised the cosine similarity test to objectively assess the repeatability of responses provided by the DeepSeek-R1 and ChatGPT-4o. Both LLMs demonstrated satisfactory consistency in responses across two different days, with no significant difference in repeatability between them. Even responses with the lowest cosine similarity still responded to the question without altering the core response. Cosine similarity score reductions were primarily caused by additional elaboration. This suggests that DeepSeek-R1 and ChatGPT-4o are reliable in maintaining response consistency regarding UTUC. Due to the high repeatability of responses across the two days, an expert assessment of the responses received on February 22 was deemed unnecessary.

Evaluating experts emphasised that LLMs often provided fully correct answers to the posed questions, but continued to elaborate unnecessarily. Most misinformation was found within these additional explanations, rather than in the core response itself. Notably, DeepSeek-R1 generated significantly longer responses than ChatGPT-4o. Extended replies could reduce readability and introduce misinformation that might otherwise be avoided.

DeepSeek-R1's responses were generated using the DeepThink function. This feature enables the LLM to first engage in advanced reasoning, allowing it to analyse the issue before formulating its final answer. The final response is structured based on this prior analysis. In our study, we assessed only the final responses, which directly addressed the questions. While the reasoning process is not essential for understanding the final answer, it gives the response a more human-like tone. Furthermore, the step-by-step analysis may enhance patients' comprehension of complex medical information by providing additional context and explanation.

Beyond UTUC, ChatGPT has been assessed by patient-centered queries related to oncologic urology, including kidney, bladder, prostate, testicular cancers [12–15]. Additionally, ChatGPT has been evaluated in other urological conditions, such as benign prostate hyperplasia (BPH), urolithiasis, paediatric urology and andrology [13, 16–18].

In a study by Choi et al. [12], 24 urologists assessed ChatGPT-3.5's responses regarding kidney cancer, providing an overall positive rating of 77.9%. However, they found that 70.8% of respondents thought that ChatGPT could not replace explanations provided by urologists.

Coshun et al. [13] compared ChatGPT's responses with a reference source of patient information on prostate cancer and found suboptimal performance of the chatbot. In this study, ChatGPT achieved a mean score of 3.62 ±0.49 on a 5-point scale. However, these results may now be considered outdated, due to the rapid advancements in AI since January 2023.

More recent research [14] evaluated ChatGPT-4's responses on prostate, bladder, kidney and testicular cancers. In this study, the majority of responses for each malignancy received a score of 5 on a 5-point scale, with mean scores ranging between 4.4 and 4.5.

Szczesniewski et al. [15] assessed the quality of information provided by ChatGPT on bladder, prostate, renal cancers, BPH and urolithiasis. ChatGPT provided well-balanced general information across all five conditions. Responses for all conditions achieved a DISCERN score of 4 out of 5, except for BPH with the lowest score of 3 out of 5.

Studies published by Cakir et al. [16] and Calgar et al. [17, 18] assessed ChatGPT's performance in providing information on urolithiasis, pediatric urology and andrology. ChatGPT provided 94.6%, 92% and 87.9% of correct answers for these topics, respectively. Additionally, the repeatability of ChatGPT's responses, defined as receiving the same score for identical queries over time, was positively evaluated in these three studies.

Crucially, the direct comparison of these studies is not possible, due to heterogeneous methodology, evaluation criteria and assessment scales. Furthermore, results in these studies were influenced by subjective factors, such as evaluator opinions and differences in question difficulty. Nonetheless, these analyses collectively show ChatGPT's potential as a valuable preliminary source of information before specialist consultation, which is consistent with the results of our research.

Beyond the field of urology, many studies have explored the ChatGPT's potential as a source of patient information. Bayley et al. [19], compared ChatGPT to the traditional search engine Google in providing patient information about breast cancer. ChatGPT demonstrated superior performance, achieving a mean score of 4.3 (0.8), in contrast with Google's 2.8 (1.1). In addition, Johnson et al. [20] evaluated ChatGPT's ability to address common cancer myths and misconceptions, reporting an 96.9% accuracy rate. Finally, Abreu et al. [5] demonstrated that ChatGPT-4o could significantly improve the readability of professional oncology-related content while preserving content quality. These findings suggest that patients may increasingly rely on LLMs, rather than on conventional search engines. Looking ahead, a chatbot trained exclusively on verified data could provide fully reliable and comprehensible information to the general population.

Notwithstanding, this study has several limitations that should be acknowledged. First, the evaluation relied on two urologists, who assessed LLMs responses subjectively and their scores varied significantly. Second, our study focused on UTUC, a rare malignancy, which limits the applicability of these findings to more prevalent conditions. Finally, given the rapid advancements in AI technology, our results may quickly become outdated.

CONCLUSIONS

Both DeepSeek-R1 and ChatGPT-4o predominantly provide satisfactory responses to patient-important questions about UTUC. The chatbots demonstrate potential as the first-line information sources for patients. However, LLMs struggle with highly specialised inquiries. Therefore they may serve as a helpful preliminary reference, but cannot replace expert medical advice.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ETHICS APPROVAL STATEMENT

The ethical approval was not required.

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024; 74: 12-49.
2. EAU Guidelines. Edn. presented at the EAU Annual Congress Paris 2024.
3. Krajewski W, Łaszkiewicz J, Nowak Ł, Szydełko T. Current methods facilitating diagnosis of upper tract urothelial carcinoma: a comprehensive literature review. Curr Opin Urol. 2023; 33: 230-238.
4. Pikul MV, Stakhovsky EO. Efficacy of combined organ-sparing management of invasive upper urinary tract urothelial carcinoma. Cent European J Urol. 2023; 76: 162-166.
5. Abreu AA, Murimwa GZ, Farah E, et al. Enhancing Readability of Online Patient-Facing Content: The Role of AI Chatbots in Improving Cancer Information Accessibility. J Natl Compr Canc Netw. 2024; 22: e237334.
6. van de Poll-Franse LV, van Eenbergen MC. Internet use by cancer survivors: current use and future wishes. Support Care Cancer. 2008; 16: 1189-1195.
7. Nedbal C, Bres-Niewada E, Dybowski B, Somani BK. The impact of artificial intelligence in revolutionizing all aspects of urological care: a glimpse in the future. Cent European J Urol. 2024; 77: 12-14.
8. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond). 2023; 3: 141.
9. Łaszkiewicz J, Krajewski W, Tomczak W, et al. Performance of ChatGPT in providing patient information about upper tract urothelial carcinoma. Contemp Oncol (Pozn). 2024; 28: 172-181.
10. Tsai CY, Cheng PY, Deng JH, Jaw FS, Yii SC. ChatGPT v4 outperforming v3.5 on cancer treatment recommendations in quality, clinical guideline, and expert opinion concordance. Digit Health. 2024; 10: 20552076241269538.
11. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in Healthcare: Revealing Opportunities and Steering Challenges of a New Open-Source Artificial Intelligence Frontier. Cureus. 2025; 17: e79221.
12. Choi J, Kim JW, Lee YS, et al. Availability of ChatGPT to provide medical information for patients with kidney cancer. Sci Rep. 2024; 14: 1542.
13. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer? Urology. 2023; 180: 35-58.
14. Ozgor F, Caglar U, Halis A, et al. Urological Cancers and ChatGPT: Assessing the Quality of Information and Possible Risks for Patients. Clin Genitourin Cancer. 2024; 22: 454-7.e4.
15. Szczesniewski JJ, Tellez Fouz C, Ramos Alba A, Diaz Goizueta FJ, García Tello A, Llanes González L. ChatGPT and most frequent urological diseases: analysing the quality of information and potential risks for patients. World J Urol. 2023; 41: 3149-3153.

16. Cakir H, Caglar U, Yildiz O, Meric A, Ayranci A, Ozgor F. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. Int Urol Nephrol. 2024; 56: 17-21.

17. Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2024; 20: 26.e1-.e5.

18. Caglar U, Yildiz O, Ozervarli MF, et al. Assessing the Performance of Chat Generative Pretrained Transformer (ChatGPT) in Answering Andrology-Related Questions. Urol Res Pract. 2023; 49: 365-369.

19. Bayley EM, Liu HY, Bonetti MA, Egro FM, Diego EJ. ChatGPT as Valuable Patient Education Resource in Breast Cancer Care. Ann Surg Oncol. 2025; 32: 653-655.

20. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr. 2023; 7: pkad015. ■