Comparative assessment of AI models in addressing questions on priapism: An evaluation of response quality and clinical utility

Halis Ahmet[1], Hacibey Ibrahim[2]
[1]Department of Urology, Yedikule Chest Diseases and Chest Surgery Training and Research Hospital, Istanbul, Turkey; [2]Department of Urology, Bagcilar Training and Research Hospital, Istanbul, Turkey

Cite as: Ahmet H, Ibrahim H. Comparative assessment of AI models in addressing questions on priapism: An evaluation of response quality and clinical utility**.** *Can Urol Assoc J* 2025 November 25; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.9302

Published online November 25, 2025

Corresponding author: Dr. Halis Ahmet, Department of Urology, Yedikule Chest Diseases and Chest Surgery Training and Research Hospital, Istanbul, Turkey; dr.ahmethls@gmail.com

***

ABSTRACT

Introduction: This study aimed to evaluate the performance of three artificial intelligence (AI) models — ChatGPT, Gemini, and Copilot — in addressing priapism-related inquiries. The accuracy, comprehensiveness, and clinical applicability of AI-generated responses were systematically analyzed.
Methods: Frequently asked questions (FAQs) regarding priapism were collected from medical guidelines, literature, and online health platforms. Each AI model generated responses, which were independently assessed by two experts based on accuracy, fluency, and clinical relevance. The Global Quality Score (GQS) was used for evaluation. Statistical analysis was performed using one-way ANOVA, with a significance threshold of $p<0.05$.
Results: ChatGPT and Gemini demonstrated comparable performance across all thematic categories, with mean scores ranging from 4.5–4.9, while Copilot showed significantly lower scores (3.2–4.2, $p<0.001$). Both ChatGPT and Gemini provided clinically relevant and accurate information, whereas Copilot's responses frequently lacked guideline-based recommendations.
Conclusions: ChatGPT and Gemini were statistically comparable in generating reliable, clinically useful responses, making them valuable tools for medical education and patient counseling. Copilot, however, exhibited lower accuracy and applicability. These findings highlight the need for continuous refinement of AI models to enhance their role in clinical decision-making while ensuring human expertise remains central to patient care.

INTRODUCTION

Access to medical information is rapidly evolving through digital platforms and artificial intelligence (AI)-powered applications (1). AI-based text generation models are playing an increasingly significant role in healthcare communication by providing users with quick and accessible answers to medical inquiries (2). However, the accuracy, reliability, and clinical applicability of the information generated by these applications remain inadequately explored (3).

Priapism is a rare urological emergency defined as a prolonged, pathological erection lasting more than four hours, which may be painful or painless(4). This condition can occur in both pediatric and adult populations and may lead to severe complications such as permanent erectile dysfunction if not treated promptly (4). Priapism is generally classified into three main subtypes: ischemic (low-flow) priapism, non-ischemic (high-flow) priapism, and recurrent or stuttering priapism. Access to accurate information about priapism is crucial for both patients and healthcare professionals (5). Although AI applications can provide responses to frequently asked questions about priapism, the quality and clinical validity of these responses have not been sufficiently investigated (3).

This study aims to evaluate the responses of three widely used artificial intelligence platforms—Gemini Advanced, ChatGPT Plus, and Microsoft Copilot (paid versions)—to questions related to priapism, and to compare these responses in terms of quality, factual accuracy, and clinical applicability. The objective is to assess the current capabilities of these subscription-based models in delivering reliable medical information and to identify domains requiring further improvement.

METHODS

This study was designed to evaluate the responses of AI models (Gemini Advanced, ChatGPT Plus, and Microsoft Copilot) to frequently asked questions (FAQs) regarding priapism. The study aims to compare the quality and clinical applicability of responses provided by each model. Frequently asked questions about priapism were gathered from social media platforms, health forums, hospital websites, academic guidelines, and medical literature. The collected questions were categorized into key topics such as etiology, symptoms, diagnosis, treatment, and follow-up. Additionally, guideline-based questions were formulated based on the recommendations of the European Association of Urology (EAU).

The selected questions were presented individually to each AI model (Gemini Advanced, ChatGPT Plus, and Microsoft Copilot). The prepared questions and the responses provided by the artificial intelligence applications are presented in Supplementary File 1. The responses generated by each model were recorded and stored for further analysis. AI-generated responses were assessed by two independent experts, both of whom have over five years of experience in the field of andrology. The evaluation was based on the following criteria: quality (accuracy, comprehensiveness, and fluency of the response), correctness (alignment with clinical knowledge), and usability (practical utility for patients and healthcare professionals). The evaluation was conducted using the Global Quality Score (GQS), which employs a rating scale ranging from 1 (low quality) to 5 (high quality). In

cases of discrepancy, a consensus was reached through discussion. If consensus could not be achieved, a third urologist with more than five years of experience in andrology was consulted to make the final decision.

Statistical analysis

Statistical analysis was performed using IBM's Statistical Package for the Social Sciences, version 27 (SPSS, Armonk, NY, USA). Normality assessment was checked with the Shapiro–Wilk test. Scores of FAQ subcategories are presented as percentages. Scores were compared between ChatGPT, Google Gemini, and Copilot using the One Way ANOVA test. Data were analyzed at 95% confidence level, and a p-value less than 0.05 was considered statistically significant. The agreement between the scores assigned by the two reviewers was assessed using the intraclass correlation coefficient (ICC).

RESULTS

The comparative analysis of ChatGPT, Gemini, and Copilot in addressing frequently asked questions (FAQs) related to priapism demonstrated significant variations in performance across different thematic categories (Table 1).

ChatGPT consistently outperformed the other AI models, achieving the highest mean scores across all categories, with statistically significant differences ($p < 0.001$) compared to Gemini and Copilot (Table 2).

ChatGPT exhibited a near-perfect performance, particularly in the domains of etiology, symptoms, diagnosis, treatment, prognosis, and follow-up, with mean scores ranging from 4.7 to 4.9 out of 5. The model provided highly accurate and comprehensive responses, particularly excelling in treatment-related inquiries (4.7±0.4) and prognosis (4.9±0.1), suggesting a strong ability to synthesize and convey relevant clinical information.

Gemini, while demonstrating moderate performance, consistently scored lower than ChatGPT, with mean values between 3.2 and 4.2. Although it provided relatively reliable responses, especially in diagnosis (4.2±0.4) and prognosis (4.2±0.7), its overall accuracy and depth of responses were inferior to those of ChatGPT.

Copilot demonstrated the lowest performance among the three models, with mean scores ranging from 3.2 to 3.9, significantly trailing behind ChatGPT and Gemini ($p < 0.001$). The model exhibited the greatest limitations in follow-up-related inquiries (3.2±0.6), indicating potential deficiencies in delivering evidence-based recommendations regarding long-term management.

These findings highlight ChatGPT's superior ability to provide accurate, detailed, and clinically relevant responses to priapism-related inquiries, making it the most reliable AI model among the three for medical information retrieval in this domain.The analysis of the agreement between the reviewers' scores revealed an ICC of 0.91. This result indicates an excellent level of concordance between the ratings (Figure 1).

DISCUSSION

The application of artificial intelligence (AI) in retrieving medical information has seen substantial advancements in recent years, with various models demonstrating the potential to

deliver accurate and accessible health-related answers (6). However, the reliability, comprehensiveness, and clinical applicability of AI-generated responses remain key considerations, particularly in the context of medical emergencies such as priapism. Our study provides a direct comparative evaluation of the three leading AI models—ChatGPT, Gemini, and Copilot—assessing their ability to address priapism-related inquiries from multiple platforms, including guideline-based questions.

Our results indicate that ChatGPT consistently outperformed both Gemini and Copilot across all thematic domains, achieving the highest mean scores in General Understanding, Clinical Application, Patient Outcomes, and Safety & Complications. This superiority suggests that ChatGPT provides the most reliable, clinically accurate, and comprehensive responses, making it a valuable tool for healthcare professionals and patients seeking evidence-based information.

Previous studies have highlighted the challenges of misinformation in online health content, particularly on social media platforms (7). For example, Alsyouf et al. demonstrated that misleading and inaccurate information regarding urological malignancies was significantly more prevalent than accurate content (8). Unlike social media sources, AI models derive information from structured datasets, which may contribute to improved reliability. In our study, ChatGPT's responses consistently received the highest Global Quality Scores (GQS), significantly surpassing both Gemini and Copilot, reinforcing its ability to synthesize and present high-quality medical information.

The ability to generate responses aligned with clinical guidelines is crucial for the applicability of AI models in healthcare. Cakır et al. previously reported that ChatGPT correctly answered 80% of questions based on the European Association of Urology (EAU) guidelines on urolithiasis (9). Similarly, our study demonstrated that ChatGPT excelled in answering priapism-related questions derived from EAU guidelines, with a near-perfect accuracy rate. In contrast, Gemini performed moderately well, while Copilot's responses were the least reliable, frequently lacking critical clinical recommendations.

Another key factor influencing the usability of AI-generated medical content is readability and comprehensibility. Moons and Van Bulck's research has shown that AI models can effectively summarize patient education materials at an accessible level (10). Our findings corroborate this, as ChatGPT not only provided the most clinically accurate responses but also structured them in a way that facilitated better patient understanding. Gemini performed reasonably well in this regard, whereas Copilot's responses were often overly general, limiting its utility in complex medical inquiries.

Despite ChatGPT's superiority, certain limitations persist regarding AI models' applicability in medical decision-making. Patel et al. previously noted that AI-generated content might exaggerate the urgency of medical conditions, leading to increased patient anxiety and healthcare costs (11). In our study, Copilot, in particular, demonstrated a lack of specificity in distinguishing between ischemic and non-ischemic priapism, which could contribute to misinterpretation by patients. Additionally, while ChatGPT provided the most detailed and accurate responses, some responses still favored generalization over nuanced clinical detail, as reflected in the proportion of Grade 2 responses.

A major strength of our study is the direct comparison of multiple AI models using standardized evaluation criteria, providing an objective assessment of their performance. The inclusion of both patient-oriented and guideline-based questions enhances the generalizability of our findings, while the use of an independent expert panel for response evaluation minimizes bias. However, a potential limitation is the relatively small number of evaluators, which could introduce inter-observer variability. Although ChatGPT demonstrated superior performance compared to Gemini and Copilot in terms of accuracy, comprehensiveness, and clinical relevance, it is important to acknowledge that not all questions were answered with the same level of precision. In certain instances, even the highest-performing model (ChatGPT) provided responses that lacked sufficient clinical nuance or deviated from guideline-based recommendations. Therefore, while these AI tools hold promise for enhancing access to medical information, the findings of this study underscore that their outputs should not yet be solely relied upon to guide clinical decision-making. Patients and healthcare providers must remain cautious, and human oversight remains essential to ensure safe and effective care. Future research with larger expert panels, additional AI models, and real-world clinical validation could provide deeper insights into the evolving role of AI in medical information dissemination.

CONCLUSIONS

Our study highlights ChatGPT's clear superiority in providing clinically relevant, comprehensive, and high-quality information on priapism. ChatGPT significantly outperformed Gemini and Copilot across all evaluated domains, particularly excelling in General Understanding, Clinical Application, Patient Outcomes, and Procedural Guidance. Gemini demonstrated moderate performance, while Copilot lagged behind, with lower accuracy and comprehensibility. These findings emphasize the need for continuous refinement of AI models to improve their reliability and clinical applicability. As AI technologies continue to advance, ChatGPT stands out as the most reliable model for healthcare-related inquiries. However, its integration into medical practice should be approached with a balanced perspective—leveraging its strengths while ensuring that human expertise remains central to patient care and decision-making.

REFERENCES

1. Davenport T, Kalakota, R. The potential for artificial intelligence in healthcare. Future Healthcare J 2019;6:94-8. https://doi.org/10.7861/futurehosp.6-2-94
2. OpenAI. 2022. ChatGPT: Optimizing language models for dialogue
3. Szczesniewski JJ, Telez Fouz C, Ramos Alba A, et al. ChatGPT and most frequent urological diseases: Analyzing the quality of information and potential risks for patients. *World J Urol* 2023;41:3149-53. https://doi.org/10.1007/s00345-023-04563-0
4. European Association of Urology. 2023. EAU Guidelines
5. Bernard A, Morgan L, Hughes S, et al. A systematic review of patient inflammatory bowel disease information resources on the world wide web. *Am J Gastroenterol* 2007;102:2070-7. https://doi.org/10.1111/j.1572-0241.2007.01325.x
6. Ge, L, Agrawal R, Singer M, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. *Syst Rev* 2024;13:269. https://doi.org/10.1186/s13643-024-02682-2
7. Loeb S, Taylor J, Borin JF, et al. Fake news: Spread of misinformation about urological conditions on social media. *Eur Urol Focus* 2020;6:437-9. https://doi.org/10.1016/j.euf.2019.11.011
8. Alsyouf M, Jaber HM, Abufaraj M, et al. The quality of online information regarding urological malignancies: A comparative analysis. *Urol Oncol* 2021;39:,670.e1-8. https://doi.org/10.1016/j.urolonc.2021.05.013
9. 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. https://doi.org/10.1007/s11255-023-03773-0
10. Forsyth F, Van Bulck L, Daelman B, et al. When the computer says yes, but the healthcare professional says no: Artificial intelligence and possible ethical dilemmas in health services. *Eur J Cardiovasc Nurs* 2024;23:e165-6. https://doi.org/10.1093/eurjcn/zvae059
11. Patel SB, Lam K. ChatGPT: The future of discharge summaries? *Lancet Digital Health* 2023;5:e107-8. https://doi.org/10.1016/S2589-7500(23)00021-3

FIGURES AND TABLES

Figure 1. This radar chart illustrates the comparative performance of three AI models — ChatGPT, Gemini, and Copilot — across five key evaluation criteria: relevance, factual accuracy, clarity/coherence, structure, and utility. ChatGPT consistently demonstrated the highest scores across all parameters, particularly excelling in factual accuracy and clarity. Gemini followed with moderate performance, while Copilot scored the lowest, showing notable deficiencies in structure and clarity. These findings highlight ChatGPT's superior ability to provide structured, accurate, and clinically relevant responses in priapism-related medical inquiries .

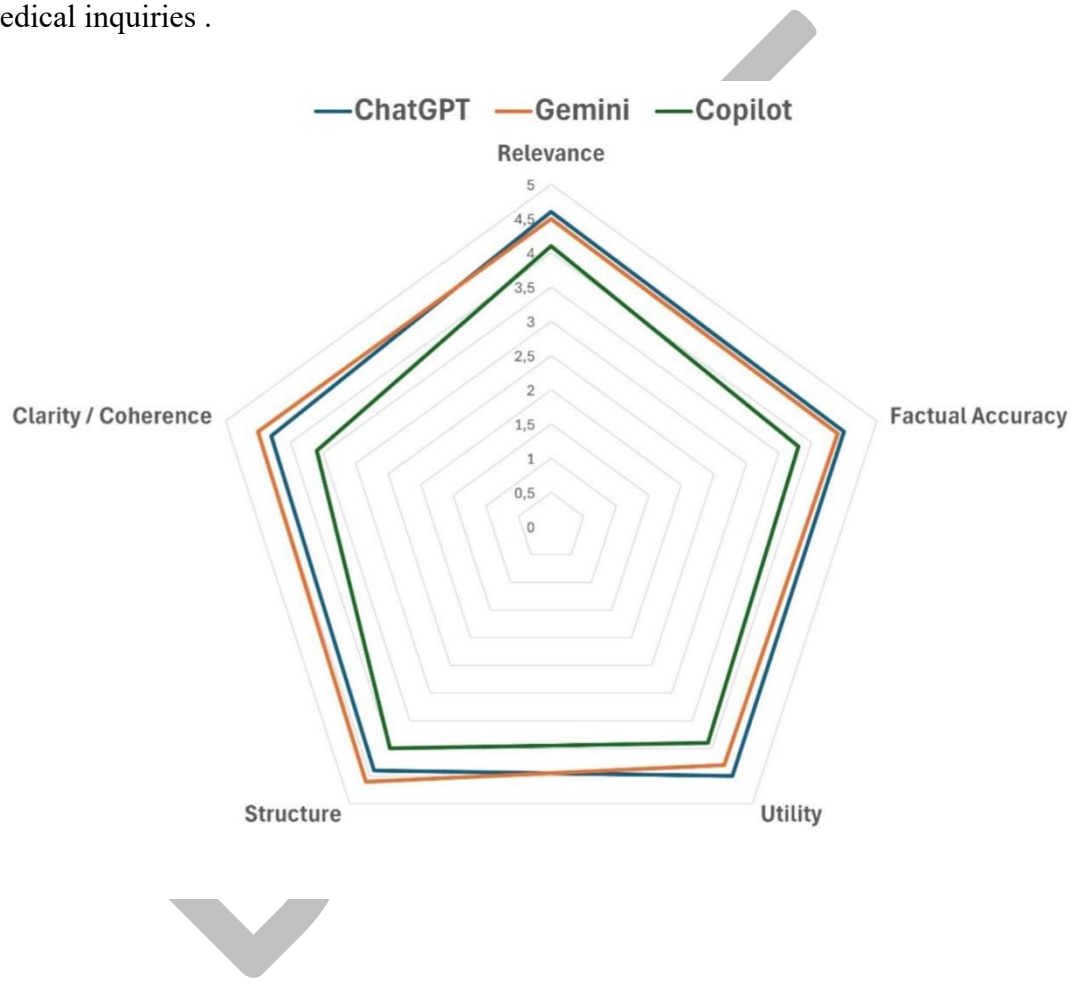

Table 1. Performance of ChatGPT, Gemini, and Copilot in answering frequently asked questions abou priapism across thematic categories						
	ChatGPT		**Gemini**		**Copilot**	
	1–4	**5**	**1–4**	**5**	**1–4**	**5**
FAQs (n=78)	10 (12.8%)	68(87.2%)	12 (15.4%)	66 (84.6%)	54 (69.2%)	24 (30.8%)
Etiology (n=14)	3 (21.4%)	11 (78.6%)	3 (21.4%)	11 (78.6%)	9 (64.3%)	5 (35.7%)
Symptom (n=7)	– (0.0%)	7 (100%)	2 (28.6%)	5 (71.4%)	3 (42.9%)	4 (57.1%)
Diagnosis (n=13)	– (0.0%)	13 (100%)	1 (7.7%)	12 (92.3%)	8 (61.5%)	5 (38.5%)
Treatment (n=19)	3 (15.8%)	16 (84.2%)	4 (21.1%)	15 (78.9%)	13 (68.4%)	6 (31.6%)
Prognosis (n=9)	– (0.0%)	9 (100%)	– (0.0%)	9 (100%)	5 (55.6%)	4 (44.4%)
Follow up (n=16)	4 (25.0%)	12 (75%)	2 (12.5%)	14 (87.5%)	16 (100%)	– (0.0%)

FAQs: frequently asked questions.

Table 2. Comparative mean scores and statistical analysis of AI models in addressing priapism-related questions				
	ChatGPT	**Gemini**	**Copilot**	**p**
FAQs (n=78)	4.8±0.3[a]	4.7±0.5[a]	3.9±0.7[b]	0.001
Etiology (n=14)	4.7±0.4[a]	4.6±0.6[a]	4.0±0.6[b]	0.001
Symptom (n=7)	4.8±0.3[a]	4.5±0.4[a]	4.2±0.4[b]	0.001
Diagnosis (n=13)	4.8±0.2[a]	4.9±0.3[a]	4.0±0.6[b]	0.001
Treatment (n=19)	4.7±0.4[a]	4.5±0.7[a]	4.0±0.5[b]	0.001
Prognosis (n=9)	4.9±0.1[a]	4.8±0.3[a]	4.2±0.7[b]	0.001
Followup (n=16)	4.8±0.4[a]	4.9±0.3[a]	3.2±0.6[b]	0.001

Lower-case letters are used to identify the group that makes the difference. The same letters (such as a-a) indicate that there is no difference, different letters (such as a-b) indicate that there is a difference. Mean±standard deviation, FAQ: frequently asked questions