ARAŞTIRMA YAZISI | ORIGINAL ARTICLE

Androl Bul 2025;27:150–154
https://doi.org/10.24898/tandro.2025.56244

Erkek Cinsel Sağlığı

Artificial intelligence responses to penile fracture: assessing accuracy and clinical utility

Penil fraktüre yönelik yapay zekâ platformlarının verdiği yanıtların doğruluğu ve klinik kullanılabilirliğinin değerlendirilmesi

Ibrahım Hacıbey[1], Ahmet Halis[2]

ABSTRACT

OBJECTIVE: This study aims to assess the accuracy and clinical utility of artificial intelligence (AI) platforms in responding to questions related to penile fracture, a rare but urgent urological emergency.

MATERIAL and METHODS: Twenty-five questions addressing key clinical aspects of penile fracture were submitted to four AI platforms: ChatGPT, Copilot, Gemini, and Perplexity. Two expert urologists evaluated each response across five domains –relevance, clarity, structure, utility, and factual accuracy– using a 5-point Likert scale. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC), and statistical comparisons were made using one-way ANOVA and Tukey's post-hoc tests.

RESULTS: Copilot and ChatGPT scored highest overall, with mean scores of 4.90 and 4.89 respectively, while Perplexity scored significantly lower (4.68; p <0.001). Copilot also achieved the highest ratings in clarity and factual accuracy. Inter-rater reliability was high, and dimensional analysis confirmed the consistent superiority of Copilot and ChatGPT in clinical relevance and clarity.

CONCLUSION: While AI platforms –especially Copilot and ChatGPT– show promise in generating medically relevant content about penile fracture, limitations in factual accuracy and clinical specificity remain. Caution is advised in using these tools in urgent care settings without professional oversight.

Keywords: artificial intelligence, clinical accuracy, emergency medicine, large language models, penile fracture, urology

ÖZ

AMAÇ: Bu çalışmanın amacı, yapay zekâ (YZ) platformlarının nadir ancak acil bir ürolojik durum olan penis fraktürü ile ilgili sorulara verdikleri yanıtların doğruluğunu ve klinik yararlılığını değerlendirmektir.

GEREÇ ve YÖNTEMLER: Penis fraktürüyle ilgili temel klinik konuları kapsayan 25 soru, dört YZ platformuna (ChatGPT, Copilot, Gemini ve Perplexity) yöneltilmiştir. Her yanıt, iki uzman ürolog tarafından beş puanlık Likert ölçeğiyle "ilgililik", "anlaşılırlık", "yapı", "klinik yararlılık" ve "gerçeklik" başlıklarında değerlendirilmiştir. Değerlendiriciler arası uyum intraclass korelasyon katsayısı (ICC) ile ölçülmüş ve istatistiksel analiz için tek yönlü ANOVA ile Tukey post-hoc testleri kullanılmıştır.

BULGULAR: Genel ortalama puanlara göre Copilot (4,90) ve ChatGPT (4,89) en yüksek puanları almıştır. Perplexity'nin skoru anlamlı şekilde daha düşük bulunmuştur (4,68; p <0,001). Copilot, özellikle anlaşılırlık ve gerçeklik kriterlerinde en iyi puanları almıştır. Değerlendiriciler arası uyum yüksek bulunmuş ve boyutsal analiz, Copilot ve ChatGPT'nin klinik açıdan tutarlı şekilde üstün performans sergilediğini göstermiştir.

SONUÇ: Yapay zekâ platformları –özellikle Copilot ve ChatGPT– penis fraktürü hakkında tıbbi açıdan anlamlı içerikler oluşturma potansiyeline sahiptir. Ancak, tüm modellerde gözlenen gerçeklik ve klinik detay eksiklikleri, bu araçların acil klinik kararlarda profesyonel gözetim olmaksızın kullanılmaması gerektiğini göstermektedir.

Anahtar Kelimeler: yapay zeka, klinik doğruluk, acil tıp, büyük dil modelleri, penis fraktürü, üroloji

INTRODUCTION

Penile fracture is a rare but emergent condition that requires immediate intervention. It typically occurs during sexual intercourse or as a result of blunt trauma.[1] The rupture of the corpus cavernosum leads to this condition, which necessitates urgent medical intervention to prevent complications. If not treated promptly, penile fracture can lead to permanent erectile dysfunction and psychological trauma.[1] Therefore, proper diagnosis and management of the condition are critical components of urological practice.

In recent years, the use of artificial intelligence (AI) technologies in medicine has rapidly increased. AI is being employed in a wide range of applications, from clinical decision support systems to patient monitoring, assisting in disease diagnosis and treatment processes.[2,3] The effectiveness of AI in medical diagnosis, particularly in decision

[1]*Basaksehir Çam and Sakura City Hospital, Department of Urology, Istanbul, Türkiye*
[2] *Yedikule Chest Diseases and Chest Surgery Training and Research Hospital, Department of Urology, Istanbul, Türkiye*

Yazışma Adresi/ *Correspondence:*
Ibrahım Hacıbey
Başakşehir Mahallesi G-434 Caddesi No: 2l Basaksehir / Istanbul, Türkiye 34100
Istanbul - Türkiye
Tel: +90 543 728 77 87
E-mail: drihacibey@gmail.com

Geliş/ *Received:* 06.04.2025
Kabul/ *Accepted:* 29.07.2025

support systems and healthcare optimization, has shown significant progress (2,3). However, the clinical knowledge and reliability of AI, especially in emergency situations, have yet to be systematically evaluated in many studies.[4] Evaluating the level of knowledge AI possesses regarding urological emergencies such as penile fracture is an important research area that may help identify the potential applications of these technologies.[4]

Recent evaluations of AI tools in emergency medicine and urological contexts have highlighted both their potential and limitations in clinical decision-making, especially in urgent care settings.[5,6]

The aim of this study is to ask various AI platforms specific questions related to penile fracture and evaluate the accuracy and scope of their responses. The responses will be scored by expert urologists and compared statistically.

MATERIALS and METHODS

Study Design

This study is a cross-sectional analysis aimed at evaluating the accuracy of responses provided by AI-based systems to questions related to penile fracture. The responses are assessed by expert urologists, and the clinical accuracy of these responses is statistically compared. This study aims to objectively measure how accurate, comprehensive, and up-to-date the AI responses are regarding penile fracture.

Data and Questions

The 25 open-ended questions used in this study cover topics related to the pathophysiology, diagnostic methods, emergency treatment options, and postoperative processes of penile fracture. The questions were formulated based on critical processes frequently encountered in the medical literature. Each question was designed to test the accuracy of clinical knowledge related to penile fracture.

Artificial Intelligence Applications

The AI platforms evaluated in this study were ChatGPT-4 (OpenAI, accessed January 10, 2025), Copilot (Microsoft Bing AI, accessed January 10, 2025), Gemini Pro (Google, accessed January 11, 2025), and Perplexity AI (Pro version, accessed January 11, 2025). These AI platforms were selected to provide responses to the 25 questions related to penile fracture. This 25 clinical questions were developed through a comprehensive review of current urology guidelines and peer-reviewed literature on penile fracture, including AUA and EAU recommendations**.** The responses provided by the AIs were evaluated by two expert urologists, who scored the answers based on the following criteria:

- **Relevance:** How closely the response relates to the question.
- **Clarity:** How clear and understandable the response is.
- **Structure:** Whether the response has a well-organized and logical structure.
- **Utility:** How useful the response is in clinical practice.
- **Factual accuracy:** How accurate the response is in terms of medical facts.

Each parameter was scored on a scale of 1 to 5 (1: very poor, 5: excellent).

Evaluator Agreement and Scoring Process

The responses provided by each AI platform were evaluated independently by two urologists. Each response was scored based on the five criteria using a 5-point Likert scale. To assess the consistency between the two evaluators, inter-rater reliability was calculated using the intraclass correlation coefficient (ICC). Specifically, a two-way random-effects model with absolute agreement was used, which is suitable for measuring consistency between multiple raters evaluating the same set of items. A high ICC value would indicate strong agreement between evaluators, supporting the robustness and reproducibility of the scoring process.

Blinding Protocol

To minimize evaluator bias, the identity of the AI platform responsible for each response was concealed from the urologists during the scoring process. All responses were anonymized and presented in a randomized order. Since the study was conducted by two researchers who also served as evaluators, responses were anonymized and manually randomized prior to scoring. Each researcher independently rated the responses, and all scores were finalized and entered into the statistical analysis software without any post-hoc modifications.

Statistical Analysis

The statistical analysis of the study compared the accuracy and consistency of the AI responses. The responses were compared based on the scores given for each criterion. SPSS software was used for data analysis, and one-way ANOVA was performed for inter- parameter comparisons. Additionally, Pearson correlation analysis was applied to

assess the relationship between the accuracy and structure of the responses. A p-value of <0.05 was considered statistically significant.

Ethical Considerations

This study did not involve human participants or identifiable personal health information. Therefore, institutional ethical approval was not required. The study was conducted in accordance with the principles outlined in the Declaration of Helsinki and adheres to accepted ethical standards for research integrity and transparency.

RESULTS

The accuracy and clinical relevance of responses provided by four AI platforms –ChatGPT, Copilot, Gemini, and Perplexity– were evaluated across five distinct criteria: Relevance, Clarity, Structure, Utility, and Factual Accuracy**.** Each response was rated by two independent expert urologists using a 5-point Likert scale, and mean scores were calculated for comparative analysis.

Overall Performance Comparison

The mean overall scores, representing general understanding across all questions, are summarized in Fig. 1. Copilot achieved the highest average score (4.90), followed closely by ChatGPT (4.89), Gemini (4.82), and Perplexity (4.68). A one-way ANOVA revealed a statistically significant difference among the AI platforms ($p < 0.001$). Post-hoc pairwise comparisons using Tukey's HSD test are presented in Table 1. These comparisons indicated that Perplexity scored significantly lower than the other three models ($p < 0.05$), whereas no significant differences were found among Copilot, ChatGPT, and Gemini.

Dimensional Performance Across Evaluation Criteria

To further evaluate the models across specific clinical dimensions, mean scores were calculated for each AI model based on the five assessment criteria. These results are presented in Table 2. Both Copilot and ChatGPT scored consistently high across all dimensions, with Copilot obtaining the highest scores in Clarity and Factual Accuracy. Perplexity showed comparatively lower performance, particularly in Utility and Factual Accuracy.

A radar chart is provided in Fig. 2, which offers a visual representation of the average performance of each AI model across the five evaluation dimensions. The chart demonstrates the relatively uniform performance of Copilot and ChatGPT, whereas Perplexity displays reduced scores across multiple dimensions.

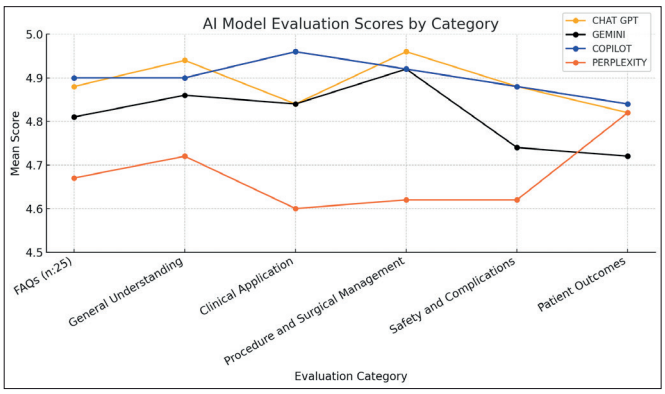

Figure 1. Mean scores of AI models based on combined evaluations from two reviewers.

Table 1. Pairwise comparison of LLMs' accuracy scores (Tukey HSD p-values)

Models	*ChatGPT-4*	*Gemini*	*Copilot*	*Perplexity*
ChatGPT-4	-	0.24	0.98	<0.001*
Gemini	0.24	-	0.16	0.01*
Copilot	0.98	0.16	-	<0.001*
Perplexity	<0.001*	0.01*	<0.001*	-

**p<0.05*

Table 2. Mean evaluation scores of four AI models across five clinical assessment dimensions

AI_Model	*Relevance (95% CI)*	*Clarity (95% CI)*	*Structure (95% CI)*	*Utility (95% CI)*	*Factual Accuracy (95% CI)*
ChatGPT-4	5.00 (4.94–5.06)	4.98 (4.92–5.04)	4.90 (4.84–4.96)	4.81 (4.75–4.87)	4.76 (4.70–4.82)
GEMINI	4.96 (4.90–5.02)	4.96 (4.90–5.02)	4.86 (4.80–4.92)	4.58 (4.52–4.64)	4.72 (4.66–4.78)
COPILOT	5.00 (4.94–5.06)	5.00 (4.94–5.06)	4.84 (4.78–4.90)	4.78 (4.72–4.84)	4.88 (4.82–4.94)
PERPLEXITY	4.88 (4.82–4.94)	4.94 (4.88–5.00)	4.70 (4.64–4.76)	4.40 (4.34–4.46)	4.46 (4.40–4.52)

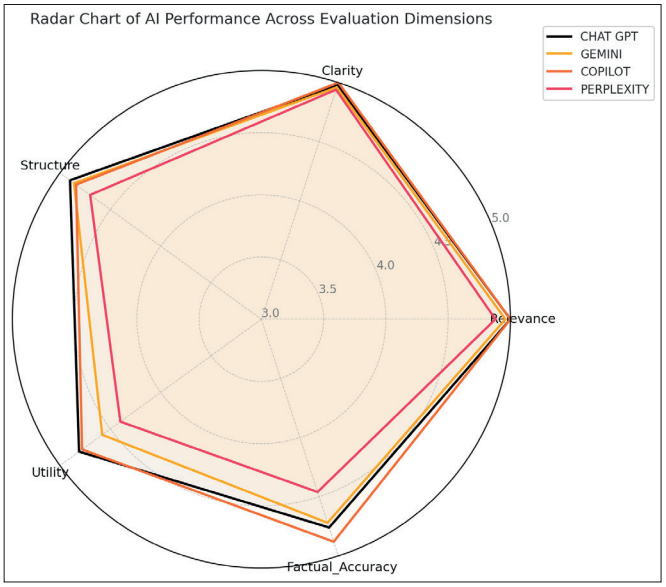

Figure 2. Radar chart illustrating the average scores of AI models across five evaluation dimensions: Relevance, Clarity, Structure, Utility, and Factual Accuracy. Higher values indicate better performance.

DISCUSSION

This study presents a novel evaluation of large language models (LLMs) in the context of a rare but urgent urological emergency: penile fracture. The condition, which involves rupture of the tunica albuginea of the corpus cavernosum, often requires immediate surgical intervention to prevent long-term complications such as erectile dysfunction and penile curvature.[7] While there is an expanding body of literature on penile fracture and its management, no prior study has assessed the reliability of AI-generated medical content on this specific topic.

Our findings demonstrate that the evaluated AI platforms –ChatGPT, Copilot, Gemini, and Perplexity– vary significantly in their clinical accuracy, clarity, and utility. Copilot and ChatGPT outperformed the others, achieving the highest overall scores across five evaluation dimensions: relevance, clarity, structure, utility, and factual accuracy. This aligns with prior studies suggesting that newer LLMs such as ChatGPT-4 provide improved medical reasoning and knowledge depth compared to earlier AI models.[8,9]

Despite the strong overall performance, even the best-performing models exhibited limitations in factual accuracy and clinical utility. This is consistent with recent investigations showing that while LLMs excel in generating linguistically polished responses, they may still hallucinate facts or omit critical clinical nuances.[10] For example, some AI responses generalized treatment approaches without emphasizing the need for immediate surgical exploration, which remains the gold standard for managing penile fracture [1,7]. Such omissions could be misleading if these tools were used unsupervised in clinical environments.

Perplexity, the lowest-performing model in our study, scored significantly lower in factual accuracy and utility, echoing previous reports about variable model performance depending on training corpus, architecture, and context sensitivity.[11] These differences highlight the importance of benchmarking AI tools individually and not assuming uniform reliability across platforms.

Our use of blinded expert scoring provides a robust and reproducible method for assessing AI-generated content. Blinding minimized cognitive bias, and the use of multiple evaluation dimensions reflects recent recommendations for AI appraisal in medicine.[12] This methodology can be adapted for evaluating AI performance across other urological conditions or emergencies.

The implications of this study are multifaceted. On one hand, AI holds promise as a supplementary educational and triage tool for urologists and patients alike. LLMs can help answer common clinical questions, generate initial explanations, and support shared decision- making in low-risk scenarios.[6] On the other hand, the variability in response quality and the presence of clinically important omissions necessitate caution in integrating these tools into emergency decision-making workflows.[13] If used without expert oversight, AI-generated responses may inadvertently reinforce common misconceptions or oversimplified interpretations of clinical conditions, posing risks in urgent care settings where nuance and accuracy are essential.[13] Despite these limitations, AI tools may offer value in low-risk scenarios such as preliminary patient education, health literacy initiatives, or as adjuncts in telemedicine consultations where clinicians remain actively involved in decision-making.[14]

Limitations of our study include its focus on a single disease entity. Although penile fracture is a high-stakes clinical topic that demands prompt recognition, the generalizability of our findings to other urological emergencies or specialties remains to be established. Furthermore, the rapid evolution of LLMs means that performance is dynamic; future updates to these models may significantly alter their capabilities, underscoring the need for periodic reassessment to ensure continued clinical reliability.[15] Ongoing, systematic validation will be necessary to monitor these changes.

CONCLUSION

This study provides a comprehensive evaluation of large language models (LLMs) in the context of a rare urological emergency –penile fracture. Among the evaluated platforms, Copilot and ChatGPT demonstrated superior performance across all dimensions, including relevance, clarity, structure, utility, and factual accuracy. However, notable shortcomings in clinical precision were observed across all models, highlighting the importance of professional oversight when integrating AI-generated content into clinical workflows.

Although these tools hold promise for educational support and low-risk clinical inquiries, their current limitations necessitate cautious use in emergency decision-making. Future studies should explore their applicability across other urological emergencies and assess performance improvements as models evolve.

Ethics Committee Approval
The study was conducted in accordance with the principles outlined in the Declaration of Helsinki and adheres to accepted ethical standards for research integrity and transparency.

Peer-review
Externally peer-reviewed.

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

Financial Disclosure
No financial support has been received.

REFERENCES

1. Amer T, Wilson R, Chlosta P, AlBuheissi S, Qazi H, Fraser M, Aboumarzouk OM. Penile fracture: a meta-analysis. Urol Int. 2016;96(3):315–29. [CrossRef]
2. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol. 2019;20(3):405–10. [CrossRef]
3. Bulten W, Kartasalo K, Chen PC, Ström P, Pinckaers H, Nagpal K, et al.; PANDA challenge consortium. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med. 2022;28(1):154–63. [CrossRef]
4. van Leeuwen KG, Schalekamp S, Rutten MJCM, Huisman M, Schaefer-Prokop CM, de Rooij M, et al.; Project AIR Working Group. Comparison of commercial AI software performance for radiograph lung nodule detection and bone age prediction. Radiology. 2024;310(1):e230981. [CrossRef]
5. Hartman V, Zhang X, Poddar R, McCarty M, Fortenko A, Sholle E, et al. Developing and evaluating large language model-generated emergency medicine handoff notes. JAMA Netw Open. 2024;7(12):e2448723. [CrossRef]
6. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. medRxiv. 2023. [CrossRef]
7. Rivas JG, Dorrego JM, Hernández MM, Portella PF, González SP, Valle JA, Barthel JJ. Traumatic rupture of the corpus cavernosum: surgical management and clinical outcomes. A 30 years review. Cent European J Urol. 2014;67(1):88-92. Epub 2014 Apr 17. PMID: 24982791; PMCID: PMC4074715. [CrossRef]
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, 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]
9. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. 2023;2023.02.02.23285399. [CrossRef]
10. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96. [CrossRef]
11. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. [CrossRef]
12. Shah M, Naik N, Somani BK, Hameed BMZ. Artificial intelligence (AI) in urology-Current use and future directions: An iTRUE study. Turk J Urol. 2020 Nov;46(Supp. 1):S27-S39. Epub 2020 May 27. PMID: 32479253; PMCID: PMC7731952. [CrossRef]
13. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, Vielhauer J, Makowski M, Braren R, Kaissis G, Rueckert D. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024 Sep;30(9):2613-2622. Epub 2024 Jul 4. PMID: 38965432; PMCID: PMC11405275. [CrossRef]
14. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43. [CrossRef]
15. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and other large language models are double-edged swords. Radiology. 2023;307(2):e230163. [CrossRef]