ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/urology.2025-111903

Assessing the Accuracy, Readability, and Clinical Applicability of Artificial Intelligence Chatbots in Primary Bladder Pain Syndrome Management: A Cross-Sectional Methodological Study

Primer Mesane Ağrı Sendromu Tedavisinde Yapay Zekâ Chatbot'larının Doğruluğu, Okunabilirliği ve Klinik Uygulanabilirliği: Kesitsel Metodolojik Araştırma

Metin KILIÇ[a], Anıl ERKAN[a], Akif KOÇ[b], Abdullah GÜL[b], Salim ZENGİN[a]

[a]Bursa Yüksek İhtisas Training and Research Hospital, Clinic of Urology, Bursa, Türkiye
[b]University of Health Sciences Bursa Faculty of Medicine, Bursa Yüksek İhtisas Training and Research Hospital, Department of Urology, Bursa, Türkiye

ABSTRACT Objective: This study aimed to evaluate the quality, readability, actionability, and guideline adherence of medical information provided by artificial intelligence chatbots (AICs) regarding treatment options for primary bladder pain syndrome (PBPS). **Material and Methods:** Four AICs were queried with the question: "What treatments are available for bladder pain syndrome?". Their responses were evaluated by 5 expert urologists using DISCERN Patient Education Materials Assessment Tool for Print Materials (PEMAT-P), the Web Resource Rating (WRR), the Coleman-Liau Index, and a guideline adherence Likert scale based on the European Association of Urology (EAU) guidelines. Data were analysed and reported using median (minimum-maximum) values for subjective scores and mean values for word count and readability. **Results:** Perplexity and Gemini achieved the highest median DISCERN scores (52), followed by Copilot and Chat Generative Pretrained Transformer (ChatGPT). Understandability was highest for Perplexity (75%), while actionability remained low across all platforms. Perplexity achieved the best WRR score (44.2), while ChatGPT scored the lowest (14.3). Readability analysis showed that AIC responses required a university-level education for comprehension, with Coleman-Liau Index scores ranging from 16.02 to 19.35. Guideline adherence according to EAU was moderate, with ChatGPT and Perplexity scoring highest (4/5). **Conclusion:** Although AICs demonstrated moderate to good reliability and understandability in providing information about PBPS treatment, concerns regarding high reading complexity and low actionability remain. AICs offer promising supplementary tools for patient education, but significant improvements in readability, actionable guidance, and clinical accuracy are needed before broader implementation in urological practice.

Keywords: Artificial intelligence; primary bladder pain syndrome; chatbots; quality assessment

ÖZET Amaç: Bu çalışma, yapay zekâ destekli chatbotların [artificial intelligence chatbots (AIC)] primer mesane ağrı sendromu (PMAS) tedavi seçenekleriyle ilgili sağladıkları tıbbi bilgilerin kalite, okunabilirlik, uygulanabilirlik ve kılavuz uyumluluğunu değerlendirmeyi amaçladık. **Gereç ve Yöntemler:** Dört AIC'ye "ağrılı mesane sendromu için hangi tedavi seçenekleri mevcuttur?" sorusu yöneltildi. Cevapları 5 uzman ürolog tarafından DISCERN, PEMAT-P (Patient Education Materials Assessment Tool for Print Materials), Web Kaynağı Derecelendirmesi [the Web Resource Rating (WRR)], Coleman-Liau İndeksi ve Avrupa Üroloji Derneği kılavuzuna uyumluluğu Likert ölçeği kullanılarak değerlendirildi. Veriler, öznel puanlar için medyan (minimum-maksimum) değerler ve kelime sayısı ile okunabilirlik için ortalama değerler kullanılarak analiz edildi ve raporlandı. **Bulgular:** Perplexity ve Gemini en yüksek medyan DISCERN puanını (52) elde etti, ardından Copilot ve "Chat Generative Pretrained Transformer" (ChatGPT) geldi. Anlaşılırlık en yüksek Perplexity'de (%75) bulunurken, uygulanabilirlik tüm platformlarda düşük kaldı. Perplexity en iyi WRR puanını (44,2) alırken, ChatGPT en düşük puanı (14,3) aldı. Okunabilirlik analizi, AIC yanıtlarının anlaşılması için üniversite düzeyinde bir eğitim gerektirdiğini gösterdi; Coleman-Liau İndeksi puanları 16,02-19,35 arasında değişti. Avrupa Üroloji Derneği [European Association of Urology (EAU)] kılavuzlarına uyumluluk orta düzeyde bulunurken; ChatGPT ve Perplexity en yüksek puanı aldı (Likert 4). **Sonuç:** AIC'ler, PMAS tedavi bilgisi sağlama konusunda orta ile iyi düzeyde güvenilirlik ve anlaşılırlık göstermiş olsa da, yüksek okuma zorluğu ve düşük uygulanabilirlik konularında endişeler devam etmektedir. AIC, hasta eğitimi için umut verici tamamlayıcı araçlar sunsa da, ürolojik uygulamalarda daha geniş kullanım öncesinde okunabilirlik, uygulanabilir rehberlik ve klinik doğruluk açısından önemli iyileştirmelere ihtiyaç duymaktadır.

Anahtar Kelimeler: Yapay zekâ, Chatbot; Primer mesane ağrı sendromu; Chatbot; Kalite değerlendirmesi

Correspondence: Metin KILIÇ
Bursa Yüksek İhtisas Training and Research Hospital, Clinic of Urology, Bursa, Türkiye
E-mail: kilicmetin@hotmail.com

Peer review under responsibility of Journal of Reconstructive Urology.

Received: 06 May 2025 ***Received in revised form:*** 30 Jun 2025 ***Accepted:*** 01 Jul 2025 ***Available online:*** 07 Jul 2025

Primary bladder pain syndrome (PBPS) is a condition characterized by an unpleasant sensation, such as pain, pressure, or discomfort, perceived to be related to the bladder and accompanied by lower urinary tract symptoms lasting for more than 6 weeks, without any identifiable infection or other underlying causes.[1,2] The cause of PBPS is not well understood, leading to a complex treatment approach that includes behavioral modifications, medications, intravesical treatments, and, in severe cases, surgery.[3,4]

With the rapid development of artificial intelligence (AI) chatbots (AICs), various AI models such as Chat Generative Pretrained Transformer (ChatGPT), Perplexity, Gemini, and Copilot have become widely used tools for obtaining health-related information. These models provide quick and accessible answers to medical queries. However, concerns remain regarding the accuracy, reliability, and clinical applicability of chatbots' generated medical information. Misinformation and incomplete content on health websites can lead to confusion for both patients and healthcare professionals. It's important to ensure that information from AIC is accurate, reliable, and relevant. This is especially crucial in fields where clear and evidence-based recommendations are necessary. This study aims to assess the quality, consistency, and adherence of AICs' generated information to established clinical guidelines. By comparing AI responses with expert evaluations and guideline recommendations, we seek to determine the potential role and limitations of AI tools in providing medical information in urology.

MATERIAL AND METHODS

The study involved querying four AIC (ChatGPT, Perplexity, Gemini, and Copilot) with the question: "What treatments are available for bladder pain syndrome?" The responses from each AIC model were evaluated by 5 expert urologists. This study is a cross-sectional methodological evaluation of AIC generated medical information, aiming to assess the quality, readability, and guideline adherence of the provided content. As this study did not involve human subjects, animal subjects or identifiable patient data therefore ethical approval was not required. However, the research was conducted in line with ethical principles consistent with the Declaration of Helsinki.

Five validated tools were used to assess the AICs' responses: DISCERN Patient Education Materials Assessment Tool for Print Materials (PEMAT-P), the Web Resource Rating (WRR) tool, the Coleman-Liau Index (CLI) and Likert scale.

DISCERN is a 16-question tool scored on a Likert scale (1-5) to assess reliability and quality. Questions 1-8 measure reliability, questions 9-15 assess the quality of treatment information, and question 16 provides an overall rating. Total scores were categorized as follows: Excellent (63-75), Good (51-62), Moderate (39-50), Poor (27-38), Very Poor (15-26). PEMAT-P evaluates understandability (questions 1-19) and actionability (questions 20-26) of patient education materials. Scores were calculated as percentages ranging from 0 to 100%. WRR consists of 13 questions: questions 1-6 assess evidence-based criteria. Questions 7-13 assess transparency and usability criteria. Scores were evaluated on a 100-point scale.

CLI is a readability formula that assesses the difficulty of a text by analyzing average word and sentence length. It calculates a readability score, indicating the U.S. grade level required to comprehend the passage. Guideline Compatibility Assessment Treatment options were evaluated using a 5-point Likert scale: a score of 1: serious or extensive deficiencies, and a score of 5: minimal deficiencies.

The 16-question DISCERN assessment underwent Intraclass Correlation Coefficient (ICC) analysis to determine inter-rater reliability. The ICC result was 0.801 [95% confidence interval (CI): 0.708-0.871; p<0.0001], indicating strong agreement among evaluators. Other assessment tools were not subjected to ICC testing, as their scoring was based on objective data using agree/disagree criteria. Scoring systems and the consistency of treatment recommendations were reported as median values (minimum-maximum), while word count and CLI scores were presented as mean values. Data analyses were performed using SPSS Statistics (version 25, IBM Corp., Armonk, NY, USA).

RESULTS

The evaluation of AICs' responses using multiple assessment tools provided insights into the quality, readability, and clinical applicability of the information provided by ChatGPT, Perplexity, Gemini, and Copilot regarding PBPS treatments.

DISCERN score, which measures the reliability and quality of consumer health information, showed variation among the AI models. Perplexity and Gemini achieved the highest median total DISCERN scores, both with a score of 52 (Perplexity: 49-46; Gemini: 51-58). They were followed by Copilot with a score of 48 (range: 45-51) and ChatGPT with 44 (range: 41-49). While Perplexity and Gemini demonstrated better adherence to quality and reliability metrics, none of the AI-generated responses achieved an "Excellent" rating based on the DISCERN scoring system (Table1).

According to the PEMAT-P tool used to assess understandability, Perplexity had the highest score with a median of 75 (range: 71-86.6), followed by Gemini at 71 (69-75), Copilot at 64 (62-66.6), and ChatGPT at 60 (58-66.6). This indicates that Perplexity and Gemini produced content that was more easily comprehensible to a general audience. However, actionability scores were consistently low across all AI models, with ChatGPT, Perplexity, and Gemini scoring 20, and Copilot scoring 0. This suggests that while the responses provided information, they lacked clear guidance on actionable steps that patients could take based on the content.

The WRR tool evaluates transparency, usability, and evidence-based quality of information. The highest WRR score was observed in Perplexity (44.2), followed by Gemini (38.2), Copilot (28.9), and ChatGPT (14.3). These results suggest that Perplexity responded with the best structural transparency and source credibility, while ChatGPT scored the lowest in this domain.

The CLI, which assesses readability by estimating the grade level required to comprehend a text, showed that Perplexity's responses were the most difficult to read (19.35), followed by Gemini (17.74), Copilot (16.65), and ChatGPT (16.02). These scores indicate that AI-generated content is written at a university-level reading complexity, making it less accessible for the general public. Regarding word count, ChatGPT generated the longest responses (357 words), followed by Perplexity (337 words), Gemini (276 words), and Copilot (234 words).

The expert panel evaluated the compatibility of AI-generated treatment recommendations with European Association of Urology (EAU) guidelines on a 5-point Likert scale. ChatGPT and Perplexity achieved the highest compatibility score (Likert 4), whereas Gemini and Copilot scored slightly lower (Likert 3). These results suggest that while AI-generated content aligns moderately well with established guidelines, discrepancies and gaps remain.

DISCUSSION

The findings of this study highlight both the potential and limitations of AICs in providing medical infor-

TABLE 1: Comparison of AI Chatbot Responses on PBPS Treatment

	ChatGPT	Perplexity	Gemini	Copilot
Total DISCERN score	44 (41-49)	52 (49-46)	52 (51-58)	48 (45-51)
PEMAT-P understandability	60 (58-66.6)	75 (71-86.6)	71 (69-75)	64 (62-66.6)
PEMAT-P actionability	20 (0-40)	20 (0-33)	20 (20-40)	0 (0-40)
WRR	14.3	44.2	38.2	28.9
Word count	357	337	276	234
CLI	16.02	19.35	17.74	16.65
Likert 1-5 (compatibility of the treatment options offered with the EAU guidelines 1-5)	4	4	3	3

ChatGPT: Chat Generative pretrained transformer; PEMAT-P: Patient education materials assessment tool for print materials; WRR: The Web resource rating tool; CLI: Coleman-Liau Index; EAU: European Association of Urology mation about PBPS. Our evaluation using validated assessment tools demonstrated that while AI-generated responses exhibit moderate to good quality and reliability, significant concerns remain regarding their readability, actionability, and adherence to clinical guidelines.

The DISCERN scores indicate that Perplexity and Gemini produced the most reliable and high-quality responses, followed by Copilot and ChatGPT. These findings align with previous research evaluating AI-generated medical information, which has noted that while AI models can provide generally accurate and well-structured content, variability in quality exists across different platforms.[5] In studies assessing AIC responses to cancer-related queries, similar trends were observed, where AI-generated responses were generally accurate but lacked the depth and context required for nuanced clinical decision-making.[6]

Our results demonstrate that AI-generated content is often written at a university-level reading complexity, with the CLI scores indicating that most responses require an advanced education level to fully comprehend. This is consistent with previous studies that have shown that AI-generated medical information is frequently too complex for the general public.[7] The readability barrier is a critical concern, as effective patient education materials should ideally be written at a 6th to 8th-grade reading level to ensure broad accessibility. Our findings regarding the low actionability and high reading complexity of AI-generated content align with those of Erkan et al. who demonstrated that AICs produced information that was difficult to read and lacked personalized, stage-specific treatment guidance for patients with urogenital cancers.[8]

Despite reasonable reliability and readability, actionability scores were notably low across all AI models, suggesting a lack of clear, patient-centered guidance on next steps. This finding corroborates prior research that has found AI-generated responses to be informative but lacking in practical, actionable recommendations.[6] Given that patient comprehension and engagement are essential for effective disease management, the inability of chatbots to provide structured, step-by-step guidance limits their usefulness in clinical settings.[9] Similarly, recent reviews have highlighted that generative AICs offer valuable support for patient education and administrative tasks in urology but remain limited in accuracy and in handling complex clinical decisions.[10]

Our Likert scale assessment of guideline adherence showed moderate agreement between AI-generated treatment recommendations and established guidelines, with ChatGPT and Perplexity scoring the highest. Previous studies evaluating AI-driven medical responses have similarly reported that while AI models can produce evidence-based content, they may omit important nuances or fail to prioritise guideline-recommended treatments.[11] The integration of real-time guideline updates into AI models could enhance their clinical relevance and reliability in medical decision-making.

This study has several limitations. First, AICs are continually evolving, and newer versions may demonstrate improved accuracy, readability, and compliance with clinical guidelines. Second, although our expert panel provided a thorough evaluation, subjective bias might have influenced the interpretation of responses. Third, AICs are unable to offer personalized medical advice, which remains a significant limitation in their use as patient education tools tools.[12]

Future research should focus on improving AIC outputs by incorporating real-time medical updates, enhancing readability through patient-centered language optimization, and refining responses to improve actionability. Additionally, regulatory oversight and AI-specific clinical guidelines will be necessary to ensure safe and effective use of AICs in urology and other medical specialties.

CONCLUSION

The rapid expansion of AICs in healthcare has introduced new possibilities for providing medical information to patients and healthcare professionals. This study assessed the quality, readability, and guideline adherence of AI-generated responses regarding treatment options for PBPS. While AI Cs such as ChatGPT, Perplexity, Gemini, and Copilot demonstrated moderate to good reliability, readability concerns and a lack of actionable guidance highlight key limitations in their application.

Perplexity and Gemini exhibited the highest overall quality and understandability, whereas ChatGPT and Copilot lagged slightly behind. However, all chatbots struggled with actionability, meaning their responses lacked clear, step-by-step recommendations for patients. Additionally, readability scores indicated that AI-generated content is often too complex for a general audience, making accessibility a concern. Although AI models showed moderate compatibility with established clinical guidelines, discrepancies and omissions remain, emphasizing the need for further refinement.

Moving forward, integrating AI Cs with real-time medical databases, improving readability through patient-centered language, and enhancing actionability with structured recommendations will be crucial to increasing their clinical utility. While AI Cs offer promising supplementary tools for disseminating medical information, they should not replace professional medical guidance. Future research and regulatory oversight will be essential to ensuring AI-driven healthcare information is accurate, understandable, and aligned with evidence-based guidelines.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

***Idea/Concept:** Metin Kılıç, Anıl Erkan; **Design:** Metin Kılıç, Akif Koç; **Control/Supervision:** Metin Kılıç, Salim Zengin; **Data Collection and/or Processing:** Metin Kılıç, Abdullah Gül; **Analysis and/or Interpretation:** Metin Kılıç, Abdullah Gül, Salim Zengin, Anıl Erkan, Akif Koç; **Literature Review:** Metin Kılıç, Salim Zengin; **Writing the Article:** Metin Kılıç, Anıl Erkan; **Critical Review:** Abdullah Gül, Akif Koç; **References and Fundings:** Metin Kılıç; **Materials:** Metin Kılıç.*

REFERENCES

1. Hanno P, Dmochowski R. Status of international consensus on interstitial cystitis/bladder pain syndrome/painful bladder syndrome: 2008 snapshot. Neurourol Urodyn. 2009;28(4):274-86. PMID: 19260081.
2. Hanno PM, Erickson D, Moldwin R, Faraday MM; American Urological Association. Diagnosis and treatment of interstitial cystitis/bladder pain syndrome: AUA guideline amendment. J Urol. 2015;193(5):1545-53. PMID: 25623737.
3. Clemens JQ, Erickson DR, Varela NP, Lai HH. Diagnosis and treatment of interstitial cystitis/bladder pain syndrome. J Urol. 2022;208(1):34-42. PMID: 35536143.
4. Marcu I, Campian EC, Tu FF. Interstitial cystitis/bladder pain syndrome. Semin Reprod Med. 2018;36(2):123-35. PMID: 30566978.
5. Cornelison BR, Erstad BL, Edwards C. Accuracy of a chatbot in answering questions that patients should ask before taking a new medication. J Am Pharm Assoc (2003). 2024;64(4):102110. PMID: 38670493.
6. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;9(10):1437-40. PMID: 37615960; PMCID: PMC10450581.
7. Kattih M, Bressler M, Smith LR, Schinelli A, Mhaskar R, Hanna K. Artificial intelligence-prompted explanations of common primary care diagnoses. PRiMER. 2024;8:51. PMID: 39569087; PMCID: PMC11578395.
8. Erkan A, Koc A, Barali D, Satir A, Zengin S, Kilic M, et al. Can patients with urogenital cancer rely on artificial intelligence chatbots for treatment decisions? Clin Genitourin Cancer. 2024;22(6):102206. PMID: 39236508.
9. Musheyev D, Pan A, Loeb S, Kabarriti AE. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? Eur Urol. 2024;85(1):13-6. PMID: 37567827.
10. Khawaja Z, Adhoni MZU, Byrnes KG. Generative artificial intelligence powered chatbots in urology. Curr Opin Urol. 2025;35(3):243-9. PMID: 40104869.
11. Talyshinskii A, Naik N, Hameed BMZ, Juliebø-Jones P, Somani BK. Potential of AI-driven chatbots in urology: revolutionizing patient care through artificial intelligence. Curr Urol Rep. 2024;25(1):9-18. PMID: 37723300; PMCID: PMC10787686.
12. Gajjar AA, Kumar RP, Paliwoda ED, Kuo CC, Adida S, Legarreta AD, Deng et al. Usefulness and accuracy of artificial intelligence chatbot responses to patient questions for neurosurgical procedures. neurosurgery. 2024. PMID: 38353558.