The New Journal of Urology
eISSN 3023-6940
doi: https://doi.org/10.33719/nju1929548

RESEARCH ARTICLE

Evaluation of Patient Information Provided by Artificial Intelligence Chatbots on Intravesical BCG Therapy for Bladder Cancer: A Cross-Sectional Methodological Study

Salim Zengin [1 *], Caglar Boyaci [2], Ahmet Olgun [1], Onur Kutluay [3], Abdullah Gul [4]

[1] Department of Urology, Yenikent State Hospital, Sakarya, Türkiye
[2] Department of Urology, Darende Hulusi Efendi State Hospital, Malatya, Türkiye
[3] Department of Urology, Cekirge State Hospital, Bursa, Türkiye
[4] Department of Urology, University of Health Sciences, Bursa Yuksek Ihtisas Training and Research Hospital, Bursa, Türkiye

Submitted: 2026-04-13
Accepted: 2026-09-01

Corresponding Author;
Salim Zengin, MD
Address: Department of Urology, Yenikent State Hospital, Sakarya, Türkiye
E-mail: salimmzengin@gmail.com

ORCID
S.Z. 0000-0002-7837-6767
C.B. 0000-0002-3158-8834
A.O. 0000-0003-0917-750X
O.K. 0009-0008-2877-3095
A.G. 0000-0003-4002-4659

Abstract

Objective: Artificial intelligence (AI) chatbots are increasingly used as accessible sources of medical information. However, concerns remain regarding the quality, reliability, readability, and clinical applicability of AI-generated health content. Intravesical Bacillus Calmette-Guérin (BCG) therapy is a standard adjuvant treatment for intermediate- and high-risk non-muscle invasive bladder cancer (NMIBC) and requires adequate patient understanding of treatment procedures and potential adverse effects. This study aimed to evaluate the quality, reliability, readability, and guideline adherence of information provided by commonly used AI chatbots regarding intravesical BCG therapy.

Methods: This cross-sectional methodological study evaluated responses generated by four AI chatbots (ChatGPT, Copilot, Gemini, and Perplexity) to a standardized question regarding intravesical BCG therapy. Responses were independently assessed by five urologists using validated tools including DISCERN, PEMAT-P, WRR, and the Coleman-Liau Index (CLI). Compliance with European Association of Urology (EAU) guidelines was evaluated using a 5-point Likert scale.

Results: Responses ranged from 454 to 550 words and required a university-level reading ability (CLI range: 16.02–19.35). DISCERN scores indicated moderate information quality across all chatbots, with Gemini demonstrating the numerically highest median score (51 [47–59]). PEMAT-P actionability scores were generally low (median 40%). Guideline adherence was moderate, with all chatbots receiving a median Likert score of 3.

Conclusion: AI chatbots provide moderate-quality information regarding intravesical BCG therapy but require high reading-grade levels and offer limited actionable guidance. AI-generated information may support patient education but should not replace physician counseling.

Keywords: artificial intelligence, bladder cancer, chatbots, intravesical BCG, patient education, readability

Cite; Zengin S, Boyaci C, Olgun A, Kutluay O, Gul A. Evaluation of Patient Information Provided by Artificial Intelligence Chatbots on Intravesical BCG Therapy for Bladder Cancer: A Cross-Sectional Methodological Study. New J Urol. 2026;21(3):131–137. **doi:** https://doi.org/10.33719/nju1929548

This work is licensed under a Creative Commons Attribution 4.0 (CC-BY) International License. © *Copyright remains with the authors.*

INTRODUCTION

With the increasing use of the internet and digital technologies, patients increasingly seek health-related information through online sources. In recent years, advances in artificial intelligence (AI) have led to the development of chatbot systems based on large language models, allowing users to obtain rapid and easily accessible medical information without direct interaction with healthcare professionals. Although these systems provide fast and convenient access to medical knowledge, concerns remain regarding the accuracy, reliability, readability, and clinical applicability of the information they generate. Previous studies evaluating AI chatbots in healthcare have demonstrated that the quality of AI-generated medical information is generally moderate, and responses may contain incomplete or non-referenced information, potentially leading to misunderstandings among patients (1,2). Furthermore, readability analyses have shown that many AI-generated texts require a high level of education, which may limit their usefulness for patient education (3).

Bladder cancer is one of the most common malignancies of the urinary tract, accounting for a significant proportion of cancer-related morbidity worldwide. Approximately 70–75% of newly diagnosed bladder cancer cases are classified as non–muscle invasive bladder cancer (NMIBC), which is characterized by a high risk of recurrence and progression despite local treatment (4). The standard initial management of NMIBC consists of transurethral resection of bladder tumor (TURBT) followed by risk-adapted intravesical therapy aimed at reducing recurrence and disease progression (5). Among available intravesical treatment options, Bacillus Calmette–Guérin (BCG) immunotherapy remains the gold standard adjuvant therapy for patients with intermediate- and high-risk disease (6).

Intravesical BCG therapy involves the administration of a live attenuated strain of *Mycobacterium bovis* into the bladder, stimulating a local immune response that reduces tumor recurrence and progression rates compared with TURBT alone (6,7). Current clinical guidelines recommend induction BCG therapy followed by maintenance treatment in selected patients in order to achieve optimal oncological outcomes (5). However, intravesical BCG therapy may cause local and systemic side effects, including dysuria, urinary frequency, hematuria, fever, and rarely, systemic BCG infection (7). In addition, the prolonged treatment schedule and potential adverse effects may lead to anxiety and uncertainty among patients undergoing therapy.

Recent studies evaluating AI chatbot performance in urological diseases have demonstrated that chatbot-generated responses generally provide moderate-quality information but often exhibit high reading-grade levels and limited actionable recommendations (8–10). Moreover, variability between chatbot responses and incomplete adherence to clinical guidelines have been reported as important limitations (9,10). These findings suggest that AI chatbots may serve as supportive tools for patient education; however, their reliability in providing accurate and guideline-consistent information remains uncertain.

Intravesical BCG therapy is a complex treatment requiring adequate patient understanding of treatment schedule, expected benefits, and possible side effects. Inadequate or misleading patient information may negatively affect treatment adherence and increase treatment-related anxiety. Therefore, evaluation of the quality and reliability of AI-generated patient information regarding intravesical BCG therapy is of particular importance.

The aim of this study was to evaluate the quality, reliability, readability, and guideline adherence of information provided by commonly used AI chatbots regarding intravesical BCG therapy for bladder cancer using validated assessment tools, including DISCERN, the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P), the Web Resource Rating (WRR), and the Coleman–Liau Index (CLI), together with an assessment of compliance with European Association of Urology (EAU) guidelines. Responses were independently evaluated by five urologists in order to determine the potential role and limitations of AI chatbots in patient education.

MATERIALS AND METHODS

This study was designed as a cross-sectional methodological investigation to evaluate the quality of patient education texts generated by artificial intelligence (AI) chatbots regarding intravesical Bacillus Calmette–Guérin (BCG) therapy for bladder cancer. As the study did not involve human

participants, animal subjects, or identifiable patient data, ethics committee approval was not required. The research was conducted in accordance with the ethical principles of the Declaration of Helsinki.

Four commonly used AI chatbots were included in the study: ChatGPT (GPT-5.3 Instant; OpenAI, USA), Copilot (GPT-5.3 Instant, Quick response; Microsoft, USA), Gemini (Gemini 3 Flash; Google DeepMind, USA/UK), and Perplexity (standard/default response setting; Perplexity AI Inc., USA). All chatbots were asked the same standardized question:

"What is intravesical BCG therapy for bladder cancer, why is it used, how is it administered, what are the possible side effects, and is it safe?"

For each chatbot, a new chat session was initiated in order to prevent previous interactions from influencing the responses. All chatbot responses were obtained on April 4, 2026, and recorded for analysis. The generated texts were included in the evaluation process without any editing or modification.

The chatbot responses were independently evaluated by five urologists. During the assessment process, widely used and validated evaluation tools for patient education materials were employed.

The quality and reliability of the information were assessed using the DISCERN instrument (11). DISCERN consists of 15 scored quality criteria and a separate overall quality rating (Question 16). Questions 1–8 evaluate the reliability of the information, whereas questions 9–15 assess the quality of treatment-related information. Each of the first 15 questions is scored from 1 to 5, yielding a cumulative score ranging from 15 to 75, with higher scores indicating better quality. Question 16 provides a separate overall quality rating based on the preceding 15 items and was not included in the cumulative DISCERN score.

The understandability and actionability of the texts were evaluated using the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) (12). PEMAT-P consists of two subdomains, understandability and actionability, and the results are reported as percentages.

The reliability and transparency of the information were evaluated using the Web Resource Rating (WRR) scale (13). WRR is a 13-item tool developed to assess the evidence base, transparency, and usability of online health information, and scores range from 0 to 100.

The readability level of the chatbot responses was assessed using the Coleman–Liau Index (CLI) (14). This index estimates the educational level required to understand a text. A CLI score above 11 indicates that a university-level reading ability is required for comprehension.

Compliance of the chatbot responses with the European Association of Urology (EAU) guidelines was evaluated using a 5-point Likert scale. In this assessment, a score of 1 indicated serious deficiencies, whereas a score of 5 indicated minimal deficiencies. Guideline compliance was assessed with reference to the main principles of intravesical BCG therapy, including indications, method of administration, expected therapeutic role, adverse effects, safety considerations, and general treatment course.

Scoring systems and guideline compatibility were reported as median (minimum-maximum) values, whereas word count and CLI scores were reported as mean values. Given the descriptive nature of the study and the evaluation of a single response from each chatbot, no inferential statistical comparisons were performed between chatbot platforms. Therefore, all findings were summarized using descriptive statistics. Statistical analyses were conducted using IBM SPSS Statistics (Version 25, IBM Corp., Armonk, NY, USA).

RESULTS

The standardized question regarding intravesical BCG therapy for bladder cancer was directed to four different artificial intelligence chatbots, and the generated responses were evaluated by five independent urologists.

The length of chatbot responses ranged from 454 to 550 words. The longest response was generated by Gemini (550 words), whereas the shortest response was provided by Copilot (454 words). Readability analysis using the Coleman–Liau Index (CLI) demonstrated that all chatbot responses required a university-level reading ability (CLI range: 16.02–19.35).

Evaluation of information quality using the DISCERN instrument showed variability among chatbot responses. Gemini demonstrated the numerically highest median DISCERN score of 51 (47–59), indicating moderate-to-good quality information. ChatGPT and Copilot demonstrated similar performance, with DISCERN scores of 42 (29–52) and 41 (32–44), respectively. Perplexity had the lowest DISCERN score, with a median value of 31 (23–41). Overall, these findings indicate that AI chatbots provide generally moderate-quality information regarding intravesical BCG therapy.

Assessment of understandability using PEMAT-P demonstrated that Gemini demonstrated the numerically highest understandability score at 75% (71–86.6), followed by ChatGPT at 71% (69–75). Copilot demonstrated moderate understandability at 64% (62–66.6), whereas Perplexity showed the lowest score at 60% (58–66.6).

Regarding actionability, PEMAT-P scores were generally low across all chatbots. Gemini demonstrated the numerically highest actionability score at 60% (0–100), while ChatGPT, Copilot, and Perplexity each demonstrated a median actionability score of 40%, indicating limited guidance for patients regarding practical decision-making or behavioral steps related to treatment.

The Web Resource Rating (WRR) scores indicated variability in reliability and transparency of information across chatbots. Gemini demonstrated the numerically highest WRR score (44.2), followed by ChatGPT (38.2) and Copilot (28.9). Perplexity demonstrated the numerically lowest WRR score (14.3), suggesting relatively lower transparency and evidence-based support compared to other chatbot responses.

Evaluation of guideline adherence using a 5-point Likert scale demonstrated moderate compatibility with European Association of Urology (EAU) guidelines across all chatbot responses. All chatbots received a median Likert score of 3 (range: 1–3), indicating moderate consistency with guideline recommendations but also the presence of clinically relevant deficiencies.

Overall, chatbot responses demonstrated moderate-quality information, high reading-grade levels, and limited actionable recommendations. Variability between chatbot responses was observed across multiple evaluation domains.

DISCUSSION

In this study, we evaluated the quality, readability, reliability, and guideline adherence of patient information provided by commonly used artificial intelligence (AI) chatbots regarding intravesical Bacillus Calmette–Guérin (BCG) therapy for bladder cancer. The results suggested that AI chatbot responses generally provided moderate-quality information, with considerable variability among different platforms. However, the texts required high reading-grade levels, and actionable recommendations were limited. In addition, adherence to European Association of Urology (EAU) guidelines was moderate across all chatbots. These findings suggest that AI chatbots may serve as supportive tools for patient education, but their outputs should be interpreted cautiously and should not replace physician counseling.

Table 1. Comparison of the quality, readability, and guideline adherence of information provided by artificial intelligence chatbots regarding intravesical BCG therapy for bladder Cancer

	ChatGPT	Copilot	Gemini	Perplexity
Total DISCERN score, median (min–max)	42 (29–52)	41 (32–44)	51 (47–59)	31 (23–41)
PEMAT-P Understandability	71 (69–75)	64 (62–66.6)	75 (71–86.6)	60 (58–66.6)
PEMAT-P Actionability	40 (40–80)	40 (20–80)	60 (0–100)	40 (20–80)
WRR	38.2	28.9	44.2	14.3
Word Count	476	454	550	496
Coleman Liau Index	16.65	16.02	19.35	17.74
Likert score (compatibility of the treatment options offered with the EAU guidelines 1–5)	3 (1–3)	3 (1–3)	3 (1–3)	3 (1–3)

The use of AI-based chatbots in healthcare has increased rapidly in recent years. Large language models are capable of generating fast and accessible medical information, which may facilitate patient education and improve access to healthcare knowledge (1,2). Studies evaluating AI chatbot performance in medical examinations and clinical knowledge have demonstrated that these systems can produce generally accurate information; however, concerns remain regarding incomplete responses, lack of referencing, and variability in the quality of medical content (2,15). The reliability of AI-generated medical information has therefore become an important topic of investigation, particularly in specialties such as urology where treatment decisions may be complex.

Intravesical BCG immunotherapy remains the gold standard adjuvant treatment for intermediate- and high-risk non–muscle invasive bladder cancer (NMIBC) and plays a crucial role in reducing tumor recurrence and progression rates (5,16). The mechanism of BCG therapy involves stimulation of the local immune response through activation of both innate and adaptive immune pathways, resulting in anti-tumor activity within the bladder mucosa (17). Despite its established efficacy, intravesical BCG therapy is associated with local and systemic side effects such as dysuria, urinary frequency, hematuria, and fever, which may lead to treatment-related anxiety among patients (16,17). Adequate patient education is therefore essential in order to improve treatment adherence and reduce uncertainty related to potential complications.

Previous studies evaluating AI chatbot performance in urological conditions have demonstrated similar findings to those observed in the present study. Investigations assessing AI-generated information regarding urological cancers have shown that, although responses to general patient questions may be of relatively high quality, performance can be lower when addressing guideline-based clinical questions (10). Similarly, studies evaluating chatbot performance in benign urological conditions have demonstrated moderate DISCERN scores, high reading-grade levels, and low actionability scores (8,9). These findings are consistent with the results of the present study, in which DISCERN scores indicated moderate information quality, while PEMAT-P results demonstrated limited actionable guidance for patients.

One of the most important findings of the present study was the consistently high reading-grade level of chatbot-generated texts. Coleman–Liau Index scores indicated that all responses required a university-level education for comprehension. Previous studies have shown that patient education materials should ideally be written at a reading level corresponding to the 6th–8th grade in order to ensure accessibility for the general population (18). High reading-grade levels may reduce patient comprehension and negatively affect treatment adherence, particularly in complex therapies such as intravesical BCG immunotherapy.

Another important finding of this study was the limited actionability of chatbot responses. Although AI-generated texts generally provided descriptive information about BCG therapy, they often lacked clear practical guidance regarding treatment expectations, management of side effects, and indications for seeking medical attention. Previous studies have similarly reported that AI chatbot responses frequently provide general medical explanations but insufficient practical recommendations for patients (10,15). This limitation may reduce the clinical usefulness of AI-generated information in real-world patient decision-making.

Evaluation of guideline adherence demonstrated moderate compatibility with EAU recommendations across all chatbot responses. Although chatbot outputs generally included basic information regarding treatment purpose and administration, important clinical details such as patient selection criteria, maintenance therapy protocols, and contraindications were not consistently addressed. Previous studies have similarly demonstrated variability in chatbot performance and reduced quality when addressing guideline-based clinical questions (9,10). This finding highlights the importance of physician supervision when patients use AI-generated medical information.

Intravesical BCG therapy represents a relatively complex treatment requiring patient understanding of treatment schedule, expected benefits, and possible adverse effects. Insufficient or inaccurate patient information may lead to increased anxiety and reduced treatment adherence. AI chatbots may provide easily accessible preliminary information for patients prior to clinical consultation; however, variability in response quality suggests that these

systems should be considered complementary rather than primary sources of medical information.

This study has several limitations. First, AI chatbot responses may change over time as models are continuously updated. Second, the evaluation was performed using a single standardized prompt, and different question formulations may produce different responses. Third, although responses were evaluated by five independent urologists, subjective scoring systems may introduce interobserver variability. Furthermore, although understandability and actionability were assessed using validated instruments, actual patient comprehension and real-world usability of the chatbot-generated information were not directly evaluated. In addition, the study was designed as a descriptive methodological evaluation and was not intended to determine statistically significant differences between chatbot platforms. Therefore, the findings should be interpreted as descriptive observations rather than evidence of superiority of any specific chatbot. Finally, AI chatbot responses may vary depending on the language of the prompt and the specific version of the chatbot used.

CONCLUSION

To conclude, AI chatbots provide moderate-quality and generally guideline-consistent information regarding intravesical BCG therapy; however, high reading-grade levels and limited actionable guidance remain important limitations. AI-generated information may serve as a supportive educational tool for patients but should not be considered a substitute for professional medical consultation. Future studies should focus on improving readability, enhancing guideline adherence, and increasing the clinical applicability of AI-generated patient education materials.

Conflict of Interest: The authors declare that they have no competing financial or non-financial interests that could have influenced the preparation or publication of this manuscript.

Informed Consent: As this study did not include human participants, patient data, or any identifiable personal information, the requirement for informed consent was waived.

Funding: The authors confirm that no financial support or external funding was obtained for this study.

Ethical Approval: This study did not involve human participants, patient data, or animal subjects. The research was conducted using responses generated by artificial intelligence chatbots and publicly available information. Therefore, ethics committee approval was not required in accordance with institutional and international research guidelines.

Author Contributions:
Conception and design: SZ, CB, AO, OK, AG
Data acquisition: SZ, CB, AO, OK, AG
Data analysis and interpretation: SZ, AG
Drafting the manuscript: SZ, AO,OK
Critical revision of the manuscript for the content: CB, AO
Statistical analysis: SZ, AG
Supervision: CB, AO, OK

REFERENCES

1. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6):887–907. https://doi.org/10.3390/healthcare11060887
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
3. McInnes N, Haglund BJ. Readability of online health information: implications for health literacy. Inform Health Soc Care. 2011;36(4):173-189. https://doi.org/10.3109/17538157.2010.542529
4. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-249. https://doi.org/10.3322/caac.21660
5. Gontero P, Birtle A, Capoun O, et al. European Association of Urology Guidelines on Non-muscle-

invasive Bladder Cancer (TaT1 and Carcinoma In Situ)-A Summary of the 2024 Guidelines Update. Eur Urol. 2024;86(6):531–549. https://doi.org/10.1016/j.eururo.2024.07.027

6. Liatsos GD, Mariolis I, Hadziyannis E, Bamias A, Vassilopoulos DJCMR. Review of BCG immunotherapy for bladder cancer. Clin Microbiol Rev. 2025;38(1):e00194-23. https://doi.org/10.1128/cmr.00194-23

7. Ehdaie B, Sylvester R, Herr HWJEu. Maintenance bacillus Calmette-Guérin treatment of non–muscle-invasive bladder cancer: A critical evaluation of the evidence. Eur Urol. 2013;64(4):579–585. https://doi.org/10.1016/j.eururo.2013.05.027

8. Chen A, Jacob J, Hwang K, Kobashi K, Gonzalez RR. AUA Guideline Committee Members Determine Quality of Artificial Intelligence-Generated Responses for Female Stress Urinary Incontinence. Urol Pract. 2024;11(4):693–698. https://doi.org/10.1097/UPJ.0000000000000577

9. Santucci J, Stapleton P, Ibrahim J, Johns-Putra L, Elmer S, Sathianathen N. Quality of patient information on interstitial cystitis from artificial intelligence chatbots. BJU Int. 2026;138(Suppl 2):S57–S63. https://doi.org/10.1111/bju.70035

10. 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(2):454–457.e4. https://doi.org/10.1016/j.clgc.2023.12.017

11. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105–111. https://doi.org/10.1136/jech.53.2.105

12. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns. 2014;96(3):395–403. https://doi.org/10.1016/j.pec.2014.05.027

13. Dobbins M, Watson S, Read K, A Tool That Assesses the Evidence, Transparency, and Usability of Online Health Information: Development and Reliability Assessment. JMIR Aging. 2018;1(1):e3. https://doi.org/10.2196/aging.9216

14. Coleman M, Liau TLJJoAP. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2):283–284. https://doi.org/10.1037/h0076540

15. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023;15(2):e35179. https://doi.org/10.7759/cureus.35179

16. Redelman-Sidi G, Glickman MS, Bochner BH. The mechanism of action of BCG therapy for bladder cancer--a current perspective. Nat Rev Urol. 2014;11(3):153–162. https://doi.org/10.1038/nrurol.2014.15

17. Jiang S, Redelman-Sidi G. BCG in Bladder Cancer Immunotherapy. Cancers (Basel). 2022;14(13):3073. https://doi.org/10.3390/cancers14133073

18. Eltorai AE, Ghanian S, Adams CA, Jr., Born CT, Daniels AH. Readability of patient education materials on the american association for surgery of trauma website. Arch Trauma Res. 2014;3(2):e18161. https://doi.org/10.5812/atr.18161