Original Article

DOI: 10.4274/haseki.galenos.2026.93585
Med Bull Haseki 2026;64(5):442-448

Guideline Concordance, Information Quality, and Readability of Large Language Model Responses to Female Urethral Stricture Questions: A Comparative Cross-sectional Study

Cagri Sevik, Metin Savun, Suleyman Sahin, Yigit Can Filtekin, Halil Lutfi Canat

University of Health Sciences Türkiye, Basaksehir Cam and Sakura City Hospital, Clinic of Urology, Istanbul, Türkiye

Abstract

Aim: Large language models (LLMs) are increasingly employed to acquire medical information. However, their reliability in addressing rare and technically intricate conditions, such as female urethral stricture disease (FUSD), remains uncertain. This study evaluated the guideline concordance, information quality, reliability, and readability of responses generated by three widely accessible LLMs in response to clinical questions related to FUSD.

Methods: This comparative cross-sectional study was conducted in January 2026. Thirty standardized questions across five clinical domains were developed based on the 2025 European Association of Urology Guidelines for Urethral Stricture Disease and refined through expert consensus. These questions were submitted to ChatGPT 5.0, Gemini 3.0, and Microsoft Copilot (Basic) under identical conditions. Responses were independently evaluated by three blinded board-certified urologists using a 5-point expert-rated accuracy scale. Information reliability and quality were assessed using modified DISCERN and ensuring quality information for patients (EQIP) independently by the blinded reviewers, while readability was evaluated using flesch reading ease score (FRES) and flesch-kincaid grade level (FKGL). Between-model comparisons were performed using the Friedman test, followed by Wilcoxon signed-rank tests with Bonferroni correction.

Results: Gemini achieved the highest expert-rated accuracy (4.80±0.41) and modified DISCERN score (19.0±1.7) among the three models. The Copilot achieved the highest EQIP (89.60±4.01) and FRES (42.23±3.89) scores and the lowest FKGL (12.67±1.04). Significant differences between the models were identified across all evaluated metrics. Gemini had significantly higher modified DISCERN scores than ChatGPT and Copilot (both $p<0.001$), whereas ChatGPT and Copilot did not differ significantly.

Conclusion: The evaluated LLMs demonstrated distinct performance profiles rather than uniform superiority. Gemini showed the highest expert-rated accuracy and information reliability, whereas Copilot provided the most favorable patient-oriented information quality and readability. Large language model-generated information on FUSD should be verified against current clinical guidance before clinical or patient education.

Keywords: Artificial intelligence, health literacy, large language models, patient education as topic, urethral stricture

Introduction

Female urethral stricture disease (FUSD) is an uncommon yet clinically significant condition that contributes to lower urinary tract symptoms in women. Its clinical presentation is often non-specific, including reduced urinary flow, incomplete bladder emptying, frequency, urgency, recurrent urinary tract infections, and urinary incontinence, which may contribute to delayed diagnosis (1). The diagnostic evaluation of FUSD requires careful clinical assessment and may include uroflowmetry, post-void residual measurement, cystourethroscopy, and radiographic or urodynamic evaluation. Treatment options range from urethral dilation and intermittent

Corresponding Author: Cagri Sevik, MD, University of Health Sciences Türkiye, Basaksehir Cam and Sakura City Hospital, Clinic of Urology, Istanbul, Türkiye
E-mail: cagri.sevik@hotmail.com **ORCID:** orcid.org/0000-0003-0288-704X
Received: 16.07.2026 **Accepted:** 30.10.2026 **Publication Date:** 17.11.2026

Cite this article as: Sevik C, Savun M, Sahin S, Filtekin YC, Canat HL. Guideline concordance, information quality, and readability of large language model responses to female urethral stricture questions: a comparative cross-sectional study. Med Bull Haseki. 2026;64(5):442-448

©Copyright 2026 The Author(s). Published by Galenos Publishing House on behalf of Istanbul Haseki Training and Research Hospital.
This is an open access article under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

self-dilation to reconstructive urethroplasty, with the choice of intervention depending on previous treatments, recurrence, stricture characteristics, and patient-related factors (1,2). Despite increasing clinical awareness, variability in definitions, diagnostic criteria, treatment strategies, and outcome reporting remains an important challenge in FUSD.

Large language models (LLMs) are increasingly used by both patients and healthcare professionals to obtain medical information and support patient education (3). Recent studies in urology have demonstrated that LLM-generated responses may provide generally accurate and informative content; however, important differences have been reported among models in terms of information quality, reliability, readability, and guideline concordance (4,5). In reconstructive urology, ChatGPT has shown potential for answering common patient questions, although limitations in readability and information quality remain (6,7). Similarly, recent comparative studies evaluating artificial intelligence (AI) chatbots in other functional urological conditions have demonstrated substantial model-dependent variation in both the quality and accessibility of generated health information (8,9). These findings suggest that the performance observed in common urological conditions cannot necessarily be extrapolated to less prevalent and technically complex disorders.

However, evidence regarding the performance of contemporary LLMs in FUSD remains limited. This is particularly relevant because FUSD combines diagnostic uncertainty, multiple treatment pathways, technically demanding reconstructive procedures, and the need for clear patient counseling, creating distinct challenges for AI-generated medical information. We hypothesized that the evaluated LLMs would differ significantly in terms of guideline concordance, information quality, and readability when responding to standardized questions related to FUSD. Therefore, the present study aimed to compare the responses generated by three widely available LLMs across the major clinical domains of FUSD, including diagnosis, treatment, surgical management, follow-up, and patient counseling, using expert assessments, validated information-quality instruments, and established readability metrics.

Materials and Methods

Study Design

This comparative cross-sectional study was conducted in January 2026 to evaluate the performance of the three LLMs in responding to standardized clinical questions regarding FUSD. The evaluated models were ChatGPT 5.0 (OpenAI), Gemini 3.0 (Google), and Copilot Basic (Microsoft). This study exclusively evaluated responses generated by publicly accessible AI platforms and did not involve human participants, animals, biological materials, patient records, or personally identifiable data. Therefore, ethics committee approval and informed consent were not required for this study.

Questionnaire Development

A set of 30 standardized clinical questions was used to assess the performance of the LLMs across major aspects of FUSD management. The questions were developed based on the 2025 European Association of Urology Guidelines for Urethral Stricture Disease (1). They were categorized into five predefined clinical domains: diagnostic evaluation (eight questions), treatment and initial management (eight questions), surgical considerations (six questions), follow-up and recurrence (four questions), and patient counseling (four questions). For transparency and reproducibility, the complete set of questions and their corresponding guidelines or reference sources are provided in Supplementary Table 1.

Large Language Model Testing Conditions

Freely accessible versions of all three LLMs were used through their official web interfaces. An identical, standardized prompt structure and the same clinical questions were used for all models. Each question was submitted as a separate conversation. The chat history and memory functions were disabled during testing to minimize contextual carryover between queries. No follow-up prompts were used, and the responses were left unchanged. Therefore, only the first response generated for each question was included in the analyses. The generated responses were subsequently collected for a blinded expert evaluation.

Expert Evaluation

Three board-certified urologists independently evaluated the responses. Before the evaluation, all model identifiers were removed, and the reviewers were blinded to the identity of the LLM that generated the responses. Each reviewer independently assessed the responses without discussion or consensus scoring during the initial evaluation.

Outcome Measures

Expert-rated accuracy and guideline concordance were assessed using a 5-point Likert scale. Responses were evaluated against the guidelines or reference standards corresponding to each clinical question, as detailed in Supplementary Table 1. The scores were interpreted as follows: 1, completely incorrect or guideline-discordant; 2, mostly incorrect with major errors or clinically relevant omissions; 3, partially correct with relevant omissions; 4, mostly correct and guideline-concordant with only

minor omissions; and 5, fully correct, complete, and guideline-concordant. For each response, the final expert-rated accuracy score used in the statistical analysis was calculated as the mean of the scores assigned by all three reviewers.

The reliability and quality of the generated health information were additionally assessed using the modified DISCERN instrument (comprising 5 core items scored on a 1-to-5 scale, yielding a total score ranging from 5 to 25) and the ensuring quality information for patients (EQIP) tool (consisting of 36 items evaluated to calculate a standardized percentage score). These tools were independently applied and scored by the blinded expert reviewers to ensure objective and standardized evaluation. Higher scores indicated higher reliability and higher information quality, according to the respective instrument. The scoring procedure and score range for each instrument were applied consistently across all model responses. The readability of each generated response was assessed using the flesch reading ease score (FRES) and flesch-kincaid grade level (FKGL). Higher FRES values indicated easier readability, whereas higher FKGL values indicated a higher educational grade level required to understand the text and therefore greater textual complexity.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics, version 27 (IBM Corp., Armonk, NY, USA). Continuous variables are presented as mean ± standard deviation. For each response, the scores assigned by the three reviewers were averaged before the between-model comparisons were performed. Inter-rater agreement among the three reviewers was assessed using the intraclass correlation coefficient (ICC) based on a two-way random-effects model with absolute agreement.

Because the same set of 30 questions was evaluated across all three LLMs, the observations constituted paired repeated measurements. Overall differences among ChatGPT, Gemini, and Copilot were therefore assessed using the Friedman test. When a significant overall difference was identified, pairwise comparisons were performed using the Wilcoxon signed-rank test. To account for the three pairwise comparisons, statistical significance for post-hoc analyses was evaluated using a Bonferroni-adjusted threshold of $p<0.0167$. Kendall's W was calculated as an effect-size measure for the Friedman test. A two-sided p-value of <0.05 was considered statistically significant for the overall analyses.

ChatGPT 5.0, Gemini 3.0, and Microsoft Copilot served as subjects in this study. Generative AI was additionally used during the manuscript revision to assist with the language, clarity, and structural editing. All AI-assisted content was critically reviewed and approved by the authors, who retained full responsibility for the scientific content and conclusions of the manuscript.

Results

A total of 90 responses generated for 30 standardized clinical questions across the three LLMs were included in the analysis. Inter-rater agreement among the three reviewers for expert-rated accuracy was excellent (ICC=0.90); similarly high consensus was observed across all evaluated quality and reliability instruments (modified DISCERN and EQIP). Descriptive statistics for expert-rated accuracy, modified DISCERN, EQIP, FRES, and FKGL scores are presented in Table 1. Gemini had the highest mean expert-rated accuracy and modified DISCERN scores, whereas Copilot had the highest EQIP and FRES scores and the lowest FKGL score.

Expert-rated accuracy was also examined across the five predefined clinical domains. Gemini showed the highest or joint-highest mean accuracy scores across most domains. The domain-specific accuracy scores for diagnostic evaluation, treatment and initial management, surgical considerations, follow-up and recurrence, and patient counseling are presented in Figure 1.

Overall comparisons using the Friedman test demonstrated statistically significant differences among the three models for all the evaluated metrics (Table 2).

Table 1. Descriptive statistics of model performance across evaluation metrics

Metric	ChatGPT	Gemini	Copilot
Expert-rated accuracy*	4.50±0.57	4.80±0.41	4.43±0.50
Modified DISCERN score*	17.0±1.0	19.0±1.7	17.0±1.5
EQIP score*	85.23±4.43	78.20±4.57	89.60±4.01
FRES*	40.29±3.91	36.20±4.07	42.23±3.89
FKGL*	13.16±1.18	13.43±1.21	12.67±1.04

*: Mean ± standard deviation
EQIP: Ensuring quality information for patients, FRES: Flesch reading ease score, FKGL: Flesch-kincaid grade level

Table 2. Comparison of model performance using the Friedman test

Metric	χ^2 (Friedman)	p-value	Effect size (Kendall W)
Expert-rated accuracy	9.36	0.009	0.16
Modified DISCERN	27.58	<0.001	0.46
EQIP	53.97	<0.001	0.90
FRES	16.62	0.00025	0.29
FKGL	12.0	0.0025	0.23

EQIP: Ensuring quality information for patients, FRES: Flesch reading ease score, FKGL: Flesch-kincaid grade level

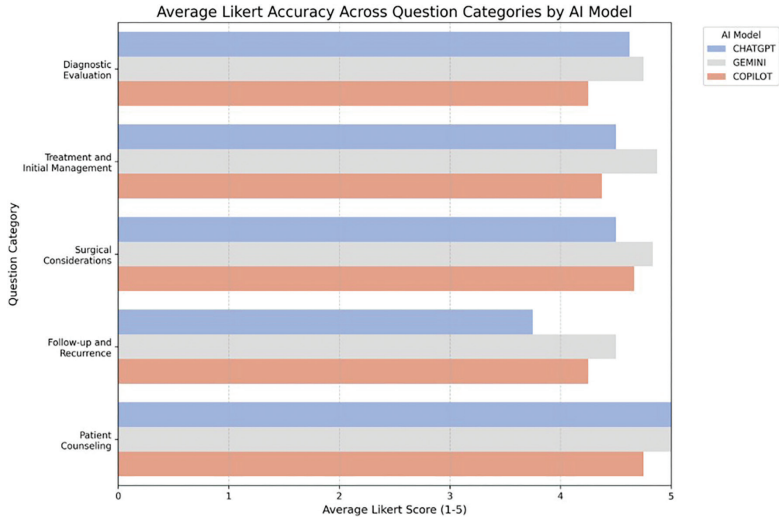


Figure 1. Mean expert-rated accuracy scores of ChatGPT 5.0, Gemini 3.0, and Microsoft Copilot across the five clinical domains evaluated

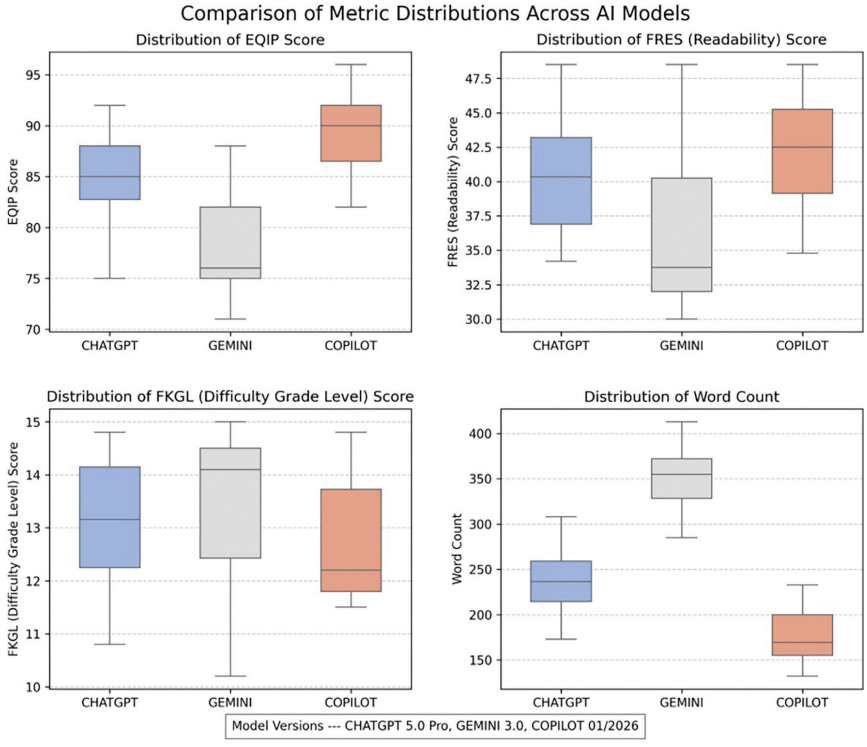


Figure 2. Distribution of modified Word Count, EQIP, FRES, and FKGL scores among ChatGPT 5.0, Gemini 3.0, and Microsoft Copilot
EQIP: Ensuring quality information for patients, FRES: Flesch reading ease score, FKGL: Flesch-kincaid grade level, AI: Artificial intelligence

For the modified DISCERN, a significant between-model difference was observed (χ^2=27.58, p<0.001), with a Kendall's W of 0.46. Significant overall differences were also observed in expert-rated accuracy and the EQIP, FRES, and FKGL scores.

The post-hoc pairwise comparisons are presented in Table 3. Gemini had significantly higher expert-rated accuracy scores than ChatGPT (p=0.012) and Copilot (p=0.0076), whereas no significant difference was observed between ChatGPT and Copilot (p=0.593). The modified DISCERN scores were significantly higher for Gemini than for both ChatGPT and Copilot (both p<0.001), whereas ChatGPT and Copilot did not differ significantly (p=0.924).

Table 3. Pairwise comparisons between models using the Wilcoxon signed-rank test

Metric	ChatGPT vs. Gemini	ChatGPT vs. Copilot	Gemini vs. Copilot
Expert-rated accuracy	0.012	0.593	0.0076
Modified DISCERN	<0.001	0.924	<0.001
EQIP score	<0.001	<0.001	<0.001
FRES	0.00077	0.0041	<0.001
FKGL	0.0095	0.059	0.00096

EQIP: Ensuring quality information for patients, FRES: Flesch reading ease score, FKGL: Flesch-kincaid grade level

For patient-oriented information quality, Copilot had the highest EQIP scores, with significant differences observed among all three model pairs (all p<0.001). Readability also differed among the models. Copilot had the highest FRES and the lowest FKGL values, whereas Gemini had the lowest FRES and the highest FKGL values. Pairwise FRES comparisons were significant for all three model pairs. For FKGL, significant differences were observed between ChatGPT and Gemini and between Gemini and Copilot, whereas the difference between ChatGPT and Copilot was not statistically significant after Bonferroni correction (p=0.059). The distributions of modified DISCERN, EQIP, FRES, and FKGL scores across the three models are shown in Figure 2.

Discussion

The principal finding of the present study was that the three evaluated LLMs demonstrated distinct rather than uniformly superior performance profiles in FUSD-related questions. Gemini achieved the most favorable results for expert-rated accuracy and information reliability, whereas Copilot performed better in patient-oriented information quality and readability. This divergence is clinically relevant because factual or guideline-concordant content does not necessarily translate into information that is easily understood by patients. Accordingly, the evaluation of medical LLMs should extend beyond a single measure of accuracy and incorporate complementary dimensions such as reliability, information quality, and linguistic accessibility.

Our findings are broadly consistent with recent studies that have demonstrated substantial model-dependent variability in AI-generated urological information. Alshak et al. (7) evaluated ChatGPT responses to 125 frequently asked questions across several reconstructive urology domains and reported generally useful information while identifying limitations in information quality and readability. Similarly, a recent review of AI-based patient education in urology concluded that chatbot responses frequently achieve acceptable levels of accuracy but often remain above recommended reading levels (3). More recently, Deger et al. (8) compared multiple AI chatbots for overactive bladder-related patient questions and identified significant differences among models in information quality and readability. These observations support our finding that the relative performance of an LLM depends on the outcome being assessed rather than reflecting a single hierarchy of overall model quality.

A recent guideline-based assessment of contemporary LLMs in preoperative counseling for robot-assisted radical prostatectomy provides an additional comparison. In that study, first-generation responses were independently evaluated by blinded urologists using several complementary measures, including factual accuracy, DISCERN, EQIP, and readability indices. Significant differences were observed across models, and no single quality dimension adequately characterized the overall performance (10). Although the disease context and evaluated models differed from those in the present study, the multidimensional pattern is comparable to our findings and reinforces the importance of evaluating accuracy, information quality, and readability separately when assessing LLM-generated medical information.

This issue may be particularly important in FUSD. Female urethral stricture is an uncommon and heterogeneous condition characterized by non-specific symptoms, variable diagnostic criteria, and several competing therapeutic approaches. Recent reviews have emphasized the continued lack of standardization in its definition, diagnostic pathways, selection of reconstructive techniques, and outcome reporting (2,11). Such heterogeneity creates a demanding context for LLMs because an apparently plausible response may still omit clinically important qualifications regarding diagnostic confirmation, previous interventions, stricture characteristics, or reconstructive options. The higher expert-rated accuracy and modified DISCERN scores observed with Gemini therefore suggest stronger concordance with the reference framework used in this study; however, these findings should not be interpreted as evidence that the model can independently provide individualized clinical decision-making.

Readability represented a separate limitation across the evaluated models. Although Copilot showed the most favorable readability profile, the mean FKGL values for all three models remained approximately at the 13th-grade level, indicating that the generated responses would require relatively advanced reading ability. This observation is consistent with previous urological AI literature, which shows that medically accurate chatbot responses may still be too linguistically complex for routine patient education (3,8). Therefore, optimization of LLM-based patient information should not focus exclusively on factual correctness. Explicit prompting for plain language, shorter sentences, explanation of technical terminology,

and adaptation to the patient's health literacy level may be necessary before such content is used for counseling.

The differences observed between expert-rated accuracy, modified DISCERN, EQIP, and readability also have implications for how these tools might be used in clinical practice. A model that performs well as a professional information resource may not necessarily provide the most suitable patient-facing explanation, and conversely, highly readable responses should not be assumed to have greater guideline concordance. Rather than identifying a universally "best" model, our findings support a task-specific approach to LLM use in which generated information is selected according to its intended purpose and verified by a clinician when applied to a rare and technically nuanced condition, such as FUSD. Given the rapid evolution of LLMs, these results should also be regarded as a comparison of the specific model versions evaluated at the time of the study rather than as a permanent ranking of the platforms.

Study Limitations

This study had several limitations. First, only three LLMs were evaluated at a single time point using versions available in January 2026. Given the rapid evolution of generative AI systems, model updates may affect their performance. Therefore, the present findings should be interpreted as a temporal snapshot rather than a permanent ranking of the platforms. Second, all responses were generated in English using standardized prompts, which may limit the generalizability of the findings to other languages, alternative prompt formulations, and different user-interaction styles. Third, although the responses were independently evaluated by three blinded board-certified urologists and the inter-rater agreement was high, expert-based scoring retained an inherent degree of subjectivity. Fourth, only the first response generated for each question was evaluated, and repeated generations were not performed for the same question. Because LLM outputs may vary across repeated queries, this study did not assess within-model response variability or reproducibility. Fifth, this study evaluated the informational characteristics of AI-generated responses rather than their effects on real-world patient comprehension, clinical decision-making, treatment selection, or patient outcomes. Finally, the analysis focused exclusively on FUSD, and the findings may not be directly generalizable to other areas of reconstructive and functional urology.

Despite these limitations, this study has several strengths. All three LLMs were evaluated using an identical set of standardized questions covering the major clinical domains of FUSD, which allowed direct paired comparisons between models. Responses were independently assessed by blinded board-certified urologists, and inter-rater agreement was formally evaluated. In addition, model performance was examined using complementary measures of expert-rated accuracy, information reliability, patient-oriented information quality, and readability rather than relying on a single outcome measure. Finally, inclusion of the complete question set, standardized prompts, and guideline/reference mapping as supplementary material enhances the transparency and reproducibility of the study.

Conclusion

In this comparative cross-sectional study, the evaluated LLMs demonstrated distinct performance profiles in responding to standardized questions on FUSD. Gemini achieved the highest expert-rated accuracy and modified DISCERN scores, whereas Microsoft Copilot demonstrated the most favorable patient-oriented information quality and readability. These findings indicate that no single model was uniformly superior across all evaluated domains. Therefore, LLM-generated information on FUSD should be interpreted according to its intended use and verified against current clinical guidance before being incorporated into patient education or professional information support.

Ethics

Ethics Committee Approval: This study did not involve human participants, animals, patient data, or biological materials. The study exclusively evaluated responses generated by publicly available artificial intelligence platforms. Therefore, according to institutional and international ethical standards, ethics committee approval was not required.

Informed Consent: Not applicable. The manuscript does not contain any individual person's data, images, or videos; hence, consent for publication was not required.

Authorship Contributions

Surgical and Medical Practices: M.S., S.S., H.L.C., Concept: C.S., Y.C.F., H.L.C., Design: C.S., S.S., Data Collection or Processing: Y.C.F., Analysis or Interpretation: C.S., M.S., Y.C.F., Literature Search: C.S., M.S., H.L.C., Writing: C.S., M.S., S.S.

Conflict of Interest: No conflicts of interest were declared by the authors.

Financial Disclosure: This study received no financial support.

Declaration on the Use of Artificial Intelligence (AI): No artificial intelligence tools were used in the preparation of this manuscript.

Supplementary Table 1 Link: https://d2v96fxpocvxx.cloudfront.net/87bf8cdb-e327-4a57-b2b5-56e4c99c141f/content-images/bea30df0-bdbc-4434-8ceb-508033591400.pdf

References

1. Lumen N, Campos-Juanatey F, Dimitropoulos K et al. European Association of Urology Guidelines on Urethral Stricture Disease. ISBN 978-94-92761-32-5. EAU Guidelines Office, Arnhem, the Netherlands, 2025.
2. Karapanos L, Beck M, Porres D; Trauma and Reconstructive Urology Working Party of the European Association of Urology Young Academic Urologists. Female urethral stricture disease: advances in diagnosis, surgical management, and standardized outcome reporting. Curr Opin Urol. 2026;36:354-62.
3. Paluszek O, Loeb S. Artificial intelligence and patient education. Curr Opin Urol. 2025;35:219-23.
4. Kim ES, Eun SJ, Youn S. The current state of artificial intelligence application in urology. Int Neurourol J. 2023;27:227-33.
5. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol. 2023;55:2717-32.
6. Ayrancı A, Çağlar U, Yıldız O, Meriç A, Sarılar Ö, Özgör F. Assessing the performance of ChatGPT in addressing questions regarding urethral strictures: a cross-sectional study. J Reconstr Urol. 2024;14:16-22.
7. Alshak MN, Cecelic J, Florissi I, Alam R, Cohen AJ. Assessing ChatGPT responses to frequently asked patient questions in reconstructive urology. Urol Pract. 2025;12:451-8.
8. Deger M, Ates T, Yilmaz IO, et al. Readability and quality assessment of AI-powered chatbot responses to overactive bladder patient questions: a comparative study. BMC Urol. 2026;26:31.
9. Youn S, Park BJ. Artificial intelligence for predicting treatment failure in neurourology: from automated urodynamics to precision management. Int Neurourol J. 2025;29 (Suppl 2):S55-S64.
10. Wang L, Liu H, Yang J, et al. Guideline-based evaluation of large language models for patient counselling before robot-assisted radical prostatectomy: accuracy, safety, and readability. J Robot Surg. 2026;20:888.
11. Turchi B, Lumen N, Verla W, Waterloos M. Female urethral stricture disease: a narrative review on diagnosis, surgical techniques and outcomes. Int J Impot Res. 2026;38:286-95.