SIUJ Société Internationale d'Urologie Journal

Article

External Validation of the EAU Guidelines Bot for Urethral Stricture: Accuracy, Completeness, and Clarity Analysis

Pietro Spatafora [1,‡], Riccardo Lombardo [2,*,‡], Manfredi Bruno Sequi [2], Marta Santioni [2], Eleonora Rosato [3], Matteo Romagnoli [2], Sabrina De Cillis [4], Enrico Checcucci [4], Daniele Amparore [4], Mauro Ragonese [5], Nazario Foschi [5], Valerio Santarelli [2], Giorgia Tema [2], Antonio Franco [2], Antonio Luigi Pastore [2], Bernardo Rocco [5], Mauro Gacci [5], Sergio Serni [5], Giacomo Gallo [6], Vincenzo Pagliarulo [6], Cristian Fiori [4], Enrico Finazzi Agrò [3], Francesco del Giudice [2], Alessandro Sciarra [2], Andrea Tubaro [2] and Cosimo De Nunzio [2] on behalf of Next-Gen Research Group

1 Department of Urology, Careggi Hospital, Univesity of Florence, 50134 Florence, Italy; pietro.spatafora@hotmail.it
2 Department of Urology, Sapienza University of Rome, 1035/1039, 00189 Rome, Italy; manfredibruno.sequi@uniroma1.it (M.B.S.); marta.santioni@uniroma1.it (M.S.); matteo.romagnoli@uniroma1.it (M.R.); va.santarelli@uniroma1.it (V.S.); giorgiat88@hotmail.it (G.T.); antonio.franco@uniroma1.it (A.F.); antonioluigi.pastore@uniroma1.it (A.L.P.); francesco.delgiudice@uniroma1.it (F.d.G.); alessandro.sciarra@uniroma1.it (A.S.); andrea.tubaro@uniroma1.it (A.T.); cosimo.denunzio@uniroma1.it (C.D.N.)
3 Department of Urology, University of Tor Vergata, 00133 Rome, Italy; eleonoraros92@gmail.com (E.R.); finazzi.agro@med.uniroma2.it (E.F.A.)
4 Division of Urology, Department of Oncology, San Luigi Gonzaga Hospital, University of Turin, 10043 Turin, Italy; sabrinatitti.decillis@unito.it (S.D.C.); checcu.e@hotmail.it (E.C.); daniele.amparore@unito.it (D.A.)
5 Department of Urology, Gemelli University, 00136 Rome, Italy; mauro.ragonese@policlinicogemelli.it (M.R.); nazario.foschi@policlinicogemelli.it (N.F.); bernardo.rocco@policlinicogemelli.it (B.R.); mauro.gacci@unifi.it (M.G.); sergio.serni@unifi.it (S.S.)
6 Department of Urology, Vito Fazzi Hospital, 73100 Lecce, Italy; giacomo.gallo@uniroma1.it (G.G.); vincenzo.pagliarulo@unisalento.it (V.P.)
* Correspondence: rlombardo@me.com
‡ These authors contributed equally to this work.

Abstract

Background/Objectives: Recently the European Association of Urology (EAU) guidelines presented the EAU Guidelines bot to assist urologists in the reading of the guidelines; however, there is a lack of up-to-date external validation. The aim of our study is to assess the accuracy, completeness, and clarity of the guidelines bot in urethral strictures. **Methods:** A total of 117 questions based on the EAU urethral strictures guidelines recommendations were developed. Each question was input to the EAU guidelines bot and the response was assessed by two expert urologists to assess the accuracy, completeness, and clarity. Moreover, 10 simple clinical cases were input. A 5-point Likert scale was used as a score and, in case of discrepancies, a third urologist was queried. Accuracy, completeness and clarity were assessed per chapter and per grade of recommendation. All questions and answers were recorded in an Excel file. **Results:** Overall 117 questions were developed. In terms of accuracy, 111/117 (95%) were defined as accurate (scores 4–5), 4/117 (3%) presented a fair accuracy (score 3), and 2/117 (2%) were deemed not accurate. In terms of completeness, 93/117 (80%) were defined as complete (scores 4–5), 22/117 (19%) presented a fair completeness (score 3), and 2/117 (2%) were deemed not complete. Finally, in terms of clarity, 104/117 (89%) were defined as clear (scores 4–5), 13/117 (11%) presented a fair clarity (score 3), and 0/109 (0%) were deemed not clear. When comparing strong and weak recommendations, no differences were recorded. Overall the answers to simple clinical cases were in line with the guidelines with good accuracy, completeness and clarity scores.

Check for updates

Academic Editor: Lysanne Campeau

Received: 18 February 2026
Revised: 4 April 2026
Accepted: 14 April 2026
Published: 21 April 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the Société Internationale d'Urologie. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Conclusions: The EAU guidelines bot represents an accurate tool for urethral stenosis guidelines. Some fine-tuning is needed to improve readability and clarity.

Keywords: urethral strictures; guidelines; chatbot

1. Introduction

The application of artificial intelligence (AI) in medicine has expanded rapidly across multiple domains. Among these, large language model (LLM)-based chatbots have emerged as a distinct class of AI systems, demonstrating promising performance in areas such as medical education, patient counseling, and clinical decision-making [1]. More broadly, AI applications in healthcare now extend from diagnostic evaluation and treatment planning to surgical training and intraoperative support. The development of well-trained and validated LLMs has shown potential to enhance clinical practice [2]. Although final decisions regarding diagnosis and patient management remain clinician-driven, chatbots facilitate more efficient access to medical knowledge and evidence-based recommendations. Nevertheless, their integration into routine clinical practice continues to raise important concerns regarding reliability, accuracy, and overall applicability [3].

Despite the high reliability of the European Association of Urology (EAU) guidelines, their substantial length and frequent updates may limit their practical use in daily clinical settings, particularly given the time constraints faced by healthcare professionals [4,5]. To address these challenges, the EAU Guidelines Office has introduced an official AI-powered chatbot, accessible via the guidelines website. This tool may enable rapid consultation of guideline recommendations during clinical encounters, with the potential to improve adherence to evidence-based practices [6].

Urethral stricture disease represents a challenging field, characterized by limited high-quality evidence due to the scarcity of randomized controlled trials. Although several recommendations are available, many clinical decision-making pathways still rely on evidence of varying strength and require subjective interpretation [2]. The use of the guidelines bot may avoid the wrong management of patients with urethral strictures, improving the management of such patients, which is often challenging.

In this context, the aim of our study was to assess the accuracy, completeness, and clarity of the EAU guidelines bot in the management of urethral stricture disease.

2. Materials and Methods

2.1. The Chatbot and the Queries

This cross-sectional study evaluated the responses provided by the EAU Guidelines chatbot regarding urethral strictures across all guideline topics. Each query corresponded to a specific guideline recommendation, and the strength of each recommendation was recorded.

The chatbot was introduced alongside the 2025 version of the EAU Guidelines and was developed using the OpenAssistantGPT (version 5.3) platform, a no-code, open-source framework that enables the creation of customized AI chatbots powered by OpenAI's Application Programming Interface (API, 2024 version). The system operates through web crawling and retrieval-augmented generation (RAG) technology. However, the specific underlying language model has not been disclosed. The chatbot is configured to generate responses exclusively based on official guideline content.

Queries were formulated by expert urologists and covered all topics within the EAU guidelines on urethral strictures, with one question per recommendation. To ensure repro-

ducibility, questions were designed to closely reflect the wording of the original recommendations (Supplementary Materials). All queries were submitted via the chatbot interface available on the EAU guidelines website, without additional prompts, clarifications, or modifications. For each query, a single AI-generated response was recorded and used for analysis.

Subsequently, ten real-life clinical scenarios were also submitted to the chatbot (Supplementary Materials). Informed consent was obtained from all patients. The study was approved by the local ethics committee of Sant'Andrea Hospital (approval number CE 6376_2021; date: 14 July 2021).

2.2. Response Evaluation and Statistical Analysis

Each response was assessed across three domains: accuracy, completeness, and clarity. Two board-certified urologists independently evaluated all answers using a 5-point Likert scale (1 = very poor; 2 = poor; 3 = fair; 4 = good; 5 = excellent). In cases of disagreement, discrepancies were resolved through discussion with a third senior urologist.

Accuracy was defined as the degree of alignment between the chatbot's response and established guideline recommendations. Completeness referred to the extent to which all relevant aspects of the query were addressed. Clarity assessed the comprehensibility and articulation of the response, considering the intended audience.

Overall performance for each domain was determined by analyzing the distribution of scores across all questions. Scores of 4 or 5 were considered indicative of high performance. For each domain, the number and percentage of responses rated as high (scores 4–5), fair (score 3), or poor (scores 1–2) were calculated.

Additionally, responses were stratified according to the strength of the corresponding guideline recommendations (strong vs. weak) to evaluate whether recommendation strength influenced chatbot performance. Separate analyses were conducted for each group and compared accordingly. The proportion of high-performance scores across domains was compared between strong and weak recommendations.

Categorical variables were reported as frequencies and percentages. Associations between variables were assessed using the Chi-square test, with statistical significance set at $p < 0.05$. Inter-rater agreement was evaluated using Cohen's kappa coefficient.

All statistical analyses were performed using SPSS version 27.0 (IBM Corp., Armonk, NY, USA).

3. Results

Overall, 117 guideline-based answers and 10 clinical cases were evaluated.

With regard to accuracy, 111 responses (95%) were rated as accurate (Likert score 4–5). Four responses (3%) demonstrated fair accuracy (score 3), while two responses (2%) were deemed not accurate (score 1–2).

In terms of completeness, 93 responses (80%) were rated as complete (score 4–5). Twenty-two responses (19%) were judged as fair completeness (score 3), typically due to omission of secondary but clinically relevant details. Only two responses (2%) were considered not complete.

Regarding clarity, 104 responses (89%) were rated as clear and well-structured (score 4–5). Thirteen responses (11%) were rated as fair clarity (score 3). Notably, no responses were considered unclear or misleading enough to receive a score of 1–2. Inter-reader agreement for accuracy, completeness and clarity was respectively 85%, 84% and 79.6%.

When stratified by strength of recommendation, no relevant differences in accuracy, completeness, or clarity were observed ($p > 0.05$).

Overall, 10 different clinical cases were submitted to the guidelines bot. The recommendations of the guidelines were evaluated accurately in all cases. In terms of completeness, 80% reached a score of 4–5. As well, clarity scores reached 4–5 in 90% of the cases.

4. Discussion

LLM-based chatbots represent a distinct category of artificial intelligence composed of two core elements: a general-purpose AI system and a conversational interface [7–11]. The LLM acts as the underlying technological engine, whereas the chatbot is the software application that receives user inputs and generates responses through human-like dialogue [1].

The present study provides an important external validation, demonstrating that the guidelines-based chatbot is accurate, complete, and clear in most cases when addressing recommendations related to urethral strictures. A major strength of this work lies in the use of a rigorous methodology, which evaluated each individual guideline recommendation as well as real-world clinical scenarios. Overall, our findings are consistent with the current literature, indicating that adequately trained AI systems achieve high accuracy and should be preferred over general-purpose models for clinical applications [6,12].

Several scientific studies have evaluated the performance of general-purpose chatbots in medical contexts (e.g., ChatGPT…) [13–16]. However, they are not designed for a specific assigned clinical task and have been reported to produce clinically relevant errors or to fail in recognizing urgent clinical scenarios, potentially compromising patient safety [4]. An additional and well-recognized risk is the phenomenon of "hallucination", where LLMs generate information that appears credible but lacks factual accuracy or evidence-based support [5].

Herein, the EAU developed the EAU guidelines bot in June 2025, introduced as a tool directly integrated into the guidelines, with the specific aim of providing rapid responses based on EAU-certified scientific content.

In this scenario, our study represents the first attempt at external validation of the EAU guidelines bot in the specific setting of urethral stricture management [2]. Our study provides several findings.

First, accuracy was particularly high, with 95% of responses rated as concordant with the EAU guidelines recommendations. This result is clinically relevant, as accuracy represents the most critical requirement for any guideline-support tool. Concerns regarding incorrect or misleading information remain one of the main limitations of AI-based chatbots in medicine. The low proportion of inaccurate responses observed in our analysis suggests that a chatbot specifically designed to retrieve and summarize guideline recommendations may mitigate some of the risks associated with general-purpose LLMs [4].

Second, completeness showed slightly lower performance compared with accuracy, with approximately 20% of responses rated as fair. In most cases, these responses correctly addressed the core recommendation but omitted additional contextual details or secondary considerations reported in the guideline text. While this limitation may be acceptable for rapid consultation, it reinforces the concept that the EAU guidelines bot should be considered a complementary tool rather than a substitute for direct guideline consultation, particularly in complex or borderline clinical scenarios.

Third, clarity was rated as satisfactory in the majority of responses, highlighting the chatbot's ability to present guideline-based information in a structured and readable format. This aspect is crucial in daily clinical practice, where time constraints often limit in-depth guideline review. Nevertheless, a subset of responses was judged as moderately clear, indicating that further refinement of language structure and explanatory depth may improve usability.

Finally, no relevant differences were observed when responses were stratified according to the strength of recommendations. The chatbot demonstrated comparable performance for both strong and weak recommendations, suggesting that it does not oversimplify or distort recommendations supported by lower levels of evidence. This finding is particularly relevant in urethral stricture disease, where many clinical decisions rely on evidence of varying quality and require individualized interpretation.

Our study included selected real-life clinical scenarios to preliminarily assess the chatbot's ability to address clinical questions. The chatbot demonstrated good performance in managing simple cases; however, the primary aim of the present study was not to evaluate its performance in complex clinical decision-making, but rather to validate its ability to accurately answer questions strictly related to its guideline-based content.

A prospective study is currently ongoing to assess the implementation of the chatbot in clinical practice and to evaluate, within a multi-center cohort, the impact of its routine use. Importantly, implementation in real clinical settings prior to the present validation was not feasible for ethical reasons.

Our study is not devoid of limitations. Firstly, despite the very low rate of inaccurate and incomplete responses, the presence of suboptimal answers derived from minor guideline deviations highlights the need for human supervision. Second, although responses were independently assessed by experienced urologists with a predefined consensus process, the evaluation of accuracy, completeness, and clarity remains partially subjective. Another possible limitation is that the bot was not tested on a general urologist; therefore, our results apply only to specialized urologists. However, a study on general urologists and general practitioners is ongoing, and the results will be available soon. Finally, the study focused on descriptive comparisons between strong and weak recommendations and was not designed to assess statistical differences or clinical outcomes related to chatbot use.

Despite these limitations, the present study provides a first "external validation" of the EAU guidelines bot for urethral stricture management and may serve as a framework for future evaluations of guideline-based AI tools.

5. Conclusions

The EAU guidelines bot demonstrates high accuracy when applied to the EAU guidelines on urethral stricture management. While minor refinements in completeness and clarity are warranted, the tool appears to be a reliable adjunct for guideline consultation. Further external validation in real-world clinical settings is encouraged.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/siuj7020030/s1, Supplementary files: S1: questions and answers; S2: clinical cases and answers.

Author Contributions: Conceptualization, P.S., R.L., M.B.S., M.S., E.R., M.R. (Matteo Romagnoli), S.D.C., E.C., D.A., M.R. (Mauro Ragonese), N.F., V.S., G.T., A.F., A.L.P., B.R., M.G., S.S., G.G., V.P., C.F., E.F.A., F.d.G., A.S., A.T. and C.D.N.; methodology, M.R. (Mauro Ragonese); formal analysis, V.S., G.T., A.F., A.L.P., B.R., M.G., S.S., G.G., V.P., C.F., E.F.A., F.d.G., A.S., A.T. and C.D.N.; investigation, M.R. (Mauro Ragonese), V.S., G.T., A.F., A.L.P., B.R., M.G., S.S., G.G., V.P., C.F., E.F.A., F.d.G., A.S., A.T. and C.D.N.; resources, P.S., R.L., M.B.S., M.S., E.R., M.R. (Matteo Romagnoli) and S.D.C.; data curation, V.P., C.F., E.F.A., F.d.G., A.S., A.T. and C.D.N.; writing—original draft preparation, P.S., R.L., M.B.S., M.S., E.R., M.R. (Matteo Romagnoli), S.D.C., E.C., D.A., M.R. (Mauro Ragonese) and N.F.; writing—review and editing, V.S., G.T., A.F., A.L.P., B.R., M.G., S.S., G.G., V.P., C.F., E.F.A. and F.d.G.; supervision, V.S., G.T., A.F., A.L.P., B.R., M.G., S.S., G.G., V.P., C.F., E.F.A., F.d.G., A.S., A.T. and C.D.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Ospedale Sant'Andrea, Rome, Italy (CE 6376_2021; date 14 July 2021).

Informed Consent Statement: Written informed consent has been obtained from the patients to publish this paper.

Data Availability Statement: Data is available upon request (Riccardo Lombardo).

Acknowledgments: We acknowledge the members of the Next Gen Research Group.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rinderknecht, E.; Engelmann, S.U.; Saberi, V.; Kirschner, C.; Kravchuk, A.P.; Schmelzer, A.; Breyer, J.; Goßler, C.; Mayr, R.; Gilfrich, C.; et al. Using ChatGPT-4 for Lay Summarization in Prostate Cancer Research to Advance Patient-Centered Communication: Large-Scale Generative AI Performance Evaluation. *J. Med. Internet Res.* **2025**, *27*, e76598. [CrossRef] [PubMed]
2. Hershenhouse, J.S.; Mokhtar, D.; Eppler, M.B.; Rodler, S.; Ramacciotti, L.S.; Ganjavi, C.; Hom, B.; Davis, R.J.; Tran, J.; Russo, G.I.; et al. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer Prostatic Dis.* **2025**, *28*, 394–399. [CrossRef] [PubMed]
3. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef] [PubMed]
4. Bayne, D.B.; Gaither, T.W.; Awad, M.A.; Murphy, G.P.; Osterberg, E.C.; Breyer, B.N. Guidelines of guidelines: A review of urethral stricture evaluation, management, and follow-up. *Transl. Androl. Urol.* **2017**, *6*, 288–294. [CrossRef] [PubMed]
5. Riechardt, S.; Waterloos, M.; Lumen, N.; Campos-Juanatey, F.; Dimitropoulos, K.; Martins, F.E.; Osman, N.I.; Barratt, R.; Chan, G.; Esperto, F.; et al. European Association of Urology Guidelines on Urethral Stricture Disease Part 3: Management of Strictures in Females and Transgender Patients. *Eur. Urol. Focus* **2022**, *8*, 1469–1475. [CrossRef] [PubMed]
6. Gallo, G.; Lombardo, R.; Pagliarulo, V.; Sequi, M.B.; Coppola, L.M.; Rosato, E.; Romagnoli, M.; De Cillis, S.; Checcucci, E.; Amparore, D.; et al. Accuracy, readability, and understandability of EAU guidelines bot for testicular cancer. *Int. Urol. Nephrol.* 2026, in press. [CrossRef] [PubMed]
7. Nicoletti, R.; Nicoletti, G.; Giannini, V.; Teoh, J.Y.C. Developers-Doctor-patients: The artificial intelligence's trifecta. *Prostate Cancer Prostatic Dis.* **2024**, *27*, 3–4. [CrossRef] [PubMed]
8. Rodler, S.; Kopliku, R.; Ulrich, D.; Kaltenhauser, A.; Casuscelli, J.; Eismann, L.; Waidelich, R.; Buchner, A.; Butz, A.; Cacciamani, G.E.; et al. Patients' Trust in Artificial Intelligence–based Decision-making for Localized Prostate Cancer: Results from a Prospective Trial. *Eur. Urol. Focus* **2024**, *10*, 654–661. [CrossRef] [PubMed]
9. Baydoun, A.; Jia, A.Y.; Zaorsky, N.G.; Kashani, R.; Rao, S.; Shoag, J.E.; Vince, R.A., Jr.; Bittencourt, L.K.; Zuhour, R.; Price, A.T.; et al. Artificial intelligence applications in prostate cancer. *Prostate Cancer Prostatic Dis.* **2024**, *27*, 37–45. [CrossRef] [PubMed]
10. Guven, S. Artificial intelligence as the continuum of surgical evolution. *Minerva Urol. Nephrol.* **2025**, *77*, 905–906. [CrossRef] [PubMed]
11. Vedovo, F.; Capogrosso, P.; Maruccia, S.; Simonetta, F.; Dal Moro, F.; Liguori, G. Agreement between artificial intelligence, experts, and the European Association of Urology Guidelines: Insights from a study on the management of benign prostatic hyperplasia. *Minerva Urol. Nephrol.* **2025**, *77*, 740–741. [CrossRef] [PubMed]
12. Puerto Nino, A.K.; Garcia Perez, V.; Secco, S.; De Nunzio, C.; Lombardo, R.; Tikkinen, K.A.O.; Elterman, D.S. Can ChatGPT provide high-quality patient information on male lower urinary tract symptoms suggestive of benign prostate enlargement? *Prostate Cancer Prostatic Dis.* **2025**, *28*, 167–172. [CrossRef]
13. Bentellis, I.; Guérin, S.; Khene, Z.-E.; Khavari, R.; Peyronnet, B. Artificial intelligence in functional urology: How it may shape the future. *Curr. Opin. Urol.* **2021**, *31*, 385–390. [CrossRef] [PubMed]
14. Karako, K.; Song, P.; Chen, Y.; Tang, W. New possibilities for medical support systems utilizing artificial intelligence (AI) and data platforms. *Biosci. Trends* **2023**, *17*, 186–189. [CrossRef] [PubMed]

15. Checcucci, E.; Rosati, S.; De Cillis, S.; Vagni, M.; Giordano, N.; Piana, A.; Granato, S.; Amparore, D.; De Luca, S.; Fiori, C.; et al. Artificial intelligence for target prostate biopsy outcomes prediction the potential application of fuzzy logic. *Prostate Cancer Prostatic Dis.* **2022**, *25*, 359–362. [CrossRef] [PubMed]
16. Au, R.C.; Jenjitranant, P.; Cool, D.W.; Izawa, J.; Inman, B.; Ward, A.; Chin, J.L. Artificial intelligence may enhance the role of magnetic resonance imaging in prostate cancer focal therapy. *Prostate Cancer Prostatic Dis.* 2025; in press. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.