The Journal of Sexual Medicine, 2026, **23(4)**, qdag041
https://doi.org/10.1093/jsxmed/qdag041
Published: March 16, 2026
Original Research

Accuracy, readability, and understandability of European Association of Urology guidelines bot for Sexual and Reproductive Health Guidelines

Valerio Santarelli, MD[1], Riccardo Lombardo, MD, PhD[1,*], Matteo Romagnoli, MD[1], Manfredi Bruno Sequi, MD[1], Ludovica Maria Coppola, MD[1], Eleonora Rosato, MD[2], Sabrina De Cillis, MD[3], Enrico Checcucci, MD[3], Daniele Amparore, MD[3], Mauro Ragonese, MD[4], Nazzario Foschi, MD[4], Pietro Spatafora, MD[5], Giorgia Tema, MD[1], Antonio Nacchia, MD[1], Antonio Cicione, MD[1], Antonio Franco, MD[1], Antonio Luigi Pastore, MD[1], Yazan Al Salhi, MD[1], Giacomo Gallo, MD[6], Vincenzo Pagliarulo, MD[6], Bernardo Rocco, MD[4], Mauro Gacci, MD[5], Cristian Fiori, MD[3], Enrico Finazzi Agro, MD[2], Alessandro Sciarra, MD[1], Francesco Del Giudice, MD[1], Andrea Tubaro, MD[1], Cosimo De Nunzio, MD[1], Next-Gen Research Group

[1]Department of Urology, Sapienza University of Rome, Via di Grottarossa 1035/1039, Rome, 00189, Italy
[2]Department of Urology, University of Tor Vergata, Rome, 00133, Italy
[3]Division of Urology, Department of Oncology, San Luigi Gonzaga Hospital, University of Turin, Orbassano, Turin, 10043, Italy
[4]Department of Urology, Gemelli University, Rome, 00136, Italy
[5]Department of Urology, Careggi Hospital, Univesity of Florence, Florence, 50134, Italy
[6]Department of Urology, "Vito Fazzi" Hospital, Piazza Filippo Muratore 1, Lecce, 73100, Italy

*Corresponding author: Department of Urology, Ospedale Sant'Andrea, La Sapienza University, Rome, Italy. Email: rlombardo@me.com

Abstract

Background: Recently, the European Association of Urology (EAU) Guidelines presented an official Bot to assist urologists during Guidelines navigation. However, up to date no external validation is available.
Aim: To assess accuracy, completeness, and clarity of the Guidelines Bot for Sexual and Reproductive Health.
Methods: A total of 228 questions based on the EAU Sexual and Reproductive Health Guidelines recommendations were developed. Each question was inputted to the EAU Guidelines Bot and the response was reviewed by two expert uro-andrologists. Discrepancies were resolved by discussion with a third expert. Results were further stratified per grade of recommendation.
Outcomes: Evaluate the rate of accurate, complete, and clear answers to guidelines-related questions using a 5-point Likert scale and the impact of the grade of recommendation on the quality of the answer.
Results: Overall, 228 questions were developed. In terms of accuracy 224/228 (98.3%) were defined as accurate (score-4-5), 2/228 (0.9%) presented a fair accuracy (score = 3) while 2/228 (0.9%) were deemed not accurate (score 1-2). In terms of completeness, 223/228 (97.8%) were defined as complete (score-4-5), 2/228 (0.9%) presented a fair completeness (score 3), while 3/228 (1.3%) were deemed not complete. Finally in terms of clarity, 225/228 (98.7%) were defined as clear (score-4-5), 2/228 (0.9%) presented a fair clarity (score 3) and 0/228 were not clear. When comparing strong and weak recommendations, no differences were recorded.
Clinical Implications: The EAU Guidelines Bot may serve as a reliable clinical decision support tool for urologists seeking rapid, evidence-based guidance on sexual and reproductive health management.
Strengths & Limitations: This is the first external evaluation of the EAU Guidelines Bot. Our results suggest a significant improvement in terms of reliability when compared to general AI tools. However, our queries were straightforward and developed directly from guideline recommendations and results might not apply to complex real-world clinical scenarios.
Conclusions: EAU Guidelines Bot represents an accurate and reliable tool for Sexual and Reproductive Health Guidelines navigation, but further validation is required to evaluate its applicability in clinical practice.

Keywords: AI; chatbot; eau guidelines; sexual and reproductive health.

Introduction

Artificial Intelligence (AI) has permeated every aspect of human life, thanks to its sophisticated and reliable decision-making and problem-solving capabilities.[1] The range of AI applications in the medical field spans from diagnostic work-up and treatment planning to surgical training and intraoperative guidance.[1] The development of trained and validated Large Language Models (LLMs) has been demonstrated to assist clinical practice.[2] While the final decision making on diagnostic work-up and patient management remains human-driven, chatbots have allowed easier access to medical knowledge and evidence-based medicine. However, the integration of LLMs into clinical practice still raises critical questions about reliability, accuracy, and overall applicability. Despite the promising potential, the adoption of AI-powered

Received: December 21, 2025. **Revised:** January 22, 2026. **Accepted:** January 25, 2026

chatbots in specialized medical domains currently remains limited and unstructured. European Association of Urology (EAU) Guidelines are updated yearly in order to provide the highest level of recommendation for high-quality care.[3] Although the EAU Guidelines are highly reliable, their annual updates and considerable length can represent a practical limitation for their daily application, particularly in light of the limited time available to the medical staff. To overcome these obstacles, an official EAU Guidelines AI-powered Chatbot has been released from the EAU Guidelines office, and can be directly accessed through the Guideline website.[3] EAU Guidelines on Sexual and Reproductive Health offer a thorough summary of the medical aspects related to sexual and reproductive health in adult men.[3]

Andrological problems are often under recognized by general urologists.[4] Nevertheless, they account for approximately 13%-17% of all urology outpatient visits and encompass a wide spectrum of clinical presentations requiring complex management strategies.[4] Adherence to clinical guidelines represents a critical issue in both urology and andrology, as it ensures the selection of the most appropriate management options for patients. In real-world clinical practice, consulting guidelines can be time-consuming; consequently, decision-support tools such as guideline-based bots may provide rapid access to relevant information for urologists and andrologists. However, for such tools to be clinically reliable, it is essential that the information they provide strictly aligns with established guidelines. In the present study, we aimed to evaluate the reliability of the EAU Guidelines Bot on Sexual and Reproductive Health guidelines. Secondarily, we evaluated the association between the quality of the answer and the strength of the related recommendations (weak vs. strong).

Materials and methods

The Chatbot and the queries

This is a cross-sectional study evaluating the responses provided by the EAU Guidelines Bot on Sexual and Reproductive Health across all guidelines' topics. Each question was linked to a specific recommendation and its strength was noted. The Chatbot was introduced with the 2025 version of the EAU Guidelines, and developed adopting the OpenAssistant-GPT platform, a no-code, open-source platform that enables the creation of custom AI Chatbots powered by OpenAI's Application Programming Interface through web crawling and retrieval-augmented generation technology.[5] However, the specific language model powering the Chatbot remains undisclosed. The Chatbot is configured to produce responses derived solely from the official Guidelines material.

The queries were developed by expert uro-andrologists and spanned across all topics of the EAU Guidelines on Sexual and Reproductive Health, one for each recommendation. Questions were developed sticking to the recommendation maintain reproducibility. Questions were submitted through the Bot interface available directly on the EAU Guidelines' website without additional follow-up prompts or modifications. For each query, the single AI-generated answer was recorded for the analysis and no further information was inputted.

Response evaluation and statistical analysis

Each response was evaluated across three dimensions: accuracy, completeness, and clarity. Two board-certified uro-andrologists independently and meticulously scored each answer using a 5-point Likert Scale (1 = very poor; 2 = poor; 3 = fair; 4 = good; 5 = excellent). Accuracy was defined as the alignment of the content with established guidelines recommendations. Completeness examined whether all important aspects of the question were addressed by the response. Clarity evaluated the understandability and articulation of the answer, with regard to the intended audience. Disagreements between the two reviewers were resolved through discussion with a third senior uro-andrologist. The overall performance for each attribute was calculated by examining the score distribution for the attribute across all questions. A score of 4 or 5 per attribute defined high performance. For each attribute, the number and percentage of answers rated high, fair, or poor were calculated. Finally, questions were further stratified according to the guideline recommendation strength (weak vs. strong) to assess whether the strength of the recommendation could influence the Bot's performance. To do so, the rate of high performance scores for each attribute was compared between strong and weak recommendations. Categorical variables were reported as number and percentages and the association between variables was tested using a Chi-square test with statistical significance defined as $P < .05$. Inter-reader agreement was evaluated using kappa coefficient of agreement.

Statistical analysis was carried out using SPSS version 27.0 (IBM Corp: Armonk, NY, USA).

Results

Quality of the answers

Overall, a total of 228 questions were developed and answered by the Chatbot. The EAU Guidelines Bot provided mostly high-quality answers. Regarding accuracy, 224/228 (98.3%) of the answers were rated as accurate (scores 4-5), 2/228 (0.9%) as fairly accurate (score 3) and 2/228 (0.9%) were deemed not accurate (score 1-2). In terms of completeness, 223/228 (97.8%) received a score of 4-5 (complete), 2/228 (0.9%) a score of 3 (fairly complete) and 3/228 (1.3%) received scores of 1 or 2 (not complete). Finally, clarity was defined as clear (score 4-5) in 225/228 (98.7%) of answers and fairly clear (score 3) in 3/228 (1.3%). No answer was defined as not clear by the reviewers (0/228). Inter-reader agreement for accuracy, completeness, and clarity was 88%, 81%, and 77.6%, respectively.

An example of an accurate and inaccurate answer is reported in Table 1, together with the corresponding guideline recommendation. The complete questions' list with the relative assigned scores is reported in the Supplementary Material 1 and 2.

The influence of the strength of recommendation

The rate of accurate, complete, and clear responses is reported in Table 2 stratified by the strength of recommendation. Respectively, a total of 149/228 (65.3%) and 79/228 (34.7%) strong and weak recommendations were identified. When comparing the quality of the responses across strong and weak recommendations, no statistically significant differences were found. In particular, a high accuracy, completeness, and clarity was found respectively for 98.7% (147/149), 98.0 % (146/149), and 98.7% (147/149) of question based on strong-recommendations compared to 97.5% (77/79), 97.5%

Table 1 Representative examples of high- and low-quality responses generated by the EAU guidelines bot. Examples reflect one response rated as accurate, complete, and clear (bottom), and one response rated as inaccurate, incomplete, and clear (top), based on expert evaluation of alignment with the EAU Sexual and Reproductive Health Guidelines.

Recommendation	Strength rating	Question	Bot answer	Evaluation
Do not offer intralesional treatment with steroids to reduce penile curvature, plaque size or pain.	strong	Which intralesional treatments should be avoided in Peyronie's disease?	"Intralesional verapamil and interferon alpha-2b are not recommended for the treatment of Peyronie's disease due to insufficient evidence of efficacy." Therefore, intralesional verapamil and interferon alpha-2b should be avoided in Peyronie's disease treatment.	-Inaccurate (2) -Incomplete (2) -Clear (5)
Repeat total testosterone on at least two separate occasions when <12 nmol/L and before starting testosterone therapy.	strong	What should be done before starting testosterone therapy if total testosterone is below 12 nmol/L?	"Repeat total testosterone on at least two separate occasions when <12 nmol/L and before starting testosterone therapy." This means that if the total testosterone level is below 12 nmol/L, it is recommended to confirm this finding by measuring total testosterone again on at least two separate occasions before initiating testosterone therapy. This approach accounts for biological variability and ensures accurate diagnosis.	-Accurate (5) -Complete (5) -Clear (5)

Table 2 Overall and stratified performance of the EAU guidelines bot in answering Sexual and Reproductive Health queries. The top half reports the distribution of accuracy, completeness, and clarity ratings across all 228 Bot-generated answers. The bottom part compares the rate of high-quality responses (score 4-5) between questions based on strong versus weak recommendations, with corresponding *P*-values.

Overall quality of answers			
Attribute	**Excellent/Good (Score 4-5)**	**Fair (Score 3)**	**Poor (Score 1-2)**
Accuracy	224 (98.3%)	2 (0.9%)	2(0.9%)
Completeness	223 (97.8%)	2 (0.9%)	3 (1.3%)
Clarity	225 (98.7%)	3 (1.3%)	0
Performance by guideline recommendation strength			
Attribute	**Strong recommendations ($n = 149$)**	**Weak recommendations ($n = 79$)**	***P***
Accuracy (4-5)	147 (98.7%)	77 (97.5%)	0.83
Completeness (4-5)	146 (98.0%)	77 (97.5%)	0.78
Clarity (4-5)	147 (98.7%)	78 (98.8%)	0.99

(77/79), and 98.8% (78/79) of questions based on weak-recommendations ($P = .83$, $P = .78$, and $P = .99$).

Discussion

In recent years, many studies have evaluated the applications of LLMs in medical practice and particularly in Urology.[2,6] With the expanding adoption of these AI-driven tools in healthcare, their impact on the medical community and most importantly on patient care must be carefully addressed. While medical practitioners at different levels of seniority have admitted to adopting some form of LLMs for medical information, ethical concerns associated with the adoption of these tools remain, such as the need to verify the accuracy of the responses and careful supervision by qualified attendings.[6,7] This is particularly true when adopting general readily available LLMs. Recently, Asker et al.[8] found an accuracy of 80% for ChatGPT-40, DeepSeek-R1 and Grok-3 in answering multiple-choice questions derived from the European Board of Urology in-service examination booklets. However, a comparative study by Eckrich et al.[9] found that consultants produced higher-quality responses than those generated by LLMs. To summarize, LLMs cannot serve as reliable substitutes for peer-reviewed publications or clinical guidelines. Hypothetically, the EAU overcame this issue by providing a validated Chatbot generated directly and solely from published EAU Guidelines.[3,5]

In the present study, we evaluated the performance of the EAU Guidelines Chatbot on Sexual and Reproductive Health. After having developed 228 queries on different topics all relevant to the Guidelines of interest, two expert uro-andrologists reviewed the responses provided by the Chatbot in terms of accuracy, completeness, and clarity. 224 (98.3%), 223 (97.8%), and 225 (98.7%) of the answers were rated as accurate, complete, and clear. The study involved the evaluation of a large number of questions which could result in fatigue and satisfying bias effect of the reviewer. To mitigate such bias we decided to include two reviewers which evaluated 30 questions per day. The bot was very precise often which explains the observed near ceiling performance. These results suggest a significant improvement in terms of reliability when compared to general AI tools.[8-10] Particularly, when using 2023 EAU guidelines on prostate cancer (PCa) as reference standard, chatGPT responses to PCa-related queries were

rated as completely correct only in 26% of cases.[10] The EAU Guidelines Bot's stronger performance can likely be attributed to its design, which uses verified EAU Guidelines content as its primary knowledge source, thus minimizing factual errors. Accordingly, while general AI systems draw from vast and varied training data, this Bot bases its answers specifically on evidence-based, peer-reviewed recommendations.[3,8,10] Moreover, we aimed to evaluate the impact of the strength of recommendation on the quality of the answer. The EAU guidelines refer to strong recommendations as reflecting high-quality evidence and/or clear advantages of benefits over risk, while weak recommendations indicate limited evidence, unclear benefit/harm, or variability in patient preferences.[11] The Chatbot demonstrated consistent reliability across questions generated on both strong and weak recommendations. In particular, it showed to take into account the grade of recommendation and information available, suggesting caution or highlighting the presence of conflicting evidence. This is a likely advantage when compared to other general LLMs, which have been criticized to provide too assertive and overconfident responses not justified by the available evidence.[12-16] In the future, similar reliable chatbots, with verified medical sources and carefully constrained outputs could be developed in simplified versions specifically designed for patient use. Such chatbots could address non-medical users to a referral center for their specific condition and simplify healthcare access, finally optimizing waitlist management and resource allocation.

Our results suggest the reliability and accuracy of the EAU guidelines Chatbot on Sexual and Reproductive Health. However, several limitations must be addressed, 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. Secondly, our queries were based directly on Guidelines recommendations and clearly expressed. Our study was intentionally designed as a structured evaluation using standardized guideline-based questions. While real-world queries were not assessed, the chatbot is capable of scenario-based reasoning, which is currently being evaluated in an ongoing case-based study. Finally, despite the high experience level of the two reviewers, the quality assessment remained inherently subjective.

Conclusions

In the present study, the EAU Guidelines Bot for Sexual and Reproductive Health demonstrated high accuracy, reliability and consistency when replying to guidelines-based queries. While our results do not directly apply to real-world scenarios, they suggest the potential adoption of the Bot for a rapid access to updated recommendations. The present study opens the road to further studies evaluating real-life scenarios and complex clinical cases.

Aknowledgement

The project is under the initiative of Next-Gen research group.

Author contributions

V.S.: Writing—original draft-lead. R.L.: Methodology-equal, Project administration-equal. M.R.: Conceptualization-equal, Data curation-equal. M.B.S.: Investigation-equal, Methodology-equal. L.M.C.: Investigation-equal, Writing—original draft-equal. E.R.: Conceptualization-equal, Data curation-equal. S.D.C.: Data curation-equal, Formal analysis-equal. M.R.: Formal analysis-equal, Writing—original draft-equal. P.S.: Data curation-equal, Writing—original draft-equal. A.F.: Conceptualization-equal, Data curation-equal. A.L.P.: Project administration-equal, Supervision-equal. G.G.: Writing—original draft-equal. V.P.: Investigation-equal, Project administration-equal. M.G.: Validation-equal, Visualization-equal. C.F.: Methodology-equal, Supervision-equal, E.F.A.: Conceptualization-equal, Software-equal. A.S.: Validation-equal, Writing—review & editing-equal. F.D.G.: Conceptualization-equal, Writing—original draft-equal. A.T.: Supervision-equal, Writing—review & editing-equal. C.D.N.: Conceptualization-equal, Supervision-equal, Writing—review & editing-equal.

Supplementary material

Supplementary material is available at *The Journal of Sexual Medicine* online.

Funding

None declared.

Conflicts of interest

The authors have nothing to disclose.

References

1. Younis HA, Eisa TAE, Nasser M, *et al.* A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. *Diagnostics (Basel)*. 2024;**14**(1):109. https://doi.org/10.3390/diagnostics14010109
2. Barreda M, Cantarero-Prieto D, Coca D, *et al.* Transforming healthcare with chatbots: uses and applications-a scoping review. *Digit Health*. 2025;**11**:20552076251319174
3. Salonia A, Capogrosso P, Boeri L, *et al.* European Association of Urology guidelines on male sexual and reproductive health: 2025 update on male hypogonadism, erectile dysfunction, premature ejaculation, and Peyronie's disease. *Eur Urol*. 2025;**88**(1):76–102. https://uroweb.org/guidelines/sexual-and-reproductive-health. https://doi.org/10.1016/j.eururo.2025.04.010
4. Duran MB, Yildirim O, Kizilkan Y, *et al.* Variations in the number of patients presenting with Andrological problems during the coronavirus disease 2019 pandemic and the possible reasons for these variations: a Multicenter study. *Sex Med*. 2021;**9**:100292
5. *Build your Own AI Agent with no Code*: [cited 22 Nov 2025]. https://www.openassistantgpt.io/en.
6. Eppler M, Ganjavi C, Ramacciotti LS, *et al.* Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol*. 2024;**85**(2):146–153. https://doi.org/10.1016/j.eururo.2023.10.014
7. Su H, Sun Y, Li R, *et al.* Large language models in medical diagnostics: scoping review with bibliometric analysis. *J Med Internet Res*. 2025;**27**:e72062. https://doi.org/10.2196/72062
8. Asker OF, Recai MS, Genc YE, Dogan KA, Sener TE, Sahin B. Chatbots in urology: accuracy, calibration, and comprehensibility; is DeepSeek taking over the throne? *BJU Int*. 2025;**136**:937–945. https://doi.org/10.1111/bju.16873
9. Eckrich J, Ellinger J, Cox A, *et al.* Urology consultants versus large language models: potentials and hazards for medical advice in urology. *BJUI Compass*. 2024;**5**(5):438–444. https://doi.org/10.1002/bco2.359

10. Lombardo R, Gallo G, Stira J, *et al*. Quality of information and appropriateness of open AI outputs for prostate cancer. *Prostate Cancer Prostatic Dis*. 2025;**28**(1):229–231. https://doi.org/10.1038/s41391-024-00789-0
11. Guyatt GH, Oxman AD, Kunz R, *et al*. Going from evidence to recommendations. *BMJ*. 2008;**336**(7652):1049–1051. https://doi.org/10.1136/bmj.39493.646875.AE
12. Bentegeac R, Le Guellec B, Kuchcinski G, Amouyel P, Hamroun A. Token probabilities to mitigate large language models overconfidence in answering medical questions: quantitative study. *J Med Internet Res*. 2025;**27**:e64348. https://doi.org/10.2196/64348
13. Geretto P, Lombardo R, Albisinni S, *et al*. Quality of information and appropriateness of ChatGPT outputs for neuro-urology. *Minerva Urology and Nephrology*. 2024;**76**(2):138–140. https://doi.org/10.23736/S2724-6051.24.05807-5
14. Hershenhouse JS, Mokhtar D, Eppler MB, *et al*. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer Prostatic Dis*. 2024;**28**(2):394–399. https://doi.org/10.1038/s41391-024-00826-y
15. Puerto Nino AK, Garcia Perez V, Secco S, *et al*. Can ChatGPT provide high-quality patient information on male lower urinary tract symptoms suggestive of benign prostate enlargement? *Prostate Cancer Prostatic Dis*. 2024;**28**(1):167–172. https://doi.org/10.1038/s41391-024-00847-7
16. Collin H, Keogh K, Basto M, Loeb S, Roberts MJ. ChatGPT can help guide and empower patients after prostate cancer diagnosis. *Prostate Cancer Prostatic Dis*. 2024;**28**(2):513–515. https://doi.org/10.1038/s41391-024-00864-6

The Journal of Sexual Medicine, 2026, 23(4), qdag041
https://doi.org/10.1093/jsxmed/qdag041
Original Research