Foreword

For reprint orders, please contact: info@becaris.com



Journal of **Comparative Effectiveness Research**

Expert panel agreement with large language model responses to uro-oncology guideline-based questions: insights from seven Delphi consensus panels across major urological cancers

Ángel Borque-Fernando*[,1], Denis Navarro[2,3], Luis M Esteban[4], Daniel Pérez-Fentes[5], José Luis Álvarez-Ossorio[6], Carmen González Enguita[7] & Manuel Doblaré[3,8,9]

[1]Department of Urology, Hospital Universitario Miguel Servet, IIS-Aragón (GIIS-071), Zaragoza, Spain
[2]Department of Electronic Engineering & Communications, Universidad de Zaragoza, Zaragoza, Spain
[3]Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza, Zaragoza, Spain
[4]Institute for Biocomputation & Physics of Complex Systems, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, Zaragoza, Spain
[5]Department of Urology, Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela, Spain
[6]Department of Urology, Hospital Universitario Puerta del Mar, Asociación Española de Urología, Fundación para la Investigación en Urología, Cádiz, Spain
[7]Department of Urology, Hospital Universitario Fundación Jiménez Díaz (HUFJD), Hospitales Públicos Quirónsalud, Madrid, Spain
[8]Institute for Health Research Aragón (IIS Aragón), Zaragoza, Spain
[9]CIBER-BBN, ISCIII, Spain
*Author for correspondence: dr.borque@gmail.com

> "The frequent expert panel agreement observed supports further evaluation of guideline-conditioned LLMs."

First draft submitted: 14 April 2026; Accepted for publication: 30 September 2026; Published online: 4 November 2026

The integration of artificial intelligence (AI) into clinical practice represents a paradigm shift in medicine, offering novel tools that may facilitate access to and interpretation of evidence-based information [1,2]. Among these emergent technologies, large language models (LLMs) have demonstrated remarkable aptitude for processing and synthesizing complex textual information, such as clinical practice guidelines [3–5]. The field of uro-oncology, with its diverse tumor types, rapidly evolving standards, and nuanced therapeutic pathways, is particularly poised to benefit from such computational advances.

This supplement presents coordinated Delphi studies that rigorously assess how well three leading LLMs, Claude [6], Gemini [7] and ChatGPT [8], interpret the 2025 European Association of Urology (EAU) guidelines and respond to related clinical questions across seven major urological cancers [9–15]. The work is organized into four complementary articles, each focusing on a distinct group of cancers: penile and testicular cancers, renal cell carcinoma, prostate cancer and muscle-invasive and nonmuscle-invasive bladder cancer together with upper tract urothelial carcinoma. For each cancer type, the studies assessed expert panel agreement with LLM responses to 25 structured clinical questions using disease-specific multidisciplinary panels of 7 to 11 specialists.

This multidisciplinary initiative was developed under the coordination of the Spanish Association of Urology (AEU), with technical implementation carried out by the Aragón Institute of Health Research (IIS-A) and the Aragón Institute of Engineering Research (I3A). The EAU signed a collaboration agreement with the AEU, granting permission to use the 2025 EAU Guidelines within the project. This collaboration supports the evaluation and future development of guideline-conditioned AI tools.

The project's multidisciplinary nature is reflected in the ongoing involvement of key Spanish scientific societies across all seven tumor-specific panels. These include the AEU, Spanish Society of Radiation Oncology (SEOR), Genitourinary Alliance of Research and Development (GUARD), Spanish Society of Nuclear Medicine and Molecular Imaging (SEMNIM), Spanish Society of Anatomical Pathology (SEAP), Spanish Society of Hospital

Pharmacy (SEFH) and Spanish Society of Medical Radiology (SERAM). Their participation ensures comprehensive expert input and oversight throughout the project.

A foundational step for this project was a comprehensive comparative evaluation of six LLM configurations, using both simple and rephrased prompts of Claude, Gemini and ChatGPT, across seven major uro-oncological cancers [16]. For each tumor type, 25 carefully designed clinical questions based on the 2025 EAU guidelines were presented to the LLMs, which generated corresponding responses. These responses were independently assessed by 9 to 11 experienced specialists who were blinded to the model identities. Ratings were assigned using a 5-point Likert scale, where 1 indicated 'incorrect/unacceptable' and 5 indicated 'optimal.' Across a total of 10,500 expert evaluations, Claude's simple prompt configuration achieved the highest proportion of top scores, with 45.9% rated as 'optimal' (score 5) and 87% rated as either 'optimal' or 'acceptable' (scores 4 or 5). Gemini and ChatGPT demonstrated lower frequencies of optimal responses but maintained acceptable performance, with 68–70% of responses rated in the combined optimal or acceptable categories. Interestingly, rephrased prompts did not consistently enhance performance; in some cases, simple prompts yielded better results. This rigorous evaluation provided insight into both the strengths and limitations of these models, highlighting the critical importance of thorough expert validation and meticulous data preparation for the rigorous evaluation of guideline-conditioned LLM responses.

A crucial step in the project's development was the meticulous adaptation of complex, multimodal guideline content, including all figures, tables and flowcharts from the EAU guidelines on seven major uro-oncological cancers, into clear, structured, plain-language textual summaries. The guidelines were organized into thematic sections and the transformation into plain language was independently reviewed by members of the specialized tumor boards. Any discrepancies or uncertainties arising during this process were resolved by consensus among the guideline coordinators. This process was designed to reduce information loss and ensure that the transformed guideline content remained faithful to the original materials.

In the four studies presented in this supplement, three specific model versions were evaluated: Claude 3.5 Haiku (Anthropic; version claude-3-5-haiku-20241022, launched 22 October 2024) [6], Gemini 2.5 Flash (Google; version gemini-2.5-flash-preview-05-20, launched 20 May 2025) [7], and GPT-4.1-mini (OpenAI; version gpt-4.1-mini-2025-04-14, launched 14 April 2025) [8]. The original English-language 2025 EAU guideline content was provided to each model, with figures and tables converted into structured, plain-language textual summaries. The same guideline materials were supplied to all three models. For each of the seven clinical entities, 25 structured questions were submitted in Spanish without modification, independently and one at a time, and each model was evaluated separately rather than within a continuous conversation. One response was generated per question and model, resulting in 525 model responses. Disease-specific multidisciplinary panels of 7 to 11 specialists evaluated the responses using a 9-point scale and a two-round modified Delphi process. Panelists were blinded to model identity and could abstain from questions outside their area of expertise; only submitted ratings were included in the calculations. In the first round, all responses were rated. Responses not achieving consensus were reassessed in a second round after panelists reviewed anonymized, aggregated feedback summarizing the group ratings. The LLM response text remained unchanged between rounds, and only panelist ratings could be revised. Consensus was defined by three criteria: a median score within the agreement (7–9) or disagreement (1–3) range; fewer than a third of panelists rating outside these ranges and an interquartile range below 4.

Across the seven clinical entities, consensus in agreement ranged from 56% to 96% for Claude, 92% to 100% for Gemini and 80% to 96% for ChatGPT. These results describe expert panel agreement under the specific study conditions and do not establish factual accuracy, clinical safety or independent clinical decision-making ability. Comparisons apply only to the specific model versions tested. Because these versions were released at different times, the observed differences cannot be attributed solely to vendors or interpreted as stable rankings of model families.

Across the studies, lower panel agreement or the need for further discussion was more frequent in scenarios involving advanced staging, prognostic nuances, limited clinical evidence or complex clinical decisions. These findings reinforce the need for specialist review, particularly in complex or atypical scenarios. Similar caution has been recommended in other studies of LLMs in oncology and urology [4,5,17,18].

The frequent expert panel agreement observed supports further evaluation of guideline-conditioned LLMs. However, the studies used limited sets of predefined questions, variable panel participation, and one response per question and model. They did not assess factual accuracy against an independent reference standard, patient outcomes, workflow integration, clinical safety or performance in real clinical settings. Therefore, the findings do

not establish real-world utility or readiness for autonomous decision support. The structured adaptation of guideline materials may nevertheless provide a useful basis for developing more standardized approaches to evaluating LLM-generated responses.

The development of the urooncoguias.aeu.es chatbot represents a potential future application of this initiative. Its clinical use will require broader validation, repeated testing, assessment in real clinical settings and continued expert oversight. Together, the studies in this supplement provide a coordinated foundation for further research into the interpretation of uro-oncology guideline information by LLMs.

Author contributions

A Borque-Fernando led the project and contributed to the study's conceptualization, methodology, project administration and supervision. All authors contributed jointly to the investigation, formal analysis, data curation, validation, resources, visualization, writing of the original draft, and review and editing of the manuscript, and approved the final version for publication.

Acknowledgments

The authors thank Dr Pablo Rivas for his support with the medical writing on behalf of Content Ed Net. The authors also express their gratitude to Manuel Espárrago for his technical collaboration on the project's evaluation tools, and to Andrés Mena for his contribution to the design of the user interface and the app for accessing the AI tool.

Financial disclosure

This project was funded by the Spanish Association of Urology. The authors received support from Instituto de Investigación Sanitaria Aragón (Spain) for the project "Uro-Onco GuÍAs: Desarrollo de un Chatbot basado en IA Generativa para la Interpretación de Guías Clínicas en Uro-Oncología".

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Writing disclosure

Medical writing of the manuscript through Content Ed Net was funded by the Spanish Association of Urology.

Ethical conduct of research

This article is the introduction of a supplement that presents coordinated Delphi studies that involved only expert panelists who participated voluntarily in the Delphi process. No patient data were used. Participation implied informed consent, and formal ethical approval was not required.

Consent to participate

Participation by expert panelists was voluntary and implied informed consent.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/

References

1. Basubrin O. Current status and future of artificial intelligence in medicine. *Cureus* 17(1), e77561 (2025).
2. Mizna S, Arora S, Saluja P, Das G, Alanesi WA. An analytic research and review of the literature on practice of artificial intelligence in healthcare. *Eur. J. Med. Res.* 30(1), 382 (2025).
3. Hager P, Jungmann F, Holland R *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* 30(9), 2613–2622 (2024).
4. Longwell JB, Hirsch I, Binder F *et al.* Performance of large language models on medical oncology examination questions. *JAMA Netw. Open* 7(6), e2417641 (2024).
5. Carl N, Schramm F, Haggenmüller S *et al.* Large language model use in clinical oncology. *NPJ Precis. Oncol.* 8(1), 240 (2024).
6. Anthropic. Models overview. Claude Platform Docs. (Accessed 23 June 2025). https://platform.claude.com/docs/en/about-claude/models/overview

7. Google AI for Developers. Gemini 2.5 Flash. Gemini API. (Accessed: 23 June 2025). https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash
8. OpenAI. GPT-4.1 mini model. OpenAI API. (Accessed 23 June 2025). https://developers.openai.com/api/docs/models/gpt-4.1-mini
9. European Association of Urology. EAU guidelines on penile cancer. (Accessed 19 May 2025). https://uroweb.org/guidelines/penile-cancer
10. European Association of Urology. EAU guidelines on testicular cancer. (Accessed 19 May 2025). https://uroweb.org/guidelines/testicular-cancer
11. European Association of Urology. EAU guidelines on renal cell carcinoma. (Accessed 19 May 2025). https://uroweb.org/guidelines/renal-cell-carcinoma
12. European Association of Urology. EAU guidelines on prostate cancer. (Accessed 19 May 2025). https://uroweb.org/guidelines/prostate-cancer
13. European Association of Urology. EAU guidelines on muscle-invasive and metastatic bladder cancer. (Accessed 19 May 2025). https://uroweb.org/guidelines/muscle-invasive-and-metastatic-bladder-cancer
14. European Association of Urology. EAU guidelines on non-muscle-invasive bladder cancer. (Accessed 19 May 2025). https://uroweb.org/guidelines/non-muscle-invasive-bladder-cancer
15. European Association of Urology. EAU guidelines on upper urinary tract urothelial cell carcinoma. (Accessed 19 May 2025). https://uroweb.org/guidelines/upper-urinary-tract-urothelial-cell-carcinoma
16. Borque-Fernando Á, Navarro D, Doblare M *et al.* Accuracy of large language models in interpreting urological clinical guidelines: a comparative study with expert evaluation. *Ther. Adv. Urol.* 18, 17562872261436905 (2026).
17. Ah-Thiane L, Heudel PE, Campone M *et al.* Large language models as decision-making tools in oncology: comparing artificial intelligence suggestions and expert recommendations. *JCO Clin. Cancer Inform.* 9, e2400230 (2025).
18. Eckrich J, Ellinger J, Cox A *et al.* Urology consultants versus large language models: potentials and hazards for medical advice in urology. *BJUI Compass* 5(5), 438–444 (2024).