Research Article

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



Journal of **Comparative Effectiveness Research**

Evaluation of large language models for clinical decision support in penile and testicular cancer: a multidisciplinary Delphi study

José Rubio-Briones[‡,1], Laura Izquierdo Reyes[‡,2], Jorge García Rodríguez[3], Urbano Anido Herranz[4], Antonio Rosino Sánchez[5], Juan Alonso Cabo González[6], Antonio Salinas Sánchez[7], Mario Domínguez Esteban[8], Alfonso Gómez de Liaño Lista[9], Pilar González-Peramato Gutiérrez[10], Rafael Luque Barona[11], Oscar Gorria Cardesa[12], María Antonia Gómez-Aparicio[13], José Lorenzo Muñoz Iglesias[14], Macarena Rodríguez-Fraile[15], Abrahams Ocanto Martínez[16], Josep M Gaya[17], Patricia Ramírez Rodríguez-Bermejo[18] & Ángel Borque-Fernando*[,19]

[1]Urology Department, Hospital VITHAS 9 de Octubre, Valencia, Spain
[2]Urology Department, Hospital Clínic Barcelona, Barcelona, Spain
[3]Urology Department, Hospital Universitario Central de Asturias, Oviedo, Spain
[4]Oncology Department, Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela, Spain
[5]Urology Department, Hospital General Universitario Morales Meseguer, Murcia, Spain
[6]Urology Department, Complejo Hospitalario Universitario de Badajoz, Badajoz, Spain
[7]Urology Department, Complejo Hospitalario Universitario de Albacete, Universidad de Castilla La Mancha, Albacete, Spain
[8]Urology Department, Hospital Universitario Marqués de Valdecilla, IDIVAL, Santander, Spain
[9]Oncology Department, Complejo Hospitalario Universitario Insular-Materno Infantil de Gran Canaria, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
[10]Pathology Department, Hospital Universitario La Paz, Universidad Autónoma de Madrid, Madrid, Spain
[11]Pathology Department, Hospital Universitario de Jaén, University of Jaén, Jaén, Spain
[12]Reconstructive Urology & Andrology Department, Hospital Universitario de Navarra, Pamplona, Spain
[13]Radiation Oncology Department, Hospital Universitario de Toledo, Toledo, Spain
[14]Nuclear Medicine Department, Complejo Hospitalario Universitario de Vigo, Vigo, Spain
[15]Nuclear Medicine Department, Clínica Universidad de Navarra, Pamplona, Spain
[16]Radiation Oncology Department, Hospital San Francisco de Asís, Madrid, Spain
[17]Urology Department, Fundació Puigvert, Barcelona, Spain
[18]Urology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain
[19]Urology Department, Hospital Universitario Miguel Servet, IIS-Aragón (GIIS-071), Zaragoza, Spain
*Author for correspondence: dr.borque@gmail.com
‡Authors contributed equally as co-first authors

Aim: Large language models (LLMs) may help interpret clinical guideline information, but their performance in penile and testicular cancer remains poorly defined. We evaluated agreement between multidisciplinary expert panels and responses generated by Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini to guideline-based questions on these cancers. **Materials & methods:** Specialized tumor boards developed 25 questions for each cancer based on the 2025 European Association of Urology guidelines. The original guideline text, with figures and tables converted into structured, plain-language textual descriptions, was provided to each LLM. Separate panels of nine experts evaluated the responses using a two-round modified Delphi process with predefined consensus criteria. **Results:** For penile cancer, consensus in agreement was reached on 21/25 responses (84%) for Claude, 24/25 (96%) for Gemini and 23/25 (92%) for ChatGPT, with no statistically significant differences in overall or pairwise comparisons. For testicular cancer, consensus in agreement was reached on 24/25 items (96%) for each model. One Claude response and one ChatGPT response reached consensus in disagreement. No significant inter-model differences were observed. In both cancers, most responses not reaching consensus or requiring a second round related to prognosis or staging. Across both cancers combined, consensus in agreement was 90% for Claude, 96% for Gemini and 94% for ChatGPT, with no significant inter-model differences. Median scores ≥7 occurred in 94–98% of responses for all LLMs. **Conclusion:** Selected lightweight variants from widely used LLM families showed substantial agreement with expert consensus when provided

with guideline content, although findings reflect performance in a controlled, guideline-fed setting and require specialist validation before clinical application.

Plain language summary: Artificial intelligence tools for interpreting penile & testicular cancer guideline questions

What is this article about? Clinical guidelines help doctors make decisions about penile and testicular cancer, but they can be long and complex. Artificial intelligence tools such as Claude, Gemini and ChatGPT may help interpret guideline information. However, their answers need to be checked by medical experts before they are used in practice.

What methodology was used in this study? Specialists created 25 questions for penile cancer and 25 for testicular cancer based on European Association of Urology guidelines. The original guideline text was provided to the three artificial intelligence tools, with figures and tables converted into clear, structured text. Separate panels of nine specialists reviewed the answers and rated whether they agreed with them in a two-round process.

What were the results? Across both cancers, expert panels agreed with 90% of Claude responses, 96% of Gemini responses and 94% of ChatGPT responses. There were no clear overall differences between the three tools. Difficulties were more common in questions about cancer staging and prognosis. In testicular cancer, expert panels disagreed with two responses: one from Claude and one from ChatGPT, meaning that most specialists considered those answers inconsistent with the guideline information.

Why are these results important? These tools may help interpret penile and testicular cancer guideline information. However, apparently convincing answers may still be unreliable. Their responses should therefore be checked by healthcare professionals with expertise in these cancers.

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

Keywords: artificial intelligence • consensus • Delphi technique • guideline adherence • large language models • penile neoplasms • testicular neoplasms

Penile and testicular cancers, while rare in developed regions, present unique clinical and psychosocial challenges within uro-oncology. In Europe, approximately 6700 new cases of penile cancer are diagnosed annually, with most cases occurring in men over 60 years of age, and approximately 2000 deaths are reported each year [1]. Despite its low incidence, penile cancer is associated with substantial psychological and quality-of-life consequences due to the intimate nature of the disease and the stigma often experienced by patients [2]. Testicular cancer represents about 1% of all adult neoplasms and 5% of urological tumors, with most cases occurring in men between 15 and 40 years of age [3]. In Europe, approximately 24,000 new cases and approximately 1600 deaths are reported each year [4]. Despite its relatively low frequency, testicular cancer has a considerable impact due to its occurrence during men's most productive years. Notably, its incidence has increased in recent decades, particularly in industrialized countries, and continues to rise [3].

Clinical practice guidelines, such as those developed by the European Association of Urology (EAU), are fundamental for standardizing the management of these malignancies, synthesizing the latest evidence and expert consensus to support optimal patient care [2,3]. By contrast, the complexity of these guidelines and the diversity of clinical scenarios can make their consistent application challenging in real-world practice. Moreover, the low frequency of these malignancies in real clinical practice may lead to uncertainties and inaccuracies, particularly regarding therapeutic management [2,3]. In recent years, large language models (LLMs) have emerged as promising digital tools to facilitate access to up-to-date recommendations and support the interpretation of guideline-based information [5–7]. Our previous study suggested that posing simple, unmodified questions to LLMs, rather than using reformulated or rewritten prompts, may result in more guideline-consistent responses for uro-oncological scenarios [8]. The results also suggest that, among the models evaluated, Claude demonstrated a higher degree of alignment with guideline content and expert consensus compared with Gemini and ChatGPT [8]. However, further research is needed to systematically evaluate the consistency and reliability of these AI-generated responses across a broader range of guideline-based scenarios and tumor types. Additionally, it remains essential to assess their alignment with expert consensus in multidisciplinary settings.

This study evaluated the guideline-based interpretative performance of responses generated by three selected lightweight LLM variants from widely used model families (Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini) for penile and testicular cancer. Specifically, we aimed to determine whether AI-generated responses to clinical guideline-based questions, generated after providing the models with the original EAU guideline text, with figures and tables converted into structured, plain-language textual descriptions, aligned with expert consensus and current guideline standards. To achieve this, we employed a Delphi methodology, engaging a multidisciplinary panel of specialists to independently evaluate the degree of agreement with the responses generated by these LLMs to guideline-based questions relevant to uro-oncology. This article focuses specifically on penile and testicular cancers, while three additional articles in this supplement address the remaining major urological malignancies: one covers upper tract urothelial carcinoma, muscle-invasive bladder cancer, and non-muscle-invasive bladder cancer; another is dedicated to prostate cancer; and a third addresses renal cell carcinoma.

Materials & methods

Study design

This multidisciplinary study focused on penile and testicular cancers and employed a two-round modified Delphi methodology, aligned with the RAND/UCLA Appropriateness Methodology [9,10], to assess the degree of agreement with responses generated by three LLMs: Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini.

The optimization of LLMs to produce guideline-based responses in accordance with the 2025 EAU guidelines for these two uro-oncological malignancies, coupled with their favorable validation through multidisciplinary expert consensus, is part of a broader strategic initiative. This project aims to deliver a registered, AI-driven chatbot (urooncoguias.aeu.es) designed to support access to and interpretation of guideline-based information across the full spectrum of uro-oncological conditions.

This project was developed under the coordination of the Spanish Association of Urology (AEU). Its technical implementation was 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 its 2025 Guidelines for the development of dedicated LLM-based models.

Organizational structure & workflow

The project was supported by one technical and one scientific board, as well as an expert panel for each malignancy. The scientific board provided strategic oversight and ensured representation across all urological cancer subtypes. The technical board contributed methodological and digital expertise. Specialized tumor boards for penile and testicular cancers, composed of multidisciplinary experts from relevant Spanish societies, were responsible for developing the 25 guideline-based questions for each malignancy, ensuring alignment with the 2025 EAU guidelines. The members of these specialized tumor boards served as the expert panel for each cancer type. The panel comprised nine multidisciplinary clinicians (urologists, radiation oncologists, medical oncologists, nuclear medicine specialists and pathologists) representing the AEU, the Spanish Society of Radiation Oncology (SEOR), the Genitourinary Alliance for Research and Development (GUARD), the Spanish Society of Nuclear Medicine and Molecular Imaging (SEMNIM) and the Spanish Society of Anatomical Pathology (SEAP). All were experts in penile or testicular cancer and independently evaluated the LLM-generated responses.

Adaptation of guideline figures & tables to plain language

All figures and tables from the EAU clinical guidelines for penile and testicular cancers [2,3] were previously and systematically adapted into structured, plain-language textual descriptions by the panel members. Each guideline was divided into thematic sections and underwent collaborative peer review by independent reviewers from the relevant tumor boards. Discrepancies were resolved by consensus among guideline coordinators, with final approval requiring unanimous validation.

Questionnaire development

A set of 25 structured questions was developed for each malignancy (penile and testicular cancer) by the specialized tumor boards, based on the original content of the EAU guidelines [2,3]. These questions covered key clinical areas such as diagnosis, risk stratification, indications for diagnostic testing and therapeutic management, and clinical cases were also proposed for evaluation.

LLM selection & response generation

Three widely accessible lightweight/efficient LLM variants were tested: Claude 3.5 Haiku (Anthropic; version claude-3-5-haiku-20241022, launched 22 October 2024) [11], Gemini 2.5 Flash (Google; version gemini-2.5-flash-preview-05-20, launched 20 May 2025) [12] and GPT-4.1-mini (OpenAI; version gpt-4.1-mini-2025-04-14, launched 14 April 2025) [13].

Each model was tested exclusively using the simple (direct) prompt approach, in which the original set of 25 structured questions was submitted without modification in Spanish, with the original guideline text and the plain-language textual adaptations of figures and tables provided to facilitate accurate interpretation by the LLMs. Questions were submitted independently and one at a time, and each model was evaluated separately rather than in a single conversation. The final LLM response documents were completed on 6 May 2025, for penile cancer and on 27 May 2025, for testicular cancer.

Delphi validation process & consensus criteria

Expert panelists independently rated each LLM response using a 9-point Likert scale (1 = strong disagreement, 9 = strong agreement) to assess alignment with the EAU guideline content. The 9-point scale was selected in accordance with the RAND/UCLA methodology [10]. Panelists were blinded to the identity of the LLM responsible for each answer. If a panelist felt unqualified to assess a specific question because of a mismatch between their professional expertise and the question's topic, they could abstain from voting; only ratings from those who participated were included in the analysis.

The Delphi process comprised two rounds. In the first round, all responses were scored. For questions where consensus was not achieved, a second round was conducted after providing anonymized, aggregated feedback summarizing the group's ratings. Consensus was defined by three criteria: the median score had to fall within either the agreement (7–9) or disagreement (1–3) range; fewer than a third of panelists could provide ratings outside these ranges and the interquartile range had to be less than 4. Questions not meeting these criteria in the first round were reconsidered in the second round. Qualitative comments were also documented to contextualize the quantitative findings.

Statistical analysis

Two binary outcomes were considered for model comparison: achievement of consensus in agreement versus lack of consensus in agreement, which included both no consensus and consensus in disagreement, and a median score ≥ 7 versus <7. Since all three LLMs were evaluated using the same set of 25 clinical questions, the analyses accounted for the paired structure of the data. Differences across the three models were examined using Cochran's Q test, while pairwise model comparisons were conducted with the exact McNemar test. Bonferroni adjustment was applied to the three pairwise comparisons within each outcome, yielding a significance threshold of $p < 0.0167$. Comparisons between penile and testicular cancer were performed using Fisher's exact test. Statistical significance was set at $p < 0.05$ for global model comparisons and between-cancer comparisons.

The full set of responses generated by each LLM to all 25 questions is available for consultation at https://www.contentednet.com/proyecto-urooncoguias-aeu/

Ethical considerations

This study involved only expert panelists who participated voluntarily in the Delphi process, and no patient data were used. Participation implied informed consent, and formal ethical approval was not required.

Results

Penile cancer

Nine panelists participated in the Delphi process for penile cancer. Sixteen LLM responses were assessed by all nine panelists, seven by eight panelists and two by seven panelists. The questions, details of the consensus results, and the main summarized panelist comments are shown in Supplementary Table 1.

Median scores assigned by panelists for each response and LLM are shown in Figure 1. The proportion of panelists scoring in each category (1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement) per response and LLM is shown in Supplementary Figure 1.

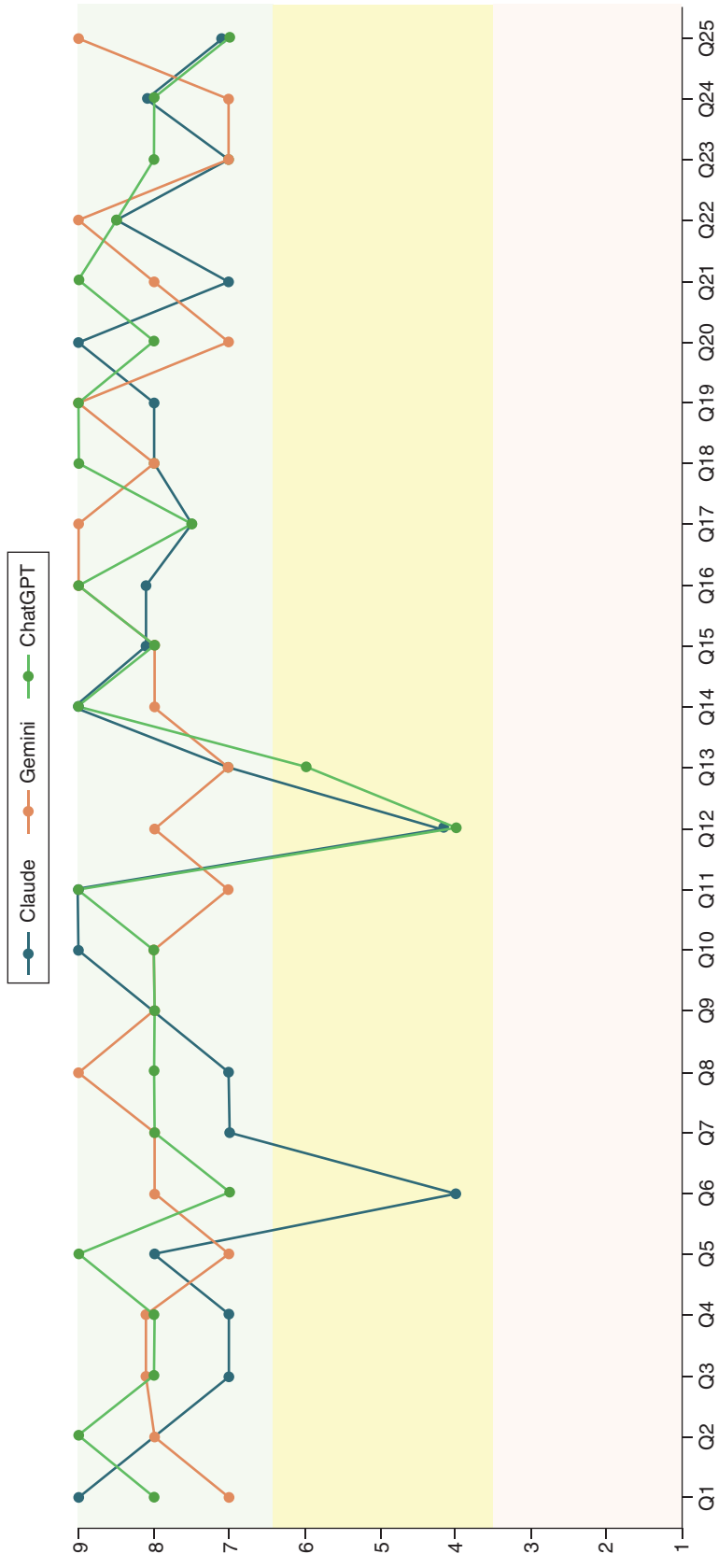


Figure 1. Median scores assigned by panelists for each response and large language model in penile cancer. Median scores on a 9-point scale are shown for each clinical question and large language model (Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini). Score ranges were defined as: 1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement.

10.57264/cer-2026-0080

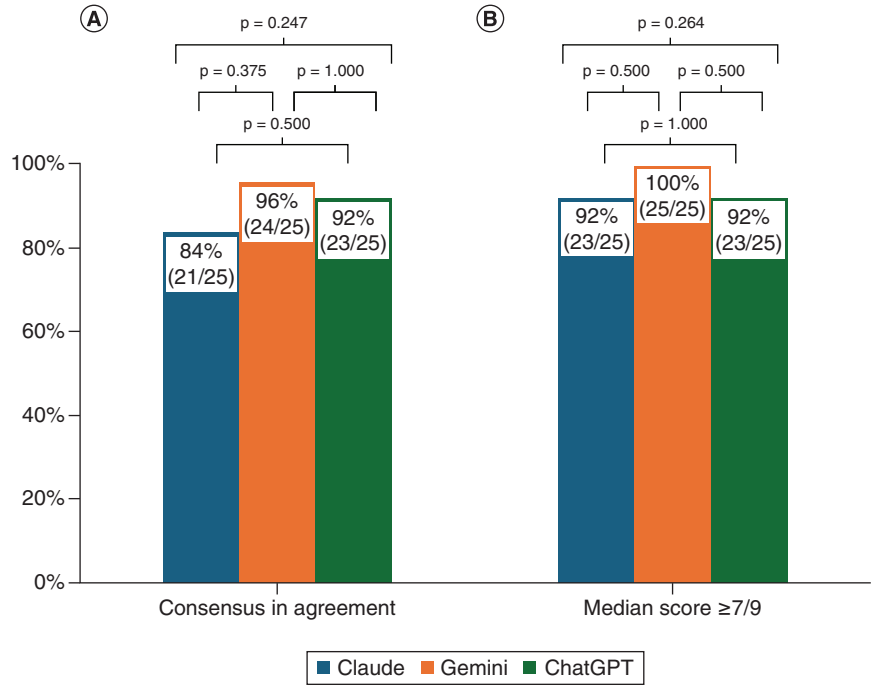


Figure 2. Comparison of expert panel agreement across large language models in penile cancer. (A) Proportion of consensus in agreement and **(B)** high-scoring items (median ≥7/9) across 25 responses for each large language model in penile cancer.

For Claude, consensus in agreement was reached in the first round for 20 responses (80%), in the second round for 1 response without prior consensus (4%), and consensus was not reached for 4 responses (16%). In total, consensus in agreement was reached for 21 out of 25 items (84%).

For Gemini, consensus in agreement was reached in the first round for 23 responses (92%), in the second round for 1 response without prior consensus (4%), and consensus was not reached for 1 response (4%). In total, consensus in agreement was reached for 24 out of 25 items (96%).

For ChatGPT, consensus in agreement was reached in the first round for 20 responses (80%), in the second round for 3 responses without prior consensus (12%), and consensus was not reached for 2 responses (8%). In total, consensus in agreement was reached for 23 out of 25 items (92%).

No statistically significant differences were observed in the proportion of items reaching consensus in agreement among the three LLMs, either overall or in pairwise comparisons (Figure 2). Similarly, the proportion of items with a median score ≥7 did not differ significantly across the three models or in pairwise analyses (Figure 2).

Six of the seven responses that either reached consensus in the second round or failed to reach consensus addressed topics related to prognosis or staging.

Testicular cancer

Nine panelists participated in the Delphi process for testicular cancer, and all nine responded to all 25 questions. The questions, details of the consensus results, and the main summarized panelist comments are shown in Supplementary Table 2.

Median scores assigned by panelists for each response and LLM are shown in Figure 3. The proportion of panelists scoring in each category (1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement) per clinical question and LLM is shown in Supplementary Figure 2.

For Claude, consensus in agreement was reached in the first round for 24 responses (96%), and consensus in disagreement was reached in the second round for one response (4%). There were no responses without consensus.

For Gemini, consensus in agreement was reached in the first round for 23 responses (92%), in the second round for one response (4%) and consensus was not reached for one response (4%). In total, consensus in agreement was reached on 24 out of 25 items (96%).

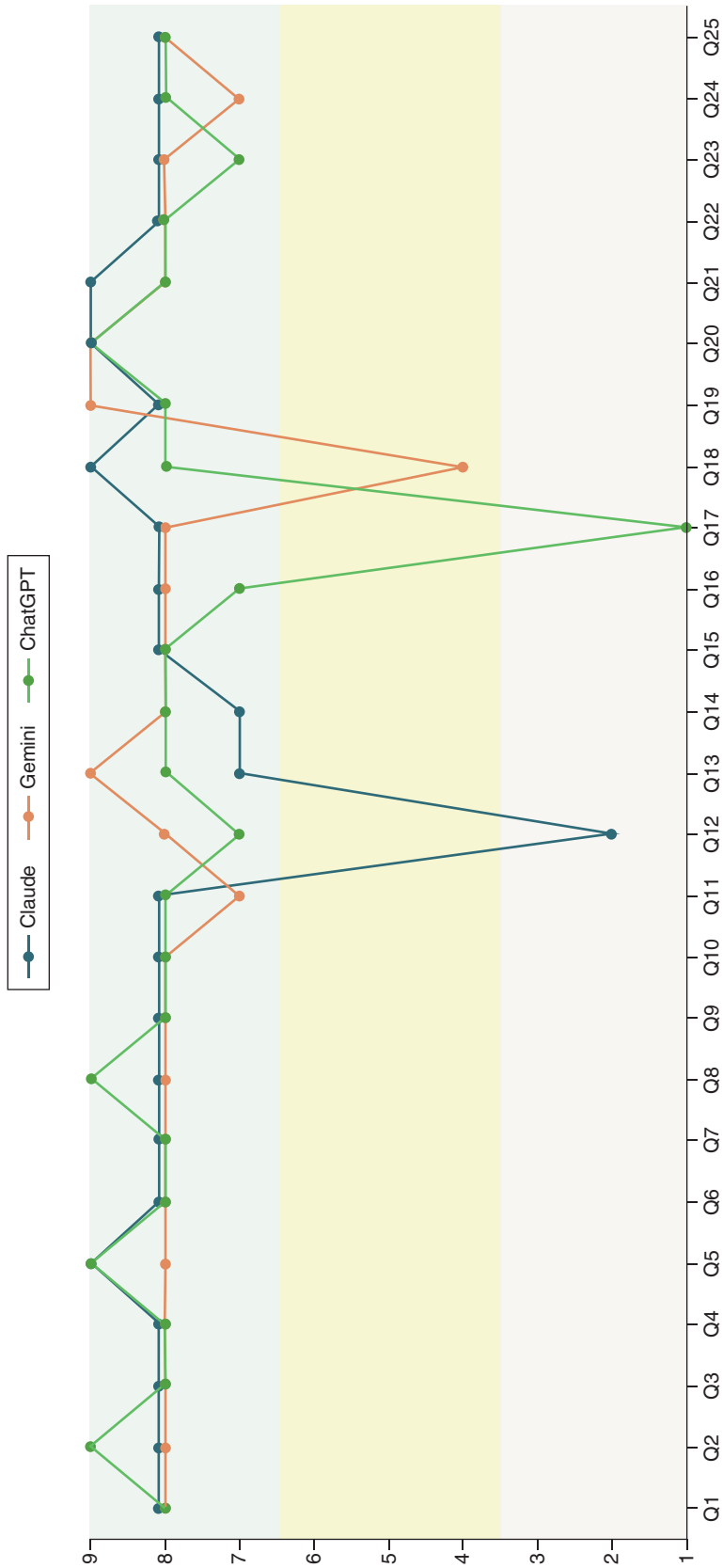


Figure 3. Median scores assigned by panelists for each response and large language model in testicular cancer. Median scores on a 9-point scale are shown for each clinical question and large language model (Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini). Score ranges were defined as: 1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement.

10.57264/cer-2026-0080

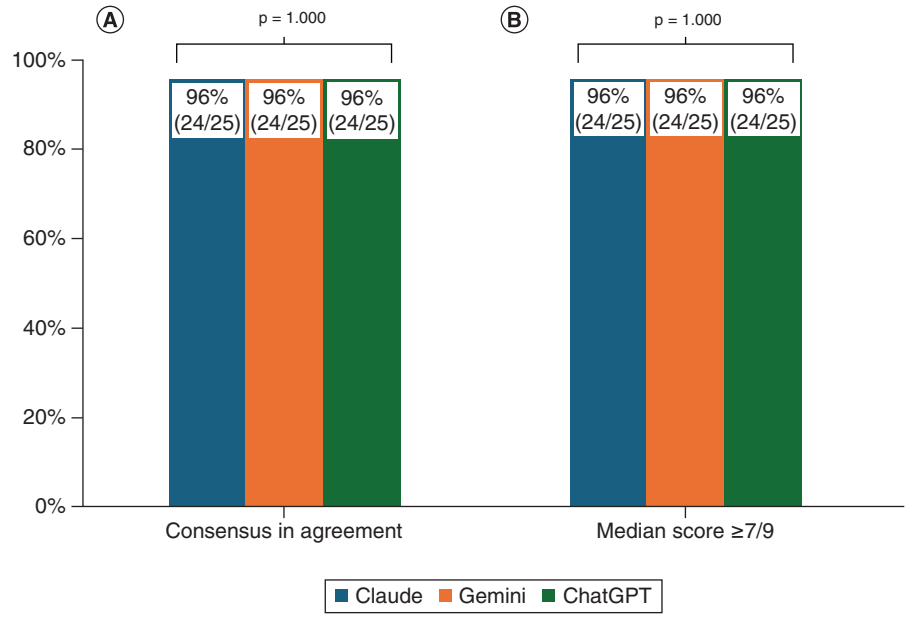


Figure 4. Comparison of expert panel agreement across large language models in testicular cancer. (A) Proportion of consensus in agreement and **(B)** high-scoring items (median ≥7/9) across 25 responses for each large language model in testicular cancer.

For ChatGPT, consensus in agreement was reached in the first round for 23 responses (92%) and in the second round for one response (4%). In total, consensus in agreement was reached on 24 out of 25 items (96%). Consensus in disagreement was reached in the first round for one response (4%). There were no responses without consensus.

No statistically significant differences were found in the proportion of items reaching consensus in agreement among the three LLMs, either overall or in pairwise comparisons (Figure 4A). Likewise, the proportion of items with a median score of ≥7 showed no significant differences across the models or in pairwise analyses (Figure 4B).

Four of the five responses that either reached consensus in the second round or failed to reach consensus addressed topics related to prognosis or staging. The two responses reaching consensus in disagreement concerned testicular cancer. Claude's response to Q12 involved errors in β-hCG interpretation, staging and treatment, whereas ChatGPT's response to Q17 was incomplete because it did not address pre-chemotherapy serum marker values.

Combined results for penile & testicular cancer

Globally, considering both tumors, the proportion of responses that reached consensus in agreement for Claude, Gemini and ChatGPT was 45/50 (90%), 48/50 (96%) and 47/50 (94%), respectively, with no statistically significant differences found in the comparison among the three LLMs, either overall ($p = 0.417$) or in pairwise analyses (all $p \geq 0.453$). Similarly, the proportion of responses scored with a median ≥7 was 47/50 (94%), 49/50 (98%) and 47/50 (94%) for Claude, Gemini and ChatGPT, respectively, again with no significant differences among the three LLMs, either overall ($p = 0.513$) or in pairwise comparisons (all $p \geq 0.625$). Only 1/50 (2%) responses for Claude and 1/50 (2%) responses for ChatGPT reached consensus in disagreement.

There were no significant differences in the total number of responses reaching consensus in agreement when combining the results from all three LLMs for penile cancer (68/75; 90%) compared with testicular cancer (72/75; 96%) ($p = 0.327$), nor in the number of responses rated with a median score of ≥7: 71/75 (95%) versus 72/75 (96%) ($p = 1.000$).

Discussion

In this study, three selected lightweight LLM variants showed high concordance with a multidisciplinary panel of experts when responding to guideline-based questions after being provided with guideline materials compris-

ing the original 2025 EAU guideline text and plain-language textual descriptions of figures and tables for penile and testicular cancer. More than 90% of the responses reached consensus in agreement, with no statistically significant differences among models, indicating broadly comparable performance in standard scenarios. This finding is particularly relevant given the rarity of these malignancies and the limited representation of such tumor types in previous research on AI applications in uro-oncology. Importantly, only a small number of responses required a second Delphi round or failed to reach consensus, and these were typically the most complex or controversial scenarios. Most disagreements were concentrated in aspects related to staging and prognostic assessment. These findings suggest that, in a controlled, guideline-fed setting, selected lightweight LLM variants can consistently interpret guideline-derived information across most standard scenarios, although limitations remain in areas of greater clinical ambiguity, where decision-making requires nuanced interpretation and accumulated expertise.

LLMs are increasingly being explored as tools to support access to evidence-based medical information in medicine [5,14–16]. These advanced AI systems are trained on vast biomedical literature and clinical guidelines, enabling them to generate detailed, contextually relevant responses to complex clinical queries [17,18]. In oncology and other specialties, LLMs have the potential to bridge the gap between rapidly evolving guidelines and clinical practice, offering clinicians timely access to synthesized knowledge and supporting access to diagnostic and therapeutic information [14,18]. However, the reliability and safety of LLM-generated responses must be rigorously validated, particularly in high-stakes fields like cancer care, to ensure effective and responsible future integration into clinical workflows [15,17].

A detailed analysis of our results indicates that the degree of agreement among the panelists was similar for testicular cancer and penile cancer responses, with no statistically significant differences observed. However, it is noteworthy that only in testicular cancer did two responses result in consensus disagreement with some of the LLMs, suggesting that these models occasionally provided responses that did not align with expert panel expectations. Across both tumor types, responses related to standard management generally reached high agreement. By comparison, responses related to disease staging and prognosis, such as the application of staging systems and the interpretation of prognostic factors, were more likely to generate disagreement or require a second Delphi round. Common panelist criticisms in these domains included omission or misinterpretation of key staging criteria and a lack of nuance in the assessment of prognostic factors. These findings highlight that, while LLMs can consistently interpret guideline-derived information in most routine scenarios, their performance appears less robust when faced with the complexity of cancer staging and prognosis, reinforcing the need for expert oversight in these critical aspects of patient management. Moreover, the apparent simplicity of such tools carries the inherent risk that even general practitioners or non-subspecialized specialists may underestimate the importance of involving dedicated experts when interpreting clinical information or considering therapeutic decisions. The two responses reaching consensus in disagreement are particularly relevant from a safety perspective. In both cases, the models generated plausible responses despite clinically relevant errors or omissions. Such inaccuracies may be difficult for non-specialist users to identify, reinforcing the need for specialist validation of LLM-generated responses.

Despite these limitations, overall acceptance of the LLM-generated responses was remarkably high, with median scores of approximately 8 out of 9 and 94–98% of LLM responses achieving a median score ≥ 7 across the three models. These high panel scores support the consistency of the LLM-generated responses with the guidelines, although validation remains necessary before clinical implementation.

Several recent studies provide important context for interpreting our findings, both in general oncology and specifically within urological oncology. In the field of medical oncology, Longwell *et al.* [18] conducted a cross-sectional evaluation of LLMs using standardized examination-style questions from the American Society of Clinical Oncology (ASCO) and the European Society for Medical Oncology (ESMO). They found that the best-performing proprietary LLM, which was not named in the publication due to legal restrictions, correctly answered 85% of questions, with most explanations containing no or only minor errors. However, the study also highlighted that a significant proportion of incorrect answers could potentially lead to moderate or severe clinical harm if acted upon, emphasizing the need for careful oversight even when overall accuracy is high.

Focusing on urological oncology, Eckrich *et al.* [16] compared the performance of several LLMs with that of urology consultants using case-based clinical questions. Although LLMs consistently generated coherent and comprehensible answers, their medical adequacy was rated lower than that of consultants. Of note, the study identified potential misinformation hazards in 2.8–18.9% of LLM-generated answers, compared with less than 1% for consultant responses. The authors concluded that, while LLMs can produce coherent medical advice in

urology, their adequacy remains inferior, and their answers may contain potentially hazardous misinformation. Therefore, they should not be used in clinical practice without expert supervision. In our study, consensus in disagreement occurred in 2 of 150 model responses (1.3%); however, this outcome is not directly comparable to the potential misinformation hazard assessed by Eckrich *et al.* due to differences in the evaluation frameworks. A similar need for expert oversight was observed in pediatric urology. Hoen *et al.* assessed LLMs' ability to update guideline content and found that, despite producing scientifically accurate responses in many cases, the models demonstrated insufficient consistency and reliability to replace expert panel review, especially when integrating new evidence [19].

A significant challenge in the evaluation of LLMs in oncology is the absence of a standardized methodology. As highlighted by Carl *et al.* [17], studies use diverse approaches, such as multiple-choice questions, open-ended scenarios or comparisons with expert decisions, making direct comparison difficult. Our use of a Delphi process with a multidisciplinary panel offers advantages, like structured consensus and detailed feedback, but alternative methods also provide valuable perspectives. This diversity underscores the need for greater methodological harmonization in future research [17].

Our results, along with recent studies, suggest that LLMs may support access to guideline-based information, especially in standard scenarios or resource-limited settings. However, our findings should be interpreted as evidence of performance in a controlled guideline-fed task, rather than as evidence of clinical decision-making capability [17,18]. Therefore, their integration into clinical practice should be gradual and always under expert supervision, beginning with low-risk tasks or educational support and expanding as more evidence emerges regarding their safety and effectiveness [14,18]. Transparent reporting of model version, training data and update criteria, as well as independent evaluation of biases and limitations, is essential [14,17]. Multidisciplinary collaboration among clinicians, AI experts, and patients is also crucial to ensure these tools address real clinical needs.

Regarding practical implementation, the planned urooncoguias.aeu.es chatbot is intended exclusively for healthcare professionals and will require professional authentication. It is designed to support access to and interpretation of guideline-based information rather than replace specialist judgment. The errors identified in the present study, particularly in staging and prognosis, reinforce the need for expert oversight when interpreting chatbot-generated responses.

Finally, in our view, to maximize benefit and minimize risk, clinical guidelines – particularly those related to complex areas such as cancer staging – should increasingly adopt structured formats, with clear decision algorithms, explicit definitions and standardized tables, ensuring interpretability by both clinicians and AI systems.

This study has several limitations. First, it focused on a limited number of clinical settings and tumor types, which may affect the generalizability of the findings. In addition, the relatively small number of questions evaluated per cancer type may have limited the ability to detect differences between models; therefore, the absence of statistically significant differences should not be interpreted as evidence of equivalence. Second, the evaluation relied on expert panel consensus, which, while robust, may introduce subjective bias. In some questions, especially those with greater specialty specificity, only a subset of the panel voted, and consensus was calculated exclusively among those who responded. While this increases content expertise for each item, it may introduce potential bias due to variable sample sizes across questions, which should be considered when interpreting the results. The study did not evaluate open-ended, real-world clinical use of LLMs. Instead, the models were provided with guideline materials comprising the original text and plain-language textual descriptions of figures and tables, so the study primarily assessed their ability to interpret and reproduce the supplied guideline-derived information. Moreover, the rapidly evolving nature of LLM technology means that model performance may differ in future versions. Furthermore, only a single response was generated per question and model, and repeated runs were not performed; therefore, output stability and reproducibility across runs could not be assessed. Nevertheless, this work has important strengths. We converted EAU guideline figures and tables into clear, structured plain-language textual descriptions, ensuring that both experts and AI models had access to the same information and reducing ambiguity in interpretation. The use of a multidisciplinary Delphi process further enhanced the rigor and relevance of the evaluation.

Overall, our results indicate that selected lightweight variants from widely used LLM families, when provided with clinical guideline content, can generate responses that align with expert consensus in most standard guideline-based scenarios involving penile and testicular cancer. These findings should be interpreted within the context of a controlled, guideline-fed evaluation and require further validation in more realistic clinical settings before any integration into oncology practice.

Summary points

- Penile and testicular cancers are uncommon malignancies in which complex guideline information may be difficult to interpret consistently.
- This study evaluated Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini using 25 guideline-based questions for penile cancer and 25 for testicular cancer.
- Model responses were assessed by separate multidisciplinary expert panels through a two-round modified Delphi process.
- Across both cancers, consensus in agreement was reached for 90% of Claude responses, 96% of Gemini responses and 94% of ChatGPT responses, with no significant differences between models.
- Median scores of at least 7 out of 9 were achieved for 94–98% of responses across the three models.
- Two responses in testicular cancer reached consensus in disagreement: one from Claude and one from ChatGPT.
- Difficulties were more common in questions related to cancer staging and prognosis.
- The findings support the potential role of large language models as guideline-interpretation tools under specialist supervision, rather than as autonomous clinical decision-making tools.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at https://becarispublishing.com/doi/epdf/10.57264/cer-2026-0080

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 Pablo Rivas for his assistance with medical writing on behalf of Content Ed Net.

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 study 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.

Data sharing statement

The results of the Delphi process are presented within the manuscript. The complete set of responses generated by each large language model to all 25 questions are available for consultation at https://www.contentednet.com/proyecto-urooncoguias-aeu/

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

Papers of special note have been highlighted as: ● of interest; ●● of considerable interest

1. Ferlay J, Ervik M, Lam F *et al.* Penile cancer fact sheet. *Global Cancer Observatory: Cancer Today*. International Agency for Research on Cancer, Lyon, France (2024). (Accessed 10 July 2025). https://gco.iarc.who.int/media/globocan/factsheets/cancers/26-penis-fact-sheet.pdf
2. European Association of Urology. EAU guidelines on penile cancer. (Accessed: 19 May 2025). https://uroweb.org/guidelines/penile-cancer
3. European Association of Urology. EAU guidelines on testicular cancer. (Accessed: 19 May 2025). https://uroweb.org/guidelines/testicular-cancer
4. Ferlay J, Ervik M, Lam F *et al.* Testicular cancer fact sheet. *Global Cancer Observatory: Cancer Today*. International Agency for Research on Cancer, Lyon, France (2024). (Accessed 10 July 2025). https://gco.iarc.who.int/media/globocan/factsheets/cancers/28-testis-fact-sheet.pdf
5. Benary M, Wang XD, Schmidt M *et al.* Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* 6(11), e2343689 (2023).
6. Kaiser P, Yang S, Bach M *et al.* The interaction of structured data using openEHR and large language models for clinical decision support in prostate cancer. *World J. Urol.* 43(1), 67 (2025).
7. Li K-P, Wang L, Wan S *et al.* Enhanced artificial intelligence in bladder cancer management: a comparative analysis and optimization study of multiple large language models. *J. Endourol.* 39(5), 494–499 (2025).
8. 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).

●● **Compares Claude, Gemini and ChatGPT across major urological cancers and provides the basis for the present Delphi validation study.**

9. Jones J, Hunter D. Consensus methods for medical and health services research. *BMJ* 311(7001), 376–380 (1995).
10. Fitch K, Bernstein SJ, Aguilar MD *et al. The RAND/UCLA appropriateness method user's manual*. RAND Corporation, CA, USA (2001).
11. Anthropic. Models overview. Claude Platform Docs. (Accessed 29 June 2026). https://platform.claude.com/docs/en/about-claude/models/overview
12. Google AI for Developers. Gemini 2.5 Flash. Gemini API. (Accessed 29 June 2026). https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash
13. OpenAI. GPT-4.1 mini model. OpenAI API. (Accessed 29 June 2026). https://developers.openai.com/api/docs/models/gpt-4.1-mini
14. Ah-Thiane L, Heudel P-E, 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).
15. 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).

●● **Evaluates large language models (LLMs) in realistic clinical decision-making scenarios and shows important limitations in diagnosis, guideline adherence, interpretation of laboratory results and integration into clinical workflows, supporting the need for caution before autonomous clinical use.**

16. 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).

● **Compares LLM-generated answers with those of urology consultants and shows lower medical adequacy and a higher risk of potentially hazardous misinformation in LLM responses.**

17. Carl N, Schramm F, Haggenmüller S *et al.* Large language model use in clinical oncology. *NPJ Precis. Oncol.* 8(1), 240 (2024).

● **Synthesizes current applications, methodologies and performance of LLMs in clinical oncology and highlights substantial heterogeneity driven by differences in evaluation methods, prompting strategies and oncological domains.**

18. 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).
19. 't Hoen LA, van Uitert A, Bussmann M *et al.* Annual updates of the European Association of Urology–European Society for Pediatric Urology (EAU-ESPU) paediatric urology guidelines: are large-language models (LLM) better than the usual structured methodology? *J. Pediatr. Urol.* 21(6), 1643–1649 (2025).