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 renal cell carcinoma: a multidisciplinary Delphi study

Sergio Fernández-Pello Montes[1], Ignacio Osman García[2], Rocio Barrabino Martín[3], María José Agustín Ferrández[4], Ferrán Algaba Arrea[5], Eliseo Carrasco Esteban[6], Bernardo Herrera Imbroda[7], María Carmen Mir Maresma[8], Carlos Nicolau[9], Marc Simó Perdigó[10], Guillermo de Velasco Oria de Rueda[11] & Ángel Borque-Fernando*[,12]

[1]Urology Department, Hospital Universitario de Cabueñes, Gijón, Spain
[2]Urology & Nephrology Department, Hospital Universitario Virgen del Rocío, Instituto Biomédico de Sevilla (Ibis), Universidad de Sevilla, Sevilla, Spain
[3]Urology Department, Hospital Universitario Torrecárdenas, Almeria, Spain
[4]Pharmacy Department, Hospital Universitario Miguel Servet, Zaragoza, Spain
[5]Pathology Department, Fundación Puigvert, Barcelona, Spain
[6]Radiotherapy Department, Hospital Central de la Defensa Gomez Ulla, Madrid, Spain
[7]Urology Department, Virgen de la Victoria Hospital. IBIMA-Plataforma Bionand, Málaga, Spain
[8]Urology Department, Icahn School of Medicine at Mount Sinai, Elmhurst/Queens Hospitals, NY, USA
[9]Radiology Department, Hospital Clinic, University of Barcelona, Barcelona, Spain
[10]Nuclear Medicine Department, Hospital Universitari Dexeus, Barcelona, Spain
[11]Oncology Department, Hospital Universitario 12 de Octubre, Madrid, Spain
[12]Urology Department, Hospital Universitario Miguel Servet, Universidad de Zaragoza, IIS-Aragón, Zaragoza, Spain
*Author for correspondence: dr.borque@gmail.com

Aim: Large language models (LLMs) may facilitate the interpretation of clinical guideline information, but their performance in renal cell carcinoma (RCC) remains insufficiently characterized. We evaluated agreement between a multidisciplinary expert panel and the responses generated by Claude, Gemini and ChatGPT to RCC guideline-based questions. **Materials & methods:** A multidisciplinary tumor board developed 25 clinical questions based on the 2025 European Association of Urology RCC guidelines. Guideline figures and tables were converted into structured, plain-language textual summaries and supplied to the models. Each LLM generated one response per question using direct prompting. An 11-member multidisciplinary expert panel evaluated the responses in a two-round modified Delphi process using a 9-point scale and predefined consensus criteria. **Results:** Consensus in agreement was reached for 24/25 Gemini responses (96%) and 23/25 responses (92%) for both Claude and ChatGPT. No response reached consensus in disagreement. The proportion of responses with a median score ≥7 was 25/25 (100%) for Gemini, 24/25 (96%) for Claude and 23/25 (92%) for ChatGPT. No statistically significant differences were detected among the models overall or in pairwise comparisons. Responses requiring further discussion or failing to reach consensus mainly involved clinically uncertain or individualized scenarios. Panelists identified occasional factual inaccuracies, omitted options and limited clinical nuance. **Conclusion:** When provided with structured, guideline-derived material, selected LLMs generated responses that frequently aligned with multidisciplinary expert consensus. These findings support their potential for guideline interpretation but do not demonstrate independent clinical reasoning or real-world decision-making performance. Expert oversight remains essential, particularly in complex scenarios.

Plain language summary: Artificial intelligence tools for interpreting kidney cancer guideline questions

What is this article about? Clinical guidelines help doctors make decisions about renal cell carcinoma, the most common type of kidney cancer, but they can be long and complex. Artificial intelligence tools such as Claude, Gemini and ChatGPT may help interpret this 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 based on the European Association of Urology guidelines for kidney cancer. The original guideline text was provided to the three artificial intelligence tools, with figures and tables converted into clear, structured text. An expert panel of 11 specialists then reviewed the answers and rated whether they agreed with them in a two-round process.

What were the results? The expert panel reached agreement with 96% of Gemini responses and 92% of responses from both Claude and ChatGPT. There were no clear overall differences between the three tools. Difficulties were more common in complex situations, such as careful monitoring without immediate treatment in kidney cancer that has spread, expected outcomes in a less common type of kidney cancer, choosing treatments that destroy the tumor without removing it surgically, and considering imaging tests as possible alternatives to kidney biopsy. Experts also identified some incorrect information, missing options and limited detail.

Why are these results important? These tools may help interpret kidney cancer guideline information. However, they were tested using information supplied from the guidelines and were not shown to make independent clinical decisions. Their responses should therefore be checked by healthcare professionals with expertise in kidney cancer.

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

Keywords: artificial intelligence • clinical decision support systems • consensus • Delphi technique • guideline adherence • large language models • renal cell carcinoma

Renal cell carcinoma (RCC) ranks as the 14th most commonly diagnosed cancer worldwide, representing approximately 2% of all cancer diagnoses [1]. In 2022, there were an estimated 434,840 new cases of RCC and 155,953 related deaths, making it the 16th leading cause of cancer mortality globally [2]. Europe alone accounted for 145,721 new cases and 52,347 deaths, comprising over a third of the global RCC burden [2]. RCC demonstrates a clear male predominance, and its incidence increases significantly with advancing age [2]. Notably, while the global incidence of RCC continues to rise, particularly in developed regions due to increased use of abdominal imaging and consequent incidental detection of asymptomatic tumors, mortality rates in parts of Europe, especially in Scandinavian and Western countries, have progressively declined since the 1980s and 1990s, likely reflecting advances in early detection and treatment strategies that improve survival [1,3].

Clinical practice guidelines, such as those issued by the European Association of Urology (EAU) [1], play a crucial role in standardizing the management of renal malignancies, providing evidence-based recommendations and expert-driven consensus to optimize patient outcomes. However, their implementation in daily practice is frequently challenged by the complexity of guidelines, the heterogeneity of clinical scenarios encountered, ongoing updates to recommendations and variability in clinician judgment and institutional resources [4–6]. This challenge is particularly important in settings with significant epidemiological variability, where differences in disease incidence, clinical presentation and access to healthcare resources can markedly impact clinical decision-making [7]. In this context, AI-driven tools that accurately interpret and facilitate guideline application offer a promising approach to support clinicians in delivering consistent and precise management of RCC across diverse clinical scenarios and healthcare settings. In particular, large language models (LLMs) have recently gained attention as innovative digital tools that can enhance access to current clinical recommendations and facilitate real-time decision-making [8–10].

In our previous work, we observed that submitting simple, unmodified clinical questions to common LLMs (Claude [11], Gemini [12] and ChatGPT [13]) resulted in responses more closely aligned with current clinical guidelines than those generated from modified or reformulated prompts [14]. Among the models evaluated, Claude (version: Claude-Haiku-2024-10-22) showed the highest concordance with guideline-based content and expert consensus, outperforming both Gemini (version: Gemini-Flash-001; May 2024) and ChatGPT (version: GPT-4o-mini-2024-07-18) [14]. Further research is necessary to rigorously assess the consistency, accuracy, and clinical relevance of LLM-generated recommendations across diverse oncologic scenarios. Importantly, their alignment with expert consensus should continue to be evaluated within the framework of multidisciplinary decision-making.

This study evaluated the agreement between a multidisciplinary expert panel and responses generated by three selected LLMs (Claude, Gemini and ChatGPT) to RCC guideline-based clinical questions. The primary objective was to determine whether these responses aligned with the EAU guideline content and multidisciplinary expert consensus. To this end, a multidisciplinary panel of specialists independently rated the degree of agreement with the LLM-generated responses using a modified Delphi approach. While this manuscript focuses on RCC, three complementary articles in this supplement address other major urological malignancies: upper tract and muscle-invasive and non-muscle-invasive bladder urothelial carcinomas, prostate cancer and penile and testicular cancers.

Materials & methods

Study design

This multidisciplinary study focused on RCC and employed a two-round modified Delphi methodology, based on the RAND/UCLA Appropriateness Method [15,16], to assess the degree of agreement with responses generated by three LLMs: Claude, Gemini and ChatGPT.

The optimization of LLMs to produce evidence-based responses in accordance with the 2025 EAU guidelines for RCC, 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 clinical decision-making 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 models.

Organizational structure & workflow

The project was structured around two distinct boards and an expert panel by malignancy: The Scientific Board provided strategic oversight and ensured comprehensive representation across all major urological cancer subtypes; The Technical Board contributed essential methodological and digital expertise; The Specialized Tumor Board for RCC was composed of multidisciplinary experts, including urologists, radiation oncologists, medical oncologists, nuclear medicine physicians, pathologists, hospital pharmacists and radiologists, from relevant Spanish scientific societies: AEU, Spanish Society of Radiation Oncology (SEOR), Genitourinary Alliance for 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). This tumor board was responsible for developing the 25 guideline-based clinical questions, fully aligned with the 2025 European Association of Urology (EAU) guidelines [1]. The expert panel consisted of eleven clinicians who independently evaluated the responses generated by LLMs.

Adaptation of clinical guidelines to plain language

All figures and tables from the EAU clinical guidelines for RCC [1] were systematically transformed into structured, plain-language textual summaries. The adapted material was organized into thematic sections and independently reviewed by members of the specialized RCC tumor board. Any discrepancies were resolved by consensus among the guideline coordinators, with final approval requiring unanimous agreement.

Questionnaire development

The specialized RCC tumor board developed a set of 25 structured questions based on the original EAU guideline content. These questions covered key clinical areas including diagnosis, risk stratification, indications for diagnostic procedures and therapeutic management.

LLM selection & response generation

Three widely accessible, general-purpose proprietary LLMs from major developers were selected: Claude, Gemini and ChatGPT. These models were chosen because they were widely available through user-facing platforms at the time of the study and could be evaluated using the same guideline material and direct prompting approach, without requiring local deployment or model-specific inference configurations. The study was designed to assess models that clinicians could readily use in routine practice rather than to provide an exhaustive benchmark of all available LLM architectures. Claude (Anthropic; version claude-3-5-haiku-20241022, launched 22 October 2024) [17], Gemini (Google; version gemini-2.5-flash-preview-05-20, launched 20 May 2025) [18] and ChatGPT (OpenAI; version gpt-4.1-mini-2025-04-14, launched 14 April 2025) [19] were tested. Each model was evaluated using a direct-prompting approach, with the original 25 structured questions submitted without modification. The models received the original EAU guideline content, together with structured, plain-language textual summaries of the guideline figures and tables. One response was generated per question and model. As the models were accessed through their user-facing platforms, temperature, top-p, top-k and other sampling parameters were not user-configurable or available for reporting and were left at the platform defaults. Repeated runs were not performed; therefore, response stability and variability across generations could not be assessed.

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 EAU guideline content. The 9-point scale was selected in accordance with the RAND/UCLA Appropriateness Method, allowing ratings to be grouped into disagreement (1–3), uncertainty (4–6) and agreement (7–9) and facilitating the application of predefined consensus thresholds. Panelists were blinded to the identity of the LLM generating each response. They could abstain from rating if they felt unqualified because the topic fell outside their area of expertise. Only ratings from participating panelists were included in the analysis.

The Delphi process consisted of two rounds. In the first round, all responses were scored. Questions lacking consensus after the first round were reevaluated in a second round, following the distribution of anonymized, aggregated feedback that summarized group ratings. Consensus was defined by three criteria: median score within an agreement (7–9) or a disagreement (1–3) range; fewer than a third of panelists with ratings outside these ranges and interquartile range (IQR) less than 4. Qualitative comments were recorded to provide context for the quantitative results.

Statistical analysis

To assess differences across the three LLMs, the paired binary outcomes for consensus status and for median scores ≥ 7 versus <7 were compared using Cochran's Q test. Pairwise comparisons between models were conducted using the exact McNemar test, with Bonferroni adjustment for multiple comparisons. Statistical significance was set at $p < 0.05$.

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/

Results

Eleven expert panelists initially participated in the Delphi process. Each of the 25 clinical questions was evaluated by at least 6 panelists; specifically, 1 question was assessed by 9 panelists, 8 questions by 8 panelists, 9 questions by 7 panelists and 7 questions by 6 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.

In the first round, consensus in agreement with the responses generated by the LLMs was reached in 76% of questions for Claude, 56% for Gemini and 88% for ChatGPT. During the second round, additional consensus was achieved for 16% of questions for Claude, 40% for Gemini and 4% for ChatGPT. No question reached consensus in disagreement for any of the evaluated LLMs. Overall, after two rounds of assessment, consensus in agreement was achieved for 24 out of 25 Gemini responses (96%) and 23 out of 25 responses (92%) for both Claude and ChatGPT (Figure 2A & Supplementary Table 1).

Comparative analysis of consensus versus non-consensus items showed no statistically significant differences among the three models ($p = 0.779$) (Figure 2A). When evaluating responses based on median scores on a 9-point Likert scale, with scores ≥ 7 versus <7, Gemini achieved a median score ≥ 7 in all 25 responses (100%), Claude in 24 responses (96%) and ChatGPT in 23 responses (92%), again with no significant differences among models ($p = 0.368$) (Figure 2B). In the pairwise comparisons, no significant differences were observed between the models in either the proportion of consensus items or the proportion of items with scores ≥ 7 (all Bonferroni-adjusted $p = 1.000$) (Figure 2).

Questions requiring additional Delphi rounds or that ultimately failed to reach consensus predominantly focused on clinical scenarios marked by uncertainty and the need for individualized patient management. In particular, consensus was not reached for at least one LLM regarding the following: the appropriateness of active surveillance in metastatic RCC (question 4), the incidence of recurrence and survival outcomes in chromophobe RCC (question 10), the optimal ablative technique for localized renal cancer (question 17) and possible alternatives to renal biopsy for the differential diagnosis of small renal masses (question 19) (Supplementary Table 1).

Qualitative comments provided by expert panelists revealed some recurring limitations across the LLM-generated responses. These included occasional minor inaccuracies, omission of clinically relevant options and insufficient detail in more nuanced clinical scenarios. In some cases, the answers were perceived as overly general or not fully

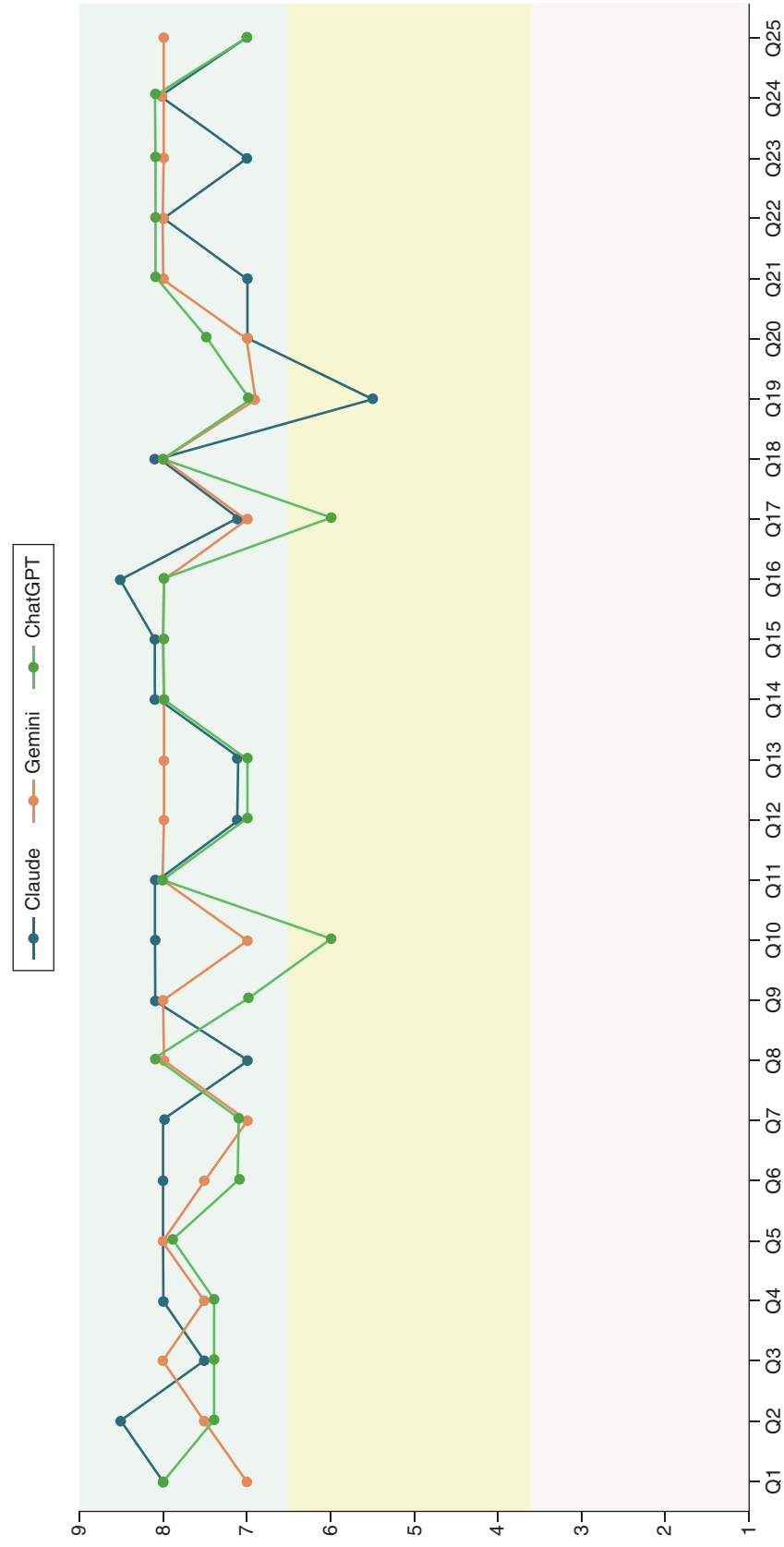


Figure 1. Median scores assigned by panelists for each question and large language model. Median scores on a 9-point scale are shown for each clinical question and large language model (Claude, Gemini, and ChatGPT). Score ranges were defined as: 1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement.

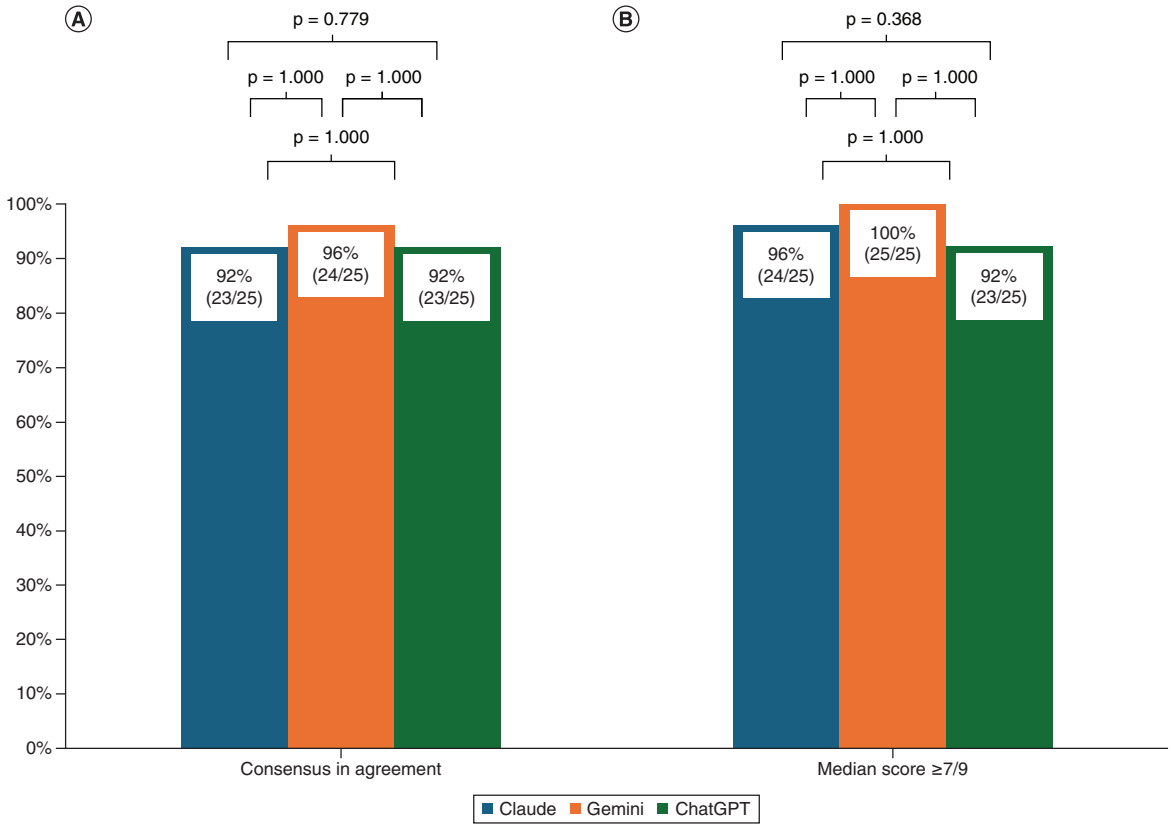


Figure 2. Comparison of expert panel agreement across large language models in renal cell carcinoma. (A) Proportion of consensus in agreement and **(B)** high-scoring items (median ≥7/9) across 25 responses for each large language model.

aligned with guideline recommendations. Such observations were more frequent in questions that underwent a second Delphi round or failed to reach consensus (Supplementary Table 1).

Discussion

This study investigated the performance of three selected LLMs, Claude, Gemini and ChatGPT, in generating guideline-based responses to clinical questions related to RCC. Using a Delphi methodology, a multidisciplinary expert panel systematically rated their agreement with each model's output. A consensus in agreement was reached in 92% to 96% of cases across the models, with median scores of 7 or higher observed in 92–100% of responses. No statistically significant differences were found between the models. However, given the limited number of matched questions and the variable number of raters across items, this finding should not be interpreted as demonstrating equivalent performance among the models. Several questions, particularly those addressing areas of clinical uncertainty or requiring nuanced, individualized decision-making, either necessitated a second Delphi round or failed to reach consensus. These results indicate that, while LLMs frequently generated guideline-concordant information and may support guideline interpretation in routine scenarios, their performance varied in more complex settings, and their outputs require expert oversight.

LLMs are increasingly being adopted as tools to support evidence-based decision-making in oncology [20,21]. These advanced AI systems are trained on extensive biomedical literature and clinical guidelines, enabling them to generate detailed, contextually relevant responses to complex clinical queries [22,23]. LLMs and other AI-driven tools can synthesize large volumes of clinical information, assist with treatment planning, and predict outcomes in selected scenarios, potentially bridging the gap between evolving guidelines and real-world clinical practice [23,24]. Nevertheless, the reliability and safety of LLM-generated recommendations require careful validation before clinical integration, as variations in content accuracy and clinical completeness persist, particularly in ambiguous or high-stakes cases [22,25].

A detailed examination of our results reveals variability in panelist agreement across the 25 clinical questions related to RCC. While several questions involving standard management achieved high levels of consensus, others addressing clinically complex scenarios were also consistently rated in agreement. However, questions that required a second Delphi round or failed to reach consensus with one or more models tended to reflect areas of greater uncertainty or scenarios requiring individualized decisions. The non-consensus scenarios involved active surveillance in metastatic RCC, prognosis in chromophobe RCC, selection of ablative techniques for localized disease and alternatives to renal biopsy. Panelists' comments additionally identified unsupported references to CARMENA and SURTIME in the active-surveillance context and omission of imaging-based alternatives to renal biopsy, including sestamibi SPECT/CT and Zr-girentuximab PET/CT (Supplementary Table 1). These findings illustrate recurring limitations such as omission of clinically relevant options, insufficient clinical nuance, and occasional factual inaccuracies. Importantly, no question resulted in consensus in disagreement, suggesting overall alignment with clinical expectations, albeit with limitations. Overall, these observations underline the potential of LLMs to support guideline interpretation, while reinforcing the need for expert validation, especially in areas where established guidance is limited.

Recent evidence in RCC supports this caution [21]. In a pilot study by Bertolo *et al.*, AI-generated treatment suggestions were compared against decisions made by a multidisciplinary tumor board across 103 RCC cases. While an overall agreement of 62% was reached, concordance varied widely depending on the disease stage, with the lowest alignment observed in advanced metastatic settings [21]. These findings suggest that although LLMs can assist clinicians in standard RCC scenarios, their integration into routine practice is limited by their current inability to fully interpret complex or individualized clinical contexts. Therefore, continuous oversight, model refinement and external validation will be crucial to ensure their safe use in urologic oncology.

Several recent studies in general and urological oncology further support and contextualize these findings. Longwell *et al.* evaluated the performance of a leading proprietary LLM using clinical questions standardized for oncology certification exams from the American Society of Clinical Oncology (ASCO) and the European Society for Medical Oncology (ESMO). These questions, similar to those used to assess the knowledge of practicing clinical oncologists, tested the model's ability to provide high-level, context-specific responses. The study reported an overall accuracy of 85.0%, with most explanations containing no or only minor errors. However, 82% of the incorrect responses were judged to have the potential to cause moderate or severe clinical harm if applied in practice, highlighting the need for expert oversight even in the presence of high accuracy [23]. In urological oncology, Eckrich *et al.* compared several LLMs with practicing urology consultants using case-based clinical questions. Although the models produced fluent and comprehensible answers, their medical adequacy was consistently rated lower than that of human experts. Importantly, potential misinformation was identified in between 2.8% and 18.9% of the LLM-generated responses, compared with less than 1% for consultant answers [26]. Similarly, Hoen *et al.* evaluated the ability of LLMs to assist in updating pediatric urology guidelines and found that, although the models were often able to generate scientifically accurate outputs, they lacked the consistency and reliability required to replace expert panel review, particularly when integrating new or emerging evidence [27]. Together with our findings in RCC, these studies reinforce that while LLMs hold promise as supportive tools in clinical decision-making, their current performance remains variable and insufficiently reliable for unsupervised use. Rigorous validation and expert oversight are crucial for ensuring their safe and effective integration into clinical urological oncology practice.

A notable challenge in evaluating LLMs in oncology is the lack of standardized methodology across studies. As noted by Carl *et al.*, existing research employs diverse strategies, ranging from multiple-choice formats and open-ended clinical prompts to comparisons with expert decisions, which complicates the direct comparison of results [22]. In this study, we employed a Delphi process involving a multidisciplinary panel of experts from urology, oncology, and other specialties, offering the advantages of structured consensus, iterative review and detailed qualitative feedback. However, other evaluation approaches may also provide valuable insights, and the current heterogeneity in methodology underscores the need for greater harmonization in future research on medical applications of LLMs.

This issue is particularly relevant when interpreting the differences between our current results and those of a previous study, which used a different evaluation framework. In a previous study evaluating LLM performance across major urological cancers, we applied a direct expert rating approach, in which 9–11 multidisciplinary uro-oncology specialists independently scored each response using a standardized 5-point Likert scale [14]. That evaluation suggested Claude's superior performance. In contrast, the present study, which focused exclusively on

RCC, employed a Delphi methodology that incorporated expert feedback and iterative consensus rounds. Under these conditions, Gemini attained the highest numerical level of agreement among panelists, although no statistically significant differences were observed among the three models. While methodological differences may account for part of the variation, the most likely explanation lies in the rapid and ongoing updates of LLM architectures and training data, which can substantially alter model behavior within short periods. These findings underscore the importance of continuous validation, as performance cannot be assumed stable over time or comparable across versions, even within the same model family.

Taken together, these findings highlight the potential of LLMs as supportive tools in RCC, particularly in standard settings or resource-limited environments, where they can facilitate quick access to guideline-based recommendations and help reduce care disparities [22,23]. By enabling rapid and consistent access to clinical recommendations, they may help reduce variability in care. However, their integration into practice should be gradual, focusing first on low-risk tasks such as education, documentation or triage, and expanding only with robust evidence of safety and clinical utility [23,24]. Transparent reporting of model characteristics and independent validation remain essential, as does close collaboration between clinicians, AI developers and patients to ensure alignment with real-world needs [22,24]. To support safe implementation, clinical guidelines, especially in complex areas, should also adopt clear and structured language interpretable by both humans and AI systems.

This study has several limitations that should be considered when interpreting the results. First, it focused exclusively on RCC and included a relatively small number of predefined clinical scenarios, which may limit the generalizability of the findings to other tumor types or more diverse clinical contexts. Second, the evaluation was based on expert consensus through a Delphi process, a robust method, but one that remains inherently subjective and may be influenced by individual interpretation or the composition of the panel. Moreover, no predefined score-specific rubric was established for each individual point on the 9-point scale, which may have contributed to variability in panelists' interpretation of intermediate scores. In addition, only one response was generated per question and model. As the models were accessed through their user-facing platforms, temperature, top-p, top-k and other sampling parameters were not user-configurable or available for reporting and were left at the platform defaults. Repeated runs were not performed; therefore, response stability and variability across generations could not be assessed. Third, 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 taken into account when interpreting the results. Moreover, the comparison was based on only 25 matched questions, with variable numbers of expert raters across items. Accordingly, the absence of statistically significant differences should not be interpreted as evidence of equivalent performance among the models. Furthermore, the rapid pace of development and frequent updates in LLMs mean that their performance is dynamic and may evolve significantly after the evaluation period. Finally, only proprietary, user-facing LLMs were evaluated, and open-weight models such as Llama, Qwen and DeepSeek were not included. Because the performance of open-weight models may vary according to model size, quantization, hardware, inference framework and implementation parameters, the present findings may not be generalizable to these architectures. Future studies should directly compare proprietary and open-weight models under standardized deployment and prompting conditions.

Despite these limitations, the study also has important strengths. In particular, figures and tables from the clinical guideline were converted into structured, plain-language textual summaries to facilitate consistent interpretation by the AI models while preserving the original guideline content. The use of a multidisciplinary Delphi methodology further strengthened the clinical relevance of the evaluation, allowing for both quantitative consensus assessment and qualitative analysis of model responses.

Our results indicate that selected LLMs, when provided with structured, guideline-derived information, can generate responses that frequently align with multidisciplinary expert consensus. However, this controlled evaluation assessed their ability to interpret supplied guideline content and should not be considered evidence of independent clinical reasoning or real-world decision-making performance. Persistent limitations were observed in more complex settings where clinical judgment remains indispensable. While LLMs show potential as tools to support guideline interpretation, their outputs should be interpreted with appropriate caution and under expert oversight. To enhance their safe integration into clinical workflows, we recommend that urological oncology guidelines adopt clear and structured formats that are readily usable by both clinicians and AI systems, and that future research continues to refine evaluation methodologies and assess performance consistency as LLMs evolve.

Summary points

- Renal cell carcinoma (RCC) guidelines can be complex, and large language models may help clinicians interpret guideline-based information.
- This study evaluated Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini using 25 clinical questions derived from the 2025 European Association of Urology RCC guidelines.
- The original guideline content was provided to the models, with figures and tables converted into structured, plain-language text.
- Model responses were assessed by an 11-member multidisciplinary expert panel using a two-round modified Delphi process.
- Consensus in agreement was reached for 24 of 25 Gemini 2.5 Flash responses and 23 of 25 responses from both Claude 3.5 Haiku and GPT-4.1-mini.
- Median scores of at least 7 out of 9 were achieved for all Gemini 2.5 Flash responses, 24 Claude 3.5 Haiku responses and 23 GPT-4.1-mini responses.
- No response reached consensus in disagreement, and no statistically significant differences were found between the three models.
- Difficulties were more common in complex or individualized scenarios, including active surveillance in metastatic disease, chromophobe RCC prognosis, ablative treatment selection and alternatives to renal biopsy.
- Panelists identified occasional factual inaccuracies, omitted options and limited clinical nuance in some responses.
- The findings support the use of large language models as tools for interpreting supplied guideline information under expert supervision, rather than as autonomous clinical decision-makers.

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-2026-0082

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 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. European Association of Urology. EAU guidelines on renal cell carcinoma. (Accessed 11 July 2026). https://uroweb.org/guidelines/renal-cell-carcinoma
2. Ferlay J, Ervik M, Lam F *et al.* Kidney fact sheet. *Global Cancer Observatory: Cancer Today*. International Agency for Research on Cancer, Lyon, France (2024).
3. Padala SA, Barsouk A, Thandra KC *et al.* Epidemiology of renal cell carcinoma. *World J. Oncol.* 11(3), 79–87 (2020).
4. Cai T, Cocci A, Gacci M *et al.* Guidelines in urology: lights and shadows. *Urologia* 87(3), 125–129 (2020).
5. Smith EJ, MacLennan S, Bjartell A *et al.* Ensuring consistent European-wide urological care by the use of evidence-based clinical practice guidelines: can we do better. *Biomed. Hub.* 2(Suppl. 1), 162–168 (2017).
6. Frazer R, McGrane JM, Challapalli A *et al.* Real-world patterns of treatment and response in metastatic renal cell carcinoma: a multicentre UK-wide review with UK Renal Oncology Collaborative (UK ROC). *ESMO Real World Data Digit. Oncol.* 3, 100027 (2024).
7. Gajdács M. The importance of reporting clinical and epidemiological data in urology: local experiences and insights from the international literature. *Medicina (Kaunas)* 56(11), 581 (2020).
8. 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).
9. 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).
10. Li KP, 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).
11. Anthropic. Meet Claude. (Accessed 11 July 2026). https://www.anthropic.com/claude
12. Google. Google Gemini. (Accessed 11 July 2026). https://gemini.google.com
13. OpenAI. GPT-4o. (Accessed 11 July 2026). https://openai.com/es-ES/index/hello-gpt-4o/
14. 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).

●● **Provides a large expert-validated comparison of Claude, Gemini and ChatGPT across seven major urological cancers, using 10,500 specialist evaluations and directly assessing the effect of prompt formulation on guideline interpretation.**

15. Jones J, Hunter D. Consensus methods for medical and health services research. *BMJ* 311(7001), 376–380 (1995).
16. Fitch K, Bernstein SJ, Aguilar MD *et al. The RAND/UCLA appropriateness method user's manual*. RAND Corporation, CA, USA (2001).
17. Anthropic. Models overview. Claude Platform Docs. (Accessed 11 July 2026). https://platform.claude.com/docs/en/about-claude/models/overview
18. Google AI for Developers. Gemini 2.5 Flash. Gemini API. (Accessed 11 July 2026). https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash
19. OpenAI. GPT-4.1 mini model. OpenAI API. (Accessed 11 July 2026). https://developers.openai.com/api/docs/models/gpt-4.1-mini
20. Zitu MM, Le TD, Duong T *et al.* Large language models in cancer: potentials, risks, and safeguards. *BJR Artif. Intell.* 2(1), ubae019 (2025).
21. Bertolo R, De Bon L, Caudana F *et al.* Role of large language models in the multidisciplinary decision-making process for patients with renal cell carcinoma: a pilot experience. *NPJ Precis. Oncol.* 9(1), 257 (2025).

●● **Evaluates AI-generated diagnostic and treatment suggestions against multidisciplinary tumor board decisions in 103 renal cell carcinoma cases, showing moderate overall concordance and lower agreement in more advanced disease settings.**

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

● **Synthesizes evidence from 34 studies on the applications and performance of large language models in clinical oncology and highlights substantial heterogeneity related to model capabilities, prompting strategies, oncological domains and evaluation methods.**

23. 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).

24. 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).

25. 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 in a realistic clinical decision-making framework based on 2400 patient cases and identified major limitations in diagnosis, guideline adherence, laboratory interpretation, instruction following and workflow integration, showing that current models are not ready for autonomous clinical use.**

26. 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).

27. '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 better than the usual structured methodology? *J. Pediatr. Urol.* 21(6), 1643–1649 (2025).