npj | precision oncology

Brief communication

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-01014-4

Role of large language models in the multidisciplinary decision-making process for patients with renal cell carcinoma: a pilot experience

Riccardo Bertolo[1] ✉, Lorenzo De Bon[1], Filippo Caudana[1], Greta Pettenuzzo[1], Sarah Malandra[1], Chiara Casolani[1], Andrea Zivi[2], Emanuela Fantinel[2], Alessandro Borsato[3], Riccardo Negrelli[3], Emiliano Salah El Din Tantawy[4], Giulia Volpi[4], Matteo Brunelli[5], Alessandro Veccia[1], Maria Angela Cerruto[1], Alessandro Antonelli[1] & AOUI Verona Uro-Oncology Multi-Disciplinary Team*

We evaluated an AI chatbot's ability to suggest diagnostic and therapeutic pathways for renal cell carcinoma (RCC) in a multidisciplinary tumor board (MDT). A retrospective analysis of 103 cases (2023–2024) found 62.1% agreement with MDT decisions ($\kappa = 0.44$, $p < 0.001$). Concordance was highest in when follow-up imaging was suggested ($p = 0.001$), with disease status influencing agreement ($p = 0.004$). These results suggest AI could assist in RCC case assessments, warranting further research.

Multidisciplinary tumor boards (MDTs) are pivotal in determining optimal diagnostic and therapeutic strategies for oncology patients[1,2]. The increasing complexity of renal cell carcinoma (RCC) management, guided by evolving recommendations such as the 2024 EAU guidelines, necessitates an efficient and evidence-based approach[3]. With the recent advancements in generative AI, particularly in natural language processing, AI-driven decision support systems may offer potential benefits in streamlining case discussions and reducing variability in clinical decision-making[4,5].

Previous research has explored AI applications in oncology, particularly in radiology and pathology[6,7]; however, its role in clinical decision support remains underexplored. This study evaluates the capability of an advanced large language model (LLM) AI chatbot to provide RCC treatment recommendations in alignment with MDT decisions, intending to assess its potential utility in enhancing oncological workflow efficiency.

Methods

All RCC cases discussed by the institutional MDT were reviewed. For each case, a summarized clinical history—including age, sex, relevant imaging findings, tumor stage, and eventual treatment and/or diagnostic procedures (i.e., biopsy) received with histology—was input into the AI chatbot after removing all identifying patient data.

All interactions were conducted using the "GPT for Slides™ Docs™ Sheets™" add-on, configured with the OpenAI GPT-4.1 model (release o1, dated 2024-05-12) with a Temperature 0.30, Top-p 1.0, and 120k token context window. UTC timestamp of the first query was on January 24, 2025, 08:43 UTC.

To automate data processing, Google Sheets was linked to ChatGPT o1 (OpenAI, San Francisco, CA, USA) using the following procedure:

1. A Google (Google LLC, Mountain View, CA, USA) account was required to access Google Sheets.
2. Within Google Sheets, the "Add-ons" menu was accessed, and the option "Get add-ons" was selected.
3. The Google Workspace Marketplace opened, where multiple applications were available to link Google Sheets to ChatGPT. The selected tool was GPT for Slides™ Docs™ Sheets™ (Qualtir Technology, Roseville, CA, USA), and the necessary permissions were granted following on-screen instructions.
4. Once installed, the new add-on appeared under the "Add-ons" menu in Google Sheets and was activated for use.

The clinical cases were derived from multidisciplinary discussions held during 2023–2024 at our institution. Only patients presented for their first

[1]Department of Surgery, Dentistry, Pediatrics and Gynecology, Urology Unit, Azienda Ospedaliera Universitaria Integrata, University of Verona, Verona, Italy. [2]Department of Medicine, Section of Oncology, Azienda Ospedaliera Universitaria Integrata, Verona, Italy. [3]Unit of Radiology, Azienda Ospedaliera Universitaria Integrata, Verona, Italy. [4]Unit of Radiation Oncology, Azienda Ospedaliera Universitaria Integrata Verona, Verona, Italy. [5]Department of Diagnostic and Public Health, Section of Pathology, University of Verona, Verona, Italy.
✉e-mail: riccardogiuseppe.bertolo@univr.it

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

discussion were included to ensure that the AI's recommendations were compared against an unbiased MDT decision-making process. The medical team summarized each case in a concise text (30–50 words) following a standardized format. These summaries were entered into the first Column of a Google Sheets worksheet. To generate automated therapeutic and diagnostic suggestions based on the 2024 European Association of Urology (EAU) guidelines[3], a custom function was implemented in Google Sheets:

fx = GPT ("*Can you help us suggest the pathway according to the 2024 EAU guidelines for the following patients? Please give us only the first choice of the next step you would recommend based on the clinical scenario and the patient's age. You must use a maximum of 30 words*"). All relevant clinical factors—such as disease stage, prior treatments, comorbidities, and imaging findings—were incorporated within the free-text clinical summary provided to the LLM.

This function allowed the patient summaries from the first column to be processed automatically, generating a response from ChatGPT within seconds.

The generated outputs were compared to the official MDT recommendations, and concordance rates were analyzed after revision by a third party (A.A).

Statistical analysis

To compare the output of the human MDT with that of the AI chatbot in suggesting the diagnostic and/or therapeutic pathway for a series of patients, the following statistical methods were used: the Cohen's Kappa (κ) measured the level of agreement between the AI and MDT while adjusting for chance agreement, stratified by the clinical stage of the disease; the Fisher's exact test was applied after categorizing the AI and MDT recommendations into different decision-making settings to assess whether a significant difference existed in the distribution of suggestions. Multivariable logistic regression analysis was performed to identify factors predicting a greater discrepancy between AI and MDT recommendations.

A total of 103 RCC cases were included. The patients' demographics and clinical characteristics are summarized in Supplementary Table 1. The analysis of agreement between the AI chatbot and the MDT in suggesting the next diagnostic and/or therapeutic pathway showed an overall agreement of 62.1%, with an expected agreement of 32.6%, resulting in a Cohen's Kappa (κ) of 0.44 ($p < 0.001$), indicating moderate agreement.

Stratifying by disease stage, agreement was highest in the Nx/N0 M0 group (73.8% observed vs. 48.9% expected, $\kappa = 0.48$, $p < 0.001$), reflecting moderate agreement. In the Nx/N0 M+ subgroup, agreement was 60% observed vs. 28.9% expected, $\kappa = 0.44$, $p = 0.001$, also suggesting moderate agreement. Conversely, a lower agreement was observed in patients with N + M0 disease (45.4% observed vs. 28.1% expected, $\kappa = 0.24$, $p = 0.03$). It was particularly weak in the N + M+ subgroup (31.2% observed vs. 22.3% expected, $\kappa = 0.11$, $p = 0.09$), where no significant agreement was detected (data detailed in Table 1).

Significant differences were found in the categories of recommendations into different decision-making settings between the AI chatbot and the MDT ($p = 0.001$).

Higher discordance was found in cases where a biopsy was suggested. Lower discordance was noted in the cases where follow-up imaging was indicated (Table 2).

The multivariable analysis identified several factors influencing the concordance between the multidisciplinary team and the AI chatbot: ongoing systemic therapy showed a potential association with higher concordance, with an OR of 4.54 (95% CI: 0.82–25.05, $p = 0.08$). However, it did not reach statistical significance. Disease status had a notable impact on concordance: compared to patients with Nx/N0 M0 disease (reference category), those with both nodal and metastatic involvement (N + M +) had significantly lower odds of concordance (OR = 0.11, 95% CI: 0.03–0.5, $p = 0.004$). Conversely, patients with nodal involvement but no metastases (N + M0) showed a signal toward reduced concordance (OR = 0.26, 95% CI: 0.06–1.11, $p = 0.07$), though this did not reach statistical significance (Table 3).

Table 1 | Cohen's Kappa (κ) measuring the level of agreement between the LLM and the MDT while adjusting for chance agreement, stratified by the clinical stage of the disease

Patient's disease	Agreement	Expected agreement	Cohen's K	*p* value
Overall	62.1	32.6	0.44	<0.001
Nx/N0 M0	73.8	48.9	0.48	<0.001
N + M+	31.2	22.3	0.11	0.09
N + M0	45.4	28.1	0.24	0.03
Nx/N0 M+	60	28.9	0.44	0.001

N0 no regional lymph node involvement, *N+* lymph node metastasis present, *Nx* regional lymph nodes not assessed, *M0* no distant metastasis, *M+* distant metastasis present.

Table 2 | Fisher's exact test was applied after categorizing the AI and MDT recommendations into different decision-making settings to assess whether a significant difference existed in the distribution of suggestions

	No agreement (*N* = 31)	Agreement (*N* = 72)	*p* value
Biopsy	11 (35.5)	4 (5.6)	<0.001
Cht/IO	12 (38.7)	12 (16.7)	
Surgery	2 (6.4)	11 (15.3)	
Palliation	-	1 (1.4)	
Imaging/follow-up	6 (19.3)	44 (61.1)	

Cht Chemotherapy, *IO* Immunotherapy.

Table 3 | Multivariable logistic regression analysis performed to identify factors predicting a greater discrepancy between AI and MDT recommendations

Variables	OR	95% CI	*p* value
Age	1.01	0.97–1.05	0.6
Gender (Female)	1.58	0.58–4.29	0.4
Previous surgery	1.82	0.63–5.26	0.3
Chemotherapy	4.54	0.82–25.05	0.08
Disease			
Nx/N0 M0	Ref		
N + M+	0.11	0.03–0.5	0.004
N + M0	0.26	0.06–1.11	0.07
Nx/N0 M+	0.48	0.12–1.98	0.3

OR odds ratio, *CI* confidence interval, *N0* no regional lymph node involvement, *N+* lymph node metastasis present, *Nx* regional lymph nodes not assessed, *M0* no distant metastasis, *M+* distant metastasis present.

Our study demonstrates that AI-driven decision support systems have the potential to align with expert MDT decision-making in a proportion of RCC cases. The AI and MDT agreements varied across disease stages, with weaker agreements in more advanced disease settings. Disagreement was more common in cases where invasive diagnostic and therapeutic procedures were recommended instead of simple follow-up with imaging.

Previous researchers have recently published pilot experiences analogous to ours but in other fields. Most attempts have been published about breast cancer[8]. In an observational study, Griewing et al. compared the concordance of treatment recommendations from ChatGPT 3.5 with those of a breast cancer multidisciplinary tumor board. Overall concordance between the LLM and the MDT was reached for half of the patient profiles[9]. Sorin et al. asked the LLM to recommend the next most appropriate step in the management of their patients, providing the LLM with detailed patient

history as a basis for the decision. Recommendations of the LLM were retrospectively compared to the decisions by the MDT: in seven out of ten cases, LLM recommendations overlapped those by the MDT. The authors underlined that the LLM tended to overlook important patient information[10]. These results are very similar to what we observed in our experience. It is entirely understandable that discrepancies may arise between the verdicts from a LLM and those by an MDT. These discrepancies may stem from unique clinical presentations that are not sufficiently addressed by the guidelines, or from the fact that the LLM lacks full awareness of the patient's frailty status and cannot view and interpret radiological imaging, for example. This is why managing atypical cases presents the greatest room for improvement when aiming at integrating the workflow of an MDT with AI.

In colorectal cancer field, Choo et al. discussed colorectal cancer cases in the MDT board at a single tertiary institution. The treatment recommendations made by the LLM ChatGPT were analyzed to ensure adherence to oncological principles. The recommendations by LLM were compared with the decision plans made by the MDT. As a result, the oncological management recommendation concordance rate between the LLM and the MDT was 86.7%, which is very optimistic compared to what observed in our experience[11].

Lechien et al. evaluated ChatGPT-4 performance in oncological board decisions regarding 20 medical records of patients with head and neck cancer. GPT-4 was accurate in 13 cases (65%)[12].

In another field, Haemmerly et al. prompted ChatGPT with detailed patient histories to recommend treatments. Like the other reported experiences, the output by the LLM was evaluated by a rater, and inter-rater agreement was assessed. The performance of the LLM was poor at classifying glioma types, but good for recommending adjuvant treatments. Overall, expert agreement was moderate, as indicated by an intraclass correlation coefficient of 0.7[13].

It is clear, early experiences with testing LLMs in multidisciplinary decision-making for oncology patients are beginning to emerge. With some exceptions, such studies consistently report similar findings, with agreement rates around 60-70%. As expected, the ability of a LLM to replicate the verdict of a MDT varies depending on case complexity, with higher concordance observed in less intricate cases.

To our knowledge, ours is the first study to assess a generative AI model's ability to propose guideline-based MDT recommendations in the field of kidney cancer. However, we are still far from the day when an LLM could fully replace a human MDT. Machine learning algorithms trained with big data about decisions made by MDT teams could progressively improve the accuracy of AI.

While the study presents an original and timely concept, it is accompanied by several limitations and controversial aspects that should be carefully considered when interpreting the results.

As concerning methodology flaws, the study focused on a single, widely used LLM, and did not include a comparative evaluation across different generative models. While this choice was intentional for a pilot feasibility analysis, it limits the generalizability of our findings across the broader and rapidly evolving landscape of LLMs. Comparing different models or even the same model with varying hyperparameters would require careful consideration of factors such as prompt design, temperature, and random seeds, which can introduce significant variability in performance. Another limitation is the non-systematic approach to prompt design: the prompt was written in a straightforward and practical manner without formal testing or comparison against alternative formulations. Prompt engineering strategies—such as testing variants, iterative refinement, or formal validation—would be recommended. While case summaries provided to the LLM were generated using a standardized format, their internal consistency was not formally assessed; subtle variability in how clinical scenarios were framed may have influenced downstream comparisons—much like different angles can alter the perception of the same object. Future studies should consider quantifying this representational variability using methods such as cosine similarity[14] or Jaccard index[15], especially in settings where LLM outputs are highly sensitive to input phrasing. When evaluating LLM outputs and MDT decisions, textual similarity analyses using BLEU[16], ROUGE[17], and cosine metrics[14] (Supplementary Material, Supplementary Discussion 1) revealed limited lexical and structural overlap, highlighting the need for refined prompting and more semantically-aware evaluation methods; however, the reader should note that low scores do not necessarily indicate clinical inaccuracy and warrant qualitative case-by-case assessment, as was done in this study.

Regarding the flaws in data evaluated, the study focused exclusively on first-time MDT discussions to ensure unbiased comparisons, thereby excluding follow-up cases where decisions are often guided by prior therapeutic steps; while this enhances internal validity, it limits insights into scenarios where LLMs might eventually offer greater workflow support. In fact, it may be precisely the more straightforward setting of re-discussions—where clinical pathways are already partially defined—that represents the most immediate opportunity for meaningful LLM integration into MDT workflows. A major limitation is the lack of granular data on performance status, the absence of a systematic frailty assessment using geriatric scores, and the lack of standardized evaluation of comorbidities, which could have provided a more detailed and personalized influence on the decision-making process. These factors were included in the clinical case scenario only when considered essential for the LLM to make an informed decision. Additionally, the lack of direct integration of imaging or pathology data into the LLM workflow represents another drawback. Unlike human MDTs, which routinely base their decisions on direct visual inspection of radiologic and histologic images, the LLM relied solely on textual inputs. This introduces an asymmetry in the comparison, as critical diagnostic nuances may not be fully captured in narrative reports. That said, incorporating raw image data into general-purpose LLMs raises substantial ethical and cybersecurity concerns, including risks related to patient privacy and data protection, which currently preclude such integration in routine clinical research settings.

A further limitation lies in the absence of systematic follow-up data, which prevents a direct evaluation of the clinical impact of both MDT and AI-driven decisions. In fact, the real-world effectiveness of MDT recommendations is itself not always measurable, making any outcome-based comparison with AI inherently challenging and beyond the scope of this study.

Finally, the study lacks external validation, as all cases were drawn from a single institution. This limits the generalizability of our findings. Future research will focus on fine-tuning the model and conducting external validations to enhance applicability across broader settings.

With all these limitations in mind, in the context of optimizing patient care, our pilot experience suggests that LLMs could at least serve as a triage tool, helping to prioritize the most critical cases before discussion.

A continuous, exponential increase in the number of cases requiring discussion is expected in the next years, in alignment with modern clinical practices. Consider, for example, localized kidney cancer: many patients who, in the past, would have been treated exclusively with surgery—such as partial or radical nephrectomy—have now to be at least evaluated for counseling for adjuvant immunotherapy[18]. And this is just the beginning.

As for legal and regulatory considerations, we acknowledge that we are still far from a point where AI could ethically or legally replace human decision-making in high-stakes clinical contexts such as MDT discussions. We remark that this study should be interpreted as a pilot exploration of AI's supportive potential, not as an endorsement of autonomous, AI-driven care.

In conclusion, LLMs show promise as a support tool for RCC decision-making within an MDT framework, particularly for cases with lower complexity. While AI may not replace human expertise, it has the potential to optimize case discussions and improve workflow efficiency. Further validation studies and AI model enhancements will be essential to maximize its utility in real-world oncology settings.

Data availability

The dataset generated and analyzed during the current study is not publicly available due to [property of AOUI Verona] but is available from the corresponding author upon reasonable request.

Code availability

Trivial statistical codes were used for the analysis in this manuscript but are available upon request from the corresponding author.

Received: 20 March 2025; Accepted: 16 June 2025;
Published online: 24 July 2025

References

1. Specchia, M. L. et al. The impact of tumor board on cancer care: evidence from an umbrella review. *BMC Health Serv. Res.* **20**, 73 (2020).
2. Mano, M. S., Çitaku, F. T. & Barach, P. Implementing multidisciplinary tumor boards in oncology: a narrative review. *Future Oncol.* **18**, 375–384 (2022).
3. https://uroweb.org/guidelines/renal-cell-carcinoma - Accessed on March 14th (2025).
4. Carl, N. et al. Large language model use in clinical oncology. *NPJ Precis Oncol.* **8**, 240 (2024).
5. Benary, M. et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* **6**, e2343689 (2023).
6. Luchini, C., Pea, A. & Scarpa, A. Artificial intelligence in oncology: current applications and future perspectives. *Br. J. Cancer* **126**, 4–9 (2022).
7. Ahmad, Z., Rahim, S., Zubair, M. & Abdul-Ghafar, J. Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. *Diagn. Pathol.* **16**, 24 (2021).
8. Lukac, S. et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch. Gynecol. Obstet.* **308**, 1831–1844 (2023).
9. Griewing, S. et al. Challenging ChatGPT 3.5 in senology-an assessment of concordance with breast cancer tumor board decision making. *J. Pers. Med*. **13**, 1502 (2023).
10. Sorin, V. et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer* **9**, 44 (2023).
11. Choo, J. M. et al. Conversational artificial intelligence (chatGPT™) in the management of complex colorectal cancer patients: early experience. *ANZ J. Surg.* **94**, 356–361 (2024).
12. Lechien, J. R., Chiesa-Estomba, C. M., Baudouin, R. & Hans, S. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *Eur. Arch. Otorhinolaryngol.* **281**, 2105–2114 (2024).
13. Haemmerli, J. et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board?. *BMJ Health Care Inf.* **30**, e100775 (2023).
14. Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**, 513–523 (1988).
15. Manning C. D., Raghavan P. & Schütze H. *Introduction to Information Retrieval* (Cambridge University Press, 2008).
16. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics,* pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135
17. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* pp. 74–81 (Association for Computational Linguistics, 2004).
18. Campi, R. et al. Could a risk-adapted approach support shared decision-making regarding eligibility for adjuvant pembrolizumab for patients with clear cell renal cell carcinoma at high risk of recurrence? A multicentre cohort study. *Eur. Urol. Oncol*. **7**, 323–327 (2024).

Acknowledgements

This study received no funding.

Author contributions

R.B. conceived the stay, analyzed and interpreted the patient data, and wrote the manuscript. L.D., G.P., S.M., and C.C. took care of data collection. F.C. created the automated generative AI an took care of data collection. A.Z. and E.F. contributed with important intellectual content as medical oncologists and revised the manuscript. A.B. and R.N. contributed with important intellectual content as radiologists and revised the manuscript. E.S. and G.V. contributed with important intellectual content as radiation oncologists and revised the manuscript. M.B. contributed with important intellectual content as pathologist and revised the manuscript. A.V. performed the statistical analysis and revised the manuscript. M.C. and A.A. contributed with important intellectual content as urologists and revised the manuscript. A.A. supervised the project. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-01014-4.

Correspondence and requests for materials should be addressed to Riccardo Bertolo.