Wiley
European Journal of Cancer Care
Volume 2025, Article ID 9029196, 9 pages
https://doi.org/10.1155/ecc/9029196

WILEY

*Research Article*

# Exploring the Utility of ChatGPT-4o and DeepSeek-R1 for Multidisciplinary Uro-Oncology Tumor Boards: A Tumor Type–Specific Analysis

**Ertugrul Cakir**,[1] **Tumay Bekci**,[2] **Serdar Aslan**,[1] **Ural Oguz**,[3] **Ercan Ogreden**,[3] **Erhan Demirelli**,[3] **Dogan Sabri Tok**,[3] **Birgul Tok**,[4] **Sendag Yaslikaya**,[5] **Sevket Zorlu**,[6] **and Yahya Han Memis**[7]

[1]*Department of Radiology, Faculty of Medicine, Giresun University, Giresun, Türkiye*
[2]*Department of Radiology, Istanbul Medipol University, Istanbul, Türkiye*
[3]*Department of Urology, Faculty of Medicine, Giresun University, Giresun, Türkiye*
[4]*Department of Pathology, Faculty of Medicine, Giresun University, Giresun, Türkiye*
[5]*Department of Oncology, Faculty of Medicine, Giresun University, Giresun, Türkiye*
[6]*Department of Nuclear Medicine, Faculty of Medicine, Giresun University, Giresun, Türkiye*
[7]*Department of Radiation Oncology, Faculty of Medicine, Giresun University, Giresun, Türkiye*

Correspondence should be addressed to Ertugrul Cakir; drcakir@outlook.com

Received 26 July 2025; Revised 12 November 2025; Accepted 14 November 2025

Academic Editor: Keval Patel



**Introduction:** This study demonstrates the potential benefits of large language models (LLMs) such as ChatGPT-4o and DeepSeek-R1 in supporting multidisciplinary uro-oncologic tumor boards (MUTBs) discussing complex urological-oncological cases. In particular, DeepSeek-R1 has demonstrated better performance than ChatGPT-4o in specific types of urological malignancies and specific treatment protocols.
**Objective:** This study examined the agreement between the AI-based systems ChatGPT-4o and DeepSeek-R1 and the decision making of a MUTB, which is a team of experts who assess uro-oncological cases.
**Methods:** A total of 97 uro-oncological cases covering a period of one year were retrospectively evaluated. The treatment plans for the cases were determined using traditional tumor board methods and separately presented to ChatGPT-4o and DeepSeek-R1. Comprehensive medical records were created for each case, and the planned treatments were classified as surgical intervention, radiotherapy/chemotherapy (RT-CT), advanced imaging, or follow-up. Agreement between the board and AI decisions was analyzed using Cohen's kappa and with precision, recall, and F1 scores.
**Results:** In prostate cancer cases, DeepSeek-R1 showed substantial agreement ($\kappa = 0.787$, $p < 0.001$) with the board decisions, while ChatGPT-4o showed moderate agreement ($\kappa = 0.593$, $p < 0.001$). For renal malignancies, DeepSeek-R1 showed moderate agreement ($\kappa = 0.603$,$p < 0.001$), while ChatGPT-4o showed slight agreement ($\kappa = 0.172$, $p < 0.001$). In bladder cancer cases, DeepSeek-R1 showed moderate agreement ($\kappa = 0.604$, $p < 0.05$), but ChatGPT-4o did not show statistically significant agreement ($\kappa = 0.217$, $p > 0.05$). When all cases were considered together, DeepSeek-R1 showed high agreement, while ChatGPT-4o showed moderate agreement (DeepSeek-R1: $\kappa = 0.773$, $p < 0.001$; ChatGPT-4o: $\kappa = 0.608$, $p < 0.001$).
**Conclusion:** Our findings suggest that DeepSeek-R1's decision support performance is better than that of ChatGPT-4o in various oncological scenarios. The results highlight the potential for publicly available AI models to be used in complex clinical fields, such as oncology.

**Keywords:** artificial intelligence in clinical decision making; ChatGPT-4o; DeepSeek-R1; multidisciplinary uro-oncology tumor board; uro-oncology

## 1. Introduction

Multidisciplinary uro-oncology tumor boards (MUTBs) are considered the gold standard for managing patients affected by urologic and other malignancies. Healthcare professionals from various specialties collaborate to optimize multidisciplinary, patient-centered care and ensure adherence to clinical guidelines [1]. MUTBs have contributed significantly to changes in diagnostic and therapeutic strategies, improved staging accuracy and the adaptation of treatment to cancer risk, and ultimately increased survival rates. The European Partnership for Action Against Cancer, launched by the European Commission in 2009, approached multidisciplinary care from a policy perspective [2]. The positive outcomes from MUTBs depend on several factors, including qualified and effective teaching staff, proper preparation and selection of cases, the structure and format of meetings, levels of expertise, effective leadership, and the interactions among the participating physicians [3].

In recent years, artificial intelligence (AI)–based language models have emerged as significant tools in the medical field, demonstrating potential for supporting the decisions of all types of multidisciplinary tumor boards [4, 5]. Generative AI (GAI) has produced numerous advances in healthcare, and large language models (LLMs)—such as ChatGPT by OpenAI, Gemini by Google, and Copilot by Microsoft—now offer the potential for improving clinical decision support, medical education, and research. With the appearance of DeepSeek's DeepThink-R1 model—an open-source LLM introduced in January 2025—new opportunities and challenges have emerged for the integration of AI into healthcare and research. With careful implementation, ethical consideration, and international collaboration, LLMs like DeepSeek could promote innovation in healthcare and deliver cost-effective and scalable AI solutions while ensuring that human expertise remains central to patient care [6].

ChatGPT, a derivative of InstructGPT and part of the GPT-3.5 family, has been refined through human-curated responses within OpenAI's API playground. In May 2024, OpenAI launched GPT-4o, its next-generation flagship model. The "o" stands for "omni" and refers to its goal to enable more natural human–computer interaction. GPT-4o performs well not only in language and code processing but also in multimodal data interpretation in medical contexts, offering both robust response capabilities and cost efficiency. The introduction of this model represents a major leap in AI technology, offering stronger support and broader possibilities for diverse medical applications and benefiting both patients and clinicians [7]. Launched in January 2025, DeepSeek's new GAI model rapidly gained widespread adoption and became the most downloaded AI application. DeepSeek stands out due to its open-source availability under the MIT license, cost-effectiveness, and impressive reasoning capabilities enabled by technical features such as "mixture of experts," reinforcement learning, multistage training, and an FP8 mixed-precision environment [8].

Numerous studies have demonstrated the utility of AI systems in determining treatment regimens for various types of cancer [9–13]. Building on this growing body of evidence, the current study aims to evaluate whether the ChatGPT-4o and DeepSeek-R1 models can provide recommendations consistent with those of an MUTB in complex uro-oncological cases involving follow-up strategies, imaging assessments, surgical interventions, and chemotherapy (CT) or radiotherapy (RT) planning—and to determine which AI model has superior performance. To the best of our knowledge, this is among the first studies to directly compare the performance of ChatGPT-4o and DeepSeek-R1 in supporting decision making for MUTBs.

## 2. Materials and Methods

*2.1. Study Design and Patient Population.* A retrospective cohort study was conducted with 97 uro-oncological cases recorded at a university hospital between February 2024 and February 2025. Each case was thoroughly discussed by the MUTB, which made treatment planning decisions.

All uro-oncology patients evaluated by the MUTB within the specified time period were eligible for inclusion. Cases were excluded if electronic medical records (EMRs) were insufficient or incomplete, if the patient was under 18 years of age or pregnant, if the case involved nononcologic urological conditions despite being discussed by the board, or if the uro-oncological diagnosis was not pathologically confirmed. The final study cohort comprised all consecutive, eligible cases discussed by the MUTB over the 12-month period that met the inclusion and exclusion criteria. All cases were evaluated by a multidisciplinary team composed of at least one medical oncologist, radiation oncologist, nuclear medicine specialist, urologist, pathologists, and radiologist as part of the MUTB. A comprehensive medical record was compiled for each patient, and the recommended management strategy was categorized as one of the following: surgical intervention, RT-CT, need for advanced imaging, or follow-up.

This study was approved by our Institutional Review Board (IRB). All procedures involving human participants were conducted in accordance with the ethical standards of the institutional and national research committees, the 1964 Helsinki Declaration and subsequent amendments, and other applicable ethical standards.

*2.2. Coding of MUTB Recommendations.* The treatment modalities determined by our MUTB represent the optimal treatment strategy consensus reached by our specialist physicians, based on current guidelines and clinical evidence. These recommendations were coded and used as the gold standard for comparison with AI model outputs. It is acknowledged that in practical clinical practice, the final treatment modality is ultimately determined through shared decision making with the patient and their family. However, the specific aim of this study was to investigate the concordance between AI recommendations and the optimal, clinically intended treatment modality as determined by the MUTB, prior to patient-specific modifications.

Accordingly, the MUTB's final consensus recommendation was coded into one of four mutually exclusive categories for analysis: (1) surgery, (2) RT-CT, (3) new imaging studies, or (4) clinical follow-up. For localized diseases where the board's recommendation represented a treatment category (e.g., "definitive local therapy") rather than a specific modality, coding was based on therapeutic intent and the primary treatment modality discussed as standard care. For instance, a recommendation for curative-intent local therapy was coded as "surgery" if surgery was the predominant approach, or as "RT-CT" if RT was the focus. This approach acknowledges the inherent limitation in categorizing complex clinical decisions where final treatment selection often occurs through subsequent shared decision making with the patient.

*2.3. Data Collection and Case Presentation.* For each case, patient data were obtained from the EMRs maintained by the residents, fellows, and assistants working in the relevant departments, who ensure the accuracy and completeness of the records before presenting the cases to the MUTB for discussion. Although the full EMRs contained extensive clinical data, the information presented to the MUTB members—and to the AI models—was limited to essential variables to ensure consistency across cases and to protect patient confidentiality. Cases with incomplete EMRs were excluded both from board discussions and the current study.

The cases that were discussed by the MUTB were carefully restructured and formulated to allow optimal interpretation by ChatGPT-4o and DeepSeek-R1. To ensure consistency, standardized prompts were developed and submitted to both AI models. For each case, clinical information was anonymized and formatted into a structured prompt that included

- Patient demographics (age and sex),
- Current or most recent physical examination findings (e.g., digital rectal examination, general inspection, and costovertebral angle tenderness),
- Primary diagnosis and staging,
- Histopathological data and key laboratory values (e.g., complete blood count and tumor markers),
- Imaging findings (e.g., CT, MRI, and PET-CT),
- Treatment history (e.g., previous surgeries and CT regimens), and
- Comorbidities.

The cases were grouped based on underlying pathology. Each group was evaluated with prompts tailored to the specific clinical features of their respective pathologies and submitted to the AI models. Four standardized clinical decision questions were posed for each case:

- Does the patient currently require surgery (or would you recommend it)?
- Does the patient currently require any CT or RT?
- Does the patient currently require any additional imaging studies?
- Should the patient be managed under surveillance?

This standardization was intended to minimize response variability between the AI models and to facilitate statistical evaluation. The case data were formatted to be easily interpretable by AI: Laboratory results were expressed as numerical values; imaging findings were summarized by radiologists; histopathological data were described in condensed form by pathologists; and prior treatments and physical examinations were included as text-based entries (Figures 1 and 2).

For each clinical decision point, the AI models provided binary responses (yes/no) and brief explanations of the recommended treatment. These responses were analyzed for agreement with the MUTB decisions.

*2.4. Statistical Analysis.* Statistical analyses were performed using IBM SPSS Statistics for Mac, Version 27.0 (IBM Corp., Armonk, NY, USA), and all data were summarized as means and standard deviations (SDs). The primary statistical approach employed a binary classification framework to ensure standardized comparison between AI model outputs and MUTB decisions. While this methodological choice simplifies the nuanced nature of clinical decision making, it was intentionally selected to establish a foundational, reproducible analysis paradigm for this emerging field of research. The binary framework allowed for clear interpretation of Cohen's kappa and classification metrics (precision, recall, and F1 score), providing an essential baseline for initial model validation. Inter-rater agreement between the AI models and the MUTB was primarily assessed using Cohen's kappa ($\kappa$) coefficient, stratified by malignancy. For granular performance evaluation, additional classification metrics were calculated for each decision category (surgery, RT/CT, imaging, and surveillance):

- Precision (positive predictive value): \(\frac{TP}{TP + FP} \)
- Recall (sensitivity): \(\frac{TP}{TP + FN} \)
- F1 score: Harmonic mean of precision and recall: \(2 \times \frac{Precision \times Recall}{Precision + Recall} \)

TP = true positives, FP = false positives, and FN = false negatives based on binary (true/false) concordance with MUTB decisions.

Analyses of the pathology groups were performed to evaluate model recommendations for specific cancer types. Finally, in cases in which both AI models and the MUTB concurred on the clinical decision, further analyses were conducted to explore treatment approach variations between different uro-oncologic malignancies. Statistical significance was set at $p \leq 0.05$.

## 3. Results

A total of 97 uro-oncologic patients were included in the study—90 male and seven female. The gender imbalance is primarily attributable to the high number of prostate cancer

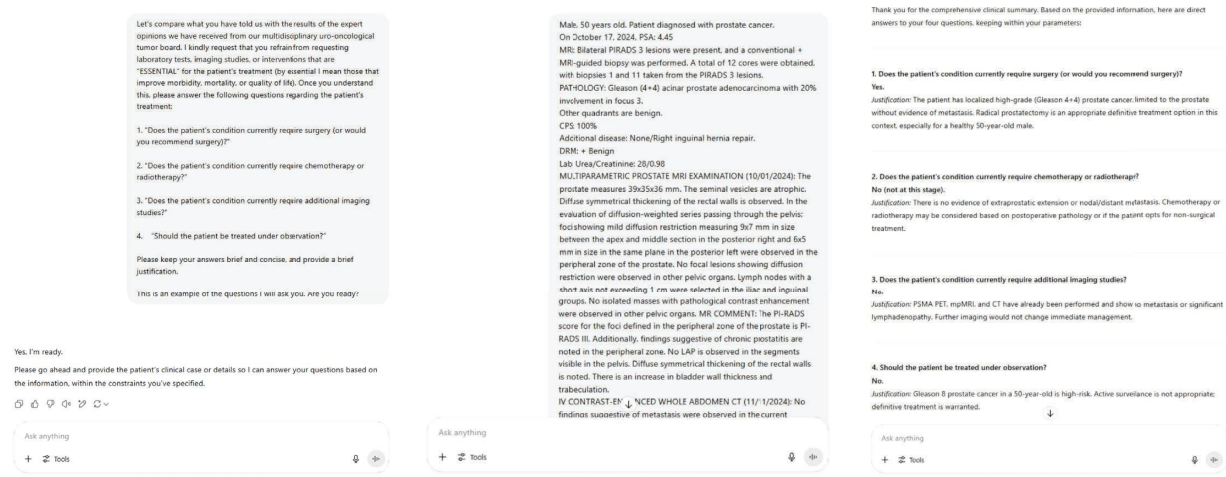


FIGURE 1: Workflow overview of ChatGPT-4o in clinical decision making. This figure presents the clinical reasoning process of ChatGPT-4o. On the left are the standardized initial instructions provided to the model, followed by sample electronic medical record (EMR) data in the center. On the right, the diagnostic and management recommendations generated by ChatGPT-4o are displayed.

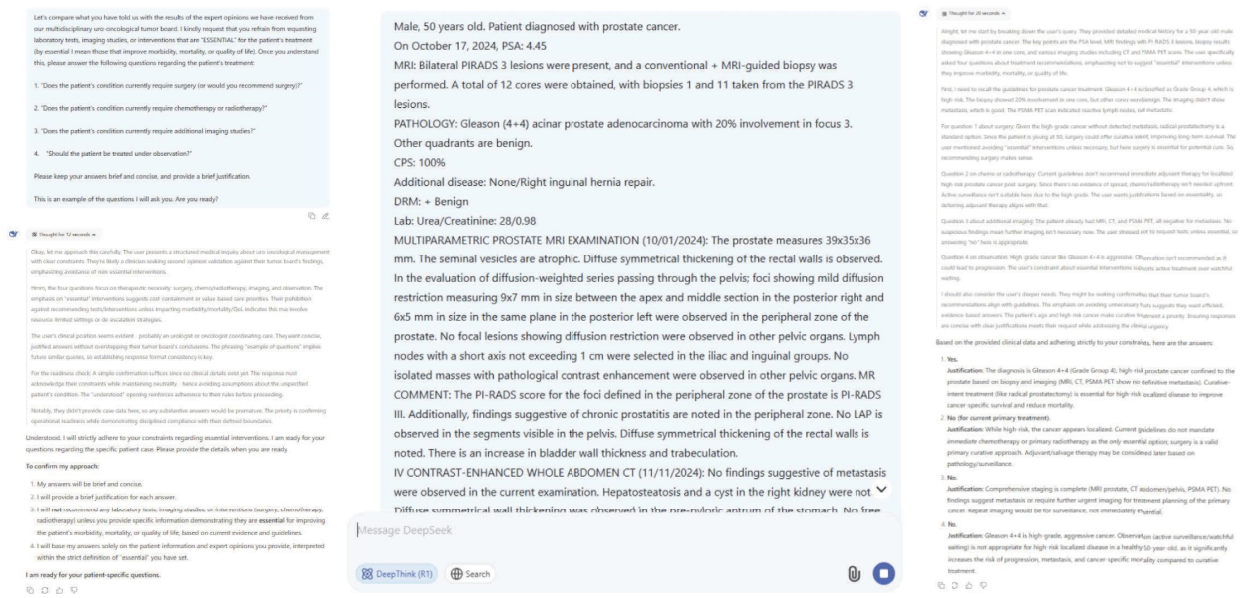


FIGURE 2: Workflow overview of DeepSeek-R1 in clinical decision making. This figure illustrates the decision-making process of DeepSeek-R1 when provided with the same clinical input as ChatGPT-4o. The structure mirrors that of Figure 1: initial instructions are shown on the left, patient EMR data in the center, and the model's clinical recommendations on the right.

cases. The mean age of the sample was 67 years (SD = 12); 58 patients had prostate cancer (mean age: 69, SD = 8), 18 had renal malignancies (mean age: 65, SD = 14), 12 had bladder cancer (mean age: 68, SD = 12), five had adrenal gland malignancies (mean age: 42, SD = 17), and four had ureteral malignancies (mean age: 65, SD = 12). The demographic characteristics of the patient subgroups and the distribution of their malignancies are summarized in Table 1.

Inter-rater agreement was assessed by malignancy type. In prostate cancer cases, the DeepSeek-R1 model exhibited a high level of agreement with the MUTB decisions ($\kappa = 0.787$, $p < 0.001$), whereas ChatGPT-4o showed only moderate agreement ($\kappa = 0.593$, $p < 0.001$). For renal malignancies, DeepSeek-R1 exhibited moderate agreement ($\kappa = 0.603$, $p < 0.001$), while ChatGPT-4o showed only low agreement ($\kappa = 0.172$, $p < 0.001$). In bladder cancer cases, DeepSeek-R1 exhibited moderate agreement ($\kappa = 0.604$, $p < 0.05$), while ChatGPT-4o did not show statistically significant agreement ($\kappa = 0.217$, $p > 0.05$). For adrenal gland malignancies, both AI models demonstrated a high level of agreement ($\kappa = 0.643$, $p < 0.05$ for both). Due to the limited number of cases, a meaningful statistical evaluation could not be performed for ureteral and adrenal gland malignancies.

Across all cases, the DeepSeek-R1 model exhibited high overall inter-rater agreement with the MUTB ($\kappa = 0.773$,

Table 1: Demographic characteristics of patient subgroups included in the study and distribution of malignancy subtypes.

| | *n* (%) | Sex (M/F) | Age mean (SD) | Primary diagnosis (*n*) |
|---|---|---|---|---|
| Prostate carcinoma | 58 (%59.8) | 58/0 | 69 (±8) | Prostatic adenocarcinoma (58)<br>G (3 + 3) (18)<br>G (3 + 4) (16)<br>G (4 + 3) (5)<br>G (4 + 4) (13)<br>G (4 + 5) (3)<br>G (5 + 4) (1)<br>G (5 + 5) (2) |
| Renal malignancies | 18 (%18.5) | 13/5 | 68 (±14) | Chromophobe RCC (2)<br>Clear cell renal carcinoma (8)<br>Papillary RCC (4)<br>Bosniak type III cyst (2)<br>Bosniak type IV cyst (2) |
| Bladder malignancies | 12 (%12.3) | 11/1 | 68 (±12) | HGPUC (6)<br>LGPUC (3)<br>Squamous cell bladder cancer (3) |
| Adrenal gland malignancies | 5 (%5.2) | 5/0 | 42 (±17) | Metastasis (3)<br>Lung metastasis (2)<br>Colorectal carcinoma metastasis (1)<br>Pheochromocytoma (2) |
| Ureteral malignancies | 4 (%4.2) | 3/1 | 65 (±12) | Transitional cell carcinoma (4) |
| Total | 97 (100) | 90/7 | 67.12 (±12) | 97 |

*Note:* M: male; F: female; G: Gleason score.
Abbreviations: HGPUC, high-grade papillary urothelial carcinoma; LGPUC, low-grade papillary urothelial carcinoma; RCC, renal cell carcinoma; SD, standard deviation.

$p < 0.001$), while the ChatGPT-4o model showed only moderate agreement ($\kappa = 0.608$, $p < 0.001$) (Figure 3). The levels of agreement between ChatGPT-4o and DeepSeek-R1 for significant oncological malignancies are shown in Table 2.

The comparative analysis revealed distinct F1 scores for ChatGPT-4o and DeepSeek-R1 across urological malignancy decision protocols. For prostate cancer, ChatGPT-4o achieved F1 scores of 0.84 (surgical treatment), 0.72 (RT/CT decision), 0.58 (new imaging studies), and 0.40 (clinical follow-up), while DeepSeek-R1 scored 0.92, 0.90, 0.66, and 0.57, respectively. In renal malignancies, ChatGPT-4o scored 0.82 (surgical treatment), 0.00 (RT/CT decision), 0.79 (new imaging studies), and 0.75 (clinical follow-up) versus DeepSeek-R1's 0.80, 0.00, 1.00, and 0.40. For bladder malignancies, ChatGPT-4o yielded 0.66 (surgical treatment), 0.54 (RT/CT decision), 0.00 (new imaging studies), and 0.00 (clinical follow-up) compared to DeepSeek's 0.75, 0.83, 0.00, and 0.00. Aggregated across all malignancies, ChatGPT-4o demonstrated scores of 0.81 (surgical treatment), 0.68 (RT/CT decision), 0.64 (new imaging studies), and 0.63 (clinical follow-up), while DeepSeek-R1 attained 0.88, 0.89, 0.79, and 0.63, respectively (Table 3).

## 4. Discussion

This study evaluated the agreement of the ChatGPT-4o and DeepSeek-R1 AI models with the decisions of the MUTB. Although both models provided recommendations consistent with MUTB decisions for specific cancer types, the overall agreement level of DeepSeek-R1 was higher than that of ChatGPT-4o, although the performance of both AI models varied by tumor type.

Our findings contribute to a growing body of literature evaluating LLMs in clinical oncology. Recent studies by Hernández-Flores et al. [4] demonstrated varying performance of ChatGPT and Gemini across challenging oncologic cases, echoing our observation that AI performance is context-dependent. Similarly, Lukac et al. [5] reported ChatGPT's utility as an adjunct in breast cancer tumor boards, noting its strength in providing guideline-based recommendations—a finding consistent with our models' performance in prostate cancer, where treatment pathways are well established. However, our study extends this literature by providing a direct comparison between two of the most advanced, recently released models—ChatGPT-4o and DeepSeek-R1—within a specialized uro-oncological context.

DeepSeek-R1 had a high level of agreement in prostate cancer cases, indicating reliable performance in this domain, while ChatGPT-4o was only moderate; this suggests that the general-purpose and multimodal architecture of GPT-4o might not be enough to help with clinical decision making for complex and diverse cancers. Nonetheless, both models achieved statistically significant results.

In renal malignancies, the moderate agreement of DeepSeek-R1 is noteworthy because it indicates a reasonable analytical capacity despite the complexity of the anatomical and prognostic variables. Conversely, the low agreement of ChatGPT-4o suggests that it struggles to adapt to highly individualized and heterogeneous clinical scenarios.

In the context of bladder cancer, only DeepSeek-R1 demonstrated statistically significant agreement,

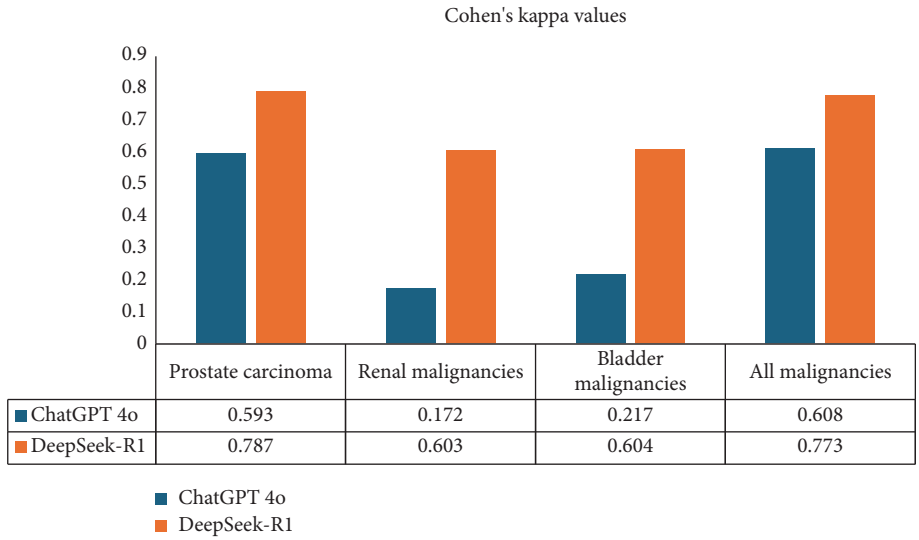


FIGURE 3: Comparison of the recommendations provided by the MUTB with those generated by ChatGPT-4o and DeepSeek-R1 across four clinical decision points: prostate carcinoma, renal malignancies, bladder malignancies, and all malignancies combined. For each decision point, Cohen's kappa ($\kappa$) values are presented to demonstrate the strength and significance of inter-rater agreement. MUTB: multidisciplinary uro-oncologic tumor board.

TABLE 2: Comparison of recommendation consistency between ChatGPT-4o and DeepSeek-R1 at important oncological decision points.

| | Chat GPT 4o | | | DeepSeek-R1 | | |
|---|---|---|---|---|---|---|
| | **Kappa value ($\kappa$)** | **Agreement level** | ***p* value** | **Kappa value ($\kappa$)** | **Agreement level** | ***p* value** |
| Prostate carcinoma | **0.593** | Moderate | < **0.001** | **0.787** | Substantial | < **0.001** |
| Renal malignancies | **0.172** | Slight | < **0.001** | **0.603** | Moderate | < **0.001** |
| Bladder malignancies | 0.217 | Fair | > 0.05 | **0.604** | Moderate | < **0.05** |
| All malignancies | **0.608** | Moderate | < **0.001** | **0.773** | Substantial | < **0.001** |

*Note:* The strength and importance of agreement between AI models related to MUTB is shown by kappa values ($\kappa$) and $p$ values for inter-rater agreement. $p < 0.05$ is considered statistically significant. Statistically significant values are highlighted in bold.Four different categories were evaluated: prostate cancer, renal malignancies, bladder malignancies, and all malignancies. Kappa ($\kappa$) and $p$ values indicating the level of agreement between models were presented for each category. This analysis aims to evaluate the consistency of AI-based clinical decision support systems.

suggesting that it can provide more consistent interpretations based on variables such as tumor invasion depth, cystoscopic findings, and responses to previous treatments. The failure of ChatGPT-4o to achieve significant agreement in this patient group raises concerns about the model's capacity to interpret the nuanced clinical characteristics unique to bladder cancer.

Both models demonstrated a high level of agreement for adrenal gland malignancies, but due to the limited number of cases, these statistically significant findings should be interpreted with caution; larger cohort studies are necessary to validate them. For ureteral malignancies, meaningful statistical analysis could not be conducted due to the small sample size; evaluating model performance in this subgroup will also require large-scale investigations.

When considering the overall patient cohort, DeepSeek-R1 achieved a high level of agreement, while ChatGPT-4o demonstrated a moderate-to-high level, suggesting that DeepSeek-R1 offers more refined and consistent performance in a clinical context, which is potentially attributable to its open-source architecture and training on domain-specific medical data. ChatGPT-4o, in contrast, may be more useful in general medical applications due to its multimodal capabilities and user-friendly design, while in scenarios requiring interdisciplinary and highly specialized clinical expertise—such as those in this study—it may perform worse. The observed variations in AI performance across the uro-oncologic tumor groups may also reflect the degree of specialization within each model or the nature of the training datasets used. For example, a model may offer more consistent recommendations for diseases such as prostate cancer, for which guidelines are well defined and standardized, while producing less reliable or more generalized outputs for malignancies such as bladder cancer, which present with complex and heterogeneous clinical variations.

The F1 score analysis (see Table 3) provides detailed insights into the models' performance across clinical decision domains. In prostate cancer, DeepSeek-R1 exhibited superior precision–recall balance for surgery and RT/CT, outperforming ChatGPT-4o. This is consistent with the $\kappa$ values, underscoring the reliability of DeepSeek-R1 for treatment-intent recommendations. However, both models

Table 3: Performance metrics (precision, recall, and F1 scores) for each malignancy type, calculated according to the clinical treatment protocols applied during AI model evaluation.

| | | Prostate carcinoma | | | Renal malignancies | | | Bladder malignancies | | | All malignancies | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Precision** | **Recall** | **F1 score** | **Precision** | **Recall** | **F1 score** | **Precision** | **Recall** | **F1 score** | **Precision** | **Recall** | **F1 score** |
| Surgery | ChatGPT-4o | 0.94 | 0.77 | 0.84 | 0.9 | 0.75 | 0.82 | 0.6 | 0.75 | 0.66 | 0.88 | 0.75 | 0.81 |
| | DeepSeek-R1 | 1 | 0.86 | 0.92 | 0.8 | 1 | 0.8 | 0.75 | 0.75 | 0.75 | 0.87 | 0.9 | 0.88 |
| RT/CT | ChatGPT-4o | 0.78 | 0.69 | 0.72 | 0 | 0 | 0 | 0.5 | 0.6 | 0.54 | 0.7 | 0.67 | 0.68 |
| | DeepSeek-R1 | 0.88 | 0.92 | 0.9 | 0 | 0 | 0 | 0.71 | 1 | 0.83 | 0.85 | 0.93 | 0.89 |
| New imaging studies | ChatGPT-4o | 0.55 | 0.62 | 0.58 | 0.66 | 1 | 0.79 | 0 | 0 | 0 | 0.61 | 0.68 | 0.64 |
| | DeepSeek-R1 | 0.71 | 0.62 | 0.66 | 1 | 1 | 1 | 0 | 0 | 0 | 0.84 | 0.68 | 0.79 |
| Clinical follow-up | ChatGPT-4o | 0.25 | 1 | 0.4 | 0.75 | 0.75 | 0.75 | 0 | 0 | 0 | 0.53 | 0.8 | 0.63 |
| | DeepSeek-R1 | 0.4 | 1 | 0.57 | 1 | 0.25 | 0.4 | 0 | 0 | 0 | 0.66 | 0.6 | 0.63 |

showed limited effectiveness in surveillance decisions, suggesting that there are challenges when assessing indolent disease trajectories.

In renal cancer, both models showed zero agreement with the MUTB in RT/CT recommendations, but DeepSeek-R1 achieved perfect agreement for imaging, suggesting contextual adaptability despite a moderate $\kappa$ value. For bladder cancer, critical deficiencies emerged: ChatGPT-4o yielded zero agreement with the MUTB for imaging and surveillance, and DeepSeek-R1 showed no agreement for imaging, indicating limitations in interpreting nuanced cystoscopic and histopathological features.

Aggregate F1 metrics support DeepSeek-R1's stronger performance in surgery and RT/CT across malignancies. Nonetheless, the suboptimal performance in surveillance decisions for both AI systems implies that current AI models cannot replace clinician judgment in follow-up planning, in which psychosocial factors and longitudinal assessment are central.

This study has several limitations. First, a potential limitation of this study is the heterogeneity of the cohort, which encompasses diverse urological malignancies with distinct treatment paradigms. This variability may limit the direct comparability of the AI models' performance across different cancer subtypes. Second, its single-center, retrospective design may limit generalizability; third, although the data provided to both AI models were standardized to a degree, the models did not interact with patients in real time; and fourth, the F1 scores revealed weaknesses in the model regarding certain decisions (e.g., kidney RT/CT), but no root-cause analysis of errors has been performed. Additionally, our binary classification framework, while methodologically necessary for standardization, simplifies the nuanced decision-making process characteristic of multidisciplinary tumor boards. Future studies should implement multilevel statistical approaches and integrate explainable AI techniques through prospective, interactive, and multicenter designs to better capture clinical complexity. Finally, further research and regulatory efforts considering ethics, legality, and patient safety will be required before AI technologies can be integrated into routine clinical practice.

## 5. Conclusions

The DeepSeek-R1 model demonstrated a high level of agreement with MUTB decisions, offering particularly consistent recommendations for more common malignancies, such as prostate cancer. ChatGPT-4o, while generally showing moderate-to-high agreement, was less consistent for certain malignancies, including kidney and bladder cancers. While DeepSeek-R1 performed well in treatment-oriented decisions (surgery and RT-CT), both models require improvement for surveillance and imaging guidance—particularly in bladder cancer.

Larger, multicenter studies are needed to confirm these findings before broader generalizations can be made. It is also crucial that AI systems are employed ethically and responsibly—not as replacements for human expertise but as support tools to aid clinical decision making. In this context, integrating GAI models into tumor board processes may speed up decision making and contribute to more personalized, case-specific treatment plans.

## Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

## Disclosure

All authors have reviewed and approved the final manuscript.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Author Contributions

Ertugrul Cakir and Tumay Bekci: writing–original draft, methodology, formal analysis, data curation, conceptualization, investigation, project administration, supervision, validation, and visualization.

Serdar Aslan, Ural Oguz, Ercan Ogreden, Erhan Demirelli, Dogan Sabri Tok, Birgul Tok, Sendag Yaslikaya, Sevket Zorlu, and Yahya Han Memis: data curation, supervision, and validation. These authors contributed equally to this work.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Acknowledgments

The authors used the AI language models ChatGPT-4o, DeepSeek-R1, and Gemini in preparing this work to improve readability and language. The authors then reviewed and edited the content as needed and take full responsibility for its publication.

## References

[1] M. R. Valerio, V. Serretta, D. Arico, et al., "A Prospective Observational Study on the Structuring Process and Implementation of a Large Regional, Inter-Hospital, Virtual Multidisciplinary Tumor Board on Prostate Cancer," *Anticancer Research* 43, no. 1 (2023): 501–508, https://doi.org/10.21873/anticanres.16187.

[2] J. M. Borras, T. Albreht, R. Audisio, et al., "Policy Statement on Multidisciplinary Cancer Care," *European Journal of Cancer* 50, no. 3 (2014): 475–480, https://doi.org/10.1016/j.ejca.2013.11.012.

[3] N. S. El Saghir, N. L. Keating, R. W. Carlson, K. E. Khoury, and L. Fallowfield, "Tumor Boards: Optimizing the Structure and Improving Efficiency of Multidisciplinary Management of Patients With Cancer Worldwide," *American Society of Clinical Oncology* no. 34 (2014): e461–e466, https://doi.org/10.14694/edbook_am.2014.34.e461.

[4] L. A. Hernández-Flores, J. B. López-Martínez, J. J. Rosales-de-la-Rosa, D. Aillaud-De-Uriarte, S. Contreras-Garduño, and R. Cortés-González, "Assessment of Challenging Oncologic Cases: A Comparative Analysis Between Chatgpt, Gemini, and a Multidisciplinary Tumor Board," *Journal of Surgical Oncology* (2025): https://doi.org/10.1002/jso.28121.

[5] S. Lukac, D. Dayan, V. Fink, et al., "Evaluating Chatgpt as an Adjunct for the Multidisciplinary Tumor Board Decision-Making in Primary Breast Cancer Cases," *Archives of Gynecology and Obstetrics* 308, no. 6 (2023): 1831–1844, https://doi.org/10.1007/s00404-023-07130-5.

[6] A. Temsah, K. Alhasan, I. Altamimi, et al., "Deepseek in Healthcare: Revealing Opportunities and Steering Challenges of a New Open-Source Artificial Intelligence Frontier," *Cureus* 17, no. 2 (2025): e79221, https://doi.org/10.7759/cureus.79221.

[7] N. Zhang, Z. Sun, Y. Xie, H. Wu, and C. Li, "The Latest Version Chatgpt Powered by GPT-4o: What Will It Bring to the Medical Field?" *International Journal of Surgery* 110, no. 9 (2024): 6018–6019, https://doi.org/10.1097/js9.0000000000001754.

[8] Y. Peng, B. A. Malin, J. F. Rousseau, et al., "From GPT to Deepseek: Significant Gaps Remain in Realizing AI in Healthcare," *Journal of Biomedical Informatics* 163, no. 104791 (2025): 104791, https://doi.org/10.1016/j.jbi.2025.104791.

[9] C. Zhang, J. Xu, R. Tang, et al., "Novel Research and Future Prospects of Artificial Intelligence in Cancer Diagnosis and Treatment," *Journal of Hematology & Oncology* 16, no. 1 (2023): 114, https://doi.org/10.1186/s13045-023-01514-5.

[10] M. A. Fink, A. Bischoff, C. A. Fink, et al., "Potential of Chatgpt and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer," *Radiology* 308, no. 3 (2023): e231362, https://doi.org/10.1148/radiol.231362.

[11] X. Liu, J. Shi, Z. Li, Y. Huang, Z. Zhang, and C. Zhang, "The Present and Future of Artificial Intelligence in Urological Cancer," *Journal of Clinical Medicine* 12, no. 15 (2023): 4995, https://doi.org/10.3390/jcm12154995.

[12] R. Rasmussen, T. Sanford, A. V. Parwani, and I. Pedrosa, "Artificial Intelligence in Kidney Cancer," *American Society of Clinical Oncology* 42, no. 42 (2022): 1–11, https://doi.org/10.1200/edbk_350862.

[13] P. Azadi Moghadam, A. Bashashati, and S. L. Goldenberg, "Artificial Intelligence and Pathomics," *Urologic Clinics of North America* 51, no. 1 (2024): 15–26, https://doi.org/10.1016/j.ucl.2023.06.001.