Cureus
Part of SPRINGER NATURE

Open Access Original Article

Agreement Between ChatGPT and Multidisciplinary Tumor Board Treatment Recommendations in Localized Muscle-Invasive Bladder Cancer: A Retrospective Exploratory Study

Review began 11/01/2026
Review ended 11/22/2026
Published 11/25/2026

© Copyright 2026
Felício et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.120263

Eduardo Felício [1], Guilherme Bernardo [1], Filipe Gaboleiro [1], André Pita [1], Diogo Carmali [1], Diogo Saraiva [1], Joao Varregoso [1], Fernando Ferrito [1]

1. Urology, Unidade Local de Saúde de Amadora/Sintra (ULS Amadora/Sintra), Lisbon, PRT

Corresponding author: Eduardo Felício, felicio.eduardo@gmail.com

Abstract

Background: Large language models (LLMs) are increasingly being explored for oncology decision support, but their ability to reproduce real-world multidisciplinary treatment decisions in muscle-invasive bladder cancer (MIBC) remains uncertain. We evaluated categorical agreement between ChatGPT (OpenAI, San Francisco, CA) and a multidisciplinary tumor board (MDT) and explored whether agreement differed across treatment modalities.

Methods: In this retrospective exploratory single-center study, 68 MDT-discussed muscle-invasive bladder cancer cases from April 2025 to April 2026 were reviewed; 24 consecutive initial treatment planning cT2-T4a N0M0 cases were included. Anonymized MDT vignettes were submitted to ChatGPT-5.5, which selected one of four predefined categories: upfront radical cystectomy, systemic therapy followed by radical cystectomy, bladder preservation/radiotherapy/trimodality treatment, or palliative/best supportive care. Agreement was assessed using observed agreement and unweighted Cohen's kappa. When available, treating physician and/or patient/family contextual input was subsequently added as an exploratory reassessment.

Results: Initial ChatGPT recommendations agreed with MDT decisions in 15 of 24 cases (62.5%, 95% confidence interval (CI): 42.7%-78.8%; kappa = 0.370). Agreement varied by MDT treatment category: systemic therapy followed by radical cystectomy, 11 of 13 cases; bladder preservation/radiotherapy/trimodality treatment, 2 of 4; upfront radical cystectomy, 1 of 6; and palliative/best supportive care, 1 of 1. Contextual input was available in 11 cases and changed the ChatGPT category in 4; final agreement was 16 of 24 cases (66.7%, 95% CI: 46.7%-82.0%; kappa = 0.447), representing one additional concordant case.

Conclusions: ChatGPT reproduced approximately two-thirds of real-world MDT treatment categories in this small localized muscle-invasive bladder cancer cohort, but agreement was uneven across treatment modalities. The model aligned most consistently with systemic therapy followed by radical cystectomy and least consistently with upfront radical cystectomy. These findings are exploratory and support larger studies of reproducibility, prompt dependence, and safety before clinical decision-support use.

Categories: Urology, Oncology, Healthcare Technology
Keywords: artificial intelligence (ai), chatgpt, large language models (llms), multidisciplinary tumor board meeting, muscle-invasive bladder cancer

Introduction

Large language models (LLMs) have rapidly entered medical research and clinical experimentation because they can synthesize complex text and generate context-sensitive responses. In oncology, proposed applications include patient communication, information retrieval, clinical documentation, and decision support [1,2]. However, LLM outputs can be incomplete, outdated, or incorrect, and their performance varies across models, prompts, and clinical contexts [1-3].

Several studies have compared LLM-generated recommendations with multidisciplinary tumor board (MDT) decisions. Reported agreement has varied substantially across cancer types and study designs [3]. Recent examples include colorectal [4], breast [5], head and neck [6], renal [7], and prostate cancer decision-support settings [8]. Evidence in bladder cancer remains limited, despite the complexity of treatment selection in localized muscle-invasive bladder cancer (MIBC).

Localized MIBC requires integration of oncological risk, patient fitness, comorbidities, renal function and

How to cite this article
Felício E, Bernardo G, Gaboleiro F, et al. (November 25, 2026) Agreement Between ChatGPT and Multidisciplinary Tumor Board Treatment Recommendations in Localized Muscle-Invasive Bladder Cancer: A Retrospective Exploratory Study. Cureus 18(11): e120263. DOI 10.7759/cureus.120263

systemic therapy eligibility, local treatment feasibility, and patient preferences. Curative pathways may include radical cystectomy, perioperative systemic therapy, or bladder-preserving trimodality treatment, while frailty, competing risks, or patient goals may favor alternative strategies [9]. MDTs integrate these factors through collaborative clinical judgment and therefore provide a relevant real-world comparator for emerging decision-support tools.

The primary objective of this retrospective exploratory study was to assess categorical agreement between treatment recommendations generated by ChatGPT (OpenAI, San Francisco, CA) and real-world MDT decisions in localized cT2-T4a N0M0 MIBC and to describe how agreement varied across treatment modalities. Secondary exploratory objectives were to assess recommendation changes after additional treating physician and/or patient/family context and to describe model-generated decision-driver scores.

This work was presented in poster form at the XIX Simpósio da Associação Portuguesa de Urologia, Figueira da Foz, Portugal, October 29-November 1, 2026, and at the 18th European Multidisciplinary Congress on Urological Cancers (EMUC26), Brussels, Belgium, November 5-8, 2026.

Materials And Methods

Study design and case selection

This retrospective observational exploratory study evaluated agreement between treatment recommendations made by an institutional urology multidisciplinary tumor board (MDT) and those generated by ChatGPT. All consecutive cases of localized cT2-T4a N0M0 muscle-invasive bladder cancer (MIBC) discussed for initial treatment planning between April 1, 2025, and April 1, 2026, were retrospectively reviewed.

MIBC was defined as histologically confirmed invasion of the detrusor muscle on transurethral resection of bladder tumor and/or clinical stage T2 or higher based on the information available at the time of MDT discussion. Cases were eligible if they represented a first MDT discussion for initial treatment planning and had no evidence of regional nodal involvement or distant metastatic disease on available staging investigations. Cases were excluded if they had been discussed previously, had clinical nodal involvement, had distant metastases, or were not being discussed for an initial treatment decision. Of 68 MIBC cases discussed during the study period, 24 met the predefined eligibility criteria.

This study involved the retrospective analysis of anonymized clinical data collected as part of routine clinical care. No additional procedures, interventions, patient contact, or changes to clinical management were performed for research purposes. According to the applicable institutional policy, formal ethics committee approval was not required for this type of retrospective observational study using anonymized routine care data, and the requirement for individual informed consent was waived. All artificial intelligence assessments were performed retrospectively and had no influence on patient management. No directly identifiable patient information was submitted to the artificial intelligence model.

At the time of the ChatGPT assessments, data sharing for model training was disabled in the ChatGPT interface. Only de-identified clinical vignettes were submitted, with direct patient identifiers removed before submission. No directly identifiable patient information was entered into ChatGPT.

Clinical vignette preparation and treatment categories

For each included case, the original clinical vignette presented at the MDT was retrieved. Direct identifiers were removed, and cases were assigned study codes before artificial intelligence assessment. The official MDT recommendation recorded at the time of discussion was used as the real-world comparator.

Each MDT recommendation and each ChatGPT recommendation was assigned to one of four predefined treatment categories: upfront radical cystectomy; systemic therapy followed by radical cystectomy; bladder preservation, radiotherapy, or trimodality treatment; or palliative treatment/best supportive care.

ChatGPT assessment

The anonymized clinical vignettes were submitted to ChatGPT (OpenAI, San Francisco, CA; GPT-5.5 model) on June 1, 2026, using the standard ChatGPT interface rather than an application programming interface (API). No API-level generation parameters were manually configured by the investigators. A separate new ChatGPT session was used for each clinical vignette to minimize cross-case carryover. The full standardized prompts are provided in the Appendices.

First, ChatGPT received only the anonymized clinical vignette as presented to the MDT and was instructed to select a single main treatment category. The model was instructed not to invent clinical information, assume missing test results, or present multiple treatment categories as equivalent. This generated the initial recommendation.

Second, when available, additional contextual information consisting of the treating physician's proposed recommendation and/or patient or family feedback, preferences, values, concerns, or goals of care was provided. ChatGPT was then asked whether this additional context changed its selected treatment category. This proposal was recorded separately and did not necessarily correspond to the final MDT treatment decision. The model was instructed not to automatically agree with the treating physician or follow patient or family preferences when these conflicted with the available clinical information, safety considerations, proportionality, or appropriateness of care. This generated the final recommendation. For cases in which no additional contextual information was available, no contextual reassessment was performed, and the initial recommendation was carried forward as the final recommendation.

Decision-driver scoring

After generating the final recommendation, ChatGPT assigned an influence score from 0 to 5 across seven predefined decision domains: tumor stage and oncological risk, histology and biological features, patient fitness and frailty, comorbidities and surgical/anesthetic risk, systemic treatment eligibility, local treatment feasibility, and patient-centered/treating physician contextual factors. These scores represented model-generated explanatory outputs and were not interpreted as direct measures of the model's internal reasoning processes.

Statistical analysis

Descriptive statistics were used to summarize baseline characteristics, treatment category distributions, recommendation changes, and decision-driver scores. Observed agreement was calculated as the proportion of cases in which ChatGPT and the MDT selected the same treatment category. Wilson 95% confidence intervals (CIs) were calculated for observed agreement. Agreement beyond chance was summarized using unweighted Cohen's kappa. Agreement was also described within each MDT treatment category using counts and proportions.

Because this was an exploratory study and subgroup sizes were small and uneven, no formal hypothesis testing was performed.

Results

Study cohort

The final cohort included 24 patients (Figure *1*). The median age was 74 years (interquartile range (IQR): 66-81), 18 patients (75%) were male, and 19 (79.2%) had an Eastern Cooperative Oncology Group (ECOG) performance status of zero. All tumors were high-grade urothelial carcinoma; variant histology or divergent differentiation was reported in 10 cases (41.7%). Baseline characteristics are summarized in Table *1*.

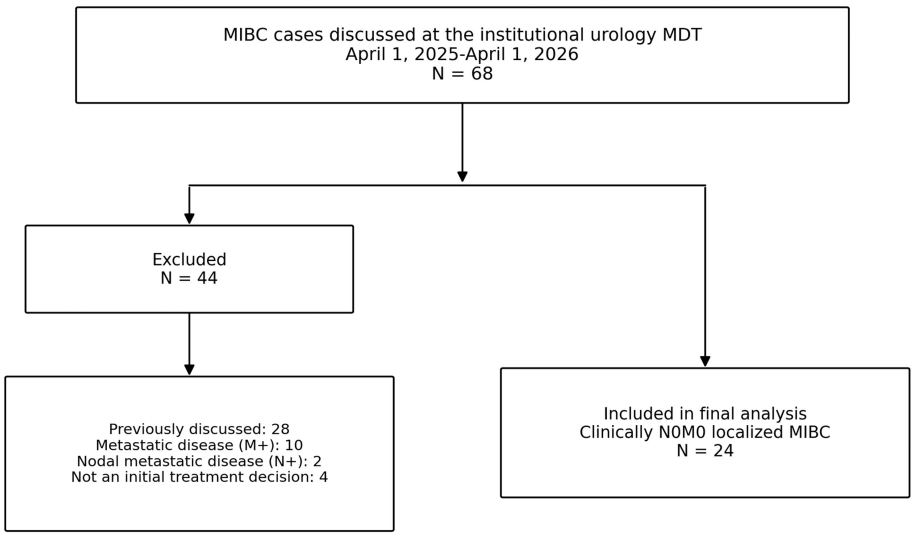


FIGURE 1: Study flow diagram of case selection

Of 68 muscle-invasive bladder cancer cases discussed by the institutional urology multidisciplinary tumor board during the study period, 24 clinically N0M0 cases met the predefined eligibility criteria and were included.

MIBC: muscle-invasive bladder cancer, MDT: multidisciplinary tumor board

Characteristic	Overall cohort (N = 24)
Age, median (IQR), years	74 (66-81)
Age range, years	49-92
Age ≥80 years	7 (29.2%)
Male sex	18 (75%)
Female sex	6 (25%)
ECOG 0	19 (79.2%)
ECOG 1	3 (12.5%)
ECOG 2	2 (8.3%)
High-grade urothelial carcinoma	24 (100%)
Variant histology/divergent differentiation reported	10 (41.7%)
Concomitant carcinoma in situ reported	6 (25%)
Hydronephrosis/upper tract dilatation reported	6 (25%)
Renal impairment potentially affecting cisplatin eligibility	2 (8.3%)

TABLE 1: Baseline characteristics of the included cohort

Values are presented as numbers (%) unless otherwise indicated.

Renal impairment potentially affecting cisplatin eligibility was defined as eGFR or creatinine clearance <60 mL/min or renal dysfunction explicitly described in the vignette as limiting systemic treatment.

ECOG: Eastern Cooperative Oncology Group, eGFR: estimated glomerular filtration rate, IQR: interquartile range

Overall and treatment-specific agreement

The MDT recommended upfront radical cystectomy in 6 cases, systemic therapy followed by radical cystectomy in 13, bladder preservation/radiotherapy/trimodality treatment in 4, and palliative treatment/best supportive care in 1. Initial ChatGPT recommendations agreed with the MDT in 15 of 24 cases, for an observed agreement of 62.5% (95% CI: 42.7%-78.8%) and a Cohen's kappa of 0.370 (Table *2*).

MDT recommendation	ChatGPT: upfront RC	ChatGPT: systemic therapy + RC	ChatGPT: bladder preservation/RT/TMT	ChatGPT: palliative/BSC	Total
Upfront RC	1	3	1	1	6
Systemic therapy + RC	2	11	0	0	13
Bladder preservation/RT/TMT	1	1	2	0	4
Palliative/BSC	0	0	0	1	1
Total	4	15	3	2	24

TABLE 2: Agreement between MDT and initial ChatGPT treatment recommendations

Observed agreement: 15/24 (62.5%)

Cohen's kappa: 0.370

BSC: best supportive care, MDT: multidisciplinary tumor board, RC: radical cystectomy, RT: radiotherapy, TMT: trimodality treatment

Agreement was not uniform across treatment modalities. When the MDT recommended systemic therapy followed by radical cystectomy, ChatGPT selected the same category in 11 of 13 cases (84.6%). Agreement was two of four cases (50%) for bladder preservation/radiotherapy/trimodality treatment and one of six cases (16.7%) for upfront radical cystectomy. The single palliative/best supportive care case was concordant but is not informative as a subgroup estimate because the denominator was one. Among the six upfront radical cystectomy decisions, ChatGPT instead recommended systemic therapy followed by radical cystectomy in three cases, bladder-preserving treatment in one, and palliative/best supportive care in one. Treatment recommendation distributions and within-category agreement are shown in Figure *2* and Table *3*.

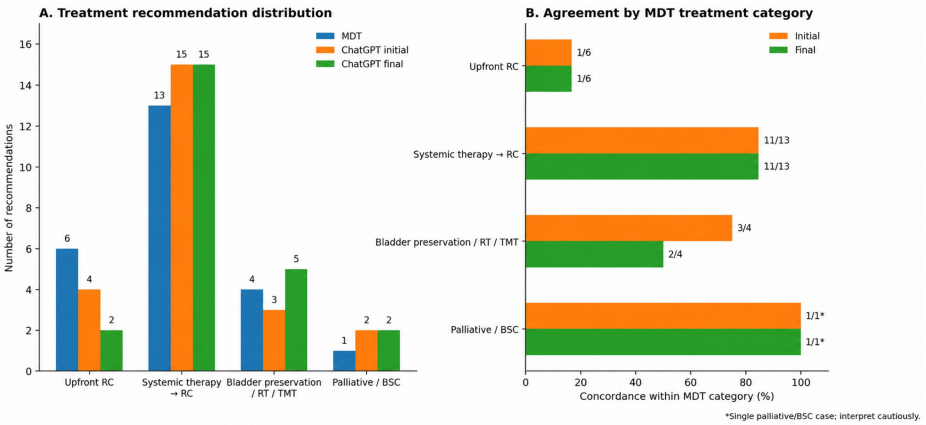


FIGURE 2: Multidisciplinary tumor board and ChatGPT treatment recommendations

Panel A shows the distribution of MDT, initial ChatGPT, and final ChatGPT treatment recommendations. Panel B shows concordance within each MDT treatment category before and after available contextual input. The palliative/best supportive care category comprised a single case and should be interpreted cautiously.

BSC: best supportive care, MDT: multidisciplinary tumor board, RC: radical cystectomy, RT: radiotherapy, TMT: trimodality treatment

MDT treatment category	MDT (number)	ChatGPT initial (number)	ChatGPT final (number)	Initial concordance	Final concordance
Upfront RC	6	4	2	1/6 (16.7%)	1/6 (16.7%)
Systemic therapy + RC	13	15	15	11/13 (84.6%)	11/13 (84.6%)
Bladder preservation/RT/TMT	4	3	5	2/4 (50%)	3/4 (75%)
Palliative/BSC	1	2	2	1/1*	1/1*

TABLE 3: Treatment recommendation distribution and within-category agreement

*The palliative/BSC category comprised one case and should not be interpreted as a stable subgroup estimate.

BSC: best supportive care, MDT: multidisciplinary tumor board, RC: radical cystectomy, RT: radiotherapy, TMT: trimodality treatment

Contextual reassessment

Treating physician and/or patient/family contextual input was available in 11 of 24 cases. ChatGPT changed its treatment category in 4 of these 11 cases (36.4%). Three changes shifted toward bladder preservation or radiotherapy, and one shifted toward systemic therapy followed by radical cystectomy (Table *4*). For the remaining 13 cases without additional contextual input, the initial recommendation was carried forward as the final recommendation.

Case	Initial recommendation	Contextual input	Final recommendation
1	Upfront RC	Patient refused cystectomy	Bladder preservation/RT
2	Systemic therapy + RC	Family prioritized autonomy; physician raised bi-/trimodal strategy	Bladder preservation/RT
3	Bladder preservation/RT	Physician proposed neoadjuvant chemotherapy followed by RC	Systemic therapy + RC
4	Upfront RC	Patient strongly refused cystectomy; physician proposed chemotherapy then reassessment for chemoradiotherapy	Bladder preservation/RT

TABLE 4: Contextual input associated with ChatGPT recommendation changes

RC: radical cystectomy, RT: radiotherapy

After contextual reassessment, the final ChatGPT recommendation agreed with the MDT in 16 of 24 cases (66.7%, 95% CI: 46.7%-82%; kappa = 0.447). This represented one additional concordant case compared with the initial assessment. By MDT treatment category, final concordance was 1 of 6 for upfront radical cystectomy, 11 of 13 for systemic therapy followed by radical cystectomy, 3 of 4 for bladder preservation/radiotherapy/trimodality treatment, and 1 of 1 for palliative/best supportive care (Table *3*). Because contextual input was available in only 11 cases, this comparison was interpreted descriptively as an assessment of recommendation sensitivity rather than evidence of improved model accuracy.

Decision-driver scores

Decision-driver scores were available for all 24 cases. Tumor stage and oncological risk were rated as strongly or decisively influential in all 24 cases. Patient fitness/frailty and local treatment feasibility were rated 4 or 5 in 23 of 24 (95.8%) and 21 of 24 cases (87.5%), respectively. Patient-centered/treating physician contextual factors had the lowest overall median score, 1 (IQR: 0-4), but were rated 4 or 5 in 7 cases (29.2%), indicating strong influence in a subset (Table *5*).

Decision domain	Median (IQR)	Range	Score 0 (number (%))	Score 4-5 (number (%))
Tumor stage and oncological risk	5 (5-5)	4-5	0 (0)	24 (100)
Histology and biological features	4 (3-4)	3-5	0 (0)	13 (54.2)
Patient fitness and frailty	4 (4-5)	3-5	0 (0)	23 (95.8)
Comorbidities and surgical/anesthetic risk	3 (2-4)	0-5	3 (12.5)	11 (45.8)
Systemic treatment eligibility	3 (3-4)	0-5	1 (4.2)	10 (41.7)
Local treatment feasibility	4 (4-5)	2-5	0 (0)	21 (87.5)
Patient-centered/treating-physician contextual factors	1 (0-4)	0-5	9 (37.5)	7 (29.2)

TABLE 5: Distribution of model-generated decision-driver scores across final ChatGPT recommendations

Legend: Scores ranged from 0 (no material influence or information not provided) to 5 (decisive influence). Scores of 4-5 indicate strong or decisive influence.

IQR: interquartile range

Discussion

Principal findings

In this exploratory cohort of localized MIBC, ChatGPT selected the same treatment category as the MDT in

15 of 24 cases. Agreement was uneven across therapeutic pathways, being highest when the MDT selected systemic therapy followed by radical cystectomy and substantially lower when the MDT selected upfront radical cystectomy. This pattern suggests that a single overall agreement percentage can obscure clinically relevant differences in how an LLM reproduces specific treatment decisions.

Comparison with previous tumor board studies

The initial 62.5% agreement observed in our study is consistent with several real-world oncology studies. Bertolo et al. evaluated 103 renal cell carcinoma cases and reported 62.1% agreement between an LLM and MDT decisions, with a kappa of 0.44 [7]. In a large real-world MDT comparison, Qu et al. analyzed 1,423 colorectal cancer cases and found concordance rates of 62.5%-65.4% across three LLMs [4]. That study also demonstrated variable intramodel stability, with repeated-run agreement ranging from 63.9% to 81%. The similarity between these results and our initial 62.5% agreement suggests that moderate concordance may be common when general-purpose LLMs are applied to heterogeneous, real-world MDT decisions, although direct comparisons are limited by differences in case mix, models, prompts, and outcome definitions [3,4,7].

Earlier proof-of-concept work also showed incomplete but clinically meaningful agreement. Sorin et al. reported that ChatGPT-3.5 recommendations were similar to breast tumor board decisions in 7 of 10 cases (70%) [5]. In primary head and neck cancer, Schmidl et al. found that ChatGPT-3.5 and ChatGPT-4.0 could generate plausible treatment discussions but occasionally suggested modalities that were not guideline-concordant [6]. Their subsequent evaluation in recurrent/metastatic head and neck cancer similarly supported an assistive rather than autonomous role because of limited personalization, lack of transparency, and occasional non-guideline recommendations [10].

More recent studies further demonstrate substantial variation across clinical domains. In 102 thoracic oncology cases, ChatGPT agreed with MDT decisions in 52% of cases (kappa = 0.432) [11]. In 242 gynecologic oncology cases, ChatGPT-5 showed only fair MDT-AI concordance (kappa = 0.258-0.334), and identical recommendations across three independent runs occurred in only 37.2% of cases [12]. By contrast, a prospective pancreatic cancer study reported 73.3% concordance with GPT-4 (kappa = 0.64) [13]. These findings place our results within a broader range of moderate agreement and reinforce that LLM performance depends on clinical context and task formulation.

Studies using more structured data or explicit guideline context have sometimes reported substantially higher agreement. Kaiser et al. provided structured clinical data and European Association of Urology guidance to ChatGPT-4 and Claude 3 Opus in 171 newly diagnosed prostate cancer cases and reported 93% overall adherence to MDT recommendations [8]. Similarly, in 99 gastrointestinal stromal tumor cases, Dehdab et al. used a disease-specific structured prompt and reported a mean normalized performance score of 0.901 for ChatGPT-5, although perfect agreement was achieved in only 52.5% of cases [14]. These studies suggest that input structure and explicit disease-specific evidence may materially affect performance.

The prospective multicenter CONCORDIA study is particularly relevant because it evaluated LLMs in urological oncology. Across 110 locally advanced or metastatic genitourinary cancer scenarios, neither ChatGPT-4 nor Claude 3.5 Sonnet met the prespecified noninferiority criterion relative to multidisciplinary tumor boards [15]. This reinforces an important distinction in interpreting our data: categorical concordance with an MDT is not equivalent to clinical noninferiority, safety, or accuracy. The MDT in our study therefore functions as a real-world comparator rather than an absolute gold standard.

Interpretation of treatment-specific discordance

The largest discrepancy in our cohort occurred when the MDT recommended upfront radical cystectomy. ChatGPT reproduced this category in only one of six cases and instead recommended systemic therapy followed by radical cystectomy in three. This pattern may reflect the relative salience of guideline-based perioperative systemic therapy pathways when individualized factors favoring direct surgery are incompletely represented in a text vignette. Current MIBC guidance requires integration of cisplatin eligibility, tumor biology, local disease characteristics, treatment feasibility, comorbidity, and patient preference [9]. The small subgroup size prevents firm conclusions, but the direction of discordance is clinically relevant because an apparently modest difference in category may represent a meaningful difference in treatment sequencing.

Pathway-specific discordance has also been reported in other malignancies. The large colorectal cancer study found lower agreement in more complex or indeterminate contexts [4]. In soft tissue sarcoma, expert review of ChatGPT-4o outputs identified treatment sequencing and chemotherapy selection as areas of persistent discrepancy despite relatively strong performance in clinical contextualization [16]. These findings are consistent with our observation that overall agreement can obscure differences between specific therapeutic pathways.

Effect of contextual input

Additional treating physician and/or patient/family information changed the ChatGPT category in 4 of 11

reassessed cases. Three changes moved toward bladder preservation or radiotherapy, while one moved toward systemic therapy followed by radical cystectomy. The direction of these changes is clinically plausible because the added information included refusal of cystectomy, preservation of autonomy, patient preferences, and treating physician judgment. Contemporary MIBC guidance emphasizes multidisciplinary discussion and shared decision-making, particularly when considering bladder-preserving treatment [9]. However, final overall concordance increased by only one case. Contextual responsiveness should therefore be interpreted as evidence that model outputs are sensitive to additional information, not as evidence that the resulting recommendation is more accurate.

The gynecologic oncology study demonstrated substantial within-model variability despite standardized inputs [12], while the sarcoma study found that contextualization was one of the model's stronger domains, while treatment sequencing remained more problematic [16]. Our model-generated decision-driver scores similarly suggested that patient-centered and treating physician context was highly influential in a subset of cases. These scores are self-attributed explanatory outputs and should not be interpreted as validated measures of internal model reasoning.

Clinical implications

Our findings do not support autonomous use of ChatGPT for MIBC treatment selection. A more defensible role for LLMs is as clinician-supervised decision-support tools that structure case information, identify missing data, summarize guideline-relevant considerations, or provide a second-pass comparison with an MDT recommendation. This interpretation is consistent with the systematic review by Prabhakaran et al. [3] and with the prospective CONCORDIA results showing that publicly available LLMs did not reach noninferiority to urological MDTs [15].

Future MIBC studies should evaluate repeated-run reproducibility, multiple model families, explicit guideline grounding or retrieval augmentation, external validation, and prospective assessment of clinically important discordance. Translational frameworks in urologic oncology similarly emphasize validation, reproducibility, generalizability, and integration into clinician-supervised, human-centered workflows before routine clinical implementation [17]. Particular attention should be paid to whether disagreement changes treatment intent, treatment sequencing, eligibility for curative therapy, or exposure to avoidable toxicity. The higher agreement reported with structured guideline-informed prostate cancer inputs [8] also supports evaluating whether standardized data representations can improve reliability without removing the patient-specific nuance central to MDT decision-making.

Limitations

This study has several limitations. It was retrospective and single-center and included only 24 cases, with small and uneven treatment category subgroups. Contextual information was available in only 11 cases, limiting interpretation of the pre-/post-contextual comparison. A single ChatGPT configuration was evaluated at one time point using one standardized prompting strategy. Because each clinical vignette was assessed only once, the reported concordance represents single-session agreement rather than reproducible model performance; the stability of treatment recommendations across repeated independent assessments remains unknown. This is particularly relevant given evidence that identical cases may yield different recommendations across repeated LLM runs [4,12].

The model was not explicitly supplied with the full EAU guideline text or a retrieval-augmented knowledge source, which limits comparison with studies using guideline-grounded inputs [8,14]. Because the treating physician's proposed category was explicitly provided to ChatGPT during contextual reassessment, this stage should be interpreted as an exploratory assessment of responsiveness to additional clinical context rather than as an independent evaluation of model accuracy. The treating physician's proposal was recorded separately and did not necessarily correspond to the final MDT treatment decision. The MDT comparator reflects real-world multidisciplinary practice but is not an objective gold standard, and discordant recommendations were not independently adjudicated for clinical appropriateness. Finally, the decision-driver scores were generated by the model itself and should be interpreted only as descriptive self-attributions rather than direct measures of internal reasoning. These limitations, together with the rapidly evolving nature of LLMs, mean that the present findings should be considered exploratory and model- and time-specific.

Conclusions

In this retrospective exploratory study, ChatGPT reproduced the MDT treatment category in approximately two-thirds of localized MIBC cases. Agreement was not uniform across treatment strategies: it was most consistent when the MDT recommended systemic therapy followed by radical cystectomy and least consistent when the MDT recommended upfront radical cystectomy.

Additional physician and patient/family context changed some model recommendations but produced only a small change in overall concordance. Larger multicenter studies with repeated model runs and prospective safety assessment are needed before LLM-generated recommendations can be considered for clinical

decision support in uro-oncology.

Appendices

The standardized prompts used for the ChatGPT assessment are presented in Table 6.

Prompt and purpose	Full standardized prompt
Prompt 1: Initial treatment recommendation from the anonymized MDT vignette	You are a medical expert in uro-oncology. I will provide an anonymized clinical case of localized muscle-invasive urothelial carcinoma of the bladder, as it was presented at a multidisciplinary tumor board. Your task is to propose one main treatment recommendation based only on the information provided. Rules: Do not invent clinical information. Do not assume test results or imaging findings that are not included in the text. Choose only ONE category. Do not present multiple options as equivalent. Possible categories: 1 = upfront radical cystectomy, 2 = systemic therapy followed by radical cystectomy, 3 = bladder preservation therapy/trimodality treatment/radiotherapy, 4 = palliative treatment/best supportive care. Please answer exactly in this format: selected category: brief justification: key factors influencing the decision: missing information: level of confidence: low/moderate/high
Prompt 2: Contextual reassessment after treating physician and/or patient/family input	The treating physician's proposed therapeutic recommendation for this case is: X. The patient's and/or family's feedback, preferences, values, concerns, or goals of care are: Y. Taking this information into account, would your selected category change? Rules: Do not automatically agree with the treating physician. Do not automatically follow the patient's or family's preference if it conflicts with the clinical information, safety, proportionality, or appropriateness of care. Reassess the case based only on the clinical information provided, the treating physician's proposed recommendation, and the patient/family feedback. Give appropriate weight to the treating physician's clinical judgment, especially when it reflects bedside assessment, context, prognosis, frailty, comorbidities, treatment burden, or goals of care. Give appropriate weight to the patient's and/or family's feedback, especially when it clarifies values, preferences, quality of life, acceptable treatment burden, or shared decision-making priorities. If your decision changes, state the new selected category and briefly explain why. If your decision does not change, keep your original selected category and briefly explain why. Choose only ONE final category. Do not present multiple options as equivalent. The final category should reflect the best integrated judgment based on clinical appropriateness, proportionality of treatment, physician input, and patient-/family-centered goals of care. Please answer exactly in this format: original selected category: treating physician's proposed category: patient/family feedback considered: would your decision change? yes/no, final selected category: brief justification: level of confidence: low/moderate/high
Prompt 3: Model-generated decision-driver scoring after the final recommendation	You initially received the MDT clinical vignette and, when applicable, additional contextual input from the treating physician and/or from the patient or family. Your final treatment recommendation was made after considering all the information provided. Now, rate how strongly each predefined category influenced your final treatment decision. Use a score from 0 to 5: 0 = did not influence the decision/not relevant/not provided, 1 = very minor influence, 2 = minor influence, 3 = moderate influence, 4 = strong influence, 5 = decisive influence. Important instructions: Use the full 0-5 range when appropriate. Do not give a high score to a category only because it is generally important in MIBC. Give a high score only if that category was directly relevant to this specific patient and influenced the selected treatment. If information related to a category was not provided in the vignette or additional input, score it as 0. Do not add new categories. After scoring all categories, identify the three most influential categories. If several categories have the same score, rank the top three according to their clinical importance for the final treatment decision. If additional input from the treating physician and/or patient/family changed your final decision, the scoring should reflect the category or categories responsible for that change whenever applicable. Treating physician input should be categorized according to its content. If it refers to frailty, comorbidities, surgical/anesthetic risk, systemic treatment eligibility, or local treatment feasibility, score it under C3, C4, C5, or C6 as appropriate. Do not automatically score all treating physician input under C7. Available categories: C1 = tumor stage and oncological risk, C2 = histology and biological features, C3 = patient fitness and frailty, C4 = comorbidities and surgical/anesthetic risk, C5 = systemic treatment eligibility, C6 = local treatment feasibility, C7 = patient-centered and treating physician contextual factors. Please answer only in the following format: C1 = tumor stage and oncological risk: __/5, C2 = histology and biological features: __/5, C3 = patient fitness and frailty: __/5, C4 = comorbidities and surgical/anesthetic risk: __/5, C5 = systemic treatment eligibility: __/5, C6 = local treatment feasibility: __/5, C7 = patient-centered and treating physician contextual factors: __/5

TABLE 6: Standardized prompts used for the ChatGPT assessment

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Eduardo Felício

Acquisition, analysis, or interpretation of data: Eduardo Felício, Guilherme Bernardo, Filipe Gaboleiro, André Pita, Diogo Carmali, Diogo Saraiva, Joao Varregoso, Fernando Ferrito

Drafting of the manuscript: Eduardo Felício

Critical review of the manuscript for important intellectual content: Guilherme Bernardo, Filipe Gaboleiro, André Pita, Diogo Carmali, Diogo Saraiva, Joao Varregoso, Fernando Ferrito

Supervision: Joao Varregoso, Fernando Ferrito

Disclosures

Human subjects: Informed consent for treatment and open access publication was obtained or waived by all participants in this study. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Carl N, Schramm F, Haggenmüller S, et al.: Large language model use in clinical oncology . NPJ Precis Oncol. 2024, 8:240. 10.1038/s41698-024-00733-4
2. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F: Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol. 2023, 13:1268915. 10.3389/fonc.2023.1268915
3. Prabhakaran S, Bell S, Lee JC, Kong JC: The utility of large language models in oncological multidisciplinary team meetings: a systematic review. Eur J Surg Oncol. 2026, 52:111741. 10.1016/j.ejso.2026.111741
4. Qu B, Cao L, Wu C, et al.: Comparison of large language models and expert multidisciplinary team decisions in colorectal cancer. BMJ Health Care Inform. 2026, 33: 10.1136/bmjhci-2025-101780
5. Sorin V, Klang E, Sklair-Levy M, et al.: Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer. 2023, 9:44. 10.1038/s41523-023-00557-8
6. Schmidl B, Hütten T, Pigorsch S, et al.: Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for primary head and neck cancer cases. Front Oncol. 2024, 14:1353031. 10.3389/fonc.2024.1353031
7. 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. 2025, 9:257. 10.1038/s41698-025-01014-4
8. 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. 2025, 43:67. 10.1007/s00345-024-05423-1
9. van der Heijden AG, Bruins HM, Carrion A, et al.: European Association of Urology guidelines on muscle-invasive and metastatic bladder cancer: summary of the 2025 guidelines. Eur Urol. 2025, 87:582-600. 10.1016/j.eururo.2025.02.019
10. Schmidl B, Hütten T, Pigorsch S, et al.: Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for recurrent/metastatic head and neck cancer cases - the first study on ChatGPT 4o and a comparison to ChatGPT 4.0. Front Oncol. 2024, 14:1455413. 10.3389/fonc.2024.1455413
11. Hasimoglu Gurun K, Cikladilmez F, Demirag G, Yilmaz B: Single-session agreement of ChatGPT and Gemini treatment recommendations with multidisciplinary tumor board decisions in thoracic oncology. BMC Cancer. 2026, 26:10.1186/s12885-026-16458-5
12. Asoglu H, Kose S, Kayim O, et al.: Feasibility and concordance of a large language model (ChatGPT-5) as a clinical decision support tool in gynecologic oncology tumor boards: a blinded, multi-observer study. J Clin Med. 2026, 15:10.3390/jcm15124451
13. Gehrisch F, Kirkgöz K, Willner A, et al.: Concordance between GPT-4 and a multidisciplinary tumor board in pancreatic cancer: a prospective pilot study. Langenbecks Arch Surg. 2026, 411:10.1007/s00423-026-04242-9
14. Dehdab R, Herrmann J, Mankertz F, et al.: Precision oncology meets generative AI: assessing large language models in multidisciplinary gist tumor boards. BMC Cancer. 2026, 26:10.1186/s12885-026-16537-7
15. Rinderknecht E, Haas M, Schnabel MJ, et al.: Benchmarking large language models against multidisciplinary tumor boards in urological oncology: results from the blinded, prospective CONCORDIA study. Eur Urol Oncol. 2026, 9:589-97. 10.1016/j.euo.2025.11.016
16. Dehdab R, Afat S, Mankertz F, et al.: When AI joins the table: evaluating large language model performance in soft tissue sarcoma tumor board decisions. J Cancer Res Clin Oncol. 2026, 152: 10.1007/s00432-026-06432-w
17. Rajaee Rizi F, Jamadi MS, Rahimi S, Bighamian M, Paidar P, Taki MJ, Bahmaie N: Multimodal artificial intelligence in urologic precision oncology: from algorithm to translational medicine (a systemized narrative review). Front Oncol. 2026, 16:1763359. 10.3389/fonc.2026.1763359