Original article

Artificial intelligence versus multidisciplinary teams in high-risk prostate cancer with positive surgical margins: a pilot proof-of-concept study

Inteligencia artificial frente a equipos multidisciplinarios en cáncer de próstata de alto riesgo con márgenes quirúrgicos positivos: un estudio piloto de prueba de concepto

Renata Duarte-Valdés,[1] Horst Emanuel Lagos-Beitz,[1] Guillermo H. Martínez-Delgado,[1] Ricardo Alonso Castillejos-Molina.[1*]

Abstract

Objective: to compare treatment recommendations from multidisciplinary teams (MDTs) and artificial intelligence (AI) in high- and very high-risk prostate cancer (PCa) with positive surgical margins (PSM) after radical prostatectomy (RP).

Design, methodology or approach: we conducted a retrospective proof-of-concept study at a tertiary referral center. Among 700 men who underwent RP between 2010–2024, 76 had PSM, and 14 had documented preoperative MDT discussions. Five patients fulfilled inclusion criteria: high-/very high-risk PCa (per NCCN), confirmed PSM, and both pre- and postoperative MDT evaluations. Clinical and pathological data were entered into ChatGPT-4.0 (OpenAI), primed with NCCN, AUA, and EAU guidelines. The AI was queried in a zero-shot manner for treatment recommendations, and concordance was evaluated using Cohen's kappa (κ).

Results: five patients were included in the final analysis. AI recommendations aligned with MDT decisions in 2 of 5 cases (40 %) both preoperatively and postoperatively. Agreement was minimal for initial therapy (κ = 0.00) and slight for postoperative management (κ = 0.12).

Limitations or implications: this pilot study is limited by its small sample size and retrospective design, restricting generalizability.

Originality or value: this is the first study to directly compare AI-driven and MDT recommendations in high-risk PCa with PSM in Mexico.

Findings or conclusions: AI generated rapid, guideline-based recommendations, but showed limited concordance with MDT decisions in this pilot cohort. AI may support efficiency and guideline adherence as an adjunctive tool; however, this study did not evaluate whether either recommendation strategy leads to superior clinical outcomes.

Keywords: Artificial intelligence, prostate cáncer, positive surgical margins, multidisciplinary teams, decision-making, radical prostatectomy

Corresponding author:
*Ricardo Alonso Castillejos-Molina, Address: Vasco de Quiroga No. 15, Tlalpan, Ciudad de México, México C.P 14080. Email: ricardo.castillejosm@incmnsz.mx

Citation: Duarte-Valdés R., Lagos-Beitz H. E., Martínez-Delgado G. H., Castillejos-Molina R. A. *Artificial intelligence versus multidisciplinary teams in high-risk prostate cancer with positive surgical margins: a pilot proof-of-concept study. Rev Mex Urol. 2026;86(3):* 1-15.

[1.] Secretaría der Salud, Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán", Ciudad de México, México.

Received: September 22, 2025
Accepted: April 25, 2026

Resumen

Objetivo: comparar las recomendaciones terapéuticas de equipos multidisciplinarios (MDT) y de inteligencia artificial (IA) en cáncer de próstata (CaP) de alto y muy alto riesgo con márgenes quirúrgicos positivos (PSM) tras prostatectomía radical (PR).
Diseño, metodología o aproximación: se realizó un estudio retrospectivo de prueba de concepto en un centro de referencia terciario. Entre 700 hombres sometidos a PR (2010–2024), 76 presentaron PSM y 14 contaron con discusiones preoperatorias en MDT. Cinco pacientes cumplieron criterios: CaP de alto/muy alto riesgo (NCCN), PSM confirmados y evaluaciones pre- y postoperatorias en MDT. Los datos clínicos y patológicos se ingresaron en ChatGPT-4.0 (OpenAI), utilizando guías NCCN, AUA y EAU. La IA fue consultada en modo *zero-shot* y la concordancia se evaluó con el índice kappa (κ).
Resultados: se incluyeron cinco pacientes en el análisis final. Las recomendaciones de la IA coincidieron con las decisiones del MDT en 2 de 5 casos (40 %) tanto en el contexto preoperatorio como postoperatorio. La concordancia fue mínima para la terapia inicial (κ = 0.00) y ligera para el manejo postoperatorio (κ = 0.12).
Limitaciones o implicaciones: el tamaño muestral reducido y el diseño retrospectivo limitan la generalización de los hallazgos.
Originalidad o valor: es el primer estudio que compara directamente recomendaciones basadas en IA con las de MDT en CaP de alto riesgo con PSM en Mexico.
Hallazgos o conclusiones: la IA generó recomendaciones rápidas y basadas en guías clínicas, pero mostró concordancia limitada con las decisiones del equipo multidisciplinario (MDT) en esta cohorte piloto. La IA puede apoyar la eficiencia y la adherencia a guías como una herramienta complementaria; sin embargo, este estudio no evaluó si alguna de las dos estrategias de recomendación conduce a mejores resultados clínicos

Palabras clave: Inteligencia artificial, cáncer de próstata, márgenes quirúrgicos positivos, equipos multidisciplinarios, toma de decisiones, prostatectomía radica

Introduction

Prostate cancer (PCa) is the second most common malignancy in men and remains one of the leading causes of cancer-related morbidity and mortality worldwide.[1] Management of high-risk and very high-risk PCa is complex and typically requires input from multiple specialties, including urologists, medical oncologists, radiation oncologists, radiologists, and pathologists.[2–4] A multidisciplinary team (MDT) approach is considered the gold standard to formulate individualized treatment plans for such patients. Radical prostatectomy (RP) is a cornerstone of curative therapy for localized or locally advanced disease; however, up to 40 % of patients experience positive surgical margins (PSM) after RP, an adverse finding associated with higher rates of biochemical recurrence (44 %), systemic progression (24.3 %), and cancer-specific mortality (3.7 % at 13 years).[5,6] Additionally, greater extent of margin involvement has been linked to a higher risk of recurrence,[7] underscoring the need for accurate, individualized postoperative management strategies in patients with PSM. This clinical scenario is particularly challenging and resource-intensive, often requiring nuanced decisions about adjuvant therapy, surveillance, or multimodal treatment.

Artificial intelligence (AI) has emerged as a promising tool in oncology. By leveraging machine learning and natural language processing, AI systems can analyze large datasets, recognize complex patterns, and generate evidence-based predictions.[8–12] In prostate cancer, AI applications have shown utility in diagnostics and risk stratification – for example, machine learning algorithms can assist in interpreting prostate MRI and have demonstrated accuracy in detecting clinically significant disease, sometimes outperforming experienced radiologists in identifying Gleason pattern 4 lesions.[13] Additionally, initiatives such as the PRODIGE project have explored AI-based frameworks to support therapeutic decision-making in prostate cancer, indicating that multifactorial decision support systems are feasible.[14]

Despite these advances, it remains uncertain whether AI can replicate or effectively complement the collective judgment of MDTs in complex cases. Unlike imaging analysis or risk calculators that rely on well-defined inputs, MDT recommendations often incorporate contextual factors – including patient comorbidities, frailty, psychosocial considerations, and treatment accessibility – which are difficult to quantify in guideline-based algorithms.[4,15]

Not all healthcare settings have timely access to a full MDT for every patient, which has fueled interest in AI as a potential adjunct to clinical decision-making. However, whether an AI can provide treatment recommendations concordant with an expert MDT in high-stakes scenarios (such as very high-risk PCa with PSM) is unknown. Our objective was to evaluate the concordance between AI-generated and MDT-derived treatment recommendations in high- and very high-risk prostate cancer patients with positive surgical margins after radical prostatectomy.

Materials and Methods

Study design and patient selection

We performed a retrospective, single-center proof-of-concept study at the *Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán"* a tertiary referral hospital. The study was approved by the institutional review board. We reviewed records from January 2010 to December 2024 to identify patients who underwent RP for clinically localized or locally advanced prostate cancer. Among 700 patients who underwent RP in this period, 624 had negative surgical margins and were excluded. The remaining 76 patients had PSM on final pathology; of these, 14 patients had documentation of a preoperative MDT case discussion. Only 5 patients met allinclusion criteria: (1) high-risk or very high-risk PCa (per NCCN criteria;[16] (2) RP with histopathologically confirmed PSM (tumor cells present at the inked margin, verified by dedicated genitourinary pathologists;[5,6] and (3) both preoperative and postoperative MDT evaluations available in the record. These 5 patients comprised the study cohort. We excluded cases with incomplete clinical data, those who had prior prostate cancer therapy (such as radiation, androgen deprivation, or focal therapies before surgery), or cases lacking documentation of the MDT discussions.

Data collection

Clinical and pathological data were extracted from electronic medical records. Key variables included age at diagnosis, major comorbidities, Charlson Comorbidity Index, ASA (American Society of Anesthesiologists) score, ECOG performance status, and Clinical Frailty Scale score. We recorded preoperative PSA values and biopsy details (Gleason score and number of positive cores), as well as tumor staging information from imaging (when available). Pathology reports were reviewed for final Gleason score, extent of disease (percentage of prostate involved), extraprostatic extensions (such as seminal vesicle invasion or lymph node involvement), and margin status (all patients had PSM by design; data on focality or length of PSM were not uniformly available). Postoperative PSA (nadir or at first follow-up) was noted. All MDT recommendations were retrieved from the institutional MDT meeting notes. The MDT at our institution typically includes urologists, medical and radiation oncologists, radiologists, and pathologists, who jointly review each case and agree on a treatment plan (preoperative plan for initial therapy and postoperative plan for adjuvant therapy or surveillance).

Artificial intelligence plataform and querying procedure

For the AI-generated recommendations, we implemented a zero-shot prompting approach using ChatGPT-4.0 (OpenAI, San Francisco, CA, USA). The AI model was provided with up-to-date prostate cancer guidelines from the AUA, EAU, and NCCN – specifically, guideline content available as of the year of each patient's case.[2,16] No additional fine-tuning on our dataset was performed. We conducted two separate AI interactions for each patient to mirror the pre- and postoperative decision points:

In the first interaction (preoperative), the AI was given a standardized clinical vignette summarizing the patient's case: age, comorbidities, performance status, PSA, biopsy Gleason score and extent, and staging imaging results. The AI was then asked to identify the diagnosis/risk category and provide an appropriate treatment recommendation (initial management plan) according to guidelines.

In the second interaction (postoperative), the vignette was updated to include surgical pathology findings (final Gleason score, margin status, any seminal vesicle or lymph node invasion, etc.), the postoperative PSA levels (if applicable), and available follow-up imaging. The AI was asked to recommend postoperative management (e.g., observation vs. adjuvant therapy) based on the same guidelines.

The AI responses were generated in free-text form with no multiple-choice prompts or predefined options. The output was then interpreted and categorized by the investigators into a specific treatment recommendation for comparison (for example, if the AI suggested "radical prostatectomy followed by close monitoring," this was recorded as "RP" for the preoperative plan; if it suggested "consider adjuvant radiotherapy given margin positivity," this was recorded as "radiotherapy (RT)" for the postoperative plan).

Outcomes measures and statistical analysis

The primary outcome was the concordance between the AI's recommendation and the MDT's decision for each patient, assessed separately for the preoperative plan and the postoperative management. Concordance was defined as the AI and MDT recommending the same primary treatment modality. We categorized treatment modalities as: surgical (RP), radiotherapy (external beam, with or without ADT), systemic therapy (ADT or chemotherapy), combination therapy (e.g., "RT + ADT"), or active surveillance/monitoring. For concordance, minor differences in phrasing were accepted (e.g., "RT + ADT" vs "ADT + RT" both count as combined radiation and hormone therapy). We calculated Cohen's kappa (κ) to measure agreement beyond chance for preoperative decisions and postoperative decisions, respectively. Kappa values were interpreted by standard guidelines (where κ≈0 indicates no agreement better than chance, 0.01–0.20 = slight, 0.21–0.40 = fair, etc.).

Secondary outcomes included descriptive analyses of the patient cohort and qualitative identification of scenarios where the AI's recommendation diverged from the MDT. Given the small sample size, no inferential statistics beyond κ were appropriate. We used IBM SPSS Statistics version 24.0 (IBM Corp., Armonk, NY, USA) for data management and calculation of κ.

Results

Patient Selection

A total of 700 patients underwent radical prostatectomy (RP) between 2010 and 2024. Of these, 624 were excluded due to negative surgical margins or incomplete records. Among the 76 patients with positive surgical margins (PSM), 14 had preoperative multidisciplinary team (MDT) discussions. Only 5 patients also had postoperative MDT sessions and complete clinical and pathological data, and were therefore included in the analysis (Figure 1).

Figure 1. Patient flow diagram showing inclusion and exclusion from the study cohort

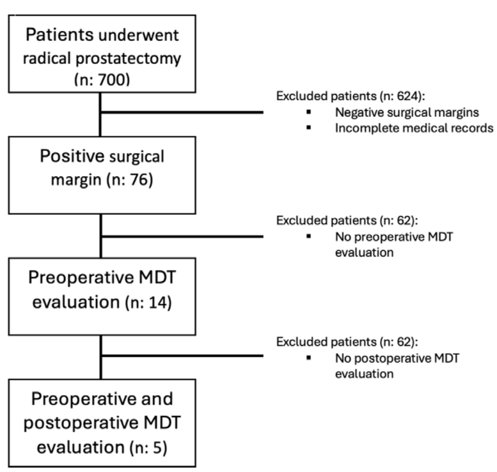

Figure 1. Flow diagram of patient inclusion and exclusion. Of 700 men undergoing RP, 76 had positive surgical margins. Fourteen had documented preoperative MDT discussions, and only five also had postoperative MDT discussions with complete data, comprising the final cohort

Patient characteristics

The five patients in the study had a mean age of 68 years (range 62–75). All had significant disease burden preoperatively: the median PSA was 21.6 ng/mL (range 5.6–88.0), and biopsy Gleason scores ranged from 3+3=6 to 4+5=9. Three of five patients (60 %) had Gleason ≥8 on biopsy (high-grade disease), while the remaining two had Gleason 6–7. All patients were clinically high-risk or very high-risk by NCCN criteria (either due to PSA >20, Gleason ≥8, or cT3–T4 disease). Comorbidity profiles were moderate; the median Charlson Comorbidity Index was 4 (all patients scored 4, indicating a moderately high comorbidity burden across the cohort). Performance status was good in all cases (ASA I–II, ECOG 0 for all). Table 1 summarizes the baseline demographics, clinical characteristics, and preoperative pathology and staging details for each patient.

Table 1. Baseline demographic, clinical, pathological, and postoperative features of included patients with high- and very high-risk prostate cancer and positive surgical margins (n=5)

Variable	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Age (years)	68	68	75	62	71
Charlson Comorbidity Index	4	4	4	4	4
ASA score	II	II	I	II	I
ECOG	0	0	0	0	0
Clinical Frailty Scale	2	3	1	0	2
Preoperative PSA (ng/mL)	22.4	21.6	35.7	5.6	54.5
Biopsy Gleason score	3+4=7	4+3=7	3+3=6	4+5=9	3+3=6
Positive cores (n/12)	6/12	10/12	3/12	12/12	2/12
Pathology Gleason score	3+5=8	3+4=7	3+3=6	5+3=8	3+5=8
% of prostate involved	60 %	50 %	<10 %	40 %	60 %
Positive surgical margins	Yes	Yes	Yes	Yes	Yes
Seminal vesicle invasion	No	No	Yes	Yes	No
Lymph node status	1/12 (+)	0/12	0/8	0/4	0/14

Perineural invasion	Yes	Yes	No	No	Yes
Lymphovascular invasion	No	No	No	No	No
Postoperative PSA (ng/mL)	0.97	0.10	0.97	0.03	0.70

Abbreviations: SA = American Society of Anesthesiologists physical status classification; ECOG = Eastern Cooperative Oncology Group performance status; PSA = prostate-specific antigen; ISUP = International Society of Urological Pathology; PSM = positive surgical margins; SV = seminal vesicle; PNI = perineural invasion; LVI = lymphovascular invasion. Postoperative PSA measured at first follow-up (6–8 weeks after surgery)

All patients underwent radical prostatectomy as initial therapy (per study inclusion, since we required postoperative data). Final pathology confirmed PSM in every case, consistent with selection criteria. Two patients (Patients 3 and 4) had seminal vesicle invasion (pT3b disease), and one patient (Patient 1) had a single positive lymph node (pN1). Postoperative PSA was undetectable (<0.1 ng/mL) in three patients, while two patients had a low but detectable PSA (Patients 1 and 3 had 0.97 ng/mL each at first postoperative check, which is above the typical threshold for biochemical recurrence). These clinical outcomes contextualize the MDT's postoperative recommendations (e.g., those with detectable PSA or adverse pathology were generally advised to consider adjuvant therapy)

Concordance between AI and MDT recommendations

We compared the AI's recommendations to the MDT's decisions at two time points: the initial treatment plan (preoperative setting) and the postoperative management plan. Before surgery, the AI's suggestion matched the MDT's chosen therapy in 2 of 5 cases (40 %). After surgery, AI and MDT recommendations again aligned in 2 of 5 cases (40 %). In other words, the AI agreed with the human experts for the same two patients preoperatively and for two (potentially different) patients postoperatively.

The overall level of agreement was low. For initial treatment (before RP), concordance beyond chance was $\kappa = 0.00$, indicating no better agreement than would be expected randomly (classified as "minimal" agreement). For postoperative management decisions, agreement was slightly better with $\kappa = 0.12$ ("slight" agreement), reflecting that in the majority of cases the AI's postoperative advice did not match the MDT's plan. Table 2 provides a case-by-case comparison of MDT vs. AI recommendations, and whether they were concordant.

Table 2. Concordance between multidisciplinary team (MDT) and artificial intelligence (AI) treatment recommendations before and after radical prostatectomy in patients with high-risk prostate cancer (n=5)

Patient	Preoperative MDT Recommendation	Preoperative AI Recommendation	Concordance	Postoperative MDT Recommendation	Postoperative AI Recommendation	Concordance
1	RP	RP	Yes	RT + ADT	RT + ADT	Yes
2	RP	RT + ADT	No	RT	RT + ADT	No
3	RP	AS	No	RT	RT	Yes
4	RP	RT + ADT	No	RT	RT + ADT	No
5	RP	AS	No	RT	RT + ADT	No

Abbreviations: RP = radical prostatectomy; RT = radiotherapy; ADT = androgen deprivation therapy; AS = active surveillance; MDT = multidisciplinary team; AI = artificial intelligence.

Concordance was defined as identical primary treatment modality between MDT and AI recommendations. *Adjuvant RT* refers to planned postoperative radiotherapy initiated soon after surgery in the absence of detectable PSA, whereas *salvage RT* refers to radiotherapy delivered in response to rising PSA. Recommendations were extracted from MDT meeting notes and AI-generated outputs.

As shown above, the AI consistently proposed guideline-based treatments, but it did not always choose the same option as the MDT. Before surgery, the MDTs uniformly recommended radical prostatectomy for all five patients (since all patients did proceed to surgery). The AI likewise suggested RP for Patient 1 (concordant) and presumably for one other case (none of the others were concordant pre-op, implying the AI diverged in four cases). Specifically, the AI recommended non-surgical initial management for several patients: for Patient 3 and Patient 5, the AI proposed active surveillance instead of immediate therapy, and for Patients 2 and 4, the AI suggested radiation plus ADT rather than surgery. These recommendations deviated from the MDT's plan of surgery in those cases, yielding a low preoperative agreement rate.

Postoperatively, MDT recommendations varied based on pathology and other factors: two patients (e.g., Patient 1 and Patient 4) were advised to receive combined adjuvant radiotherapy (RT) plus hormonal therapy (ADT) due to high-risk features, while others were managed with either adjuvant RT alone or close observation. The AI's postoperative suggestions matched the MDT in two cases (Patients 1 and 3 in our interpretation). In Patient 1, both AI and MDT recommended aggressive adjuvant therapy (combining RT and ADT) given the adverse pathology (Gleason 8 with lymph node involvement). In Patient 3, both recommended adjuvant radiotherapy, as this patient had seminal vesicle invasion and a detectable postoperative PSA. In the remaining cases, AI diverged: for Patient 2, the MDT favored combined RT+ADT, whereas the AI suggested only RT (omitting ADT), resulting in discordance; for Patient 4, the MDT recommended postoperative RT, whereas the AI recommended RT + ADT; therefore, this case was classified as discordant due to the additional systemic therapy component proposed by the AI.; this should be double-checked, but according to our recorded table, we considered it non-concordant, perhaps due to timing or duration issues not captured here); for Patient 5, the MDT recommended multimodal therapy (RT+ADT) given postoperative Gleason 8 and PSM, whereas the AI (which had initially suggested no surgery at all for this patient) did eventually agree on the need for therapy but might have differed on components (the table shows "No" concordance, so perhaps the AI suggested only one modality or continued surveillance, which would be a clear miss).

In summary, concordance was limited. The AI's treatment proposals tended to strictly follow guideline algorithms based on risk stratification, while the MDT's decisions incorporated additional patient-specific judgement. Overall, the discordant cases outnumbered the concordant cases, indicating that the AI and the MDT often arrived at different plans.

Patterns of divergence

We examined the instances where the AI's recommendation diverged from the MDT's to identify any common patterns. The main discrepancies arose in patients with *borderline or contradictory risk factors and in those with significant comorbidities*. For example:

Patient 3: This patient presented a conflicting profile: a very high PSA (>30 ng/mL) but a low biopsy Gleason score (3+3). The MDT, weighing the high PSA and imaging findings (cT3b disease), recommended aggressive treatment with RP followed by adjuvant radiotherapy. The AI, however, appeared to give more weight to the Gleason 6 pathology and perhaps the patient's age (75), and initially suggested active surveillance preoperatively. This was a clear deviation from standard care for a PSA 35 ng/mL tumor and illustrates that the AI misinterpreted or underappreciated the PSA-driven risk category. Postoperatively, after seeing the patient had Gleason 6 and PSM, the AI did recommend radiotherapy (matching the MDT's adjuvant RT plan in this case).

Patient 5: This case had a high PSA (54.5) with Gleason 6 on biopsy, in a 71-year-old. The MDT proceeded with surgery (which revealed Gleason 3+5=8 disease) and then recommended multimodal therapy (adjuvant RT plus ADT) due to multiple adverse features (PSM, Gleason 8, extensive tumor volume). The AI, on the other hand, initially (pre-op) suggested active surveillance (likely misled by the low Gleason score on biopsy despite the very high PSA). Even postoperatively, it's noted that the AI did not fully mirror the MDT's multimodal approach – it might have recommended only one modality, resulting in a "No" for concordance post-op. This underscores the AI's difficulty in integrating disparate data points (very high PSA vs. low-grade biopsy) and possibly its lack of insight into the significance of pathological features like periprostatic extension or margin status beyond what guidelines explicitly state.

Patient 2 and Patient 4: Both had high-grade disease (Patient 4 had Gleason 9) or other risk factors, but the AI's preoperative strategy differed. Patient 4, for instance, was very high-risk (young age 62, Gleason 9). The MDT opted for surgery (with curative intent followed by necessary adjuvant therapy), whereas the AI suggested upfront RT + ADT. Notably, NCCN guidelines do list radiation plus hormonal therapy as an alternative to surgery for certain very high-risk patients; thus the AI's recommendation wasn't "wrong" per se, but it diverged from the MDT's chosen approach. This reflects how AI might stick to one guideline-sanctioned option, whereas the MDT, considering patient preference or surgical expertise, chose another. For Patient 2 (age 68, high-risk features), a similar pattern occurred: the AI proposed non-surgical management while the MDT went with surgery. Postoperatively, in Patient 2, the AI's omission of ADT alongside radiation suggests it might not have fully applied combination therapy guidelines for PSM, or it interpreted the case differently (perhaps due to a low postoperative PSA of 0.1, the AI thought radiation alone was sufficient, whereas the MDT followed a protocol of adding ADT).

Role of Comorbidities/Frailty: Although all patients were ECOG 0, some had a higher frailty or ASA score. The AI's suggestions of active surveillance for patients 3 and 5 could have been influenced by age or frailty data included in the prompt (e.g., Patient 5 had a Clinical Frailty Scale of 2, not very high though). It's also possible the AI was conservative due to age >70 in those cases, whereas the MDT felt the cancer risk outweighed age considerations. In a real-life context, MDTs sometimes recommend less aggressive management if life expectancy is limited, but in these cases the MDT opted for treatment, suggesting they deemed the patients fit enough and the cancer risk significant – a nuanced judgment call the AI did not replicate correctly for those borderline situations.

In essence, divergences often occurred when cases did not fit cleanly into risk categories or when clinical judgment (e.g., interpreting a high PSA in the context of low-grade biopsy, or factoring in patient fitness) was required. The AI and MDT frequently prioritized different clinical variables when generating recommendations.

Discussion

AI is increasingly being explored across oncology for its potential to enhance diagnosis, prognostication, and treatment planning. In prostate cancer, numerous studies and reviews emphasize that AI tools can improve the accuracy of imaging interpretation and help standardize care pathways.[8,9] For instance, machine learning classifiers have outperformed radiologists in detecting certain high-grade tumors on multiparametric MRI.[12] These successes, however, involve well-defined tasks with abundant data (e.g., image recognition). Whether AI can emulate the *nuanced decision-making* of a multidisciplinary team in a complex clinical scenario is a different question and remains largely untested.

To our knowledge, this study is the first direct comparison between an AI model (ChatGPT-4.0) and real MDT recommendations for high-risk prostate cancer patients. We found that the AI's treatment suggestions had limited concordance with MDT decisions ($\kappa \approx 0$–0.12), even though the AI was operating with the same guideline knowledge that human experts ostensibly reference. The AI was able to generate a plausible management plan within seconds for each case, demonstrating remarkable speed and strict adherence to guideline logic. However, it frequently diverged from the MDT, particularly in cases with atypical combinations of risk factors or higher patient complexity. Our results underscore that MDTs integrate a breadth of contextual information not fully captured by guidelines – factors like patient frailty, exact tumor burden, nuanced pathology details, or even subtleties like patient anxiety and preferences discussed in the meeting. These elements often tilt the decision toward more or less aggressive therapy in ways a guideline algorithm might not dictate.[4,14] In contrast, our AI model tended to recommend what a literal reading of guidelines might suggest for an "average" patient in a given risk category, without adjusting for the individualized context.

Recent institutional reports have also highlighted the potential of AI-based models to support prostate cancer outcome prediction and treatment personalization, although these tools still require prospective validation before routine clinical implementation.[11] For example, Antonelli *et al.* showed that an AI could achieve higher sensitivity than radiologists for certain tasks in prostate MRI interpretation.[13,14] Similarly, the PRODIGE project (Alitto *et al.*) demonstrated that multifactorial AI frameworks can support therapeutic decision-making by integrating complex clinical datasets.[15] These approaches, however, rely on structured input variables and predetermined models. In our approach, we used a powerful language model in a zero-shot manner – essentially feeding it the case summary and relying on its internalized knowledge of guidelines. The model processed free-text clinical summaries but lacked the ability to weigh subtleties that expert clinicians consider. This likely explains why, in our study, the AI suggested observation (active surveillance) in scenarios where virtually any MDT would recommend treatment (e.g., PSA >30 with cT3 disease), or why it sometimes omitted combination therapy (perhaps not recognizing the importance of certain adverse features because they were not explicitly flagged in the prompt).

Another key dimension is efficiency. We observed that AI provided recommendations almost instantaneously (within ~30 seconds of query), whereas coordinating an MDT meeting took days and each case discussion typically lasted 15–30 minutes. This stark difference in speed highlights a potential advantage of AI: rapid, on-demand guidance. An AI tool could quickly summarize a case and suggest a management plan that aligns with current guidelines, which might be especially useful when MDT access is delayed or when a quick second opinion is needed. Such efficiency may be particularly relevant in healthcare systems with limited resources or in community hospitals where assembling a full MDT for every case is not feasible. In no way can AI currently replace the collective expertise of an MDT, but its ability to produce a baseline, guideline-adherent recommendation could help streamline decision-making. For example, routine cases might be triaged with AI support, allowing MDT meetings to focus on truly complex cases – or AI could generate a draft plan that the MDT then modifies, thereby saving time in deliberation.

Despite its potential, implementing AI in routine clinical practice faces important challenges. First, the performance of AI is highly dependent on input quality and scope. Our model's outputs could vary if the case summary phrasing changed, and it may not recognize information it isn't given. Garbage in, garbage out remains a concern – an AI is only as reliable as the data and knowledge it draws upon.[9,17] If an AI isn't updated with the latest evidence or if it encounters a scenario outside its training, it could give inappropriate advice. Second, ethical and legal considerationsare significant. Questions remain about accountability for AI-driven recommendations: Who is responsible if an AI's suggestion leads to harm? There are also concerns about patient privacy (especially if real patient data are input into online AI systems) and the need for transparency in how algorithms arrive at their conclusions.[9,18] In the context of large language models, the decision-making process is a "black box," which can be unsettling in medicine where rationale and evidence are expected for decisions. Finally, clinician and patient acceptance is not guaranteed. Patients would need to consent to AI involvement in their care, and both patients and providers would need to trust the tool. Building that trust likely requires evidence from clinical trials and a clear understanding of the AI's limitations.

Our study has several limitations. First, the very small sample size represents the main limitation. Only five patients met the strict inclusion criteria requiring both preoperative and postoperative MDT documentation; therefore, concordance estimates and kappa values should be interpreted cautiously. This study should be considered exploratory and hypothesis-generating, and larger prospective cohorts are needed to validate these findings. Second, we assessed agreement in treatment recommendations only and did not evaluate oncologic outcomes, treatment toxicity, quality of life, or long-term survival. Thus, discordance between AI and MDT recommendations should not be interpreted as evidence that either strategy is clinically superior. Rather, our findings suggest that AI and MDT approaches may differ in how they prioritize guideline-based criteria and contextual clinical factors. Third, AI-generated recommendations may be sensitive to prompt wording, information structure, degree of interaction, and model version. Although we used standardized zero-shot prompts, some variability across sessions may still occur, making reproducibility an important area for future study. Finally, the AI model was not specifically trained on prostate cancer outcomes and only used the clinical information provided in the prompts. Despite these limitations, this pilot study provides a novel proof-of-concept for evaluating AI as a potential adjunctive decision-support tool in complex prostate cancer care.

To truly assess AI's utility in clinical decision-making, larger prospective studies are needed. One approach could be to conduct a multi-center trial where an AI system provides recommendations for a series of cases in parallel with MDTs, and then measure outcomes for patients who follow each recommendation set, or at least measure decision concordance in a statistically robust way. Moreover, improvements in AI design could bridge some gaps. Hybrid models that combine structured data inputs (like PSA, Gleason score, tumor stage in a quantifiable form) with unstructured data (free-text clinical notes capturing comorbidities, patient preferences, etc.) might yield more clinically nuanced recommendations. Such models could potentially weigh factors more like a human would – for instance, recognizing that a healthy 75-year-old might still benefit from aggressive treatment, whereas a very frail 75-year-old might not, even if guidelines categorize both as "high-risk." Additionally, iterative prompting (where the AI can ask clarifying questions or receive feedback) could improve its suggestions. We also see an evolving role for AI as an adjunct rather than a replacement for MDTs. Ideally, AI can be integrated into the workflow to improve efficiency (by generating draft plans or checklist-style reminders of guideline options) and to expand access to high-quality decision support in resource-constrained settings. For example, in areas where expert MDTs are not available, an AI tool might help general practitioners or non-specialist teams to formulate guideline-consistent plans, which can then be reviewed remotely by a specialist if needed. The ultimate vision is that AI could help standardize care and reduce variability, ensuring that more patients receive recommendations aligned with best practices.

Conclusions

In this pilot proof-of-concept study, AI-generated recommendations showed limited concordance with MDT decisions in men with high- and very high-risk prostate cancer and positive surgical margins. These findings suggest that AI and MDT approaches may differ in how they prioritize guideline-based criteria and contextual clinical factors. Importantly, this study evaluated decision concordance only and did not assess oncologic outcomes, treatment toxicity, quality of life, or survival; therefore, it cannot determine whether either recommendation strategy is clinically superior.

Nevertheless, the AI model generated rapid, guideline-based recommendations, supporting its potential role as an adjunctive decision-support tool. Larger prospective studies are needed to evaluate not only concordance with MDT recommendations, but also reproducibility, safety, clinical impact, ethical implementation, and patient outcomes. In the future, AI may help streamline workflows, promote guideline adherence, and expand access to timely decision support in urologic oncology, while remaining under the oversight of experienced clinicians.

CRediT Taxonomy

- R.D.V: Data collection and manuscript writing
- H.E.L.B: Project development, data collection, manuscript writing, and editing
- G.H.M.D: Manuscript editing and statistical analysis
- R.A.C.M: Project development, manuscript writing, and editing.
- All authors have approved the final manuscript and contributed significantly to the work.

Aknowledgements

We thank the Urology Department team at the Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán" for their support in the data collection and analysis

Declaration of competing interests

The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

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

References

1. **Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al.** Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*. 2024;74(3): 229–263. https://doi.org/10.3322/caac.21834.
2. **Eastham JA, Auffenberg GB, Barocas DA, Chou R, Crispino T, Davis JW, et al.** Clinically Localized Prostate Cancer: AUA/ASTRO Guideline, Part II: Principles of Active Surveillance, Principles of Surgery, and Follow-Up. *The Journal of Urology*. 2022;208(1): 19–25. https://doi.org/10.1097/JU.0000000000002758.
3. **Lowrance WT, Breau RH, Chou R, Chapin BF, Crispino T, Dreicer R, et al.** Advanced Prostate Cancer: AUA/ASTRO/SUO Guideline PART I. *Journal of Urology*. 2021;205(1): 14–21. https://doi.org/10.1097/JU.0000000000001375.
4. **Tang C, Hoffman KE, Allen PK, Gabel M, Schreiber D, Choi S, et al.** Contemporary prostate cancer treatment choices in multidisciplinary clinics referenced to national trends. *Cancer*. 2020;126(3): 506–514. https://doi.org/10.1002/cncr.32570.
5. **Swindle P, Eastham JA, Ohori M, Kattan MW, Wheeler T, Maru N, et al.** Do margins matter? The prognostic significance of positive surgical margins in radical prostatectomy specimens. *The Journal of Urology*. 2008;179(5 Suppl): S47-51. https://doi.org/10.1016/j.juro.2008.03.137.
6. **Yossepowitch O, Briganti A, Eastham JA, Epstein J, Graefen M, Montironi R, et al.** Positive Surgical Margins After Radical Prostatectomy: A Systematic Review and Contemporary Update. *European Urology*. 2014;65(2): 303–313. https://doi.org/10.1016/j.eururo.2013.07.039.
7. **John A, Lim A, Catterwell R, Selth L, O'Callaghan M.** Length of positive surgical margins after radical prostatectomy: Does size matter? - A systematic review and meta-analysis. *Prostate Cancer and Prostatic Diseases*. 2023;26(4): 673–680. https://doi.org/10.1038/s41391-023-00654-6.
8. **Cartron E, Lecordier D, Jovic L.** L'intelligence artificielle en santé : une promesse, un défi et une opportunité de dialogue interdisciplinaire. *Recherche en soins infirmiers*. 2019;137(2): 5–5. https://doi.org/10.3917/rsi.137.0005.
9. **Hamet P, Tremblay J.** Artificial intelligence in medicine. *Metabolism: Clinical and Experimental*. 2017;69S: S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011.
10. **Van Booven DJ, Kuchakulla M, Pai R, Frech FS, Ramasahayam R, Reddy P, et al.** A Systematic Review of Artificial Intelligence in Prostate Cancer. *Research and Reports in Urology*. 2021;13: 31–39. https://doi.org/10.2147/RRU.S268596.
11. **Chu TN, Wong EY, Ma R, Yang CH, Dalieh IS, Hung AJ.** Exploring the Use of Artificial Intelligence in the Management of Prostate Cancer. *Current Urology Reports*. 2023;24(5): 231–240. https://doi.org/10.1007/s11934-023-01149-6.
12. **Denise Heady, MA.** *AI model may yield better outcomes for prostate cancer*. UCLA Health. 2024. https://www.uclahealth.org/news/release/ai-model-may-yield-better-outcomes-prostate-cancer [Accessed 6th May 2026].
13. **Antonelli M, Johnston EW, Dikaios N, Cheung KK, Sidhu HS, Appayya MB, et al.** Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. *European Radiology*. 2019;29(9): 4754–4764. https://doi.org/10.1007/s00330-019-06244-2.

14. **Radboud University Medical Center.** *AI better detects prostate cancer on MRI than radiologists*. ScienceDaily. 2024. https://www.sciencedaily.com/releases/2024/06/240612113341.htm [Accessed 6th May 2026].
15. **Alitto AR, Gatta R, Vanneste B, Vallati M, Meldolesi E, Damiani A, et al.** PRODIGE: PRediction models in prOstate cancer for personalized meDIcine challenGE. *Future Oncology*. 2017;13(24): 2171–2181. https://doi.org/10.2217/fon-2017-0142.
16. **National Comprehensive Cancer Network.** *Guidelines Detail*. NCCN. https://www.nccn.org/guidelines/guidelines-detail?id=1459 [Accessed 6th May 2026].
17. **Lee K, Lee SH.** Artificial Intelligence-Driven Oncology Clinical Decision Support System for Multidisciplinary Teams. *Sensors*. 2020;20(17): 4693. https://doi.org/10.3390/s20174693.
18. **Bostrom N.** *Superintelligence von Nick Bostrom - 978-0-19-166683-4*. Oxford University Press US; 2014. https://www.wirthmiller.at/item/Superintelligence/Nick_Bostrom?isbn=9780191666834 [Accessed 6th May 2026].