



# ChatGPT v4 outperforming v3.5 on cancer treatment recommendations in quality, clinical guideline, and expert opinion concordance

DIGITAL HEALTH  
Volume 10: 1–13  
© The Author(s) 2024  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20552076241269538  
journals.sagepub.com/home/dhj



Chung-You Tsai<sup>1,2</sup> , Pai-Yu Cheng<sup>3,4</sup>, Juinn-Horng Deng<sup>2</sup>, Fu-Shan Jaw<sup>3</sup>  
and Shyi-Chun Yii<sup>3,4</sup> 

## Abstract

**Objectives:** To assess the quality and alignment of ChatGPT's cancer treatment recommendations (RECs) with National Comprehensive Cancer Network (NCCN) guidelines and expert opinions.

**Methods:** Three urologists performed quantitative and qualitative assessments in October 2023 analyzing responses from ChatGPT-4 and ChatGPT-3.5 to 108 prostate, kidney, and bladder cancer prompts using two zero-shot prompt templates. Performance evaluation involved calculating five ratios: expert-approved/expert-disagreed and NCCN-aligned RECs against total ChatGPT RECs plus coverage and adherence rates to NCCN. Experts rated the response's quality on a 1-5 scale considering correctness, comprehensiveness, specificity, and appropriateness.

**Results:** ChatGPT-4 outperformed ChatGPT-3.5 in prostate cancer inquiries, with an average word count of 317.3 versus 124.4 ( $p < 0.001$ ) and 6.1 versus 3.9 RECs ( $p < 0.001$ ). Its rater-approved REC ratio (96.1% vs. 89.4%) and alignment with NCCN guidelines (76.8% vs. 49.1%,  $p = 0.001$ ) were superior and scored significantly better on all quality dimensions. Across 108 prompts covering three cancers, ChatGPT-4 produced an average of 6.0 RECs per case, with an 88.5% approval rate from raters, 86.7% NCCN concordance, and only a 9.5% disagreement rate. It achieved high marks in correctness (4.5), comprehensiveness (4.4), specificity (4.0), and appropriateness (4.4). Subgroup analyses across cancer types, disease statuses, and different prompt templates were reported.

**Conclusions:** ChatGPT-4 demonstrated significant improvement in providing accurate and detailed treatment recommendations for urological cancers in line with clinical guidelines and expert opinion. However, it is vital to recognize that AI tools are not without flaws and should be utilized with caution. ChatGPT could supplement, but not replace, personalized advice from healthcare professionals.

## Keywords

Artificial intelligence, ChatGPT, patient information, cancers, prostate, bladder, kidney

Submission date: 29 November 2023; Acceptance date: 20 June 2024

## Introduction

In November 2022, ChatGPT was introduced by OpenAI, aiming to develop conversational AI systems capable of comprehending and responding to human language.<sup>1</sup> Subsequent applications in the field of healthcare have demonstrated the utility of ChatGPT in diverse areas, including clinical documentation and note-taking, patient communication and support, medical education, medical literature review, and research assistance.<sup>2</sup>

<sup>1</sup>Divisions of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei

<sup>2</sup>Department of Electrical Engineering, Yuan Ze University, Taoyuan

<sup>3</sup>Department of Biomedical Engineering, College of Medicine and College of Engineering, National Taiwan University, Taipei, Taiwan

<sup>4</sup>Department of Surgery, Far Eastern Memorial Hospital, New Taipei

### Corresponding author:

Shyi-Chun Yii, Department of Biomedical Engineering, National University, Taipei City; Department of Surgery, Far Eastern Memorial Hospital, No. 21, Sec. 2, Nanya S. Rd., Banciao Dist., New Taipei City 220.  
Email: zodiac0518@gmail.com



In the field of urology, empirical evidence supports ChatGPT's role as a proficient virtual healthcare assistant for benign prostatic hyperplasia.<sup>3</sup> It has also proven valuable as an educational and preventive tool for prostate cancer,<sup>4</sup> providing support for urological residents and assisting in the composition of urological papers and academic work.<sup>5</sup> The adaptability and efficacy of ChatGPT underscore its potential to enhance various aspects of healthcare and medical research within urology.

Patients are increasingly using large language model (LLM) chatbots as a source of treatment information. A study published by Chen et al.<sup>6</sup> in *JAMA Oncology* in August 2023 found that approximately one-third of the cancer treatment recommendations made by ChatGPT-3.5 only partly aligned with the National Comprehensive Cancer Network (NCCN) guidelines.<sup>7</sup> This has led to numerous media reports that cast doubt on ChatGPT's suitability as a cancer treatment advisor (Supplementary File S1).

The NCCN guidelines are widely adopted as the standard for cancer care and are among the most frequently updated clinical practice guidelines in oncology. They provide evidence-based, consensus-driven management strategies essential for optimal patient outcomes. Aligning AI-generated recommendations (RECs) with these guidelines is critical, as deviations could lead to less effective or potentially harmful treatment RECs.

Research assessing the outcomes of AI-generated medical information has shown inconsistent results.<sup>6,8,9</sup> To date, a scant study has simultaneously conducted a quality assessment and measured the concordance of ChatGPT's responses with clinical guidelines and expert opinions regarding urological cancer-related inquiries. Thus, our study aims to analyze and assess the performance of ChatGPT-4,<sup>10</sup> the latest iteration following ChatGPT-3.5, as medical information resources for patients in providing urological cancer treatment recommendations. We evaluate its alignment with NCCN guidelines<sup>11</sup> and expert opinions, along with an examination of the quality of its responses.

## Materials and methods

### ChatGPT models

During our study period, two of the latest ChatGPT versions were employed to measure and compare model performance. The ChatGPT-3.5-turbo (gpt-3.5-turbo-0613 model) represents the pinnacle of the GPT-3.5 series and is available in the free version of ChatGPT. In contrast, ChatGPT-4 (gpt-4-0613 model) is the most sophisticated in the GPT series, known for handling complex tasks with greater depth but with a slower response time, and is utilized by the ChatGPT Plus service. Due to the rapid advancement in AI tools, both versions show significant improvements over their predecessors.<sup>12</sup>

### Prompts design

From the perspective of patients seeking treatment recommendations, they often pose open-ended and non-specific questions. However, they may also request responses based on guideline-specific recommendations. Consequently, two zero-shot prompt templates were designed to query treatment recommendations. Template 1, referred to as "Non-Specific Prompt," simply asks, "What is the recommended treatment for [xxx] cancer?" whereas Template 2, labeled as "NCCN-Specified Prompt," appends "according to NCCN" to the query, thus enhancing the specificity to the guideline source. The placeholder "[xxx]" will be replaced with 54 distinct cancer state descriptions, varying by three cancer types and their respective disease statuses. These descriptions were determined based on either NCCN guideline-specific disease status descriptions or general cancer descriptions according to cancer stage. Ultimately, this approach yielded 108 unique prompts when combined with both templates. For example, when querying high-risk localized prostate cancer:

- Non-specific prompt (Template 1): What is the recommended treatment for high-risk localized prostate cancer?
- NCCN-specified prompt (Template 2): What is the recommended treatment for high-risk localized prostate cancer according to NCCN?

The full set of 108 prompts, used for querying the ChatGPT models, is detailed in Supplementary Files S6 in the "Query prompt" column. All prompts did not involve any patient data.

Each prompt was entered into the OpenAI ChatGPT interface, and the subsequent responses from ChatGPT models were recorded. These interactions, including both the prompts and the ChatGPT responses, are comprehensively documented in Supplementary File S6.

### Study design

This exploratory cross-sectional study was carried out from 5 September to 22 October 2023. It was structured into two distinct stages. The first stage performed a preliminary comparative analysis between ChatGPT-4 and ChatGPT-3.5, with a focus on their capability to generate cancer treatment recommendations. For a fair comparison, both models were queried using the same set of 16 prompts specifically tailored to prostate cancer scenarios. A smaller pilot study was chosen due to the observed significant performance differences between the models, indicating that a limited sample size could adequately demonstrate ChatGPT-4's superiority.

The second stage formed the main body of the research, dedicated to an in-depth evaluation of ChatGPT-4 exclusively. This phase assessed the model's accuracy, compliance with NCCN guidelines, and overall utility for providing patient treatment information. We extended our

analysis to include the three most common urological cancers: prostate (36 prompts), kidney (34 prompts), and bladder (38 prompts). A broad spectrum of clinical situations was represented by creating diverse disease descriptions that considered cancer stage (1–4), the degree of invasion (muscle-invasive or not), and disease status (localized, metastatic, or recurrent). This systematic approach in prompt generation allowed for an exhaustive assessment of ChatGPT-4’s performance in offering treatment recommendations across varied cancer stages and disease statuses.

### Ethics and consent statement

No human participants nor patient data were involved in the study. According to the Regulations on Human Trials by the Ministry of Health and Welfare in Taiwan, the IRB Board only handles matters related to human subjects and does not accept applications for ethical review when the study does not involve human participants. Furthermore, as per the Common Rule, IRB approval was not required for studies without human participants.

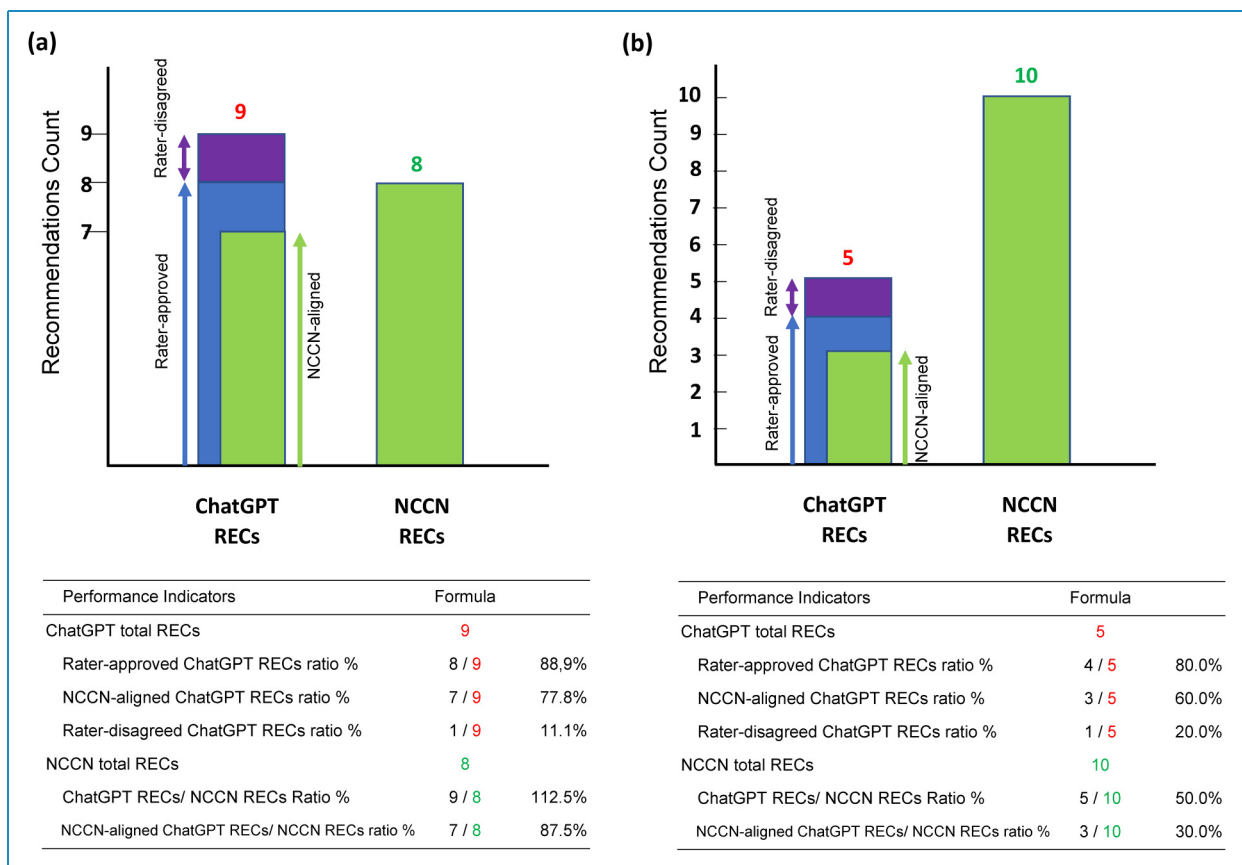
### Outcome measures

Three board-certified urologic oncologists independently scored ChatGPT responses using predefined scoring guidelines (Supplement File S2) and a scoring questionnaire (Supplement File S3).

Performance was evaluated using five distinct ratios (Figure 1). We established two primary categories of indicators to assess the urological cancer treatment recommendations (RECs) provided by ChatGPT in comparison to NCCN standards.

Category 1: ChatGPT-generated REC evaluation (ChatGPT RECs as denominator). Using the total number of ChatGPT recommendations as the denominator, this category evaluates the proportion of recommendations made by ChatGPT that are aligned with NCCN, approved or disagreed upon by experts:

1. Rater-approved ChatGPT REC ratio (%): The proportion of ChatGPT recommendations approved by domain experts, indicating the acceptance rate of the AI’s recommendations.
2. NCCN-aligned ChatGPT REC ratio (%): The proportion of ChatGPT recommendations aligned with NCCN



**Figure 1.** Performance indicator variability in assessing ChatGPT’s recommendations against NCCN guidelines across two scenarios. Panels (a) and (b) display the discrepancy in cancer treatment recommendations (RECs) by ChatGPT relative to NCCN guidelines for two distinct cancer queries as assessed by a single rater. (a) The scenario where ChatGPT’s RECs are more than NCCN RECs. (b) The scenario where ChatGPT’s RECs are fewer than NCCN RECs.

guidelines, showing the compliance rate of the AI's recommendations with medical standards.

3. Rater-disagreed ChatGPT REC ratio (%): The proportion of ChatGPT recommendations that domain experts disagreed with, identifying potential inaccuracies or deviations from standard practices.

Category 2: NCCN benchmark compliance (NCCN RECs as denominator). Using the total number of recommendations provided by the NCCN as the denominator, this category evaluates ChatGPT's coverage and alignment with the total available standard care:

1. ChatGPT REC/NCCN REC ratio (%): This ratio indicates the total ChatGPT recommendations against NCCN recommendations, showing the AI's coverage of standard treatment options.

2. NCCN-aligned ChatGPT REC/NCCN REC ratio (%): Focuses on the number of NCCN-aligned recommendations made by ChatGPT as a proportion of the NCCN's total recommendations, reflecting the accuracy and relevance of the AI's recommendations in standard treatment options.

These two categories provide complementary perspectives. The first category measures ChatGPT's recommendations' quality against its total output, while the second evaluates ChatGPT's output's comprehensiveness and alignment within the broader landscape of established medical guidelines. This dual approach offers a comprehensive assessment of the AI's performance in recommending treatment options, accounting for both the correctness of its individual recommendations and its ability to cover the breadth of standardized care.

Figure 1 presents two scenarios (a) and (b) showcasing the variance in cancer treatment recommendations by ChatGPT in response to different cancer disease descriptions, assessed by a single rater, and highlights the variability of five evaluative ratios:

Scenario (a) ChatGPT RECs > NCCN REC: ChatGPT provides more recommendations than the NCCN. This could indicate that ChatGPT is suggesting additional treatment options or possibly including less common or emerging treatments not yet fully established in the NCCN guidelines or incorporating recommendations with which experts disagree. Scenario (b) ChatGPT RECs < NCCN RECs: ChatGPT provides fewer recommendations than the NCCN. This might suggest that ChatGPT is being more selective or perhaps missing some established treatment options that are listed by the NCCN or only focusing on more common treatment options while omitting alternative therapies.

These examples emphasize the importance of expert review and the need for alignment with established guidelines, ensuring that AI-generated recommendations are both comprehensive and adhere to current medical standards. The evaluation protocol referenced the 2021 NCCN guidelines to match GPT-3.5 turbo's training cut-off in September 2021.

Quality assessments were conducted in four dimensions: correctness, comprehensiveness, specificity, and appropriateness of the RECs on a 5-point scale (1-5) based on expert judgment independent of guidelines.

### Statistical analysis

The performance ratios and quality assessment scores were averaged across the three raters to produce mean values for analysis. All ratios and scores from three raters were presented as mean (SD). In the initial phase of our analysis, we employed paired *t*-tests to compare the performance indicators between GPT-4 and GPT-3.5. To ensure the validity of our paired *t*-tests, we assessed the normality of the differences between paired observations using the Shapiro–Wilk test. For the second phase of analysis, which focused exclusively on GPT-4, we conducted subgroup analyses. Independent *t*-tests were utilized for comparisons between the two groups, and the assumption of equal variances was checked with Levene's test before applying the tests. ANOVA and subsequent post-hoc tests were applied for three-group comparisons, ensuring that the assumptions of normality and homogeneity of variances were satisfied. The reliability of the evaluations among three raters was assessed via the intraclass correlation coefficient (ICC). The analysis was conducted employing IBM SPSS Statistics, version 25.

## Results

### ChatGPT-4 versus ChatGPT-3.5

In comparing prostate cancer treatment recommendations via 32 prompts (Table 1, Figure 2, and Supplement Files S4 and S5), ChatGPT-4 significantly outperformed ChatGPT-3.5, as evidenced by a higher average word count per response (317.3 vs. 124.4,  $p < 0.001$ ) and more total RECs made (6.1 vs. 3.9,  $p < 0.001$ ), as shown in Figure 2(a) and (b). Although the difference in the ratio of NCCN-aligned RECs between the two versions was not significant (90.4% for ChatGPT-4 vs. 88.5% for ChatGPT-3.5,  $p = 0.657$ ), ChatGPT-4 had a notably higher ratio of rater-approved RECs (96.1% vs. 89.4%) and a significantly lower ratio of rater-disagreed RECs (0.8% vs. 10.6%,  $p = 0.011$ ). Furthermore, ChatGPT-4 outshined ChatGPT-3.5 in the ChatGPT/NCCN REC ratio (87.7% vs. 57.6%,  $p = 0.001$ ) and the NCCN-aligned ChatGPT/NCCN REC ratio (76.8% vs. 49.1%,  $p = 0.001$ ), indicating a superior alignment with NCCN total RECs (Figure 2(c)).

Quality assessments further underscored the superiority of ChatGPT-4, with significantly better scores in correctness (4.8 vs. 3.3), comprehensiveness (4.6 vs. 2.1), specificity (3.8 vs. 1.4), and appropriateness (4.6 vs. 2.3), all  $p < 0.001$ , as shown in Figure 2(d). These results highlight the advancements in ChatGPT-4 over ChatGPT-3.5, reflecting more robust and aligned treatment recommendations in the context of prostate cancer (Table 1 and Figure 2).

**Table 1.** Comparison of treatment recommendations for prostate cancer: ChatGPT-3.5 versus ChatGPT-4.

ChatGPT	Version 3.5		Version 4		p-value	
	Mean	(SD)	(SD)	(SD)		
Query prompts (n)	16		16			
Response word count	124.4	(52.9)	317.3	(69.9)	<0.001**	
ChatGPT total RECs <sup>a</sup>	3.9	(1.3)	6.1	(2.4)	<0.001**	
Rater-approved ChatGPT REC ratio % <sup>b</sup>	89.4	(12.5)	96.1	(7.1)	0.065	
NCCN-aligned ChatGPT REC ratio % <sup>b</sup>	88.5	(12.2)	90.4	(13.3)	0.657	
Rater-disagreed ChatGPT REC ratio % <sup>b</sup>	10.6	(12.5)	0.8	(3.1)	0.011**	
NCCN total RECs	6.0	(0.6)	6.0	(0.6)	-	
ChatGPT REC/NCCN REC ratio % <sup>c</sup>	57.6	(18.4)	87.7	(30.2)	0.001**	
NCCN-aligned ChatGPT REC/NCCN REC ratio % <sup>c</sup>	49.1	(14.3)	76.8	(26.4)	0.001**	
Correctness	(range 1-5)	3.3	(0.8)	4.8	(0.4)	<0.001**
Comprehensiveness	(range 1-5)	2.1	(0.7)	4.6	(0.9)	<0.001**
Specificity	(range 1-5)	1.4	(0.6)	3.8	(0.8)	<0.001**
Appropriateness	(range 1-5)	2.3	(0.4)	4.6	(0.7)	<0.001**

<sup>a</sup>RECs: recommendations.

<sup>b</sup>ChatGPT total RECs as the denominator.

<sup>c</sup>NCCN total RECs as the denominator.

\*\*Significant  $p < 0.01$  NCCN total RECs (query prompts  $n = 12$ ).

### ChatGPT-4's evaluations

In evaluating ChatGPT-4's ability to propose urological cancer treatments in line with NCCN guidelines and expert opinions, the model was tested across 108 prompts for prostate, kidney, and bladder cancers (Supplement Files S6 and S7). Inter-rater reliability testing revealed high ICC values among all indicators, ranging from 0.8 to 0.96 with good agreement. ChatGPT-4 produced an average of 6.0 RECs per case, with a high approval ratio of 88.5% from raters, an 86.7% alignment with NCCN guidelines, and a low rater-disagreed REC ratio of 9.5%. These figures demonstrate high rater approval and NCCN alignment among the total RECs provided by ChatGPT, reflecting the model's robustness in providing relevant and expert-endorsed treatment options (Table 2 and Figure 3).

Noteworthy was the assessment of ChatGPT RECs against total NCCN RECs, with a 100% mean ratio for all cancers, highlighting the comprehensive coverage of NCCN recommendations. The NCCN-aligned ChatGPT REC/NCCN REC ratio also indicated strong alignment,

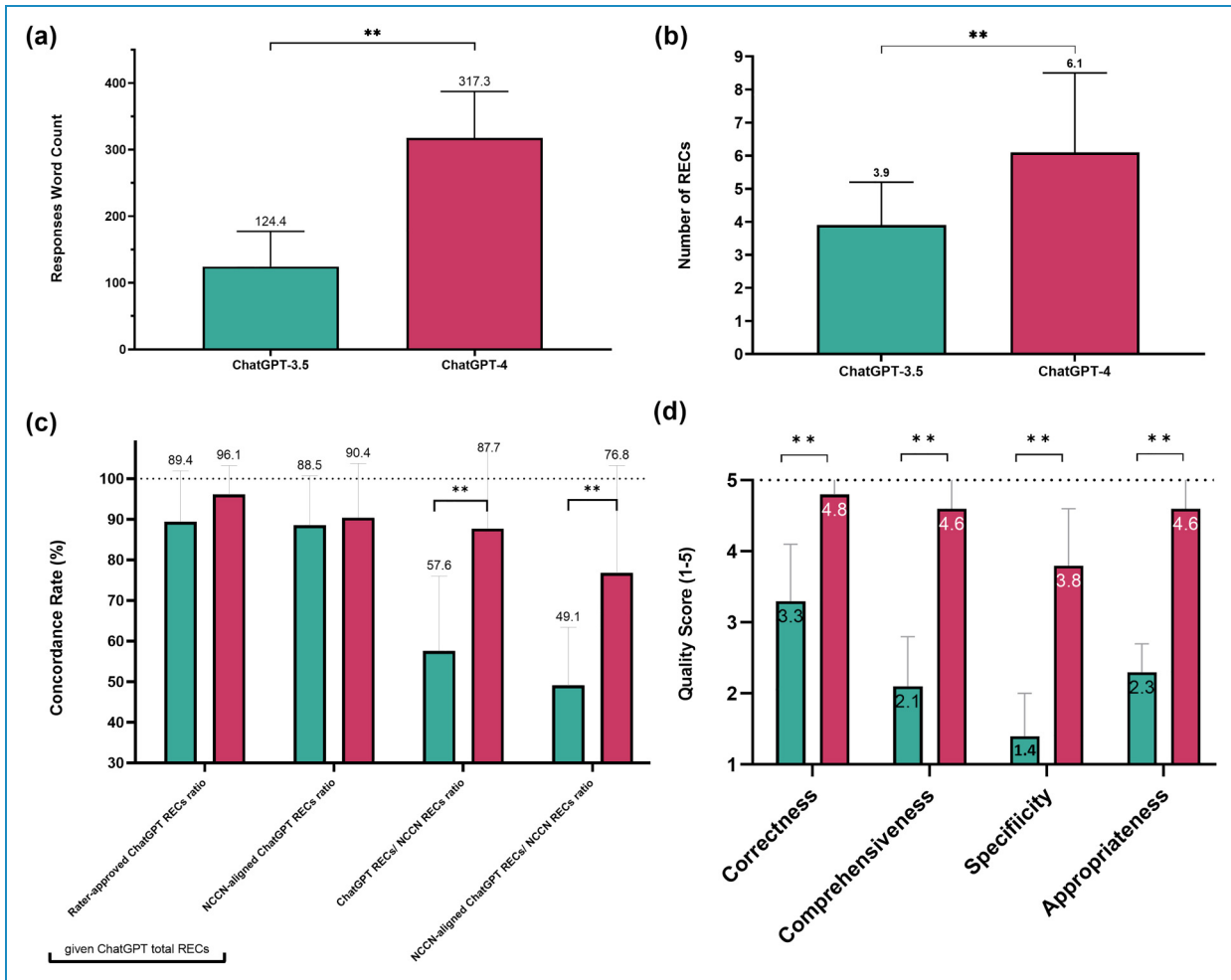
with an 81% mean across all cancers, suggesting high accuracy and adherence to NCCN standards.

Quality assessments showed high scores across the board, with correctness averaging at 4.5, comprehensiveness at 4.4, specificity at 4.0, and appropriateness averaging 4.3, all on a 5-point scale (Figure 3). These high scores across all quality dimensions highlight the precision and relevance of the ChatGPT-4 generated recommendations, confirming its potential as a valuable assist tool for cancer treatment information in urology for patients.

### Subgroup analysis of ChatGPT-4's performance

Subgroup analysis by cancer type revealed the differential performance of ChatGPT-4's treatment recommendations in alignment with NCCN guidelines and expert approval. Specifically, bladder cancer recommendations exhibited the highest rater approval with a 91.4% approval rate and a 91.7% NCCN guideline alignment, suggesting a strong concordance with established medical standards for bladder cancer treatments (Figure 4).





**Figure 2.** Comparison of Treatment Recommendations for Prostate Cancer: ChatGPT-3.5 vs. ChatGPT-4. Panels (a) to (d) illustrate the differences in performance between two ChatGPT models when queried about prostate cancer using 32 unique prompts: (a) Response word count. (b) Number of treatment recommendations (RECs) provided by ChatGPT. per query. (c) The concordance rate of recommendations was evaluated by four performance indicators. (d) Quality assessments on a 5-point scale (1-5) in four dimensions: correctness, comprehensiveness, specificity, and appropriateness. ChatGPT-4 (red) significantly outperformed ChatGPT-3.5 (green) in most measured aspects. All bar charts present mean values with standard deviations. Significant differences ( $p < 0.01$ ) between the two models are indicated by double asterisks (\*\*).

Prostate cancer recommendations were also well-received, achieving a 90.1% rater approval and 85.0% guideline alignment, indicating reliable performance in this domain. Kidney cancer treatment recommendations, while slightly lower, still demonstrated a substantial 83.4% rater approval and 82.9% NCCN concordance. The rater-approved ChatGPT REC ratio showed significant differences ( $p < 0.05$ ) across the three cancer types, and the NCCN-aligned ChatGPT REC ratio also exhibited marginal significance ( $p = 0.050$ ). Overall, the disagreement rate with expert raters was low across all cancer types, reflecting the model's reliable performance in generating medical recommendations in line with expert opinion and established guidelines (Table 2 and Figure 4).

For disease status (Figure 5), recommendations for systemic or metastatic cancers showed the highest 93.3.7% rater

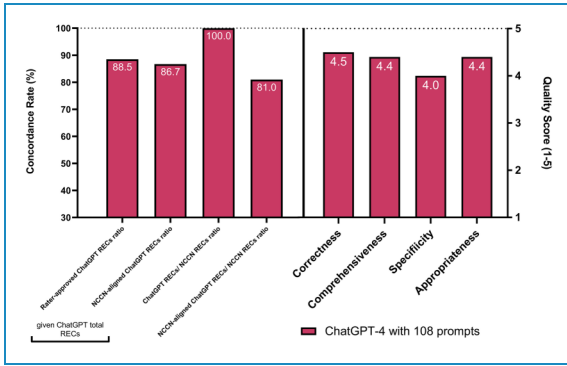
approval and 91.4% NCCN alignment among total ChatGPT recommendations. However, for recurrent cases, there was a notable decrease in 73.8% rater approval and 73.2% NCCN alignment, which was significantly lower compared to localized and systemic cases ( $p = 0.002$  and  $p = 0.009$ , respectively). This suggests that the model may require further refinement to better handle complex recurrent cases.

Table 3 illustrates the results of subgroup analysis stratified by prompt template. NCCN-specified prompt refers to appending 'according to NCCN' to the prompt, as opposed to the non-specific prompt which does not include this specification. Does the NCCN-specified prompt affect ChatGPT's performance? Indeed, the results presented in Table 3 show significant differences in many indicators, suggesting that each type of prompt has distinct advantages:

**Table 2.** Overall and subgroup analysis of ChatGPT-4 on urological cancer treatment recommendations.

ChatGPT-4	Cancer type				Disease Status				p-value	p-value
	All		Bladder Ca.		Localized		Recurrent			
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)		
Query prompts (n)	108	36	34	38	56	40	6			
ChatGPT total RECs <sup>a</sup>	6.0 (1.92)	6.4 (2.10)	5.8 (1.64)	5.8 (1.97)	5.7 (1.55)	5.6 (2.04)	7.8 (1.85)	0.017*		
Rater-approved ChatGPT REC ratio (%) <sup>b</sup>	88.5 (14.8)	90.1 (12.8)	83.4 (19.0)	91.4 (11.0)	85.7 (16.0)	93.3 (11.1)	73.8 (16.6)	0.002*		
NCCN-aligned ChatGPT REC ratio (%) <sup>b</sup>	86.7 (16.1)	85.0 (16.1)	82.9 (19.2)	91.7 (11.8)	83.7 (17.8)	91.4 (12.0)	73.2 (17.9)	0.009*		
Rater-disagreed ChatGPT REC ratio (%) <sup>b</sup>	9.5 (13.7)	7.9 (12.4)	13.9 (18.1)	7.1 (9.0)	12.4 (15.4)	4.3 (7.8)	24.5 (17.1)	<0.001**		
NCCN total RECs	6.0 (2.18)	5.9 (1.11)	6.1 (2.89)	6.0 (2.29)	5.7 (1.89)	6.1 (2.40)	7.6 (2.33)	0.131		
ChatGPT REC/NCCN REC ratio (%) <sup>c</sup>	100.0 (40.5)	102.7 (35.5)	103.4 (56.7)	95.2 (28.5)	108.5 (45.1)	86.6 (30.3)	113.8 (41.3)	0.057		
NCCN-aligned ChatGPT REC/NCCN REC ratio (%) <sup>c</sup>	81.0 (20.6)	80.8 (20.8)	78.0 (20.1)	83.6 (21.3)	84.6 (20.0)	77.3 (21.8)	77.3 (17.2)	0.314		
Correctness (range 1-5)	4.5 (0.65)	4.4 (0.69)	4.4 (0.75)	4.6 (0.49)	4.4 (0.67)	4.7 (0.51)	3.7 (0.79)	<0.001**		
Comprehensiveness (range 1-5)	4.4 (0.70)	4.6 (0.69)	4.3 (0.68)	4.5 (0.73)	4.5 (0.72)	4.4 (0.67)	4.1 (0.65)	0.207		
Specificity (range 1-5)	4.0 (0.71)	4.0 (0.67)	3.8 (0.71)	4.1 (0.75)	3.9 (0.59)	4.0 (0.80)	3.7 (0.71)	0.545		
Appropriateness (range 1-5)	4.4 (0.70)	4.3 (0.73)	4.3 (0.75)	4.5 (0.62)	4.3 (0.75)	4.5 (0.52)	3.6 (0.88)	0.006*		

<sup>a</sup>RECs: recommendations.<sup>b</sup>ChatGPT total RECs as the denominator.<sup>c</sup>NCCN total RECs as the denominator.\*Significant  $p < 0.05$ .



**Figure 3.** ChatGPT-4's overall concordance rate and quality assessments using 108 prompts. This bar chart displays the concordance rates of ChatGPT-4's treatment recommendations (RECs) when queried about prostate, kidney, and bladder cancers using 108 unique prompts. Concordance rates were evaluated across four performance indicators: (1) Rater-approved ChatGPT REC ratio (based on total ChatGPT RECs). (2) NCCN-aligned ChatGPT REC ratio (based on total ChatGPT RECs). (3) ChatGPT RECs/ NCCN REC ratio. (4) NCCN-aligned ChatGPT RECs/NCCN REC ratio. Quality assessments were evaluated on a 5-point scale (1-5) in four dimensions: correctness, comprehensiveness, specificity, and appropriateness.

#### Non-specific prompt advantages:

1. A higher total number of recommendations from ChatGPT (6.9 vs. 5.0,  $p < 0.001$ ).
2. A higher ChatGPT REC/NCCN REC ratio (116.1% vs. 84.1%,  $p < 0.001$ ).
3. A greater NCCN-aligned ChatGPT RECs/NCCN REC ratio (89.7% vs. 72.4%,  $p < 0.001$ ).
4. A higher score in comprehensiveness (4.7 vs. 4.2,  $p < 0.001$ ).
5. A better specificity (4.2 vs. 3.7,  $p < 0.001$ ).

#### NCCN-specified prompt advantages:

1. A higher rater-approved ChatGPT REC ratio (91.2% vs. 85.8%,  $p = 0.011$ ).
2. A higher NCCN-aligned ChatGPT REC ratio (89.9% vs. 83.5%,  $p = 0.006$ ).
3. A higher score in correctness (4.6 vs 4.4,  $p = 0.017$ ).

The observation of the indicator with the greatest discrepancy, the ChatGPT RECs/NCCN REC ratio, between non-specific and NCCN-specified prompts (116.1% vs. 84.1%,  $p < 0.001$ ), suggests that the non-specific prompts generate more recommendations than the total NCCN RECs, while the NCCN-specified prompts offer fewer. This implies that non-specific prompts may lead to a greater quantity and a more comprehensive and specific set of recommendations. On the other hand, the responses from NCCN-specified prompts appear to be more conservative, generating fewer treatment

options than the total NCCN RECs. However, they are more clinically relevant and closely aligned with the guidelines, receiving higher approval from raters and scoring better in terms of correctness.

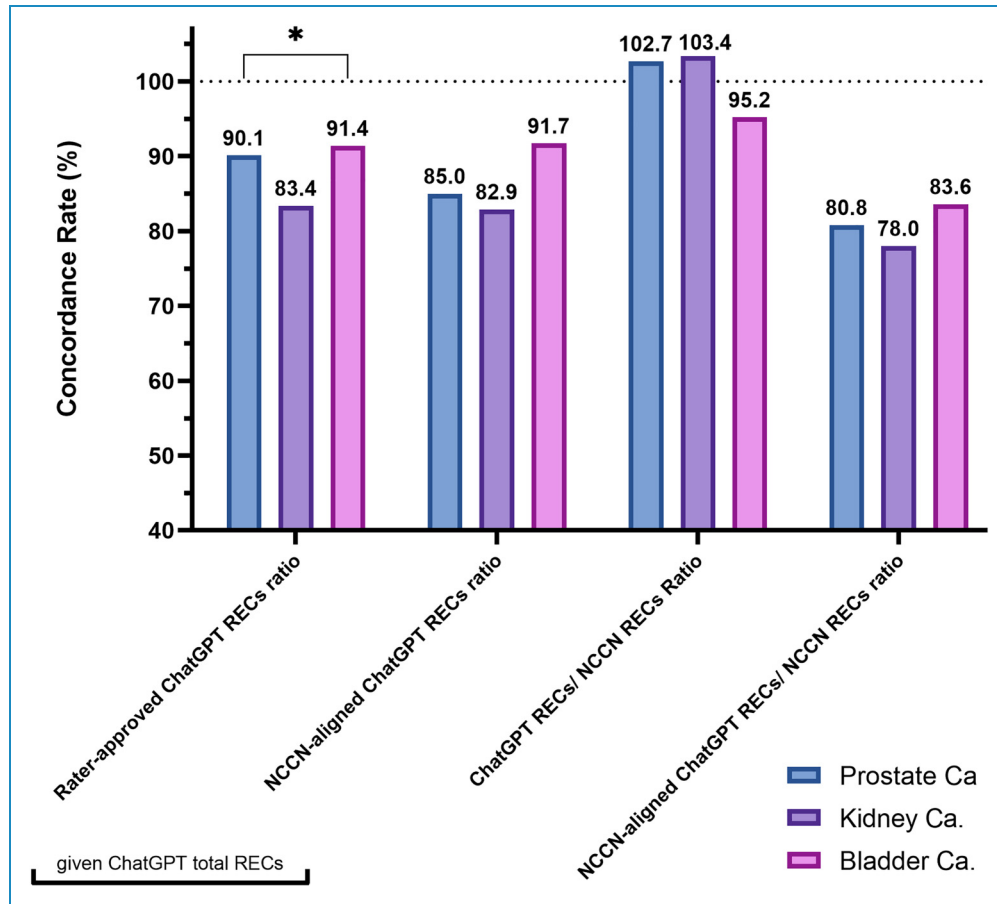
## Discussion

This study conducted a comprehensive assessment of both the quantity and quality of ChatGPT's responses, focusing on their concordance with clinical guidelines and expert opinions in the context of prostate, kidney, and bladder cancer inquiries. It addresses a research gap in the existing literature, providing insights into the capabilities and limitations of AI tools as treatment information resources for patients in the field of oncology.

A study published in August 2023 indicated that approximately one-third of cancer treatment recommendations made by ChatGPT-3.5 were only partially aligned with NCCN guidelines.<sup>6</sup> In contrast, our first-stage study, which concentrated on prostate cancer treatment recommendations generated by the AI tool, revealed that ChatGPT-4 exhibits significant improvements over ChatGPT-3.5. This advancement is evidenced by an increase in the word count of responses, a higher total RECs, and a lower ratio of disagreements among raters. Additionally, ChatGPT-4 exhibits enhanced performance in terms of alignment with NCCN guidelines. The evaluation of response quality further supports the superiority of ChatGPT-4, with higher scores in correctness, comprehensiveness, specificity, and appropriateness. These findings collectively underscore the advancements and effectiveness of ChatGPT in its evolved version.

The results from the second-stage study highlight ChatGPT-4's robustness in providing comprehensive and accurate treatment recommendations for three urological cancers, aligning well with NCCN guidelines and expert opinions, across different cancer types, prompt templates, and disease statuses. Additionally, the overall rater-disagreed REC ratio is low at 9.5%. Notably, ChatGPT-4 has demonstrated an enhancement not only in the quantity of response content but also in quality. The raters awarded scores for correctness (4.5), comprehensiveness (4.4), specificity (4.0), and appropriateness (4.4) out of a possible five points. Most responses were presented in a bullet-point format, which enhances readability and comprehension. The responses, in addition to providing treatment suggestions, frequently emphasize potential treatment risks and advise patients to seek further discussion with a healthcare professional. From the perspective of providing treatment information to patients, these recommendations are considered highly appropriate, as reflected in the high scores for appropriateness (4.4) given by raters. In conclusion, ChatGPT-4 has significantly outperformed version 3.5 in providing cancer treatment recommendations, excelling in both qualitative and quantitative assessments with an acceptable level of concordance with clinical guidelines and expert opinions.

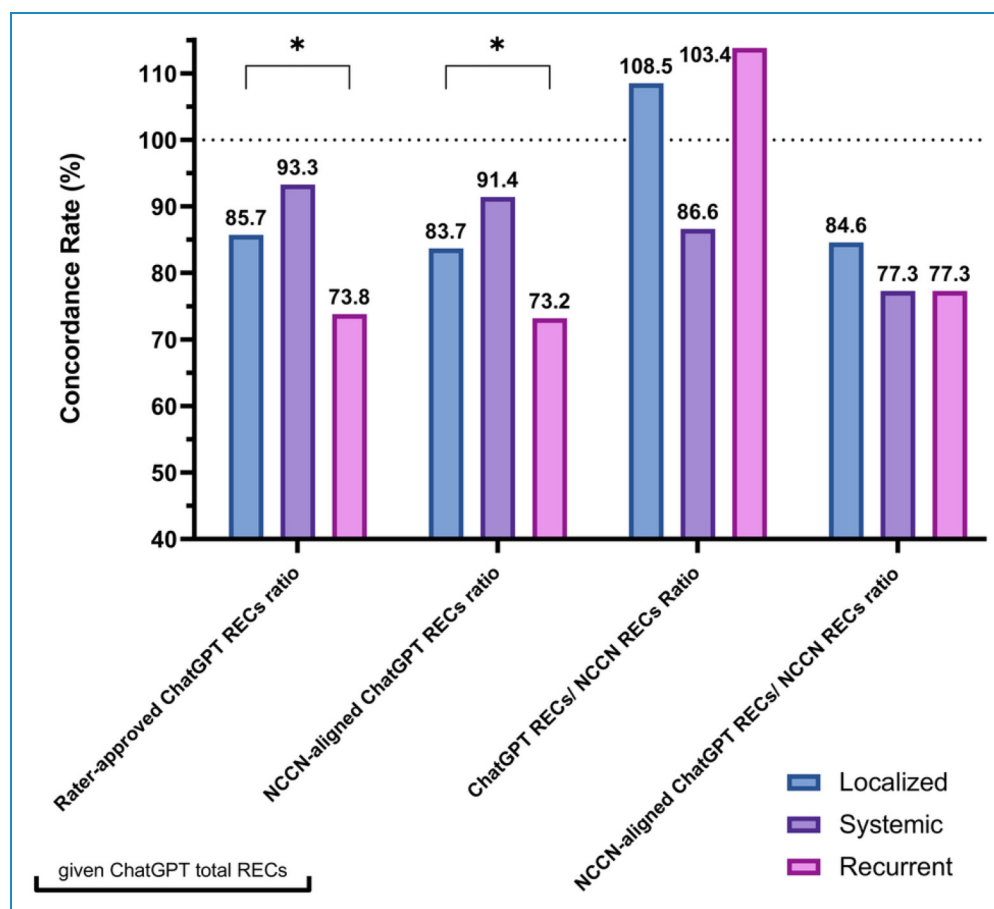




**Figure 4.** Subgroup analysis stratified by cancer type: concordance of ChatGPT-4's treatment recommendations. This bar chart displays the concordance rates of ChatGPT-4's treatment recommendations (RECs) when queried about prostate, kidney, and bladder cancers using 108 unique prompts. Concordance rates were evaluated across four performance indicators: (1) Rater-approved ChatGPT REC ratio (based on total ChatGPT RECs). (2) NCCN-aligned ChatGPT REC ratio (based on total ChatGPT RECs). (3) ChatGPT REC/ NCCN REC ratio. (4) NCCN-aligned ChatGPT REC/NCCN RECs ratio. The values in the bar chart are presented as mean values. A significant difference ( $p < 0.05$ ) between the cancer types is indicated by an asterisk (\*).

In the subgroup analysis, differential performance was observed across cancer types in the alignment of ChatGPT-4's treatment recommendations with NCCN guidelines and expert approval. Bladder cancer recommendations had the highest rater approval at 91.4% and NCCN guideline alignment at 91.7%. In contrast, kidney cancer recommendations showed a substantially lower rater approval of 83.4% and an 82.9% NCCN concordance. We hypothesize that the underlying reasons for these discrepancies may be related to the complexity of treatment recommendations in the NCCN guidelines. For kidney cancer, many treatment recommendations involve combination therapy, which requires the simultaneous use of more than one class of drugs, such as various tyrosine kinase inhibitors (TKIs) plus immunotherapy (IO) drugs, or dual IO drug treatments. These regimens indicate that treatment for kidney cancer, particularly in systemic disease states, is inherently complex and varied. This complexity could be contributing to the lower concordance of ChatGPT's recommendations for kidney cancer.

Conversely, bladder cancer systemic treatments are primarily monotherapies before 2022. In different disease statuses, treatments are sequenced from first-line to second-line and third-line therapies. This relative simplicity could explain the higher alignment of ChatGPT's recommendations with guidelines for bladder cancer. Therefore, it is suggested that as the complexity of guideline-recommended treatments increases, such as with combination therapies, the performance of ChatGPT is adversely affected. Additionally, similar patterns were observed in the subgroup analysis for disease status. Specifically, for recurrent cancer status, there was a notable decrease in rater approval at 73.8% and NCCN alignment at 73.2%, both significantly lower compared to localized and systemic cancer status. The treatment for recurrent cancer is inherently complex, necessitating considerations of second-line and third-line treatment options and sequences. It appears that ChatGPT's performance declines as the complexity and variability of treatment strategies for recurrent cancers increase.



**Figure 5.** Subgroup analysis stratified by disease status: concordance of ChatGPT-4's treatment recommendations. This bar chart displays the concordance rates of ChatGPT-4's treatment recommendations (RECs) when queried about localized, systemic, and recurrent cancers using 108 unique prompts. Concordance rates were evaluated across four performance indicators. The values in the bar chart are presented as mean values. A significant difference ( $p < 0.05$ ) between the disease statuses is indicated by an asterisk (\*).

This study was solely focused on evaluating ChatGPT as a treatment information resource for patients. Typically, this usage is initiated autonomously by patients themselves, rather than through provision by medical professional organizations. Consequently, this research did not formally assess whether such a general-purpose LLM is suitable for deployment in clinical settings. However, findings from the subgroup analysis, which indicated a decline in performance in complex treatment scenarios, suggest that these general-purpose models may not yet be suitable for use by healthcare professionals to manage complex treatments.

The overall ratio of rater-disagreed recommendations (RECs) generated by ChatGPT-4 is low at 9.5%. Upon further examination of these rater-disagreed RECs, it was discovered that some treatment options listed as 'Recommendations' were deemed inappropriate by experts; however, ChatGPT's descriptive content was not entirely incorrect but may be partially correct. Such treatment options should potentially be categorized as alternative options under special considerations, rather than as primary recommendations. For example, in response to the

K01 disease description query regarding kidney cancer (Supplement File S6):

What is the recommended treatment for stage I (T1a) kidney cancer primary treatment?

one REC listed by ChatGPT-4 was:

4. Targeted Therapy or Immunotherapy: While not typically the first line of treatment for stage I kidney cancer, in some cases where surgery is not an option or cancer has specific characteristics, these therapies might be considered.

Experts did not agree with this treatment option being categorized as a primary REC. Nevertheless, the explanation explicitly states that it is 'not typically the first line of treatment' and specifies that it might be considered under special conditions. This attached explanation is partially

**Table 3.** Comparison of ChatGPT-4's performance using prompts with and without the specification "according to NCCN."

ChatGPT-4	Prompt template						p-value	
	All		Non-specific prompt <sup>b</sup>		NCCN-specified prompt <sup>c</sup>			
	Mean	(SD)	Mean	(SD)	Mean	(SD)		
Query prompts (n)	108		54		54			
ChatGPT total RECs <sup>a</sup>	6.0	(1.92)	6.9	(1.67)	5.0	(1.68)	<0.001**	
Rater-approved cGPT REC ratio % <sup>d</sup>	88.5	(14.8)	85.8	(15.9)	91.2	(13.3)	0.011*	
NCCN-aligned cGPT REC ratio % <sup>d</sup>	86.7	(16.1)	83.5	(16.7)	89.9	(15.0)	0.006**	
Rater-disagreed cGPT REC ratio % <sup>d</sup>	9.5	(13.7)	11.7	(14.8)	7.4	(12.3)	0.020**	
NCCN total RECs	6.0	(2.18)	6.0	(2.18)	6.0	(2.18)		
ChatGPT REC/NCCN REC ratio % <sup>e</sup>	100.0	(40.5)	116.1	(39.6)	84.1	(35.0)	<0.001**	
NCCN-aligned ChatGPT REC/NCCN REC ratio % <sup>e</sup>	81.0	(20.6)	89.7	(15.5)	72.4	(21.6)	<0.001**	
Correctness	(range 1-5)	4.5	(0.65)	4.4	(0.69)	4.6	(0.59)	0.017*
Comprehensiveness	(range 1-5)	4.4	(0.70)	4.7	(0.41)	4.2	(0.81)	<0.001**
Specificity	(range 1-5)	4.0	(0.71)	4.2	(0.57)	3.7	(0.73)	<0.001**
Appropriateness	(range 1-5)	4.4	(0.70)	4.4	(0.72)	4.3	(0.68)	0.640

<sup>a</sup>RECs: recommendations.

<sup>b</sup>Prompt without "according to NCCN".

<sup>c</sup>Prompt with "according to NCCN".

<sup>d</sup>ChatGPT total RECs as the denominator.

<sup>e</sup>NCCN total RECs as the denominator.

\* Significant  $p < 0.05$ ; \*\* Significant  $p < 0.01$ .

correct rather than entirely erroneous. Therefore, raters in the study assessed these RECs using strict standards.

Inevitably, patients are drawn to seek advice from AI technologies like ChatGPT for treatment recommendations due to their immediate availability and user-friendly interface.<sup>13</sup> Despite improvements over its predecessor, ChatGPT-4 is not flawless. Responses occasionally may be a mix of correct and incorrect advice. With an error rate of 9.5%, non-experts may struggle to identify inaccurate recommendations, posing a significant ethical concern regarding patient safety and the risk of reliance on AI for critical health decisions. The potential for misinformed decisions is of particular concern for patients relying solely on AI tools for treatment recommendations. This highlights the need for systems that ensure AI tools are used in conjunction with the expertise of healthcare professionals, emphasizing the importance of careful integration

of such technologies in healthcare. Acknowledging this, while ChatGPT-4 has demonstrated its utility as an auxiliary source of information, it must not replace personalized advice from healthcare professionals. We must critically examine the ethical implications of deploying AI tools in patient treatment information, especially since they cannot yet emulate the delicate judgments of trained professionals.

Moving forward, the development of medical-specific AI models that are finely tuned to provide reliable medical advice is crucial. Future research should focus on minimizing risks and enhancing the safe application of AI in patient care, ensuring that such tools are developed with a stringent emphasis on ethical standards and patient safety. To safely integrate AI recommendations into clinical practice, it is essential to establish practical guidelines that include rigorous validation by healthcare professionals. These professionals should play a pivotal role in interpreting and applying AI advice,

ensuring that it complements their clinical expertise and aligns with individual patient needs. The commitment to ethical AI use in healthcare is paramount to advancing patient safety and ensuring responsible AI integration into treatment information for patients and clinical decision-support systems for healthcare professionals.

The limitations of this study include its focus on only three urological cancers, which may not adequately reflect the diversity of oncological diseases. Consequently, this limits the generalizability of the findings. We recommend that future research explore the performance of ChatGPT across a broader range of cancer types to validate and possibly extend these results. Secondly, during the study period, the evaluation protocol used the 2021 NCCN guidelines, corresponding to the GPT-3.5 turbo's training cutoff in September 2021. This reliance could bias the concordance measures against the most recent treatment standards. However, despite the continuous updates to the NCCN guidelines, the training dataset for ChatGPT-4 turbo was also continuously updated until 2023. Thirdly, given the rapid progression of AI tools, the findings from this cross-sectional research may only represent the model's performance at that specific time.

## Conclusions

ChatGPT-4 has significantly outperformed version 3.5 in providing cancer treatment recommendations, excelling in both qualitative and quantitative assessments with an acceptable level of concordance with clinical guidelines and expert opinions.

With the widespread application of AI, there is a pressing need to ensure that AI tools are safe, reliable, and beneficial for patients. Therefore, there is an essential need for close collaboration between AI developers and healthcare professionals to create medical-specific, finely tuned versions that could offer more dependable advice in the future.

At present, it is vital for patients to recognize that although these tools are advancing, they are not without flaws and should be utilized with caution. ChatGPT could supplement, but not replace, personalized advice from healthcare professionals. With the rapid evolution of language models, there is a pressing need to create medical-specific, finely-tuned versions that could offer more dependable advice in the future.

**Acknowledgements:** We would like to acknowledge the efforts of the three independent raters who evaluated ChatGPT's responses.

**Contributorship:** CYT contributed to the conception and design, acquisition of data, analysis, and interpretation of data, drafting of the manuscript, and statistical analysis. PYC contributed to disease prompt design, analysis, and visualization of data, and JHD and FSJ contributed to supervision. SCY contributed to disease prompt design, critical revision of the manuscript, statistical analysis, and supervision. All authors collected the data.

**Data availability statement:** The datasets generated and/or analyzed during the current study are available in the online Supplemental Files (S1 to S7).

**Declaration of conflicting interests:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethical approval:** No human participants were involved in the study. According to the Regulations on Human Trials by the Ministry of Health and Welfare in Taiwan, the IRB Board only handles matters related to human subjects and does not accept applications for ethical review when the study does not involve human participants. Also, in accordance with the IRB Review Regulations from the Ethics Center of National Taiwan University Hospital (<https://www.ntuh.gov.tw/RECO/Fpage.action?muid=5018&fid=5536>), the study outside the scope of human research does not need to be submitted to the Research Ethics Committee as per the above regulations. All experiments were performed in accordance with relevant guidelines and regulations.

**Funding:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

**Guarantor:** S.C. Yii.

**ORCID iDs:** Chung-You Tsai  <https://orcid.org/0000-0003-0527-1830>

Shyi-Chun Yii  <https://orcid.org/0000-0002-1044-9413>

**Supplemental material:** Additional Supplemental Files can be downloaded from the online version of this article. The Supplemental Material includes

- S1. Cancer Treatment advised by ChatGPT\_Google Search Results.pdf:  
Numerous media reports that cast doubt on ChatGPT's suitability as a cancer treatment advisor in Google Search results
- S2. Scoring Guidelines for Raters.pdf:  
Predefined scoring guidelines for three raters to evaluate ChatGPT responses, including quantitative and qualitative assessment of cancer treatment recommendations (RECs)
- S3. Scoring Questionnaire for Raters.pdf:  
Questionnaire for three raters to evaluate ChatGPT responses, including quantitative and qualitative assessment of recommended cancer treatment RECs
- S4. ChatGPT\_4 vs 3.5\_responses\_to\_32\_query\_prompts.xlsx:  
This file records ChatGPT-4 and ChatGPT-3.5's responses to 32 query prompts regarding prostate cancer treatment. To view the complete screenshots of ChatGPT's outputs, click the PDF file hyperlinks in the Excel file.
- S5. ChatGPT\_4 vs 3.5\_assessment\_data by 3\_Raters.xlsx:  
Three Raters' scoring results for ChatGPT-4 and ChatGPT-3.5's responses to 32 query prompts regarding prostate cancer treatment RECs
- S6. ChatGPT-4\_responses\_to\_108\_query\_prompts.xlsx

This file records ChatGPT-4's responses to 108 query prompts regarding prostate, kidney, and bladder cancers. To view the complete screenshots of ChatGPT's outputs, click the PDF file hyperlinks in the Excel file.

- S7. ChatGPT-4\_assessment\_data by 3\_Raters.xlsx  
 Three Raters' scoring results for ChatGPT-4's responses to 108 query prompts regarding prostate, kidney, and bladder cancer treatment RECs

## References

- Gordijn B and Have HT. ChatGPT: evolution or revolution? *Med Health Care Philos* 2023; 26: 1–2.
- Talyshinskii A, Naik N, Hameed BZ, et al. Expanding horizons and navigating challenges for enhanced clinical workflows: ChatGPT in urology. *Front Surg* 2023; 10: 1257191.
- Tung JY, Lim DY and Sng GG. Potential safety concerns in use of the artificial intelligence chatbot 'ChatGPT' for perioperative patient communication. *BJU Int* 2023; 132: 157–159.
- Zhu L, Mou W and Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med* 2023; 21: 1–4.
- Deebel NA and Terlecki R. ChatGPT performance on the American Urological Association (AUA) Self-Assessment Study Program and the potential influence of artificial intelligence (AI) in urologic training. *Urology* 2023; 177: 29–33.
- Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence Chatbots for cancer treatment information. *JAMA Oncol* 2023; 9: 1459–1462.
- National Comprehensive Cancer Network. <http://www.nccn.org> (2023, accessed June 4).
- Pan A, Musheyev D, Bockelman D, et al. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncol* 2023; 9: 1437–1440.
- Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *J Med Internet Res* 2023; 25: e47479.
- OpenAI. ChatGPT, <https://chat.openai.com/chat> (2023).
- National Comprehensive Cancer Network (NCCN) guidelines. <http://www.nccn.org> (accessed July 15, 2023).
- OpenAI. Models overview, <https://platform.openai.com/docs/models/continuous-model-upgrades> (2023).
- Arora VM, Madison S and Simpson L. Addressing medical misinformation in the patient-clinician relationship. *Jama* 2020; 324: 2367–2368.