Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: A Multi-Metric Assessment

Damien Gibson, Stuart Jackson, Ramesh Shanmugasundaram, Ishith Seth, Adrian Siu, Nariman Ahmadi, Jonathan Kam, Nicholas Mehan, Ruban Thanigasalam, Nicola Jeffery, Manish I Patel, Scott Leslie

Submitted to: Journal of Medical Internet Research
on: January 05, 2024

Table of Contents

Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: A Multi-Metric Assessment

Damien Gibson[1, 2, 3*] BBiomedSc(Hons), MD, MS; Stuart Jackson[4*] BEd(Health), BSc(Chiro), MD, MS; Ramesh Shanmugasundaram[1, 2] MD, MS; Ishith Seth[5] BBiomed(Hons), MD, MS; Adrian Siu[3, 6] BPharm(Hons), MD, MS; Nariman Ahmadi[7, 8] B.Sc.(Med), M.B.B.S., M.S.(Urol), F.R.A.C.S.(Urol); Jonathan Kam[8] BMed, BSci(Med)Hons, MD, F.R.A.C.S.(Urol); Nicholas Mehan[9] MD, F.R.A.C.S.(Urol); Ruban Thanigasalam[7, 8] MBBS, MS, FRACS (Urology); Nicola Jeffery[7, 8] MBBS, BSc(Med), FRACS(Urology); Manish I Patel[10] MMed, MBBS, PhD, FRACS(Urology); Scott Leslie[3, 4, 7, 8] BSc (Med), MBBS (Hons), FRACS (Urology)

[1]Department of Urology Saint George Hospital Kogarah AU
[2]Faculty of Medicine The University of New South Wales Sydney AU
[3]Surgical Outcomes Research Centre Sydney AU
[4]Faculty of Medicine University of Sydney Sydney AU
[5]Department of Surgery Peninsula Health Victoria AU
[6]Concord Institute of Academic Surgery Concord Hospital Sydney AU
[7]Department of Urology Chris O'Brien Lifehouse Sydney AU
[8]RPAH Institute of Academic Surgery Royal Prince Alfred Hospital Sydney AU
[9]Nepean Urology Research Group Nepean Hospital Sydney AU
[10]Department of Urology Westmead hospital Sydney AU
[*]these authors contributed equally

Corresponding Author:
Damien Gibson BBiomedSc(Hons), MD, MS
Department of Urology
Saint George Hospital
Gray St
Kogarah
AU

Abstract

Background: Artificial intelligence (AI) chatbots like ChatGPT have made significant progress. These chatbots, particularly popular among healthcare professionals and patients, are transforming patient education and disease experience with personalized information. They are especially beneficial for populations such as men with prostate cancer concerns. Accurate, timely patient education is crucial for informed decision-making, especially regarding Prostate-Specific Antigen screening and treatment options. AI chatbots can address the gap in quality prostate cancer information, reaching wider demographics, including remote communities. However, the accuracy and reliability of AI chatbots' medical information must be rigorously evaluated. Studies testing ChatGPT's knowledge in prostate cancer are emerging, but there's a need for ongoing evaluation to ensure the quality and safety of information provided to patients.

Objective: To evaluate the quality, accuracy, and readability of ChatGPT-4's responses to common prostate cancer questions posed by patients.

Methods: Eight questions were formulated with an inductive approach. These were based on information topics searched for and desired by prostate cancer patients in peer reviewed literature, and Google Trends data. The eight artificial intelligence (AI) outputs were judged by seven expert urologists, using an assessment framework developed to assess accuracy, safety, appropriateness, actionability and effectiveness. Adapted versions of the Patient Education Materials Assessment Tool (PEMAT-AI), Global Quality Score (GQS), and DISCERN (DISCERN-AI) tools were used by four independent reviewers to assess the quality of the AI responses. Readability of the AI responses was assessed using established algorithms (Flesch Reading Ease score, Gunning Fog Index, Flesch-Kincaid Grade Level, The Coleman-Liau Index and SMOG Index). A brief tool (REF-AI) was developed for analysis of the references provided by AI outputs, assessing for reference hallucination, relevance, and quality of references.

Results: PEMAT-AI understandability score was very good (mean 79.44%, SD 10.44), GQS was rated as high (mean 4.46/5, SD 0.50), and DISCERN-AI rating of moderate quality (mean 13.88, SD 0.93). NLAT-AI pooled means (SD) included accuracy 3.96 (0.91), safety 4.32 (0.86), appropriateness 4.45 (0.81), actionability 4.05 (1.15), and effectiveness 4.09 (0.98). Readability algorithm consensus was “difficult to read” (Flesch Reading Ease score mean 45.97 SD 8.69, Gunning Fog Index mean 14.55 SD 4.79), averaging a grade 11 reading level, equivalent to 15 – 17-year-old (Flesch-Kincaid Grade Level mean 12.12 SD 4.34, The Coleman-Liau Index mean 12.75 SD 1.98, SMOG Index mean 11.06 SD 3.20). REF-AI identified two reference hallucinations, while the majority of references appropriately supplemented the text. Most references were from reputable government organizations, while a handful were direct citations from scientific literature.

Conclusions: Our analysis found that ChatGPT-4 provides generally good responses to common prostate cancer queries, making it a potentially valuable tool for patient education in prostate cancer care. Objective quality assessment tools indicated that the natural language processing (NLP) outputs were generally reliable and appropriate, but there is room for improvement.

(JMIR Preprints 05/01/2024:55939)
DOI: https://doi.org/10.2196/preprints.55939

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper: Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: A Multi-Metric Assessment

Keywords: Prostate Cancer, Patient Education, Large Language Model, ChatGPT, AI language model

Authors: **Damien Gibson[1,2,3],** Stuart Jackson[4], Ramesh Shanmugasundaram[1,2], Ishith Seth[5], Adrian Siu[3,6], Nari Ahmadi[7, 8], Jonathan Kam[9], Nicholas Mehan[9], Ruban Thanigasalam[7, 8], Nicola Jeffery[7, 8], Manish I Patel[4, 10], Scott Leslie [3, 4, 7, 8]

Affiliations:

[1] Department of Urology, St George Hospital, Sydney, NSW, Australia
[2] The University of New South Wales, Sydney, NSW, Australia
[3] Surgical Outcomes Research Centre, Sydney, NSW, Australia
[4] University of Sydney, Sydney, NSW, Australia
[5] Department of Surgery, Peninsula Health, Victoria, Australia
[6] Concord Institute of Academic Surgery, Concord Hospital, Sydney, Australia, 2139.
[7] Chris O'Brien Lifehouse, Sydney, NSW, Australia
[8] RPAH Institute of Academic Surgery, Sydney, NSW, Australia
[9] Nepean Urology Research Group, Sydney, NSW, Australia
[10] Department of Urology, Westmead hospital, Sydney, NSW, Australia

Corresponding Author Information:
Dr. Damien Gibson, Faculty of Medicine, University of New South Wales, University of New South Wales, 2200, Australia. Email: Damien.gibson@unsw.edu.au

Abstract

Background: Artificial intelligence (AI) chatbots like ChatGPT have made significant progress. These chatbots, particularly popular among healthcare professionals and patients, are transforming patient education and disease experience with personalized information. They are especially beneficial for populations such as men with prostate cancer concerns. Accurate, timely patient education is crucial for informed decision-making, especially regarding Prostate-Specific Antigen screening and treatment options. AI chatbots can address the gap in quality prostate cancer information, reaching wider demographics, including remote communities. However, the accuracy and reliability of AI chatbots' medical information must be rigorously evaluated. Studies testing ChatGPT's knowledge in prostate cancer are emerging, but there's a need for ongoing evaluation to ensure the quality and safety of information provided to patients.

Objective: To evaluate the quality, accuracy, and readability of ChatGPT-4's responses to common prostate cancer questions posed by patients.

Methods: Eight questions were formulated with an inductive approach. These were based on information topics searched for and desired by prostate cancer patients in peer reviewed literature, and Google Trends data. Adapted versions of the Patient Education Materials Assessment Tool (PEMAT-AI), Global Quality Score (GQS), and DISCERN (DISCERN-AI) tools were used by four independent reviewers to assess the quality of the AI responses. The eight artificial intelligence (AI) outputs were judged by seven expert urologists, using an assessment framework developed to assess accuracy, safety, appropriateness, actionability and effectiveness. Readability of the AI responses was assessed using established algorithms (Flesch Reading Ease score, Gunning Fog Index, Flesch-Kincaid Grade Level, The Coleman-Liau Index and SMOG Index). A brief tool (REF-AI) was developed for analysis of the references provided by AI outputs, assessing for reference hallucination, relevance, and quality of references.

Results: PEMAT-AI understandability score was very good (mean 79.44%, SD 10.44), DISCERN-AI rating of moderate quality (mean 13.88, SD 0.93) and GQS was rated as high (mean 4.46/5, SD

0.50). NLAT-AI pooled means (SD) included accuracy 3.96 (0.91), safety 4.32 (0.86), appropriateness 4.45 (0.81), actionability 4.05 (1.15), and effectiveness 4.09 (0.98). Readability algorithm consensus was “difficult to read” (Flesch Reading Ease score mean 45.97 SD 8.69, Gunning Fog Index mean 14.55 SD 4.79), averaging a grade 11 reading level, equivalent to 15 – 17-year-old (Flesch-Kincaid Grade Level mean 12.12 SD 4.34, The Coleman-Liau Index mean 12.75 SD 1.98, SMOG Index mean 11.06 SD 3.20). REF-AI identified two reference hallucinations, while the majority of references appropriately supplemented the text. Most references were from reputable government organizations, while a handful were direct citations from scientific literature.

Conclusions: Our analysis found that ChatGPT-4 provides generally good responses to common prostate cancer queries, making it a potentially valuable tool for patient education in prostate cancer care. Objective quality assessment tools indicated that the natural language processing (NLP) outputs were generally reliable and appropriate, but there is room for improvement.

Keywords: Prostate Cancer, Patient Education, Large Language Model, ChatGPT, AI language model

Introduction

Artificial intelligence (AI) chatbots have made significant strides in recent years[1]]. This was emphatically signposted with the launch of Chat Generative Pre-Trained Transformer (ChatGPT) 3 in November 2022 (OpenAI, California), with ChatGPT becoming the most popular online tool for both patients and healthcare professionals [2, 3]. Now in its 4th iteration (ChatGPT-4), the AI language model can generate responses to a wide range of health questions and topics[4]. AI chatbots such as ChatGPT have the potential to significantly impact patient education and disease experience by providing reliable, accessible, and personalized information [4, 5]. One patient population who stands to benefit from this are men who are concerned about prostate cancer.

With the rising prevalence of prostate cancer globally, accounting for an estimated 1 414 259 new cases and over 375 304 deaths in 2020 alone—there is an urgent need for accurate and timely patient education information[6]. The rate of prostate cancer survivorship is increasing, but this comes with its own challenges such as escalating healthcare costs and large numbers of survivors requiring ongoing care [3]. In this context, shared decision-making becomes pivotal, particularly concerning Prostate-Specific Antigen (PSA) screening and prostate cancer treatment selection [3]. Given the various treatments available, management decisions can be greatly influenced by a patients' understanding of anatomical, functional and psychological impacts of treatment[7]. Side effects like urinary incontinence and erectile dysfunction can severely affect a patient's quality of life, necessitating well-informed patients, and treatment choices [8]. Furthermore, patient education has been shown to minimise psychological impacts such as depression and treatment regret[9].

There are well documented issues with unmet information needs of both men and their support networks, throughout the prostate cancer care continuum[10]. This includes challenges related to information quality and readability [11]. The assessment of online healthcare information in prostate cancer has been well described through multiple domains including webpage articles, YouTube and social media [10]. The internet is now often the first source of information for men (and their stakeholders) seeking answers about diagnosis, treatment, and prognosis [8]. Despite this trend, most long-term literature suggests that online health information is of moderate to poor quality [10-12].

AI chatbots are a potential solution to fill the prostate cancer information quality gap[2]. Given their scalability, AI chatbots can reach a wide demographic, including those in remote or underserved communities where medical resources are scarce [2]. Natural language processing technologies (NLPTs) enable these platforms to present complex jargon in patient specific terms, with the potential to address eHealth literacy variability, and to enhance patient understanding[13]. Such platforms are also able to do this across a diverse number of languages[14]. Despite these qualities, the accuracy and reliability of AI chatbot medical information must still be assessed using rigorous evaluation tools. Only a handful of studies have begun to test ChatGPT's applicability in prostate cancer, one testing its knowledge directly with questions and statements [15] and another assessing its appropriateness in screening recommendations[16]. However, a significant knowledge gap persists in understanding the quality and safety of information patients receive from ChatGPT-4 for common internet queries. Ongoing evaluation is a necessary step to build healthcare provider confidence in these new technologies, while ensuring that patients have access to vetted and safe healthcare and educational information.

In this study, we aim to demonstrate and assess the quality of ChatGPT responses to commonly asked patient education topics in prostate cancer care. By doing so, this study seeks to:

1) Illustrate for clinicians whether ChatGPT-4 is currently a reliable and safe patient education tool for prostate cancer information.
2) Provide clinicians with a greater understanding of the current strengths and limitations of health-based queries which patients are likely to encounter when using technologies such as ChatGPT-4.

A range of assessment tools will be applied to the AI generated responses to assess output quality, safety, understandability, actionability, ease of use, readability, and reliability. A parallel assessment of the outputs by prostate cancer experts will also be conducted.

Methods:

Question/keyword strategy

Questions tested with the AI chatbot model (ChatGPT 4) were selected through an iterative process of literature and Google Keyword analysis. Literature concerning information needs of men considering prostate cancer investigation and treatment were reviewed to determine the most common information topics and prostate cancer questions of interest to men[10, 17-20]. Subsequently, worldwide Google Trends data was analysed to provide a more current public measure of prostate cancer information searches[21]. Using “prostate cancer” as a keyword, both rising and top ‘related topics’ and ‘related queries’ of the past year were collected. Finally, whilst limited to training materials up to 2021, ChatGPT was itself queried, asking ‘What are the most common prostate cancer questions asked to ChatGPT?’’ (Appendix 1). Two authors thematically analysed this information to define the following eight questions to discuss with the AI model:

1. What are the symptoms of prostate cancer?
2. What are the risk factors for prostate cancer?
3. What is the survival rate of prostate cancer?
4. How is prostate cancer diagnosed?
5. What age should men start getting screened for prostate cancer?
6. What are the pros and cons of treatment options for prostate cancer?
7. How does prostate cancer affect sexual function?
8. How does prostate cancer affect bladder function?

Each question was posed to ChatGPT-4 with an additional request for references. A new ChatGPT account was established with a novel email address for each prompt in effort to reduce potential effects of each response on subsequent outputs of the AI model. Each output was recorded for individual quality and readability assessment (Appendix 2).

Quality assessment

Due to a current absence of tools to evaluate the quality of AI natural language outputs, each conversation was evaluated using modified versions of pre-existing information quality assessment tools. These included the Patient Education Materials Assessment Tool (PEMAT) and DISCERN criteria [22, 23]. These tools were iteratively modified to accommodate the text only nature and characteristics of AI natural language outputs. While DISCERN criteria has been adapted in literature, the PEMAT modification is new [24]. Internal validity testing was undertaken by four reviewers using ChatGPT outputs from similar question sets for breast cancer and bowel cancer. The reliability of each tool tested was satisfactory, with a Cronbach alpha >0.8 (DISERN 0.852, PEMAT 0.82, GQS 0.85). The Global Quality Score (GQS) was not modified[25].

PEMAT-AI

The PEMAT tool evaluates and compares the understandability and actionability of patient education materials [23]. The tool incorporates seventeen items measuring understandability and seven assessing actionability, these were reduced to eight and three, respectively, to suit the AI text-only outputs (Appendix 4). Each item was given a single point if criteria were met, the total score was measured as a total percentage. Final scores were recorded as ‘pass’ or ‘fail’ based on the ≥70% cut-off score set by the PEMAT guidelines [23].

DISCERN-AI

The DISCERN criteria is a previously validated tool which aids health care consumers and health practitioners appraise the quality of health care treatment information [22, 23]. To address the AI output, this criteria was modified to seven questions (of original 15) on a scale of one to three, utilising questions 3-9 (Appendix 3). Based on previous DISCERN quality assessment in the literature, each output was scored as very poor (6), poor (7-9), fair (8-12), good (13-15) and excellent (16-18) quality patient education material [26, 27].

Global quality scale

The GQS is a five-point Likert scale based on the quality of information, and the flow and ease of use of information presented online. The GQS encompasses a scale of 1 to 5; where 1 indicates "low quality" and 5 implies "high quality". Results that received a score of 4 or 5 were rated high quality, those with a score of 3 were assessed as medium quality, and the ones with a score of 1 or 2 were categorized as low quality[24, 28].

Readability

Readability of the AI responses was assessed using a battery of established algorithms: Flesch Reading Ease score, Gunning Fog Index, Flesch-Kincaid Grade Level, The Coleman-Liau Index and SMOG Index[29-32]. Multiple tools were used in an effort to limit bias of each respective algorithm[33]. Each AI output text was copied to Microsoft Word to maintain formatting, and then to Readable.com for analysis[34, 35]. Results from the answered questions were averaged across all outputs into a readability consensus [36]. The Flesch Reading Ease score gauges text simplicity, where score ranges from 0 to 100, with higher scores indicating easier readability. Texts with a score between 60-70 are generally considered to be at an 8th- to 9th-grade reading level and are usually easier for the average adult to read. The Gunning Fog Index and The Flesch-Kincaid Grade Level measure sentence complexity, the score represents the number of years of formal education a reader would need to understand the text on the first reading. As example, a score of 12 would mean the text is suitable for a 12th-grade reading level or higher. The Coleman-Liau Index is similar to Gunning Fog and The Flesch-Kincaid Grade Level but focuses on character count. This score also correlates with a U.S. school grade level but is calculated using the number of characters instead of syllables, making it more suited for languages where syllable count is less indicative of complexity. The SMOG Index evaluates syllable density to assess readability and is often used for checking health messages. A score of 12 would mean the text is suitable for someone with at least a 12th-grade level of reading comprehension.

Natural Language Assessment Tool for AI (NLAT-AI)

Expert review of each output was undertaken by seven independent experienced urologists, using an assessment framework (NLAT- AI) developed to assess the accuracy, safety, appropriateness, actionability, and effectiveness of information. Each domain was scored as a five-point Likert scale (1 = strongly disagree, to 5 = “strongly agree”) (Appendix 5). All results were collated and presented as descriptive statistics. Qualitative feedback on each domain was sought regarding potential improvement and overall performance.

References assessment

Due to known issues of AI hallucination: “the phenomenon of a machine, such as a chatbot, generating seemingly realistic sensory experiences that do not correspond to any real-world input”, a final brief tool (REF-AI) was developed for analysis of the references provided by AI outputs[37]. Each reference was reviewed by accessing the content via the direct link provided by the AI output, or a Google search of the reference. This tool assessed for reference hallucination (real or not),

relevance (correlation between the references and AI output), and quality of references (type of institution linked to the reference). Each criterion was assessed with a score of 1-3, with a lower summative score indicating lower reference quality, and a higher score indicating high reference quality (Appendix 6.0). Scores were averaged to yield a composite score for each axis of evaluation. Reliability of this tool tested similar question sets for breast cancer and bowel cancer was satisfactory (0.81).

Ethical Consideration

After consultation with the local institutional review board, it was determined that no formal ethical approval was required for this study as no human or animal participants were involved.

Results

ChatGPT Outputs

The responses generated by the artificial intelligence model, ChatGPT-4, provided broad, medically aligned information (Appendix 2.0). The assessment of the ChatGPT-4 output utilising PEMAT-AI, DISCERN-AI and GQS patient education material assessment tools demonstrated high results across all tools. The pooled PEMAT-AI understandability score easily passed the acceptability threshold of >70% (mean 79.44%, SD 10.44), only question three failed the >70% threshold at 66.67% while the remaining were 76% or greater (Figure 1). The pooled DISCERN-AI rating scored as “good” quality 77% (mean 13.88, SD 0.93), all individual questions rated “good” on the DISCERN-AI except for question five which scored excellent (mean 15.67) (Figure 2). The pooled GQS rated as high (mean 4.46/5, SD 0.50) (Figure 3). Assessment tool results for each question are tabulated and graphed (Table 1, Figure 1-3). Reliability testing was high with a Cronbach alpha 0.846.

Table 1 – Quality assessment tools

	PEMAT-AI mean (SD)	DISERN-AI mean (SD)	GQS mean (SD)
Q1) Symptoms	79.37 (18.03)	13.67 (0.58)	4.67 (0.58)
Q2) Risk Factors	85.51 (2.10)	12.67 (0.58)	4.33 (0.58)
Q3) Survival Rates	66.67 (8.25)	14.00 (0.00)	5.00 (0.00)
Q4) Diagnosis	84.92 (14.35)	13.67 (0.58)	4.33 (0.58)
Q5) Screening	74.60 (9.91)	15.67 (1.15)	4.67 (0.58)
Q6) Treatment	79.36 (10.99)	14.00 (0.00)	4.33 (0.58)
Q7) Sexual Function	84.92 (1.37)	13.67 (0.58)	4.33 (0.58)
Q8) Bladder Function	80.16 (7.65)	13.67 (0.58)	4.00 (0.00)
Total	79.44 (10.44)	13.88 (0.93)	4.46 (0.50)

Figure 1 – PEMAT-AI

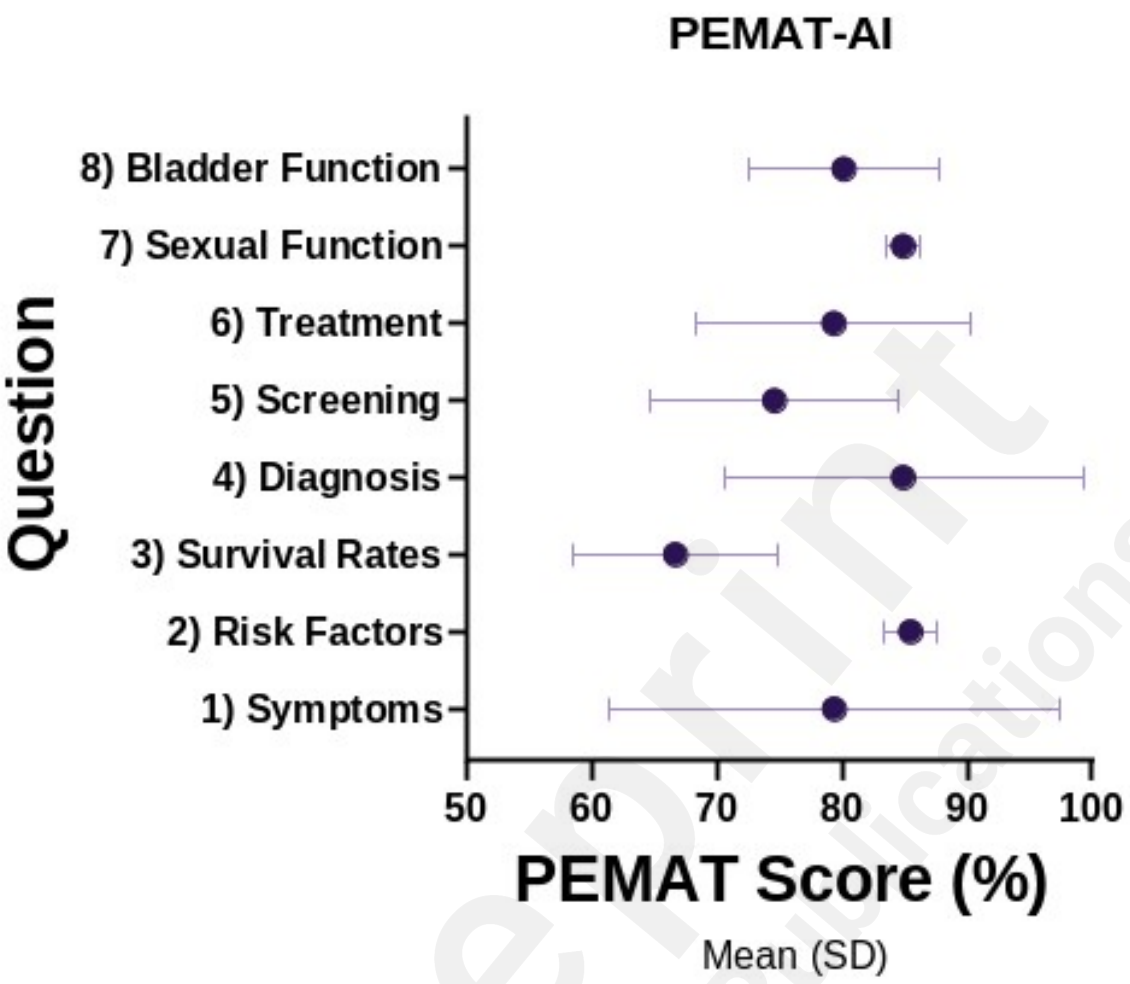


Figure 2 – DISERN-AI

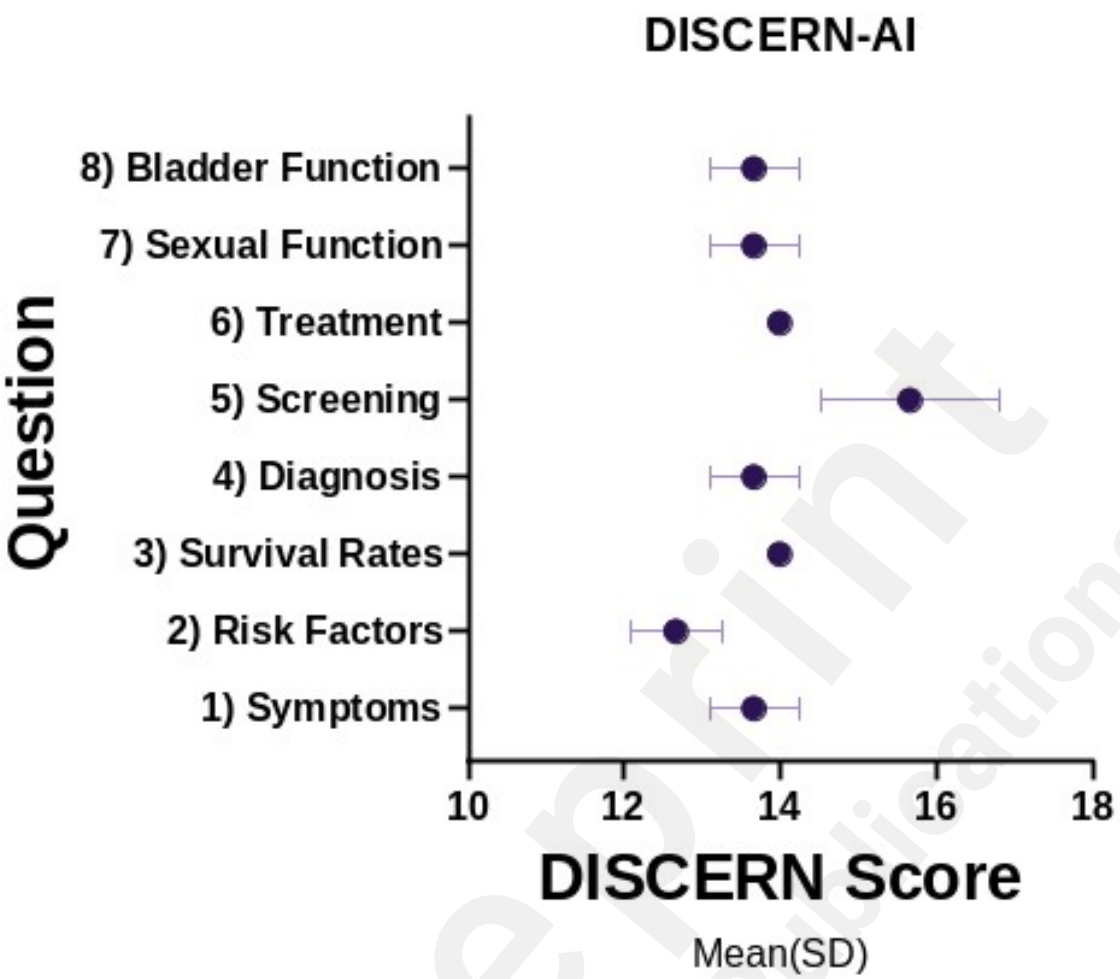
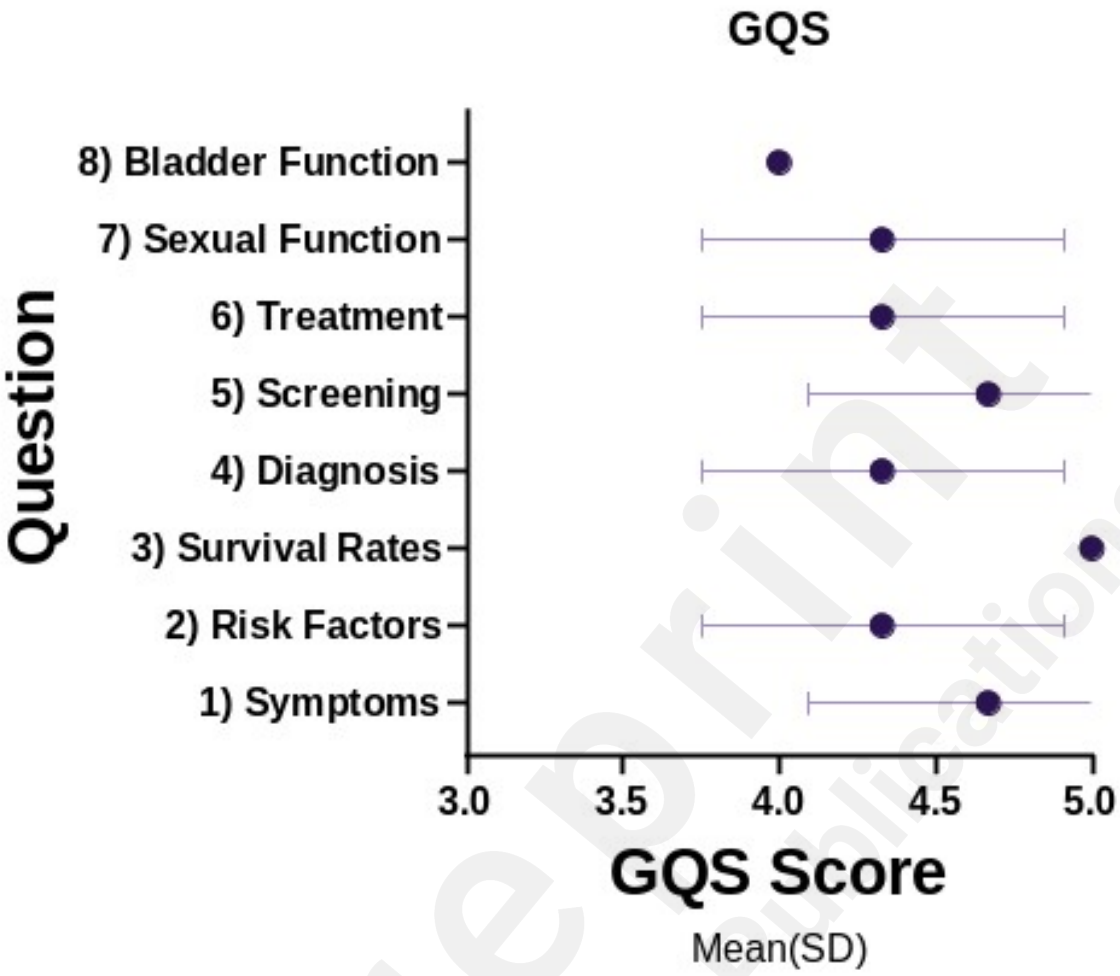


Figure 3 – GQS



NLAT-AI

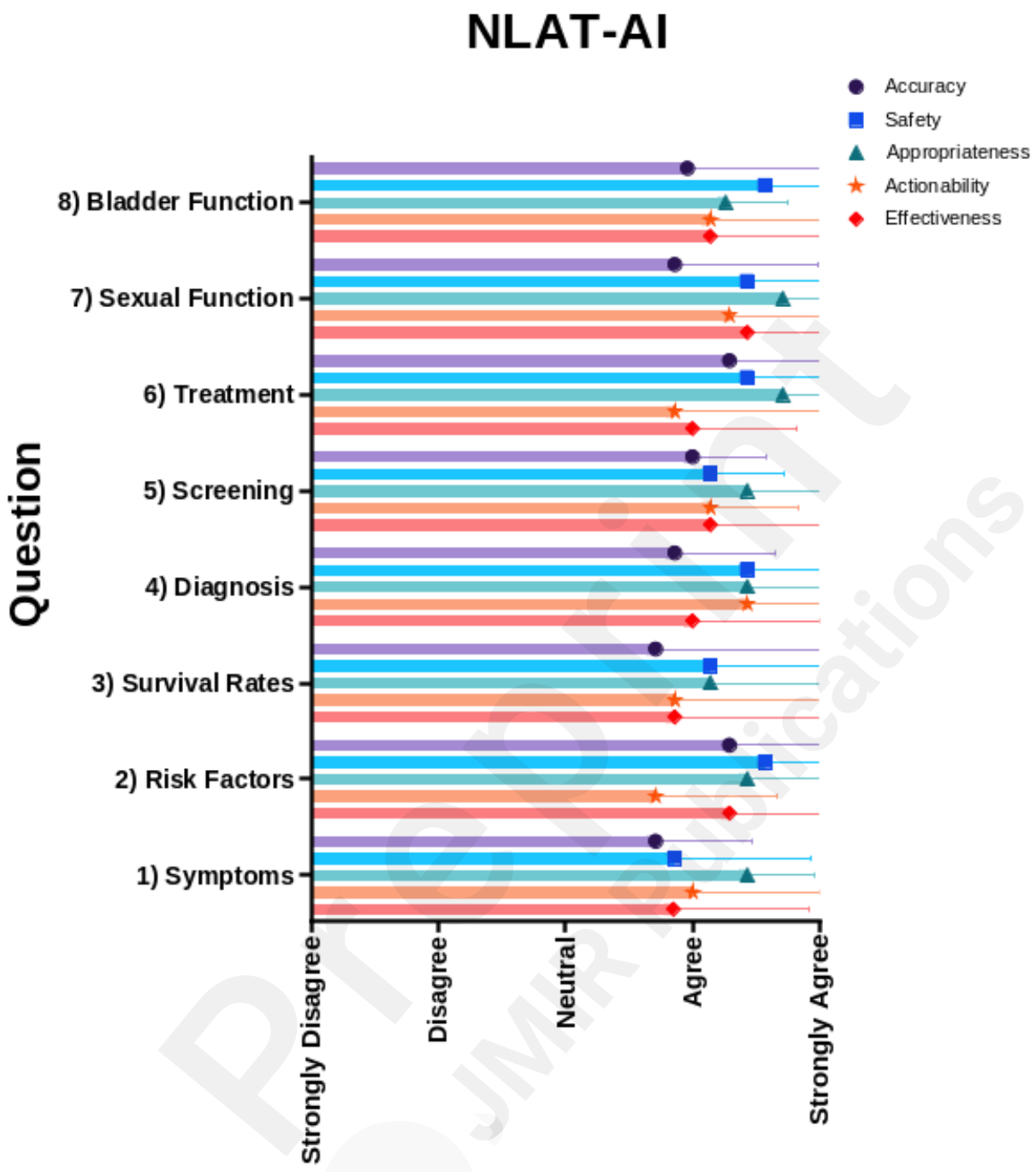
Expert assessment of the AI outputs with NLAT-AI were consistent with a mean >3.0/5.0 (Neutral) in all domains across all question replies. NLAT-AI pooled means (SD) included accuracy 3.96 (0.91), safety 4.32 (0.86), appropriateness 4.45 (0.81), actionability 4.05 (1.15), and effectiveness 4.09 (0.98). Descriptive statistics for each question are tabulated and graphed (Table 2, Figure 4). Internal validity testing demonstrated high reliability with a Cronbach alpha 0.906.

Table 2 – NLAT-AI

	Accuracy mean (SD)	Safety mean (SD)	Appropriateness mean (SD)	Actionability mean (SD)	Effectiveness mean (SD)
Q1) Symptoms	3.71 (0.76)	3.86 (1.07)	4.43 (0.53)	4.00 (1.00)	3.85 (1.07)
Q2) Risk Factors	4.29 (0.76)	4.57 (0.53)	4.43 (0.98)	3.71 (0.95)	4.29 (0.76)
Q3) Survival Rates	3.71 (1.60)	4.14 (1.60)	4.14 (1.07)	3.86 (1.68)	3.86 (1.68)
Q4) Diagnosis	3.86 (1.07)	4.43 (0.79)	4.43 (0.79)	4.43 (1.51)	4.00 (1.00)
Q5) Screening	4.00	4.14	4.43 (0.79)	4.14 (0.69)	4.14 (0.90)

	(0.58)	(0.58)			
Q6) Treatment	4.29 (0.49)	4.43 (0.79)	4.71 (0.49)	3.86 (1.46)	4.00 (0.82)
Q7) Sexual Function	4.00 (0.82)	4.43 (1.13)	4.71 (0.49)	4.29 (0.79)	4.43 (0.79)
Q8) Bladder Function	3.86 (1.07)	4.57 (1.07)	4.26 (0.49)	4.14 (0.90)	4.14 (0.90)
Total	3.96 (0.91)	4.32 (0.86)	4.45 (0.81)	4.05 (1.15)	4.09 (0.98)

Figure 4 – NLAT-AI



Qualitative feedback via NLAT-AI on questions one through eight indicate some areas for improvement despite the generally accurate and easy-to-understand nature of responses. Common themes were a need for greater specificity, updated and comprehensive information, and a more globally inclusive perspective (Table 3). Outputs were often characterised as good starting points or

overviews which could benefit patients.

Table 3 – NLAT-AI qualitative feedback

	Excepts
Q1) Symptoms	• Overall, a reasonable answer to the question… more emphasis should be put on the fact that prostate cancer is usually asymptomatic, usually detected on screening, only symptomatic when advanced. • Could have been better if discussed symptoms of locally advanced prostate cancer (LUTS, haematuria etc) and symptoms of metastatic prostate cancer (bone pain, weight loss etc) • Need to strongly emphasise that most prostate cancers are asymptomatic so PSA testing is necessary
Q2) Risk Factors	• Considering that this is tailored for Americans, it may not be actionable for others. • From a safety perspective I would emphasize the importance of seeking medical review in the event of family history. • Remove the modifiable risk factors as it makes patients think they can prevent it
Q3) Survival Rates	• There is…no mention of the impact of treatment on survival so a patient could be forgiven for thinking this was survival rates in the event of no treatment being given. • "Relative survival" is not clearly explained. • The survival rate [is] overestimated in organ confined disease as this is far more complex. It should be more clarified. • When talking about prostate cancer survival 10 years is the minimum that should be discussed • Fairly good- this is what I would tell my patients.
Q4) Diagnosis	• Overall reasonable answer from chat GPT • CT and bone scan is used for staging; but now in Australia is superseded by PSMA • Reasonable answer. Some inaccuracies in how the tests are used, as well as their sequencing. PSMA PET not mentioned which is an important part of diagnosis and staging. These deficiencies likely reflect the rapidly evolving nature of prostate cancer diagnosis. • The answer is easy to understand and general principles of diagnosis sound.
Q5) Screening	• Point 2 is very contentious and…gives a very one-sided view of prostate cancer screening. • This is only appropriate for American audience. • Point 2 is concerning as this represents one [clinician] group who is very much against prostate cancer screening… therefore may risk not giving a balanced view. • No mention of any local guidelines, and no EAU [European Association of Urology] guidelines.
Q6) Treatment	• Very useful summary for patients immediately after diagnosis. • No mention of novel tx [treatments] eg : focal therapy, cryo, HIFU • No mention of robotic surgery vs open surgery • This is a very simple table about the pros and cons.
Q7) Sexual Function	• Overall a very good answer – misses minor points • Very well written • Would also mention that erectile function improves over time. • Surgery does not damage the vessels for erection
Q8) Bladder Function	• Nice summary. • Accurate and easy to understand. • Minor issues only with the discussion on stress or urge incontinence • Hormone therapy should not causes bladder dysfunction. In fact, it might improve it

Readability

The readability algorithm consensus was "difficult to read" (Flesch Reading Ease score mean 45.97 SD 8.69, Gunning Fog Index mean 14.55 SD 4.79), averaging a grade 11 reading level, equivalent to 15 – 17-year-old (Flesch-Kincaid Grade Level mean 12.12 SD 4.34, The Coleman-Liau Index mean 12.75 SD 1.98, SMOG Index mean 11.06 SD 3.20). Questions one and two were the easiest to read scoring at grade 8 level while questions six (grade 23 level), seven (grade 12) and eight (grade 13 level) were very difficult to read (table 4).

Table 4 – Readability assessment

	Flesch Reading Ease score	Gunning Fog Index	Flesch-Kincaid Grade Level	The Coleman-Liau Index	SMOG Index
Q1) Symptoms	53.2	10.1	8.7	11	7.8
Q2) Risk Factors	59.4	8.5	7.7	11	7.6
Q3) Survival Rates	51.4	15	11	10	11.1
Q4) Diagnosis	46.2	13.4	10.9	13	10.3
Q5) Screening	49.3	13.7	11.1	12	10.6
Q6) Treatment	57.2	25.7	22.8	16	18.7
Q7) Sexual Function	39.9	14.9	11.6	14	11
Q8) Bladder Function	31.2	15.1	13.2	15	11.4
Pooled total	45.97	14.55	12.12	12.75	11.06

REF-AI

REF-AI identified two reference hallucinations from 30 total references across all questions (pooled REF-AI Real mean 2.86). Most references effectively supported the text, while four questions had one or two citations that were not directly supporting the information provided (table 5) (pooled REF-AI Supporting mean 2.13). Eighty-six percent of references (26/30) were from reputable government organizations, while two were direct citations from scientific literature (pooled REF-AI Source mean 2.13). Individual statements were provided a direct reference in only three outputs. Remaining outputs instead provided a list of references at the bottom of the text. Some direct links to references were not complete, instead delivering the user to the organisation's primary website Uniform Resource Locator (URL), likely reflecting updated website directories since the 2021 ChatGPT indexation. The two hallucinated references were present in question seven and eight, where weblinks did not connect and despite extensive google and library searches the original material was unable to be located.

Table 5 – REF-AI assessment

	Real mean (SD)	Supporting mean (SD)	Source mean (SD)
Q1) Symptoms	3.00 (0.00)	3.00 (0.00)	2.00 (0.00)
Q2) Risk Factors	3.00 (0.00)	2.67 (0.58)	2.00 (0.00)
Q3) Survival Rates	3.00 (0.00)	2.67 (0.58)	2.00 (0.00)
Q4) Diagnosis	3.00 (0.00)	3.00 (0.00)	2.00 (0.00)
Q5) Screening	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)
Q6) Treatment	3.00 (0.00)	3.00 (0.00)	2.00 (0.00)
Q7) Sexual	3.00	2.33	2.00

Function	(0.00)	(0.58)	(0.00)
Q8) Bladder Function	2.00 (0.00)	2.33 (0.58)	2.00 (0.00)
Total	2.86 (0.00)	2.75 (0.29)	2.13 (0.00)

Discussion

In the digital information age, understanding what patient health information is accessed, and the quality of this information is crucial. This study demonstrates several examples of information that patients (and their carers) may encounter when conducting searches related to prostate cancer management. In our analysis, ChatGPT-4 provided generally comprehensive answers to prostate cancer questions, mostly in line with current medical guidelines and literature. ChatGPT-4 demonstrated promise when assessed with a range of patient education and information quality assessment tools, as well as expert review. Robust scores and expert feedback indicate that the generated content was reliable, safe, and actionable for patients, albeit with room for improvement in minor nuanced details, global applicability, and readability.

Current evidence indicates that 75% of people turn to the internet for decision-making during a health crisis [38]. Despite the abundance of available patient information, studies assessing the quality of online health information indicate significant shortcomings [39]. For prostate cancer, the quality of information that reaches the patient is known to be inconsistent [10, 19, 40-42]. In example, previous assessment of the top 100 ‘prostate cancer’ webpage results identified via search engine query showed that only 11.1% of sites demonstrate an excellent on the original DISCERN criteria [10]. While our analysis has employed necessarily disparate methods, comparison of our DISCERN-AI results (good-excellent) to static webpage DISCERN scores suggests that ChatGPT prostate cancer information outputs may be of a higher quality than many traditional webpages [10]. ChatGPT4 appears capable of providing broad and largely accurate information which may further augment self-directed patient or stakeholder enquiry. Nevertheless, direct comparison of ChatGPT outputs to established gold standard information sources is necessary to clearly define the role of this new communication technology as part of patient care and education.

Our findings appear to differ from Coskun and colleagues', where ChatGPT-3 had accuracy issues using queries generated from the European Association of Urology Patient Information[15]. Interestingly, Zheng and colleagues discovered that ChatGPT-4 can offer suitable counselling on disease prevention and screening for prostate cancer patients[16]. These differences may represent the rapid evolution of the algorithm as our testing utilised the newer model. Exclusive use of US-centric guidelines raised questions of bias amongst our experts. Others have also highlighted such bias, noting that 51% of training data for major Large Language Models (LLMs) is US-sourced[43, 44]. The disparities between ChatGPT-3 and ChatGPT-4 highlight the continual advancement and refinement in the underlying technology, reinforcing the need for periodic assessment and validation as newer models emerge[4, 16]. Conversely, a lack of validated and reproducible tools to make reliable quality assessments of NPLTs is likely to play a role in varied results within this juvenile domain of clinical research[45]. While the methods employed in our study were an effort to standardise output assessment in our work, we recognise and encourage further rigorous work to develop validated and reproducible assessment tools which can be applied to a range of NPLT outputs and platforms.

Despite the NLAT-AI rating, and general appropriateness of the language across all questions, the objective readability from algorithms demonstrated a high reading level and difficulty. This is likely reflective of literacy bias present amongst our highly educated expert pool[33, 46, 47]. While the recommended reading level for patient education material varies between organisations, the consensus is that it should generally lie between grade six and eight reading levels[46, 48]. The readability algorithms thus suggest that the generated content may be challenging for some readers. These findings are of importance given that lower readability may limit accessibility for certain socioeconomic or minority groups[46]. Literacy is a known negative correlate of prostate cancer health outcomes[8, 49-51]. Compounding this concern is the effect of user’s overarching eHealth literacy, which is likely to affect chat-bot engagement behaviours and patterns of information

comprehension and utilisation[2, 15, 49, 52]. Effects of both traditional literacy and eHealth literacy on end user experience of NLPTs require urgent investigation due to the pervasiveness which these technologies are already presenting within society and in online health communication[53, 54].

The interactive nature of the ChatGPT 4 model, where users can continuously engage and seek clarifications, offers a potential advantage and solution to static patient information materials. Although beyond this study's scope, the ChatGPT 4 model permits ongoing discussions, enabling patients to seek clarifications of information. These conversations allow for personalised explanations related to patient health results, the opportunity to simplify language, and may ultimately address some concerns raised by our expert assessors. This is an extremely powerful and unique component of this new digital technology. Future iterations of such models may benefit by incorporating clear adaptability features, where the complexity and specificity of the content can be adjusted based on user preferences or needs. Further studies are required to explore how the longitudinal and dynamic features of NLPTs affect information quality and patient comprehension. This will be particularly important in comparison to traditional website and social media-based information sources which currently dominate the landscape of self-educative information sourcing in prostate cancer care[10, 19, 55]. NPLTs with pre-determined or flexible user settings attuned to patient preference, needs or literacy level are a potential futurist pathway to cost effective and scalable forms of tailored patient health education materials.

Hallucination, where information is fabricated by the NLPT and presented as valid, is a well-documented phenomenon specific to NLPT's and ChatGPT[37]. In our study we demonstrated that hallucination could occur when searching for prostate cancer with NPLT/chat bots. While only occurring in two instances of thirty, these findings continue. Designation between hallucination vs. faux-hallucination should also be considered. Faux-hallucination results from modified references after ChatGPT-4's indexation, leading to broken links or lost references. Website redesign or content no longer existing after the 2021 indexation are potential aetiologies for hallucination that have not been fully explored. Equally, such disappearance of content with time may also match the definition of hallucination in the future. While not a prominent issue in this study, these findings continue to demonstrate the potential for fabricated information, which can be easily overlooked by the unassuming clinician, patient or researcher. While still in its infancy, LLMs must continue to solve the issue of hallucination before integration into high-risk systems such as healthcare can be considered.

While hallucinations are a notable concern, there are several other limitations of current NPLTs that need to be considered. Despite malleability, it is unknown whether the ChatGPT-4 model may fully replicate the nuance of human communication necessary for effective patient health education[2]. Additionally, the most significant limitation of ChatGPT is its potential for biased, outdated or misleading content generation [1, 3, 5, 52]. Even with relatively high-quality scores, our study shows that ChatGPT can still produce misleading or biased content under discriminatory and expert scrutiny, posing some element of risk for those with poor eHealth literacy [3, 5, 52]. However, while expert reviewers identified minor inaccuracies, none of these points were considered to be significantly concerning safety issues. Nevertheless, there is currently a lack of evidence to predict the impact of these technologies on patients' understanding, decision-making, or health, without further enquiry and consideration of patients' ability to interact with these new eHealth technologies. We strongly recommend clinicians report these concerns to prostate cancer patients and their stakeholders when guiding patient use of online information in their care. Furthermore, the opaque and dynamic nature of this technology's private enterprise proprietary algorithms is also a concern[2, 3, 45]. Algorithm development will likely outpace quality assurance efforts, and raise questions about the necessity of clinician involvement in NPLT model development which aims to present health based information[2, 44]. The effectively unknown and vast array of sources from which ChatGPT's training data is derived raises ethical concerns. Without knowing the origins and credibility of such data, it's difficult for clinicians to fully trust generated content, presenting us with

a modernised but perpetual issue of distrust in online information which may ultimately hinder adoption and progress[2, 5, 52]. Finally, there are also financial considerations; the cost of using ChatGPT-4 (as opposed to the currently free ChatGPT 3.5) or other NPLTs may form a barrier to widespread adoption in healthcare settings, and has the potential to drive disparate levels of healthcare if not effectively managed and regulated.

Limitations of this study include the sample size of assessors which may skew the evaluation of the included tool's reliability and efficacy. The qualitative assessments of experts are at inherent risk of bias for or against the use of novel technology and ChatGPT-4. However, these experts are also deeply aware of the nature and quality of current prostate cancer education materials, providing additional insight which is of value to this work.

It is important to note this assessment was purposefully narrow in scope and may not reflect the myriad of interactions under the vast topics of prostate cancer. It is unknown how applicable these interactions are in wider prostate cancer education scenarios and ongoing investigation is required. Work is currently underway to assess an expanded question set with comparison to currently accepted patient education gold standards in prostate cancer.

While not an explicit purpose of this study, the exploratory assessments used in this work (DISCERN-AI, PEMAT-AI, NLAT-AI and REF-AI) demonstrate inter-reliability, and replicability across several cancer type information outputs. They may thus have potential utility for clinicians and researchers interested in reviewing the quality of other cancer-based outputs of Chat-GPT4 or other NPLTs. Nevertheless, their validity requires further testing and greater investigation is necessary to develop specific tools to assess NPLT output quality in the long term.

Conclusion:

Our analysis found ChatGPT-4's responses to common prostate cancer queries were of good quality, and a potentially useful patient education adjunct for prostate cancer care. Objective quality assessment tools were reflective of NPLT outputs which were generally reliable and appropriate, though with room for improvement. Our expert panel was impressed by the appropriateness and safety of the language and information given. However, clinicians should be aware that there are several limitations to ChatGPT-4 prostate cancer outputs, including: hallucination, specificity issues, and difficult readability. Future studies are required to assess whether more longitudinal (back-and-forth) ChatGPT-4 discourse may offset some of the concerns highlighted in this analysis, and how patients of differing eHealth literacy levels may engage with and have care affected by such technologies.

Acknowledgments

None

Conflicts of interest

None declared

Abbreviations:

Artificial intelligence (AI)
Chat Generative Pre-Trained Transformer (ChatGPT)
Large Language Models (LLMs)
Natural Language Assessment Tool for AI (NLAT-AI)
Natural language processing technologies (NLPTs)
Patient Education Materials Assessment Tool (PEMAT)
Prostate-Specific Antigen (PSA)
The Global Quality Score (GQS)
Uniform Resource Locator (URL)

References

1. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI cancer spectrum. 2023;7(2):pkad015.
2. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. Journal of Medical Systems. 2023;47(1):1-5.
3. Patel SB, Lam K. ChatGPT: the future of discharge summaries? The Lancet Digital Health. 2023;5(3):e107-e8.
4. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plastic Surgery. 2023:1-9.
5. Seth I, Cox A, Xie Y, Bulloch G, Hunter-Smith DJ, Rozen WM, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. Aesthetic Surgery Journal. 2023:sjad140.
6. Wang L, Lu B, He M, Wang Y, Wang Z, Du L. Prostate Cancer Incidence and Mortality: Global Status and Temporal Trends in 89 Countries From 2000 to 2019. Front Public Health. 2022;10:811044.
7. Catt S, Matthews L, May S, Payne H, Mason M, Jenkins V. Patients' and partners' views of care and treatment provided for metastatic castrate-resistant prostate cancer in the UK. Eur J Cancer Care (Engl). 2019;28(6):e13140.
8. Ellimoottil C, Polcari A, Kadlec A, Gupta G. Readability of websites containing information about prostate cancer treatment options. J Urol. 2012;188(6):2171-5.
9. Baunacke M, Schmidt ML, Groeben C, Borkowetz A, Thomas C, Koch R, et al. Decision Regret after Radical Prostatectomy does Not Depend on Surgical Approach: 6-Year Followup of a Large German Cohort Undergoing Routine Care. J Urol. 2020;203(3):554-61.
10. Moolla Y, Adam A, Perera M, Lawrentschuk N. 'Prostate Cancer' Information on the Internet: Fact or Fiction? Curr. 2020;13(4):200-8.
11. Alsyouf M, Stokes P, Hur D, Amasyali A, Ruckle H, Hu B. 'Fake News' in urology: evaluating the accuracy of articles shared on social media in genitourinary malignancies. BJU Int. 2019;02:02.
12. Sehn E, Mozak C, Yuksel N, Sadowski CA. An analysis of online content related to testosterone supplementation. Aging Male. 2019;22(2):141-9.
13. Ayre J, Mac OA, McCaffery KJ, McKay BR, Liu M, Shi Y, et al. New frontiers in health literacy: Using ChatGPT to simplify health information for people in the community. medRxiv. 2023:2023.07. 24.23292591.
14. Jiao W, Wang W, Huang J-t, Wang X, Tu Z. Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:230108745. 2023.
15. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology. 2023.
16. Zheng Y, Xu Z, Yu B, Xu T, Huang X, Zou Q, et al. Appropriateness of Prostate Cancer Prevention and Screening Recommendations Obtained From ChatGPT-4. 2023.
17. Haun MW, Ihrig A, Karschuck P, Thomas C, Huber J. The era of the digital natives is approaching: Insights into online peer-to-peer support for persons affected by prostate cancer. World J Urol. 2020;07:07.
18. van Eenbergen M, Vromans RD, Boll D, Kil PJM, Vos CM, Krahmer EJ, et al. Changes in

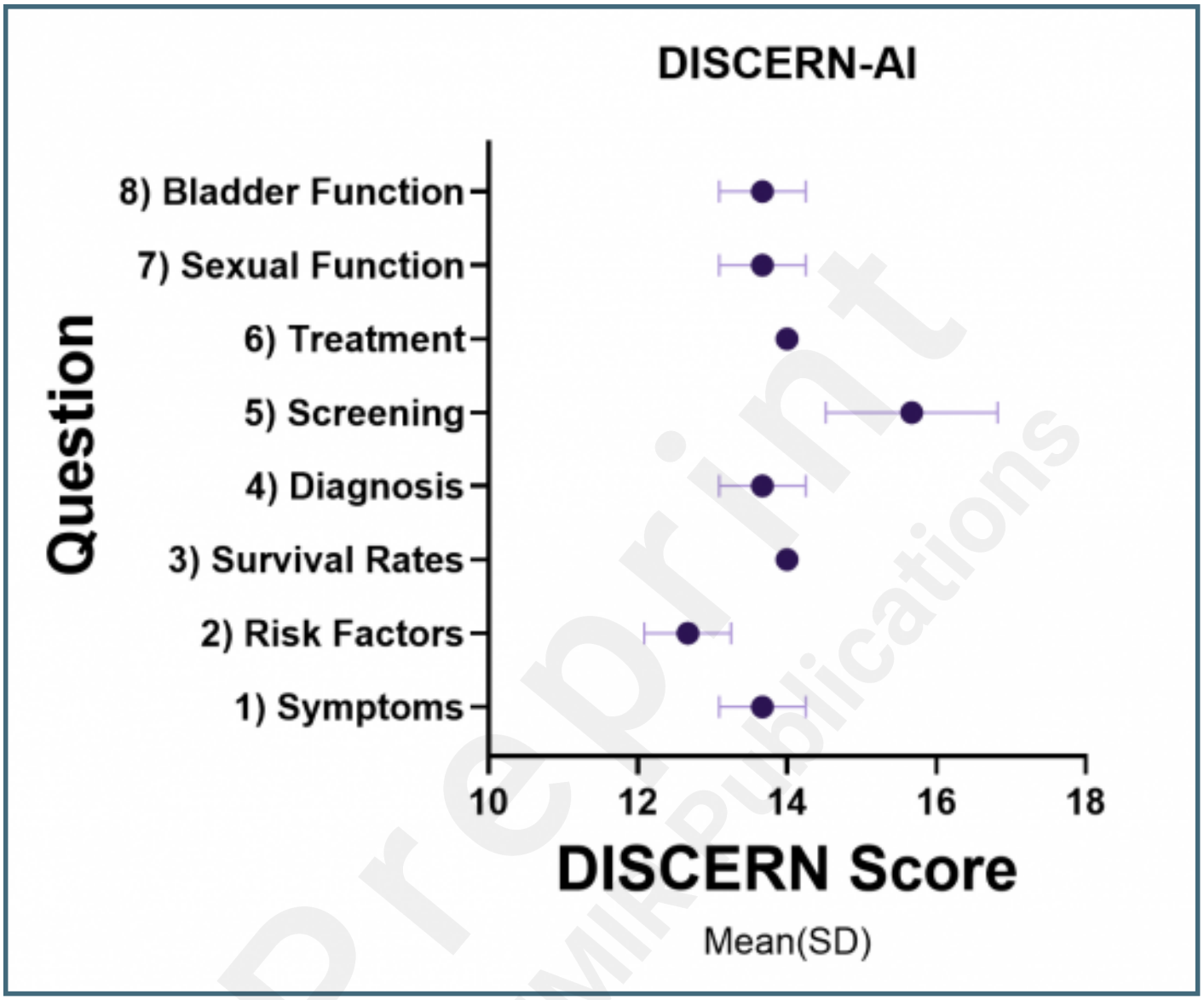
internet use and wishes of cancer survivors: A comparison between 2005 and 2017. Cancer. 2020;126(2):408-15.
19. Cacciamani GE, Bassi S, Sebben M, Marcer A, Russo GI, Cocci A, et al. Consulting "Dr. Google" for Prostate Cancer Treatment Options: A Contemporary Worldwide Trend Analysis. Eur Urol Oncol. 2019;30:30.
20. Rezaee ME, Goddard B, Sverrisson EF, Seigne JD, Dagrosa LM. 'Dr Google': trends in online interest in prostate cancer screening, diagnosis and treatment. BJU Int. 2019;17:17.
21. Trends G. Google Trends Explore 2023 [Available from: https://trends.google.com/trends/explore?geo=AU&hl=en-AU.
22. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105-11.
23. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient education and counseling. 2014;96(3):395-403.
24. SINGH AG, SINGH S, SINGH PP. YouTube for Information on Rheumatoid Arthritis — A Wakeup Call? The Journal of Rheumatology. 2012;39(5):899-903.
25. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Van Zanten SV. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Official journal of the American College of Gastroenterology| ACG. 2007;102(9):2070-7.
26. Cassidy JT, Baker JF. Orthopaedic patient information on the World Wide Web: an essential review. JBJS. 2016;98(4):325-38.
27. Weil AG, Bojanowski MW, Jamart J, Gustin T, Lévêque M. Evaluation of the quality of information on the Internet available to patients undergoing cervical spine surgery. World neurosurgery. 2014;82(1-2):e31-e9.
28. Altunisik E, Firat YE. Quality and Reliability Analysis of Essential Tremor Disease Information on Social Media: The Study of YouTube. Tremor Other Hyperkinet Mov (N Y). 2022;12:32.
29. Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.
30. Mc Laughlin GH. SMOG grading-a new readability formula. Journal of reading. 1969;12(8):639-46.
31. Gunning R. The technique of clear writing. (No Title). 1952.
32. Coleman M, Liau TL. A computer readability formula designed for machine scoring. Journal of Applied Psychology. 1975;60(2):283.
33. Hansberry DR, John A, John E, Agarwal N, Gonzales SF, Baker SR. A Critical Review of the Readability of Online Patient Education Resources From RadiologyInfo.Org. American Journal of Roentgenology. 2014;202(3):566-75.
34. Readable.com [Available from: https://readable.com/.
35. Microscoft. Word. 2023.
36. Kugar MA, Cohen AC, Wooden W, Tholpady SS, Chu MW. The readability of psychosocial wellness patient resources: improving surgical outcomes. J Surg Res. 2017;218:43-8.
37. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023;15(2).
38. Dickerson SS, Reinhart A, Boemhke M, Akhu-Zaheya L. Cancer as a problem to be solved: internet use and provider communication by men with cancer. Comput Inform Nurs. 2011;29(7):388-95.

39. Berland GK, Elliott MN, Morales LS, Algazy JI, Kravitz RL, Broder MS, et al. Health information on the Internet: accessibility, quality, and readability in English and Spanish. jama. 2001;285(20):2612-21.
40. Lange L, Peikert ML, Bleich C, Schulz H. The extent to which cancer patients trust in cancer-related online information: a systematic review. PeerJ. 2019;7:e7634.
41. Ghai S, Trachtenberg J. Internet information on focal prostate cancer therapy: help or hindrance? Nat Rev Urol. 2019;16(6):337-8.
42. Asafu-Adjei D, Mikkilineni N, Sebesta E, Hyams E. Misinformation on the Internet regarding Ablative Therapies for Prostate Cancer. Urology. 2019;133:182-6.
43. Dodge J, Sap M, Marasović A, Agnew W, Ilharco G, Groeneveld D, et al. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:210408758. 2021.
44. Healy M. Approaches to Generative Artificial Intelligence, A Social Justice Perspective. arXiv preprint arXiv:230912331. 2023.
45. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, 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.
46. Mac OA, Muscat DM, Ayre J, Patel P, McCaffery KJ. The readability of official public health information on COVID-19. The Medical Journal of Australia. 2021;215(8):373.
47. Rosenberg SA, Francis D, Hullett CR, Morris ZS, Fisher MM, Brower JV, et al. Readability of Online Patient Educational Resources Found on NCI-Designated Cancer Center Web Sites. Journal of the National Comprehensive Cancer Network. 2016;14(6):735-40.
48. Health SA. Engaging with Consumers, Carers and the Community: Guide and Resources. SA Health Adelaide; 2021.
49. Basch CH, Ethan D, MacLean SA, Fera J, Garcia P, Basch CE. Readability of Prostate Cancer Information Online: A Cross-Sectional Study. Am j. 2018;12(5):1665-9.
50. Maciolek KA, Jarrard DF, Abel EJ, Best SL. Systematic Assessment Reveals Lack of Understandability for Prostate Biopsy Online Patient Education Materials. Urology. 2017;109:101-6.
51. Borgmann H, Wolm JH, Vallo S, Mager R, Huber J, Breyer J, et al. Prostate Cancer on the Web-Expedient Tool for Patients' Decision-Making? J Cancer Educ. 2017;32(1):135-40.
52. Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer and Prostatic Diseases. 2023:1-6.
53. Zhang Z, Genc Y, Wang D, Ahsen ME, Fan X. Effect of ai explanations on human perceptions of patient-facing ai-powered healthcare systems. Journal of Medical Systems. 2021;45(6):64.
54. Zhang Z, Genc Y, Xing A, Wang D, Fan X, Citardi D. Lay individuals' perceptions of artificial intelligence (AI)-empowered healthcare systems. Proceedings of the Association for Information Science and Technology. 2020;57(1):e326.
55. Qan'ir Y, Song L. Systematic review of technology-based interventions to improve anxiety, depression, and health-related quality of life among patients with prostate cancer. Psychooncology. 2019;28(8):1601-13.

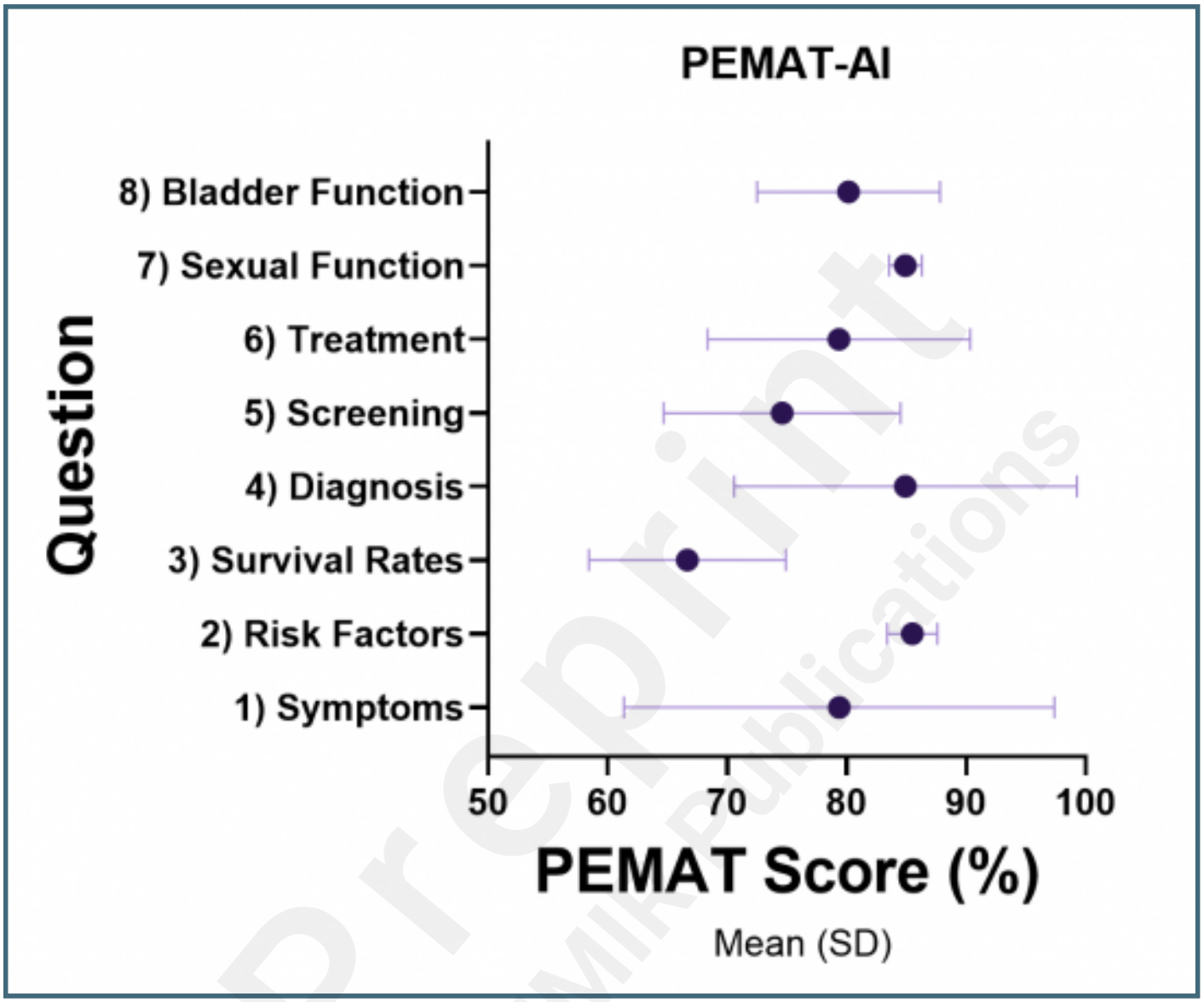
Supplementary Files

Figures

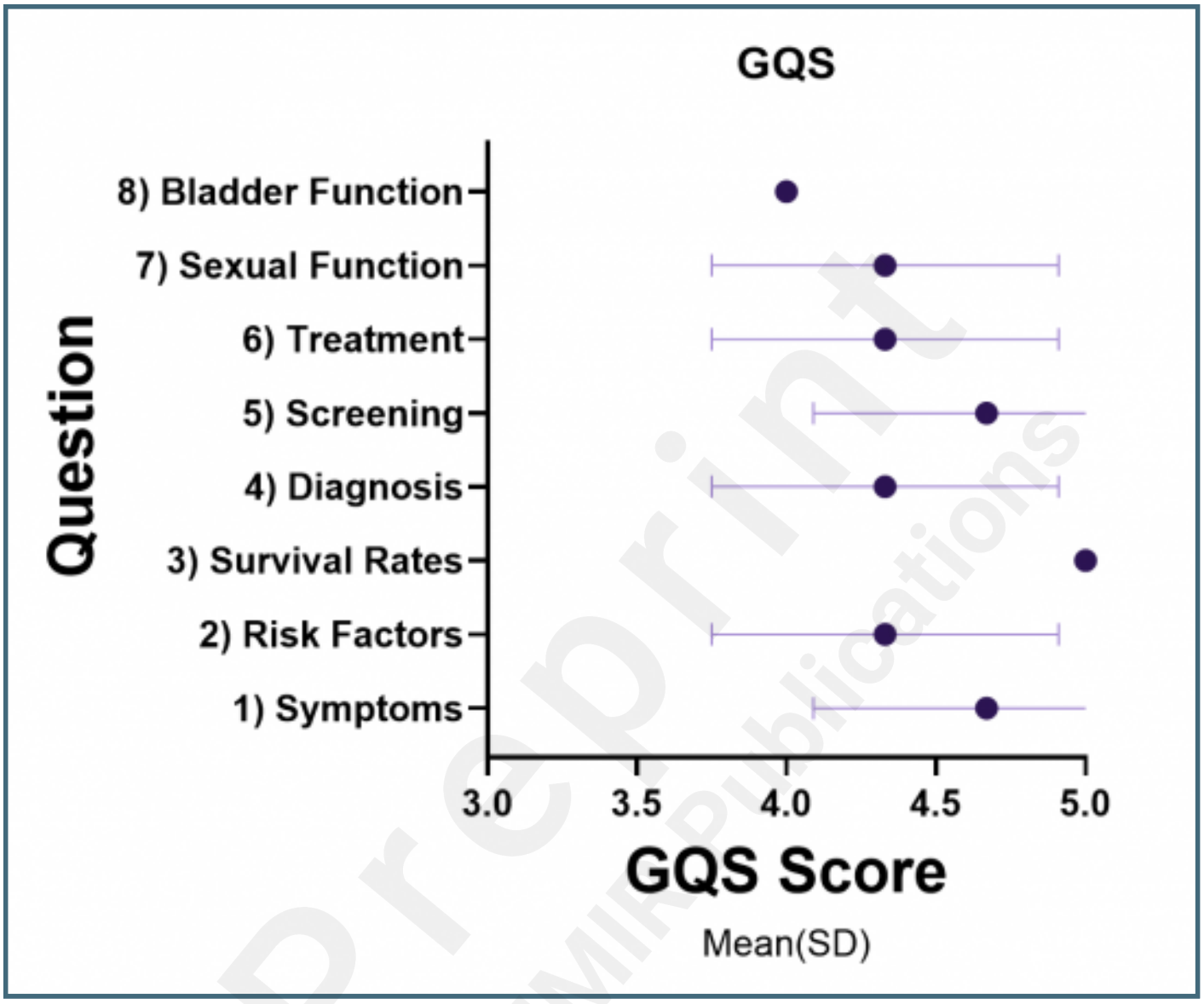
DISCERN-AI mean score by ChatGPT question output.



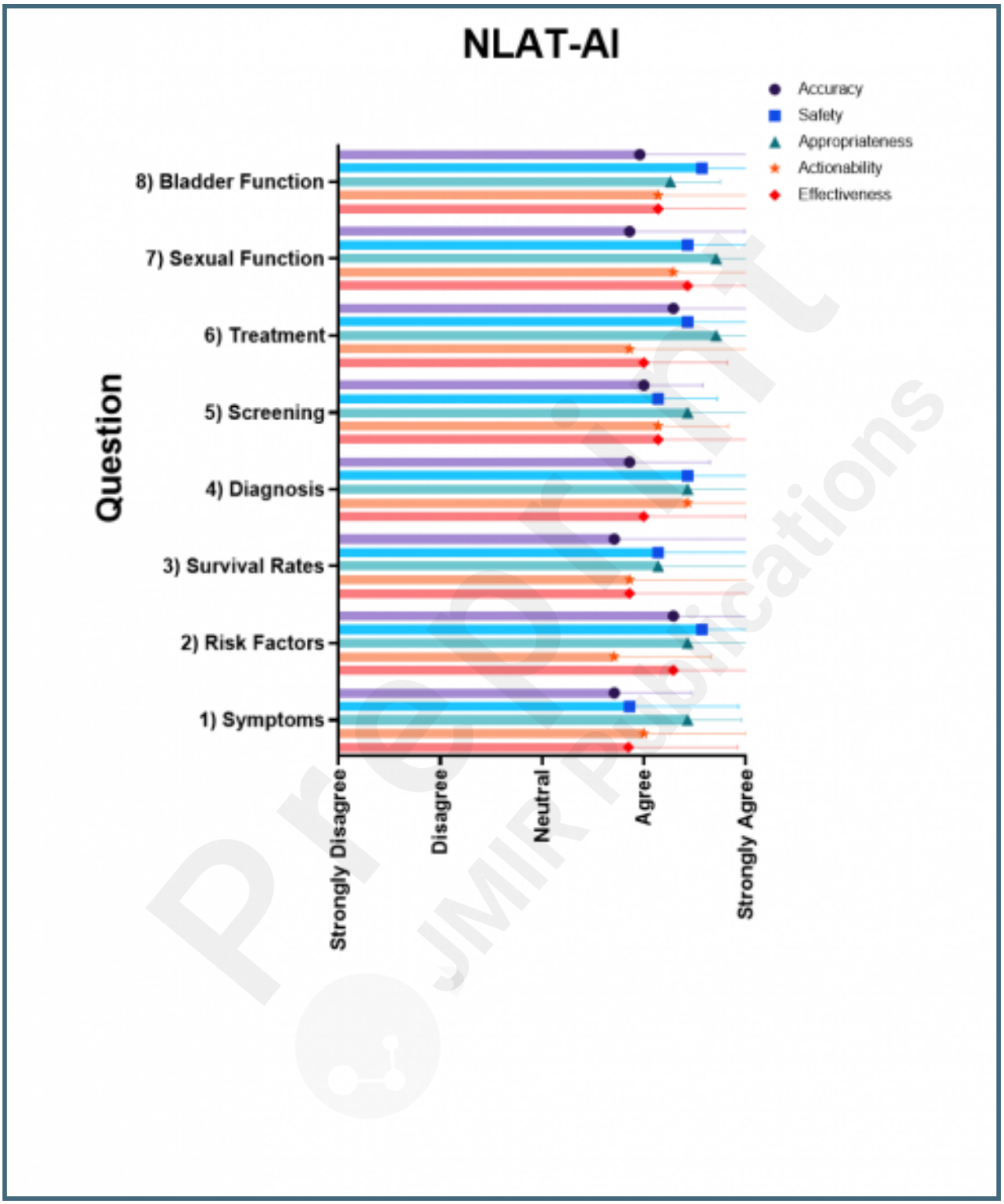
PEMAT-AI mean score by ChatGPT question output.



GQS mean score by ChatGPT question output.



NLAT-AI mean score by ChatGPT question output.



Multimedia Appendixes

Common prostate cancer questions.
URL: http://asset.jmir.pub/assets/20f3789e8de1e384a91d46bb99a5c829.docx

ChatGPT4 Output.
URL: http://asset.jmir.pub/assets/9b06594e4b2184f69dcf23817eaff9c4.docx

DISCERN-AI Tool.
URL: http://asset.jmir.pub/assets/5238c08b2f0e33dac3fe8658fe6aafbe.docx

PEMAT-AI Tool.
URL: http://asset.jmir.pub/assets/fb066a6b933138e41741cb60991b532f.docx

Natural Language Assessment Tool for AI (NLAT-AI).
URL: http://asset.jmir.pub/assets/5d5ca06007c2e30cd802764f3da4a5d2.docx

REF-AI Tool.
URL: http://asset.jmir.pub/assets/9b6e32a3aa7b771df44a794dda12b291.docx