

Evaluating an Artificially Intelligent Chatbot 'Prostate Cancer Info' for Providing Quality Prostate Cancer Screening Information: A Cross-Sectional Study

Otis Owens, Michael Leonard

Submitted to: JMIR Cancer
on: February 16, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 19

..... 19

Please Select A Types..... 20

 Please Select A Type 1..... 20

Related publication(s) - for reviewers eyes onlies 21

 Related publication(s) - for reviewers eyes only 0..... 21

Evaluating an Artificially Intelligent Chatbot ‘Prostate Cancer Info’ for Providing Quality Prostate Cancer Screening Information: A Cross-Sectional Study

Otis Owens¹ PhD, MPH; Michael Leonard¹ MS, MEd, MEng

¹ College of Social Work University of South Carolina Columbia US

Corresponding Author:

Otis Owens PhD, MPH

College of Social Work
University of South Carolina
1512 Pendelton Street
Columbia
US

Abstract

Background: Generative AI Chatbots may be useful tools for supporting shared prostate cancer screening decisions, but the information produced by these tools sometimes lack quality or credibility. ‘Prostate Cancer Info’ is a custom GPT chatbot developed to provide plain-language PrCA information only from websites of key authorities on cancer and peer-reviewed literature.

Objective: To evaluate the accuracy, completeness, and readability of ‘Prostate Cancer Info’s responses to frequently asked PrCA screening questions.

Methods: Twenty-three frequently asked PrCA questions were individually input into ‘Prostate Cancer Info.’ Responses were recorded in Microsoft Word and reviewed by two raters for their accuracy and completeness. Readability of content was determined by pasting responses into an online Flesch Kincaid Reading Ease Scores calculator.

Results: Responses to all questions were accurate and culturally appropriate. Seventeen of twenty-three questions (74%) had complete responses. The average readability of responses was 64.5 (written at an 8th-grade level).

Conclusions: Generative AI chatbots, such as Prostate Cancer Info, are great starting places for learning about PrCA screening and preparing to engage in shared decision making but should not be used as independent sources of PrCA information because key information may be omitted. Men are encouraged to use these tools to complement information received from a healthcare provider.

(JMIR Preprints 16/02/2025:72522)

DOI: <https://doi.org/10.2196/preprints.72522>

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 visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/72522>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/72522>

Original Manuscript

Evaluating an Artificially Intelligent Chatbot 'Prostate Cancer Info' for Providing Quality Prostate Cancer Screening Information: A Cross-Sectional Study

Abstract

Background: Generative AI Chatbots may be useful tools for supporting shared prostate cancer screening decisions, but the information produced by these tools sometimes lack quality or credibility. 'Prostate Cancer Info' is a custom GPT chatbot developed to provide plain-language PrCA information only from websites of key authorities on cancer and peer-reviewed literature.

Objective: To evaluate the accuracy, completeness, and readability of 'Prostate Cancer Info's responses to frequently asked PrCA screening questions.

Methods: Twenty-three frequently asked PrCA questions were individually input into 'Prostate Cancer Info.' Responses were recorded in Microsoft Word and reviewed by two raters for their accuracy and completeness. Readability of content was determined by pasting responses into an online Flesch Kincaid Reading Ease Scores calculator.

Results: Responses to all questions were accurate and culturally appropriate. Seventeen of twenty-three questions (74%) had complete responses. The average readability of responses was 64.5 (written at an 8th-grade level).

Conclusions: Generative AI chatbots, such as Prostate Cancer Info, are great starting places for learning about PrCA screening and preparing to engage in shared decision-making but should not be used as independent sources of PrCA information because key information may be omitted. Men are encouraged to use these tools to complement information received from a healthcare provider.

Keywords: Generative Artificial Intelligence; Chatbot; ChatGPT; Prostate Cancer; Cancer Screening; Shared Decision Making;

Introduction

Generative AI chatbots such as ChatGPT, Google Gemini, and Microsoft Copilot have become highly publicized for enhancing work efficiency and effectively responding to diverse queries. These sophisticated programs leverage large language models, machine learning, and natural language processing to understand and respond to a query with publicly available or third-party information [1]. Over the past 2 years, researchers have demonstrated a growing interest in evaluating generative AI chatbots for providing quality health and cancer information [2-4]. While the performance of generative AI chatbots has varied depending on the disease queried, complexity of the query, and brand of chatbot used, these tools show promise for being reliable health information resources in the future [3, 5, 6].

In terms of prostate cancer (PrCA), the second leading cause of cancer mortality among men in the U.S. [7], the American Cancer Society [8], American Urological Association [9, 10], and the United States Preventive Services Task Force [11] recommend that men make shared PrCA screening decisions with their healthcare providers. To prepare for this important decision, men need access to credible, readable, and culturally-appropriate (e.g., African-Americans have a higher mortality risk[12]) PrCA screening information [13]. Multiple studies have investigated the quality of PrCA information generated by AI chatbots [14-22]. Overall, these studies show that PrCA information produced by chatbots can be accurate/reliable and moderately comprehensive, but readability and credibility are often compromised. In a recent study by the authors, Owens and Leonard [23] discovered that soliciting plain-language responses from chatbots to PrCA screening inquiries significantly enhanced the response's readability. Conversely, credibility was difficult to ascertain because generative AI chatbots don't consistently reference authoritative information sources [23]. To create a reliable and credible plain-language resource for PrCA screening information, we have developed 'Prostate Cancer Info,'(PCI) a generative AI chatbot using Open AI's custom GPT platform [24]. PCI is unique because it only responds to inquiries information solely from credible, PrCA expert-curated websites (like the ACS). This method is different from current generative AI chatbots, which search the entire web and produce responses from a variety of expert-vetted and non-vetted sources. In addition, we have programmed PCI to always provide a source for responses, which is uncommon for current generative AI chatbots. Lastly, we have programmed PCI to provide responses that don't exceed 6th to 8th grade readability as recommended by the American Medical Association [25]. The study's purpose is to evaluate the accuracy, completeness, and readability of PCI responses to 23 frequently asked PrCA screening questions. The study will contribute insight into the safety and efficacy of using AI chatbots for shared PrCA screening decision-making and the utility of developing customized AI chatbots for PrCA decision-making.

Methods

Intervention Development

Author (M.L.) developed PCI using a multi-step process. Websites published by ACS, AUA, USPSTF, and CDC were programmed into the GPT builder [24]. The rationale for limiting our search to these websites is that these organizations are globally recognized for providing timely, evidence-based PrCA screening information and/or screening recommendations. In particular, the PrCA screening recommendations from the ACS, AUA, and USPSTF are the most widely recognized in U.S. PrCA research and clinical practice. PCI was then directed to draw responses exclusively from these web pages in the order listed. Therefore, PCI relied on the ACS website as a primary source unless the information or was requested from a non-ACS source. Strict directives were given to PCI to: (1) only retrieve information from websites provided (2) respond with language at or below 8th-grade readability (3) ignore non-PrCA queries and (4) provide sources for responses. PCI was pre-tested to confirm its adherence to these directives.

PCI answers user questions through a well-defined process: first, it limits itself to information from pre-approved websites. Then, it indexes these sites, reading and organizing their content. When a user asks a question, PCI searches its indexed data for relevant details, analyzes the information to understand the context, and creates a concise, accurate answer from approved sources. This ensures consistent, trustworthy answers.

Study Protocol

Twenty-three frequently asked PrCA questions were adopted from prior studies by Zhu et al. [26] and Owens and Leonard [23]. One author entered questions into PCI. Responses were saved in a document for rating by both authors. The authors used a coding form containing questions with key points/answers from ACS and CDC education resources [27, 28] along with screening recommendations from ACS, AUA, and the USPSTF [8-11] and checkboxes to evaluate whether a response was accurate (contained correct statements) and complete (presented all salient facts without significant omissions). For example, a response to “What is the prostate?” would be considered accurate if it stated that the prostate is a gland is a part of the male reproductive system. However, to be considered complete, the response would also need to include information on the size of the prostate, its location, and its purpose. If any parts of the response were not correct, they were rated as inaccurate. See Table 1 for key points used to determine accuracy and completeness. Each of these key points is critical to a shared PrCA decision since a patient must consider factors such as the risks, benefits, and uncertainties of screening; their age, race, family history; and their personal values and preferences. Our chatbot responses have been included in Appendix 1. Additional space was allotted to record details about inaccuracies or omissions. The authors had 100% interrater agreement. The readability of responses was determined via an online Flesch Kincaid Reading Ease Scores calculator. Each response was copied and pasted into the calculator, excluding the reference website. The Flesch Kincaid Reading Ease Score uses total words, sentences, and syllables in an excerpt of text to calculate a score between 0 and 100, which corresponds to grade-level readability. Scores of 60 to 100 are considered easy to read by someone possessing an education at or below 8th to 9th grade. Scores of 50 to 60 require a 10th to 12th-grade education (i.e., fairly difficult) and scores below 50 require a college education (i.e. very difficult) to comprehend.

Table 1: Key Points for Determining Accuracy and Completeness^a

Questions	Key Points
<i>Basic Questions</i>	
1. What is the prostate?	• Male reproductive organ • The size of the prostate increases with age but is walnut-sized in younger men • Located below the bladder and in front of the rectum • Produces some of the fluid in semen
2. How common is prostate cancer?	• About 313,780 new cases of prostate cancer (1 in 8 men) • About 35,770 deaths from prostate cancer (1 in 44 men)
3. What are the risks for prostate cancer?	• Risk increases with age • More common among African-American men • More prevalent in North America, northwestern Europe, Australia, and the Caribbean islands. • Risk is doubled if a man has a first-degree relative (e.g., father, brother, son) with prostate cancer Gene mutations can increase the risk for prostate cancer • Less common risk factors are diet, obesity, smoking, chemical exposure, inflammation of the prostate, STIs, and vasectomy
4. When and how often should a man be screened for prostate cancer?	ACS: age 50 (average risk), 45, (high risk), 40 (very high risk) AUA: age 45 to 50 (average risk), age 40 (high risk) USPSTF: age 55 to 69 (average risk)
5. What are the symptoms of prostate cancer?	• Can have no symptoms in early stages • Urinary problems • Blood in urine or semen • Erectile dysfunction • Pain in hips, back, chest, or other areas • Weakness or numbness in legs or feet • Loss of bladder or bowel control
6. What are the types of screenings for prostate cancer?	• DRE, gloved finger test, not 100% accurate • PSA, a blood test, is not 100% accurate and can produce false positives and false negatives
7. What are the benefits and harms of prostate cancer screening?	• Benefit: can find cancer early • Harms: tests, especially PSA can produce false positives or negatives which can lead to unnecessary tests or treatments which carry risks.
8. How is prostate cancer diagnosed?	Biopsy – tissue samples from the prostate
9. What are the risks of a prostate biopsy?	Pain, blood in the semen, or infection
10. How long can I live if I have prostate cancer?	The 5-year relative survival rate is 97% on average, but depends on how far the cancer has spread

<i>Difficult Questions</i>	
11. Is the PSA or DRE more effective for finding prostate cancer?	PSA is more effective
12. My father had prostate cancer. Will I have prostate cancer too?	Having a father or brother with prostate cancer can more than double a man's risk of prostate cancer
13. I have a high PSA level. Do I have prostate cancer?	The probability of having prostate cancer increases with PSA level but there is no set PSA that can definitively indicate the presence of prostate cancer
14. What does a PSA level of 4 mean?	Men with a PSA level between 4 and 10 have about a 1 in 4 chance of having prostate cancer
15. What does a PSA level of 10 mean?	The chances of having prostate cancer are 50% with a PSA of 10 or more
16. What does a PSA level of 20 mean?	The chances of having prostate cancer are more than 50% with a PSA of 20 or more
17. What newer tests for prostate cancer may be more accurate than the PSA test?	• The prostate health index (PHI) • 4Kscore test • IsoPSA test • Urine-based tests
18. If my biopsy sample is positive for cancer, should I receive genetic testing?	Some men who have a strong family history or certain inherited genes, prior cancer diagnosis, or cancer that has spread to other parts of the body, should speak to their healthcare provider about this option
19. If my biopsy sample is positive for cancer, how soon should I start treatment?	Will depend on the stage and grade of the cancer and their: • Age and expected lifespan • Other serious health conditions • Feelings about treatment • The likelihood of a cure and doctor's opinion • Feelings about treatment side effects
20. Are there any cons to taking an at-home PSA test?	At home PSA tests do not give a man an opportunity to make a shared decision with their healthcare provider about the risks, benefits, and uncertainties of the PSA test
21. I am an African-American male, aged 40, with a family history of prostate cancer, at what age should I begin receiving prostate cancer screening?	Screening should begin at age 40 based on both the ACS and AUA screening guidelines
22. I am an African-American male, aged 40, with a family history of prostate cancer, can you	Response should include all key points such as: • Prostate cancer incidence and mortality statistics • Prostate cancer risks for African-American men • Symptoms for prostate cancer • Screenings for prostate cancer for African-American men

provide me with all of the information I need to know to make a shared decision about prostate cancer screening?	• Risks and uncertainties of prostate cancer screening • Meaning of PSA results • Biopsy for diagnosis • Risk of biopsy • Steps after a positive biopsy
23. What are the differences in screening recommendations between major health organizations?	• ACS: age 50 (average risk), 45, (high risk), 40 (very high risk) • AUA: age 45 to 50 (average risk), age 40 (high risk) • USPSTF: age 55 to 69 (average risk)

^aKey points developed from web sources produced by ACS, CDC, AUA, and UPSTF

Data Analysis

Data was transferred from coding forms to Microsoft Excel spreadsheets for analysis. Descriptive statistics were calculated to determine the percentage of questions answered accurately and completely. An average mean readability score was also calculated.

Results

Accuracy and Completeness

Responses to all questions were accurate. Seventeen of twenty-three questions (74%) were answered completely. Of the six questions with less complete responses, one lacked information about geography as a risk for PrCA and the higher prevalence of PrCA in North America. Of note is that this response recognized that African Americans may be at greater risk for the disease, but statistics were not provided in any responses that substantiated the burden of incidence and mortality among African-American men. Three questions related to the meanings of PSAs of 4, 10, and 20 lacked statistics about the probability of PrCA, but did state men's greater chance of being diagnosed with PrCA at PSAs higher than 4. A fifth question about how soon a man should start treatment after a positive biopsy lacked information about how age, expected life span, comorbidities, and patient feelings about side effects factor into treatment decisions. Lastly, a sixth question about what information an African-American male, aged 40 with a family history of PrCA needs to know to make a shared screening decision yielded an answer that lacked information about what PSA results mean or the purpose and risks of a prostate biopsy.

Readability

The average readability was 64.5 ± 8.7 which indicates most responses were written at an 8th-grade level or below. However, five of twenty-three responses (22%) were written at a 10th to 12th grade reading level and one response was written at a college level. Three of the five responses addressed difficult questions. Scores ranged from 48.6 to 81.3. The lowest readability score (i.e., 48.6) was in response to a basic question about symptoms of PrCA. See Table 2.

Table 2. Accuracy, Completeness, and Readability of Prostate Cancer Info Responses to Questions About Prostate Cancer

Questions	Accurate?	Complete?	Readability Score
-----------	-----------	-----------	-------------------

<i>Basic Questions</i>			
1. What is the prostate?	Yes	Yes	81.3
2. How common is prostate cancer?	Yes	Yes	79.4
3. What are the risks for prostate cancer?	Yes	No	65.5
4. When and how often should a man be screened for prostate cancer?	Yes	Yes	70.3
5. What are the symptoms of prostate cancer?	Yes	Yes	48.6 ^b
6. What are the types of screenings for prostate cancer?	Yes	Yes	70.8
7. What are the benefits and harms of prostate cancer screening?	Yes	Yes	64.6
8. How is prostate cancer diagnosed?	Yes	Yes	71.5
9. What are the risks of a prostate biopsy?	Yes	Yes	56.1 ^a
10. How long can I live if I have prostate cancer?	Yes	Yes	63
<i>Difficult Questions</i>			
11. Is the PSA or DRE more effective for finding prostate cancer?	Yes	Yes	74.7
12. My father had prostate cancer. Will I have prostate cancer too?	Yes	Yes	60.4
13. I have a high PSA level. Do I have prostate cancer?	Yes	Yes	70.7
14. What does a PSA level of 4 mean?	Yes	No	67.5
15. What does a PSA level of 10 mean?	Yes	No	67.2
16. What does a PSA level of 20 mean?	Yes	No	67.5
17. What newer tests for prostate cancer may be more accurate than	Yes	Yes	63.1

the PSA test?			
18. If my biopsy sample is positive for cancer, should I receive genetic testing?	Yes	Yes	52.4 ^a
19. If my biopsy sample is positive for cancer, how soon should I start treatment?	Yes	No	58.8 ^a
20. Are there any cons to taking an at-home PSA test?	Yes	Yes	65
21. I am an African-American male, aged 40, with a family history of prostate cancer, at what age should I begin receiving prostate cancer screening?	Yes	Yes	51.8 ^a
22. I am an African-American male, aged 40, with a family history of prostate cancer, can you provide me with all of the information I need to know to make a shared decision about prostate cancer screening?	Yes	No	51.6 ^a
23. What are the differences in screening recommendations between major health organizations?	Yes	Yes	60.6
	Total Yes=100%	Total Yes=74%	Average Readability Score = 64.5 (SD= 8.7) Range = 48.6 – 81.3

^aReadability was fairly difficult (requires a 10th to 12th-grade education)

^bReadability was very difficult (requires a college education)

Discussion

Principal Results

PCI had pristine accuracy and average completeness/readability. On average, completeness and readability were higher on responses to basic questions as compared to difficult questions. Specifically, nine of ten or 90% of responses to basic and eight of thirteen or 62% of responses to difficult questions were complete. Also, eight of ten or 80% and nine of thirteen or 69% of readability scores for basic and difficult questions, respectively, were below an 8th to 9th grade level. Difficult questions often contained longer and more complex responses, which likely affected readability. Four of the six incomplete responses only lacked one key point which did not significantly dilute these responses. For example, three responses on PSA levels at 4, 10, and 20 did not effectively highlight differences in cancer likelihood (e.g., over 50% chance), but each response indicated a greater chance of prostate cancer. Therefore, men would be informed that a PSA over 4 is concerning and warrants counsel from a provider. Key points missed in responses about when to start treatment for any man and about shared decision-making for African-American men are more concerning as the omitted information (e.g., biopsy as a diagnosis tool) is focal to a PrCA screening decision. Not possessing this knowledge could lead to a PrCA screening decision that is not ideally informed and may not truly be shared between the patient and their healthcare provider. Specifically, knowing that the biopsy, not the PSA, is the only definitive means to diagnose PrCA may somewhat lessen the fear of an increased PSA score because another diagnostic step exists. Being informed about the biopsy could also prompt shared discussion about the relevance of a biopsy for the patient's circumstance. In addition, providing African-American men with all the information necessary to share a PrCA screening decision based on their demographic profile could be exceptionally useful for those men who may lack access to a question list, and not have time to ask multiple questions to an AI chatbot, or simply want a more tailored answer to their given circumstance. This tailored information can also facilitate a shared PrCA screening decision that is more patient-centered.

Limitations

The 23 questions we used may not reflect the full breadth of inquiries someone may have about PrCA screening. PCI was programmed to seek information from a finite set of web pages from key medical authorities, but several equally credible websites were not included (e.g., Mayo Clinic) which may have slightly improved PCI's performance. While much of the general information about PrCA on these additional websites (e.g., as signs, symptoms, and prostate anatomy) would likely be similar, there may be cutting-edge research on new PrCA screenings that may not yet be publicized on ACS or similar websites but could provide more context for more difficult questions like those related to newer tests for PrCA. Lastly, although rigorous research methods, such as interrater reliability, were used to mitigate any study bias, we acknowledge that as the developers of PCI, we may be susceptible to unconscious biases that could affect our ratings. For transparency, we have included all PCI responses in Appendix 1. Future studies to evaluate PCI and similar chatbots should include external raters and user feedback.

Comparison with Prior Work

Similar to our prior research [23] and research by others [15-19, 21, 22] generative AI chatbots like PCI can be highly accurate when responding to PrCA and PrCA screening inquiries. The completeness and readability of Prostate Info's responses to PrCA screening questions varied. PCI generally performed better than Lombardo et al. [20] and comparable to Geantă, M., et al., [20-22], both of whom investigated chatbot performance on non-U.S. PrCA guidelines. As compared to

studies using U.S. PrCA guidelines, PCI performed better than ChatGPT-3.5, ChatGPT-4, Microsoft Co-Pilot, Google Gemini, and Google Gemini Advanced, but equal to Microsoft Copilot on completeness of responses to basic PrCA screening queries posed in our prior comparative study which solicited both standard and plain-language responses [23]. However, the average readability on basic questions was lower than all, but one (i.e., Microsoft Co-Pilot) generative AI chatbot when considering plain language responses only [23]. Otherwise, PCI outperformed all, but one (i.e. Google Gemini Advanced) chatbot when we asked it to provide a standard response [23]. Compared to Zhu et al. [26], who evaluated multiple generative AI chatbots' performance on a similar combination of basic and difficult questions to our study, PCI did not perform nearly as well as ChatGPT and ChatGPT Plus (earlier versions of Chat GPT) on completeness. PCI outperformed all other chatbots evaluated by Zhu et al. [26] including Perplexity, YouChat, Chatsonic, and NeevaAI. However, PCI's average readability may have been (but not definitively) lower than all generative AI chatbots evaluated in Zhu et al.'s study. It is important to note that Zhu et al. [26] used a slightly different method for calculating completeness and readability than the current study or study by Owens et al. [23]. Zhu et al. [26] determined the percentage of comprehensiveness as opposed to indicating whether a response was simply complete or not complete. Numbers listed in Table 3 for Zhu et al. represent the percentage of questions that were "very comprehensive" (i.e. fully complete). Readability was rated by reviewers as opposed to using a validated readability measure. Percentages reported in Table 3 for Zhu et al., represent that percentage of total responses that were "very easy to read." Standard deviations for Zhu et al. were not reported.

Table 3. Comparison of Completeness and Readability of Chatbot Responses on U.S. Prostate Cancer Screening Guidelines

Chatbot Name (study authors)	Completeness (%)	Average Readability Score or % (Standard deviation)
<i>PCI (current study)</i>	74%	64.5 (8.7)
ChatGPT (Zhu et al)	95% ^a	100% (NR)
ChatGPT Plus (Zhu et al.)	92% ^a	100% (NR)
Perplexity Detailed (Zhu et al.)	65%	95% (NR)
ChatSonic (Zhu et al.)	65%	95% (NR)
YouChat (Zhu et al.)	47%	98% (NR)
Neeva AI (Zhu et al.)	40%	84% (NR)
Perplexity Concise (Zhu et al.)	30%	95% (NR)
ChatGPT 3.5 standard response (Owens et al.)	54%	38.0 (7.6)
ChatGPT 3.5 low literacy response (Owens et al.)	34%	70.3 (7.2) ^b
ChatGPT 4.0 standard response (Owens et al.)	63%	43.1 (9.2)
ChatGPT 4.0 low literacy response (Owens et al.)	63%	74.1 (9.9) ^b
Google Gemini standard response (Owens et al.)	54%	55.7 (10.4)
Google Gemini low literacy response (Owens et al.)	45%	81.0 (3.6) ^b
Google Gemini Advanced standard response (Owens et al.)	54%	66.3 (9.4) ^b
Google Gemini Advanced low literacy response (Owens et al.)	54%	79.4 (5.1) ^b
Microsoft Copilot standard response (Owens et al.)	72%	50.8 (9.3)
Microsoft Copilot low literacy response (Owens et al.)	54%	65.1 (6.6) ^b

Microsoft Copilot Pro standard response (Owens et al.)	63%	61.2 (9.5)
Microsoft Copilot Pro low literacy response (Owens et al.)	54%	78.8 (4.7) ^b

^a Chatbot had a higher completeness score than PCI

^b Chatbot had definitively higher readability scores than PCI based on the Flesch-Kincaid readability. Other scores may also be higher but were not based on a validated measure.

We expected PCI to outperform most other commercially available generative AI chatbots on completeness because of its development using the latest ChatGPT-4.0 technology and its directive to secure information from specific websites, but PCI underperformed earlier versions of ChatGPT. Therefore, additional training of the large language model that undergirds ChatGPT-4.0 will be needed for this niche area. Also unexpected was the lower average readability of responses from PCI, especially compared to our previous work, which solicited plain-language responses from multiple chatbots including ChatGPT-3.5 and ChatGPT-4. Nonetheless, the average readability of PCI is suitable for an audience with a middle school education.

Conclusions and Future Directions

Generative AI chatbots, such as PCI, are great starting places for learning about PrCA screening and preparing for shared decision-making but should not yet be used as sole sources of PrCA information because of their periodic omission of key information. Nevertheless, with further testing and validation, model training, and refinement of the source selection process, we hope PCI can be a publicly available resource for credible, evidence-based, and culturally appropriate information for PrCA screening decisions. In the future, PCI could be integrated into the decision-making workflow by prompting patients to use PCI prior to their medical visit via an emailed link or hard copy one-pager with a QR code. This same email or document could contain multiple frequently asked questions about PrCA. Men should be encouraged to pose as many of these questions as possible, but especially those on our list that are more complex (e.g., Is the PSA or DRE more effective for finding prostate cancer?) PCI questions and responses could then be saved and/or printed; notated to indicate areas of concern or need for clarity; and taken to their appointment to be used to guide the shared PrCA screening discussion. During this discussion, the healthcare provider should ensure that men understand their personal PrCA risk; screening options; risks, benefits, and uncertainties of PrCA screening. The discussion should then shift to focus on men's questions and their screening preferences. Using this method of generative AI integration into the shared decision process could fortify men's PrCA knowledge and identify patient values and preferences.

To improve the overall performance of PCI in the future, it will be necessary to iteratively fine-tune our model which will include expanding the sources from which PCI retrieves data which could include the most current peer-reviewed journal articles in addition to websites of major research hospitals and international health organizations. All sources will be curated by a PrCA expert who will review each data source to ensure it contains quality information. Equally important will be soliciting routine feedback from healthcare providers and patients through an embedded satisfaction survey that can enable them to comment on the quality of questions developed by the research team, potential questions that should be added to the database, and the quality of the responses generated by the PCI. Additional model training will come from tracking common follow-up questions from users to incorporate them into the initial responses. Based on these continuous feedback loops, PCI's performance could be improved significantly and always remain up-to-date. Future research should focus on the clinical deployment of PCI and testing to assess its acceptability, ease of use in a clinical workflow, and usefulness in the shared PrCA screening decision process.

Acknowledgments

None noted.

Conflicts of Interest

None declared.

Abbreviations

AI: Artificial intelligence

ACS: American Cancer Society

AUA: American Urological Association

CDC: Centers for Disease Control and Prevention

PrCA: Prostate Cancer

PCI: Prostate Cancer Info

USPSTF: United States Preventive Services Task Force

References

1. Open AI. How ChatGPT and our foundation models are developed. Accessed on February 11, 2025 at <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed>.
2. Milne-Ives, M., et al., *The effectiveness of artificial intelligence conversational agents in health care: systematic review*. Journal of medical Internet research, 2020. **22**(10): p. e20346.
3. Wang L, et al. A systematic review of ChatGPT and other conversational large language models in healthcare. medRxiv; 2024 Online pre-print. doi: 10.1101/2024.04.26.24306390
4. Xu L, et al. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer; 2021 7(4): e27850. doi: 10.2196/27850.
5. Potapenko I, et al. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol; 2023 101(7): 829-831. doi:10.1111/aos.15661
6. Shuaib A. Transforming healthcare with AI: promises, pitfalls, and pathways forward. Int. J. Gen. Med; 2024 1(17): 1765-1771. doi: 10.2147/IJGM.S449598
7. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*; 2024 74(1):12-49. doi: 10.3322/caac.21820
8. Wolf AM, et al. American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA Cancer J Clin*; 2010 60(2):70-98. <https://doi.org/10.3322/caac.20066>
9. Wei JT, et al. Early detection of prostate cancer: AUA/SUO guideline part I: prostate cancer screening. *J Urol*; 2023 210(1):46-53. <https://doi.org/10.1097/JU.0000000000003491>
10. Wei JT, et al. Early detection of prostate cancer: AUA/SUO guideline part II: considerations for a prostate biopsy. *J Urol*; 2023 210(1):54-63. <https://doi.org/10.1097/JU.0000000000003492>
11. US Preventive Services Task Force. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *JAMA*; 2018 319(18):1901-1913. doi:10.1001/jama.2018.3710
12. Giaquinto AN, et al. Cancer statistics for African American/Black People 2022. *CA Cancer J Clin*; 2022 72(3): p. 202-229. <https://doi.org/10.3322/caac.21718>
13. Grene M, Cleary Y, Marcus-Quinn A. Use of plain-language guidelines to promote health literacy. *IEEE Trans Prof Comms*; 2017 60(4): p. 384-400. DOI:10.1109/TPC.2017.2761578
14. Coskun B, et al. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? *Urol*; 2023 180: p. 35-58. <https://doi.org/10.1016/j.urology.2023.05.040>
15. Zhu L, Mou W, 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): p. 269. <https://doi.org/10.1186/s12967-023-04123-5>
16. Alasker A, Alsalamah S, Alshathri N, et al. Performance of Large Language Models (LLMs) in Providing Prostate Cancer Information. *BMC Urol*; 2024 24(177). <https://doi.org/10.1186/s12894-024-01570-0>

17. Musheyev D, et al., How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? *Euro Urol*; 2024 85(1): p. 13-16. <https://doi.org/10.1016/j.eururo.2023.07.004>
18. Gibson, D., et al., Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment. *J Med Internet Res*; 2024 26 (1): p. e55939. <https://doi.org/10.2196/55939>.
19. Chiarelli, G., et al., Adequacy of prostate cancer prevention and screening recommendations provided by an artificial intelligence-powered large language model. *Int Urol Nephrol*; 2024 56(4): p. 1-7. <https://doi.org/10.1007/s11255-024-03156-0>.
20. Lombardo, R., et al., Quality of information and appropriateness of Open AI outputs for prostate cancer. *Prostate Cancer Prostatic Dis*, 2024. <https://doi.org/10.1038/s41391-024-00789-0>.
21. Geantă, M., et al., The emerging role of large language models in improving prostate cancer literacy. *Bioengineering*, 2024 11(7): p. 654. <https://doi.org/10.3390/bioengineering11070654>.
22. Geantă, M., et al., The potential impact of large language models on doctor–patient communication: A case study in prostate cancer. *Healthcare*, 2024 12(15): p. 1548. <https://doi.org/10.3390/healthcare12151548>.
23. Owens, O., Leonard, MS, A Comparison of Prostate Cancer Screening Information Quality on Standard and Advanced Versions of Chat GPT, Google Gemini, and Microsoft Co-Pilot: A Cross-Sectional Study. *American J Health Promot*, 2025. doi: 10.1177/08901171251316371.
24. Open AI. GPT Builder. Accessed on February 11, 2025 at <https://help.openai.com/en/articles/8770868-gpt-builder>.
25. Weiss BD. *Health literacy and patient safety: Help patients understand. Manual for clinicians*. 2nd ed. AMA; 2007.
26. American Cancer Society. *Prostate Cancer*; 2025. Accessed on February 14, 2025 at <https://www.cancer.org/cancer/types/prostate-cancer.html>.
28. Centers for Disease Control and Prevention. *Prostate Cancer*; 2024. Accessed on February 14, 2025 at <https://www.cdc.gov/prostate-cancer/index.html>.

Supplementary Files

Untitled.

URL: <http://asset.jmir.pub/assets/4898904f5d493bce5ea9135be4d20503.docx>

Please Select A Types

Generative AI Chatbot Responses.

URL: <http://asset.jmir.pub/assets/38702c4268a5d91ee8822ce515da697a.docx>

Related publication(s) - for reviewers eyes onlies

Point by Point Response to Reviewers.

URL: <http://asset.jmir.pub/assets/c9b39550a73d416ce48174cbf020b946.pdf>