Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.journals.elsevier.com/patient-education-and-counseling

Are AI chatbots concordant with evidence-based cancer screening recommendations?

Brooke Nickel[a,b,*], Julie Ayre[a], M Luke Marinovich[c], David P. Smith[c], Karen Chiam[c], Christoph I. Lee[d], Timothy J. Wilt[e,f], Melody Taba[a], Kirsten McCaffery[a,b], Nehmat Houssami[b,c]

[a] *Sydney Health Literacy Lab, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Australia*

[b] *Wiser Healthcare, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Australia*

[c] *The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, NSW, Australia*

[d] *Department of Radiology, University of Washington School of Medicine; Fred Hutchinson Cancer Center, Seattle, WA, USA*

[e] *Minneapolis VA Center for Care Delivery and Outcomes Research (CCDOR), MN, USA*

[f] *School of Medicine and School of Public Health, University of Minnesota, MN, USA*

ARTICLE INFO

Keywords:
Breast cancer
Prostate cancer
Screening
Public health
Artificial intelligence
Informed decision making

ABSTRACT

Objective: This study aimed to assess whether information from AI chatbots on benefits and harms of breast and prostate cancer screening were concordant with evidence-based cancer screening recommendations.

Methods: Seven unique prompts (four breast cancer; three prostate cancer) were presented to ChatGPT in March 2024. A total of 60 criteria (30 breast; 30 prostate) were used to assess the concordance of information. Concordance was scored between 0 and 2 against the United States Preventive Services Task Force (USPSTF) breast and prostate cancer screening recommendations independently by international cancer screening experts.

Results: 43 of 60 (71.7 %) criteria were completely concordant, 3 (5 %) were moderately concordant and 14 (23.3 %) were not concordant or not present, with most of the non-concordant criteria (9 of 14, 64.3 %) being from prompts for the oldest age groups. ChatGPT hallucinations (i.e., completely made up, non-sensical or irrelevant information) were found in 9 of 60 criteria (15 %).

Conclusions: ChatGPT provided information mostly concordant with USPSTF breast and prostate cancer screening recommendations, however, important gaps exist. These findings provide insights into the role of AI to communicate cancer screening benefits and harms and hold increased relevance for periods of guideline change.

Practice implications: AI generated information on cancer screening should be taken in conjunction with official screening recommendations and/or information from clinicians.

1. Introduction

Cancer screening involves a complex array of potential consequences including both benefits and harms. These underpin differences in recommendations including whether, how often and at what age to be screened. The United States Preventive Service Task Force (USPSTF) is an independent panel of experts that makes evidence-based recommendations about preventive services [1], including breast and prostate cancer screening. Providing evidence-based information to the public and their clinicians about cancer screening is crucial to support informed decision making [2]. However, with the public continuing to turn online to ask health questions and seek advice [3–5], it is important to understand how health information online compares to evidence-based information about cancer screening.

While many generative Artificial Intelligence (AI) tools exist, one model for health advice increasingly being used by the public is AI chatbots, including ChatGPT. Like other AI chatbots, ChatGPT produces plausible, contextually appropriate and human-like responses to questions by learning from large volumes of data to predict text or dialogue. While ChatGPT and other AI chatbots are generally easy to use and freely available, they are not always correct despite sounding confident. This, therefore, has the potential to lead to public health risks stemming from misinformation including, but not limited to, having a disproportionate impact on vulnerable communities including those with low health literacy and from non-English speaking backgrounds [6,7]. Evidence is emerging on how ChatGPT responds to various health-related

* Correspondence to: The University of Sydney, 127A Edward Ford Building (A27), NSW 2006, Australia.
E-mail address: brooke.nickel@sydney.edu.au (B. Nickel).

https://doi.org/10.1016/j.pec.2025.108677
Received 21 October 2024; Received in revised form 15 December 2024; Accepted 20 January 2025
Available online 21 January 2025

questions [8], including in the context of commonly asked questions about cancer and treatment advice [9–12]. However, there has been no data on whether ChatGPT cancer screening responses are consistent with evidence-based recommendations. Given the lack of consensus about the benefits and harms of breast and prostate cancer screening in recent years [13], this information will help to better understand implications for cancer screening communication.

2. Methods

2.1. Study design

This cross-sectional descriptive study assessed whether information on benefits and harms of cancer screening using ChatGPT were concordant with the USPSTF recommendations for breast (2016) [14] and prostate (2018) [15] cancer. Seven unique prompts (four breast cancer and three prostate cancer – see supplement 1) were presented to ChatGPT (GPT-3.5-turbo-0301 model, freely available at the time of the study) in March 2024. These prompts were developed by experts in public health including cancer screening, health literacy and digital health, and cancer clinicians. They were presented to, discussed and refined with a community panel of diverse consumers in which breast and prostate cancer screening is of relevance. They were purposely designed to contain common and simple language. Prompts for both breast and prostate cancer included one question on the general benefits and harms of screening with the remaining including the same question with age-specific information to provide personal context. Age groups were specifically chosen both within and beyond the recommendations for screening. Ethics approval was not obtained as all data were derived from publicly available information online.

2.2. Assessment and analysis

The concordance of information provided by ChatGPT were assessed independently by international cancer screening experts (three breast; three prostate). The scoring template was designed based on information included in the USPSTF recommendations, discussed with the study team and tested by two international experts (one breast; one prostate). A total of 60 criteria (30 breast; 30 prostate – see Table 1) relating to benefits (e.g., improved breast cancer-specific mortality), harms (e.g., false-positives), and recommendations (e.g., informed/shared decision making) were used to assess the concordance of information. Concordance was scored out of 2: 0 =not concordant i.e., complete omission of information or incorrect interpretation, 1 =moderately concordant i.e., partial omission of information or moderately correct interpretation, and 2 =completely concordant i.e., consistent or in complete agreement (see supplement 2). Scores (1 and 2) were reduced by one mark if ChatGPT provided hallucinated information i.e., completely made up, non-sensical or irrelevant. A majority rule was used for agreement; where this was not achieved, experts re-scored to reach majority agreement. Data was analysed using Excel version 16.85 (Microsoft Corp).

3. Results

For 52 of the 60 criteria (86.7 %), information was present in the ChatGPT outputs. Breast criteria were present in 29 of 30 outputs (96.7 %) and prostate criteria in 23 of 30 (76.7 %), with 6 of 7 absent criteria for prostate being on benefits (Table 1).

The majority of criteria (43 of 60, 71.7 %) were completely concordant with the USPSTF recommendations, 3 (5 %) were moderately concordant and 14 (23.3 %) were not concordant or not present, with most of non-concordant criteria (9 of 14, 64.3 %) being from prompts for the oldest age group (age 75 for breast and 70 for prostate). Across all criteria, there was a mean concordance score of 1.48 out of 2 (SD=0.85). Overall concordance for all present and non-present criteria

Table 1
Breast and prostate cancer screening information and concordance against the United States Preventive Service Task Force Recommendations[*].

Information criteria	**General or Age-specific prompt used**	**Presence vs absence of information**
		Present = 1
		Absent = 0
			Concordance[**]
			0 = not concordant
			1 = somewhat/ moderately concordant
			2 = concordant

BREAST			
Benefits			
Improved breast cancer-specific mortality	General	1	2
	45	1	2
	60	1	2
	75[†]	1	0
Reduced advanced-stage breast cancer	General	1	1
	45	1	2
	60	1	2
	75[†]	1	0
Harms	General	1	2
False positives	45	1	2
	60	1	2
	75	1	0
Breast biopsies	General	1	2
	45	1	2
	60	1	1
	75	1	0
False negatives	General	1	2
	45	1	2
	60	0	0
	75	1	2
Overdiagnosis	General	1	2
	45	1	2
	60	1	2
	75	1	0
Overtreatment	General	1	2
	45	1	2
	60	1	2
	75	1	0
Psychological[‡]	45	1	2
Recommendations (e.g., informed decision making)[‡]	45	1	2
PROSTATE			
Benefits			
Improved prostate cancer – specific mortality	General[†]	0	0
	60[†]	0	0
	70[†]	0	0
Reduced metastatic prostate cancer	General[†]	0	0
	60[†]	0	0
	70[†]	0	0
Harms			
False positives	General	1	2
	60	1	2
	70	1	2
Psychological	General	1	2
	60	1	1
	70	1	2
Harms of diagnostic procedures	General	1	2
	60	1	2
	70[†]	0	0
Overdiagnosis	General	1	2
	60	1	2
	70	1	2
Overtreatment	General	1	2
	60	1	2
	70	1	2
Erectile dysfunction	General	1	2
	60	1	2
	70	1	2
Urinary incontinence	General	1	2

(continued on next page)

Table 1 (*continued*)

Information criteria	General or Age-specific prompt used	Presence vs absence of information
		Present = 1
		Absent = 0
			Concordance[**]
			0 = not concordant
			1 = somewhat/ moderately concordant
			2 = concordant

	60	1	2
	70	1	2
Recommendations (e.g., shared/informed decision making)	General	1	2
	60	1	2
	70	1	2

[*] Breast: Siu A. L. on behalf of the U.S. Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Annals of Internal Medicine*. 2016;164(4):279–96.

Prostate: U. S. Preventive Services Task Force, Grossman DC, Curry SJ, et al. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2018;319(18):1901–13.

[**] Scores (1 and 2) were reduced by one mark if hallucinated information was provided by ChatGPT.

[†] Hallucinations present.

[‡] Only one age-related (40–49 years) prompt applicable.

were similar between breast and prostate (72.4 % and 73.3 %). Noting that not all criteria were applicable for all age groups, for breast cancer there was complete concordance in 9/9 (100 %) for age 45, 5/7 (71.4 %) for age 60 and 1/7 (14.3 %) for age 75 and for prostate cancer 7/10 (70.0 %) for age 60 and 7/10 (70.0 %) for age 70.

ChatGPT hallucinations were found in 9 out of 60 criteria (15 %), mainly related to extrapolation of benefits (in the age 75 + prompt for breast and all prompts for prostate) where the information criteria were already deemed as not present or not concordant.

4. Discussion

Similar to other data on general and treatment information for cancer [9,10], ChatGPT performed reasonably well and for the majority of responses ChatGPT provided information concordant with USPSTF recommendations for breast and prostate cancer. However, important gaps exist. Almost one-third of the ChatGPT information on the benefits and harms of breast and prostate cancer screening was completely or partially non-concordant with the USPSTF recommendations, including in 15 % of the outputs where ChatGPT produced hallucinated responses. Notably, for prostate cancer, the USPSTF benefit information was not present in ChatGPT outputs. Also, the prompts for the oldest age groups, particularly breast (age 75), were the least concordant.

Other recent studies that assessed ChatGPT's output information on breast and prostate cancer, not specific to screening, found varying results in accuracy with breast results seeming to show more promise [11,12]. While findings from our study reveal that ChatGPT information may provide the public with an unbalanced view of the benefits and harms evidence, particularly for prostate cancer benefits, it highlights some of the ongoing uncertainties in both breast [16,17] and prostate [18] cancer screening. Interestingly, information on patient-related harms, which is often less likely to be reported in relation to cancer screening [19], was provided in adequate detail in the ChatGPT information. ChatGPT may therefore provide some useful information to the public within the currently recommended screening ages for breast and prostate cancer, particularly in terms of screening related harms. However, it should not be taken as an evidence-based recommendation and the public should take caution when using it alone to make screening or health decisions. In particular, in those with low health literacy, there is limited evidence showing that ChatGPT can adequately simplify information, although it may have the capacity to do this [8]. Future research could assess how the public, with a broad range of health literacy, interprets the benefits and harms of breast and prostate cancer screening in ChatGPT outputs and make decisions based on this information.

4.1. Strengths and limitations

Limitations of this study are that information was evaluated at one point in time and assessed against USPSTF recommendations only, including the 2016 breast cancer guidelines which were recently updated (May 2024) [20]. While this may limit the generalisability of the findings, the USPSTF is arguably the most widely known evidence-based recommendations for cancer screening internationally and this version of ChatGPT used information from available guidelines at the time of the study. Furthermore, while the prompts were consumer-guided, they did not include potential hallucination mitigation strategies [21] which could be explored further. Public comprehension of the ChatGPT outputs were also not assessed.

4.2. Conclusions

Findings from this study provide initial insights into how ChatGPT performs compared to published USPSTF recommendations for the public and clinicians. With growing concerns around the risks associated with the use of AI to produce false or misleading content [22], this study is the first to sheds light on current strengths and weakness of AI to communicate cancer screening benefit and harms.

4.3. Practice implications

As cancer screening guidelines continue to change it is likely that the public will increasingly look to AI chatbots, including ChatGPT, for updated guidance on screening. AI generated information on cancer screening should be taken in conjunction with official screening recommendations and/or information from clinicians. There is also a need for guardrails to this information to be monitored and improved overtime. Importantly, there is a vital need to integrate AI literacy into public health campaigns to help individuals better prompt, understand and assess AI-generated health information.

Funding

BN, JA, KM and NH are funded through an Australian National Health and Medical Research Council Investigator Grant (NHMRC: 1194108, 2017278, 2016719, 1194410). LM is funded through a National Breast Cancer Foundation Investigator Initiated Research Scheme Grant (2023/IIRS0028). DPS and KC receive grant funding from Prostate Cancer Foundation of Australia. NH is additionally funded through a National Breast Cancer Foundation Chair in Breast Cancer Prevention grant (EC-21-001). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

CRediT authorship contribution statement

Nickel Brooke: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Ayre Julie:** Writing – review & editing, Methodology, Conceptualization. **Marinovich M Luke:** Writing – review & editing, Methodology, Formal analysis. **Smith David:** Writing – review & editing, Formal analysis. **Chiam Karen:** Writing – review & editing, Formal analysis. **Lee Christoph I:** Writing – review & editing, Formal analysis. **Wilt Timothy J:** Writing – review & editing, Formal analysis. **Taba Melody:** Writing – review & editing, Methodology. **McCaffery Kirsten:** Writing – review & editing, Methodology, Conceptualization. **Houssami Nehmat:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of Competing Interest

BN and KM are current members of the International Scientific Committee of Preventing Overdiagnosis. CIL receives textbook royalties from UpToDate, Inc., McGraw Hill, Inc., and Oxford University Press; personal fees from the American College of Radiology for journal editorial board work; and research consulting fees from DeepHealth/RadNet for work outside of the submitted work. KM is the company director of Health Literacy Solutions pty, a consultancy that holds the sub-licence to the SHeLL Health Literacy Editor and takes no personal income for this role outside of the submitted work. All other authors declare no conflicts of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.pec.2025.108677.

References

[1] U. S. Preventive Services Task Force. ⟨https://www.uspreventiveservicestaskforce.org/uspstf/⟩.

[2] Hersch JK, Nickel BL, Ghanouni A, Jansen J, McCaffery KJ. Improving communication about cancer screening: moving towards informed decision making. Public Health Res Pract 2017;27(2).

[3] Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients' use of the Internet for medical information. J Gen Intern Med 2002;17(3):180–5.

[4] Bundorf MK, Wagner TH, Singer SJ, Baker LC. Who searches the internet for health information? Health Serv Res 2006;41(3 Pt 1):819–36.

[5] Wong C, Harrison C, Britt H, Henderson J. Patient use of the internet for health information. Aust Fam Physician 2014;43(12):875–7.

[6] Meyrowitsch DW, Jensen AK, Sørensen JB, Varga TV. AI chatbots and (mis) information in public health: impact on vulnerable communities. Front Public Health 2023;11:1226776.

[7] Ayre J, Cvejic E, McCaffery KJ. Who is asking ChatGPT health questions? Analysis of a nationally representative Australian community sample. medRxiv 2024;2024.10.13.24315426.

[8] Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, Rezwan A, Dunn AG. New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. J Gen Intern Med 2024;39(4):573–7.

[9] Chen S, Kann BH, Foote MB, Aerts H, Savova GK, Mak RH, Bitterman DS. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023;9(10):1459–62.

[10] Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol 2023;9(10):1437–40.

[11] Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 2023;307(4):e230424.

[12] 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;180:35–58.

[13] Carter SM. Why does cancer screening persist despite the potential to harm? Sci, Technol Soc 2021;26(1):24–40.

[14] Siu AL. on behalf of the U.S. Preventive Services Task Force., Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med 2016;164(4):279–96.

[15] U.S. Preventive Services Task Force, D.C. Grossman, S.J. Curry, D.K. Owens, K. Bibbins-Domingo, A.B. Caughey, K.W. Davidson, C.A. Doubeni, M. Ebell, J.W. Epling, Jr., A.R. Kemper, A.H. Krist, M. Kubik, C.S. Landefeld, C.M. Mangione, M. Silverstein, M.A. Simon, A.L. Siu, C.W. Tseng, Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement, JAMA: the journal of the American Medical Association 319(18) (2018) 1901-1913.

[16] Bell KJ, Nickel B, Pathirana T, Blennerhassett M, Carter S. Breast cancer screening from age 40 in the US. Bmj 2024;385:q1353.

[17] Mathieu E, Noguchi N, Li T, Barratt AL, Hersch JK, De Bock GH, Wylie EJ, Houssami N. Health benefits and harms of mammography screening in older women (75+ years)-a systematic review. Br J Cancer 2024;130(2):275–96.

[18] Carter HB. Prostate-specific antigen (PSA) screening for prostate cancer: revisiting the evidence. JAMA: J Am Med Assoc 2018;319(18):1866–8.

[19] Kamineni A, Doria-Rose VP, Chubak J, Inadomi JM, Corley DA, Haas JS, Kobrin SC, Winer RL, Elston Lafata J, Beaber EF, Yudkin JS, Zheng Y, Skinner CS, Schottinger JE, Ritzwoller DP, Croswell JM, Burnett-Hartman AN. Evaluation of harms reporting in U.S. cancer screening guidelines. Ann Intern Med 2022;175(11):1582–90.

[20] Services USPreventive, Force Task. Screening for Breast Cancer: US Preventive Services Task Force Recommendation Statement. JAMA: J Am Med Assoc 2024;331(22):1918–30.

[21] Towhidul Islam Tonmoy S, Mehedi Zaman S, Jain V, Rani A, Rawte V, Chadha A, Das A, Comprehensive A. Survey of hallucination mitigation techniques in large language models. arXiv 2401.01313v3 2024.

[22] Haupt CE, Marks M. FTC regulation of AI-generated medical disinformation. JAMA: J Am Med Assoc 2024.