BMC Women's Health

RESEARCH Open Access

An assessment of ChatGPT's responses to frequently asked questions about cervical and breast cancer

Zichen Ye[1†], Bo Zhang[1†], Kun Zhang[1], María José González Méndez[2], Huijiao Yan[1], Tong Wu[1], Yimin Qu[1], Yu Jiang[1,3*], Peng Xue[1*] and Youlin Qiao[1]

Abstract

Background Cervical cancer (CC) and breast cancer (BC) threaten women's well-being, influenced by health-related stigma and a lack of reliable information, which can cause late diagnosis and early death. ChatGPT is likely to become a key source of health information, although quality concerns could also influence health-seeking behaviours.

Methods This cross-sectional online survey compared ChatGPT's responses to five physicians specializing in mammography and five specializing in gynaecology. Twenty frequently asked questions about CC and BC were asked on 26th and 29th of April, 2023. A panel of seven experts assessed the accuracy, consistency, and relevance of ChatGPT's responses using a 7-point Likert scale. Responses were analyzed for readability, reliability, and efficiency. ChatGPT's responses were synthesized, and findings are presented as a radar chart.

Results ChatGPT had an accuracy score of 7.0 (range: 6.6-7.0) for CC and BC questions, surpassing the highest-scoring physicians ($P < 0.05$). ChatGPT took an average of 13.6 s (range: 7.6-24.0) to answer each of the 20 questions presented. Readability was comparable to that of experts and physicians involved, but ChatGPT generated more extended responses compared to physicians. The consistency of repeated answers was 5.2 (range: 3.4-6.7). With different contexts combined, the overall ChatGPT relevance score was 6.5 (range: 4.8-7.0). Radar plot analysis indicated comparably good accuracy, efficiency, and to a certain extent, relevance. However, there were apparent inconsistencies, and the reliability and readability be considered inadequate.

Conclusions ChatGPT shows promise as an initial source of information for CC and BC. ChatGPT is also highly functional and appears to be superior to physicians, and aligns with expert consensus, although there is room for improvement in readability, reliability, and consistency. Future efforts should focus on developing advanced ChatGPT models explicitly designed to improve medical practice and for those with concerns about symptoms.

Keywords Artificial intelligence, ChatGPT, Cervical cancer, Breast cancer, Frequently asked question

†Zichen Ye and Bo Zhang contributed equally to this work.

*Correspondence:
Yu Jiang
jiangyu@pumc.edu.cn
Peng Xue
xuepeng_pumc@foxmail.com

[1]School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[2]Department of Primary Healthcare and Family Medicine, Faculty of Medicine, Universidad de Chile, Santiago, Chile
[3]School of Health Policy and Management, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Introduction

Cervical cancer (CC) and breast cancer (BC) significantly threaten women's well-being, and early prevention has proven difficult because people tend to present late only when symptoms become more severe [1, 2]. The World Health Organization has called upon governments to strengthen screening practices, diagnostics, and treatments of both CCs and BCs [3, 4]. However, women (depending on the level of education and income, etc.) tend to be better healthcare seekers and are perhaps less reluctant to take proactive measures to monitor their health [5]. Therefore, women are more likely to utilize the Internet to gain insights into the early signs of conditions such as CC and BC. However, the Internet (broadly speaking) is not always a great source of reliable knowledge, as approximately one-third of all cancer-related articles on social media platforms are judged as either misleading or harmful [6, 7].

Artificial intelligence (AI) chatbots such as Siri and Cortana have become integral to our daily lives, although these tools are often limited in addressing scientific or professional matters. OpenAI's advanced chatbot, ChatGPT, stands out because it replicates human communications and can provide accurate answers to more complex problems [8–10]. However, the focus should be on the advantages and disadvantages for patients who might not be ready to present at a clinic or whose symptoms are not perceived to be severe enough to warrant an appointment with a general practitioner. Given the complexities of personal health and well-being, particularly concerning CC and BC, it is crucial to comprehensively analyze ChatGPT outputs to understand the potential effects of AI-generated advice.

There are few rigorously conducted studies of ChatGPT's responses, and there needs to be more population-based studies of chatbot advice for CC and BC. A previous study assessed ChatGPT's ability to handle clinical queries in obstetrics and gynaecology [11]. The answers ChatGPT provided were considered valuable as a source of preliminary information, but several drawbacks were identified. For example, advice may not represent the most recent information, and it is occasionally misleading. Several other issues should be considered, and the academic community must build an evidence base to help develop these technologies to ensure we provide the best, most up-to-date evidence, in a clear, comprehensible manner, to individuals with concerns about their health. This study aims to assess ChatGPT's ability to address women's inquiries about CC and BC, which could help develop chatbots like ChatGPT to reduce disparities and support women with health concerns.

Materials and methods

Study design and participants

This online, cross-sectional study conducted between March to May 2023 establishes an expert consensus as the reference standard and compares ChatGPT's responses with five physicians specializing in mammography and five specializing in gynaecology to 20 popular science questions. The ChatGPT responses were collated from instances recorded in Chile and Germany on April 26 and 29, 2023. The panel of three experts is experienced tumour epidemiologists with over 20 years of expertise in either CC or BC, and there are two gynaecologists and two mammologists working in tertiary hospitals with more than ten years of experience. The physicians involved in this study were dedicated front-line doctors who have worked in underserved communities in China for extensive periods.

This study was approved by the Institutional Review Board of the Chinese Academy of Medical Sciences and Peking Union Medical College (No. CAMS & PUMC-IEC-2022-022). All participants were required to sign an informed consent form before participating in this study.

Procedures

This study had four main phases, outlined in Fig. 1. The first phase involved formulating questions and 'reference standard' answers. Frequently-asked questions regarding CC and BC were developed based on popular science materials and current hot topics related to tertiary prevention. A set of 10 questions were designed for each cancer. Subsequently, experts were invited to review the proposed questions and formulate expert consensus answers. The final set of questions and expert consensus answers were determined through several rounds of consultation and assessment. Please refer to Table S1 in Supplementary material 1 for details.

During the second phase of this study, answer collection took place. Researchers accessed OpenAI (ChatGPT version 3.5) and posed 20 pre-defined questions using a standardized format. ChatGPT was asked for sources for each answer, explicitly asking for the basis of the response along with a link or reference plus citation time. Researchers recorded response times, answers, and sources provided by ChatGPT. Additionally, two researchers independently accessed OpenAI at different times and locations. Each of them asked the same question three times consecutively, resulting in six recordings. Two additional scenarios were introduced, and the results from these interactions were recorded. The specific method of questioning can be found in Appendix A in Supplementary material 2. Five mammographers and five gynaecologists were consulted through face-to-face consultations to obtain answers from physicians, simulating real clinical scenarios. Each physician only answered

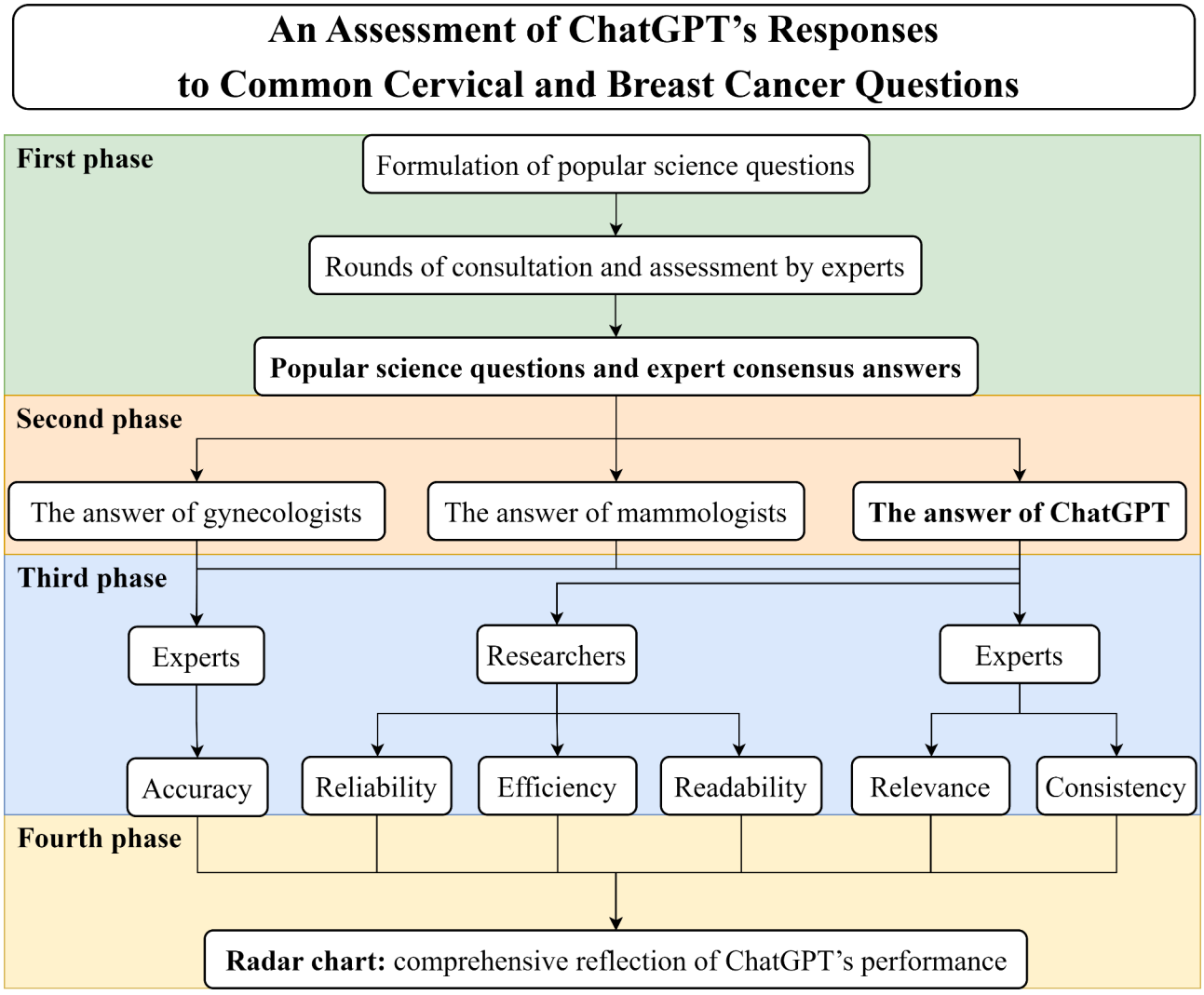

Fig. 1 Flow chart of this study. This study had four main phases. The first phase is to formulate questions and 'reference standard' answers for frequently-asked questions regarding CC and BC. The second phase is to collect the answers from ChatGPT and physicians. The third stage is to evaluate ChatGPT and physicians' answers through six dimensions. The fourth phase is to comprehensively reflect the performance of ChatGPT through radar chart. CC, cervical cancer; BC, breast cancer.

ten questions specifically related to their respective fields. Please refer to Table S1 and S3 in Supplementary material 1 for details.

In the third phase, an evaluation of responses was undertaken. Physicians' answers and ChatGPT's answers were randomly arranged while keeping identifiable information concealed. Subsequently, the experts who developed the reference answers assessed the accuracy of the six answers, one from ChatGPT and five from physicians, based on expert consensus and their professional knowledge. After assessing accuracy, experts assessed the consistency of the six answers regarding the repeated questions posed to ChatGPT, as well as the relevance of the additional scenarios. Researchers also assessed the readability, reliability, and efficiency of ChatGPT's answers. Additionally, the consistency of additional expert scores was calculated.

In the fourth phase, a comprehensive assessment was performed. ChatGPT's performance was visually assessed with a radar chart. This chart intercalated accuracy, readability, consistency, reliability, efficiency, and relevance to reflect the feasibility of applying ChatGPT to the comprehensive prevention and treatment of CC and BC.

Outcomes

The primary outcome of this study was the overall performance of ChatGPT in providing advice on the prevention and control of CC and BC. The secondary research indicators include accuracy, readability, consistency, reliability, efficiency, and relevance.

Definition of evaluation indicators

We defined the concepts of accuracy, readability, consistency, reliability, efficiency, and relevance, as follows:

- Accuracy refers to the ability of answers to provide comprehensive, omission-free, and misinformation-free answers to the given questions;
- Readability refers to the ease and clarity of ChatGPT's answers, considering the language difficulty;
- Consistency refers to the ability of ChatGPT to consistently provide the same or similar answers when asked the same questions multiple times, at different times and locations;
- Reliability refers to the credibility of ChatGPT's reference sources for the provided answers;
- Efficiency refers to the time taken by ChatGPT to provide answers to the questions asked;
- Relevance refers to ChatGPT's capacity to provide customized and adaptive responses to specific goals or needs.

Statistical analysis

A 7-point scale was utilized to measure accuracy, consistency, and relevance, with scores ranging from 1 (complete inaccuracy/ inconsistency/ irrelevance) to 7 (complete accuracy/ consistency/ relevance). Skewed data are presented using the median score, lower and upper quartiles (P_{25} and P_{75}), and range (Min and Max). Normal data are described using the mean and range. Inter-rater agreement was evaluated using Fleiss kappa [12–14], and the results of the interpretation of the kappa values are presented in Supplementary material 2, Table S2. The readability calculator utilizes the Flesch Reading Ease Score (FRES) and the Flesch-Kincaid Grade Level (FKGL) [15, 16]. Group differences are assessed using a standard t-test for normal data and non-parametric tests for non-normal data. All questions, except for consistency, were inputted into ChatGPT by a researcher. The calculation method of each indicator and radar chart can be found in Appendix B in Supplementary material 2.

Data were tabulated using Excel, and statistical analysis was conducted using SPSS (version 27.0) and Python (version 3.8). All statistical tests were two-sided, and the threshold for statistical significance was set at $P<0.05$.

Table 1 Results from Wilcoxon's signed ranks test of accuracy scores

	Median [(P_{25}, P_{75}), (range)]	Difference (P_{25}, P_{75})	Z	P
Cervical cancer				
Top scoring physician	6.1 [(5.6, 6.4), (4.4-6.8)]	0.8 (0.5, 1.4)	2.657	0.008
ChatGPT	7.0 [(6.6, 7.0), (6.6-7.0)]			
Breast cancer				
Top scoring physician	6.0 [(5.6, 6.4), (3.4-7.0)]	1.0 (0.3, 1.3)	2.527	0.010
ChatGPT	7.0 [(6.9, 7.0), (6.6-7.0)]			
CC and BC				
Top scoring physician	6.0 [(5.7, 6.4), (3.4-7.0)]	0.9 (0.5, 1.4)	3.663	< 0.001
ChatGPT	7.0 [(6.7, 7.0), (6.6-7.0)]			

Abbreviations CC, cervical cancer; BC, breast cancer

Results

After being reviewed by seven experts, ChatGPT achieved accuracy scores of 7.0 (range: 6.6-7.0) for 20 questions. For gynaecologists and mammologists, the accuracy scores were 6.1 (range:4.4-6.8) and 6.0 (range: 3.4-7.0), respectively, and demographic information for the ten physicians is shown in Table S1 of Supplementary material 2. The results of the difference test revealed that ChatGPT's accuracy in CC, BC, and both CCs and BCs were significantly better than that of the physicians with the highest score ($P<0.001$), as shown in Table 1. The inter-expert agreement for the scores was good; further details are shown in Table S3 of Supplementary material 2.

The average time for ChatGPT to answer the 20 questions was 13.6s (range: 7.6-24.0). The average sentence length was 192 words (range: 108-260). The FRES was 44.9 (range: 34.4-73.7), and the FKGL score was 11.7 (range: 4.5-15.5). The results from six repetitions were consistent with these findings, as shown in Table 2. The word count, FRES, and FKGL of standardized expert answers were 127 words (range: 50–213), 33.1 (range: 0-60.5), and 12.6 (range: 7.1–19.2). And the FRES and FKGL of physicians' answers were 18 words (range: 1–87), 39.5 (range: 0-100), and 11.2 (range: 0-30.8). For more detailed information, please see Supplementary material 1, Table S2. Furthermore, correlations between time, number of words, FRES, and FKGL were analyzed, as shown in Supplementary material 2, Table S6.

The consistency of ChatGPT answers is shown in Table 3. For CC, BC, and combined cases, the consistency scores for six answers were 5.1 (range: 4.4-6.7), 5.3 (range: 3.4-6.4), and 5.2 (range: 3.4-6.7), respectively. The scores for the first three consecutive questions, the last three consecutive questions, the scores for all six questions were 7.0 (range: 3.3-7.0), 5.6 (range: 4.1-7.0), and 5.2 (range: 3.4-6.7), respectively. Moreover, the first three scores were significantly higher than the last three scores between CC and BC ($P=0.04$). The classification of question-answer response methods revealed that the point-by-point response had the highest scores, followed by the

Table 2 Time, number of words, and readability of answers to CC & BC questions by ChatGPT

	No.	Question	Key assessment answers				Answers to six repeated questions			
			Time	Word count	FREE	FKGL	Time	Word count	FREE	FKGL
Cervical cancer	1	What are the common symptoms of cervical cancer?	9.3	151	40.6	11	7.51	138	54.9	8.7
	2	What are the risk factors for cervical cancer?	20.35	232	39.4	13.6	17.21	194	43	11.2
	3	Why should women get the HPV vaccine?	17.23	243	44.1	13.8	12.63	170	49.2	12.2
	4	How to choose different valent of HPV vaccine?	24	257	49.8	11.6	21.31	238	47.7	11.8
	5	What are the benefits and risks of cervical cancer screening?	8.05	182	30.7	12.8	17.01	248	40.9	11.3
	6	What are the common screening methods for cervical cancer?	12.69	159	53.5	10.2	17.75	209	45.6	11.2
	7	What should I do if my cervical cancer screening result is abnormal?	11.75	159	38.4	11.9	12.91	170	33.5	13.4
	8	Is cervical cancer curable?	14.43	173	35.1	13.1	10.52	118	45.2	11.7
	9	What are the treatment methods for cervical cancer?	19.89	241	40.1	11.2	14.32	178	34.1	12.5
	10	What is the prognosis for cervical cancer?	13.2	158	34.5	13.4	12.36	148	35.4	14.1
Breast cancer	1	What are the common symptoms of breast cancer?	7.58	108	61.5	7.1	7.58	122	70.1	5.9
	2	What are the risk factors for breast cancer?	15.5	238	49.6	11.7	12.52	201	50.2	10.1
	3	How to prevent breast cancer?	13.37	258	52.8	10.5	12.58	251	55.7	9.4
	4	Can breast cancer be inherited?	12.34	178	48.8	12	8.25	113	46	12.4
	5	Why is breast cancer screening important?	13.58	194	51.8	10.9	8.14	126	40.4	33.3
	6	What are the risks of breast cancer screening?	14.2	248	43.6	11.9	11.23	202	29.3	21.7
	7	What are the common screening methods for breast cancer?	9.19	160	55.2	9.5	9.4	162	51.1	10.1
	8	What is the survival rate of breast cancer?	11.18	147	46.6	12.8	8.5	123	45.2	13.4
	9	How to treat breast cancer?	11.9	220	48.1	10.2	11.66	202	39.5	11.4
	10	Can breast cancer recur?	12.39	141	34.4	15.5	7.99	96	45.4	12.3
	Mean		**13.6**	**192**	**44.9**	**11.7**	**12.07**	**170**	**45.1**	**12.9**
	Max		**24.0**	**260**	**73.7**	**15.5**	**21.31**	**251**	**70.1**	**33.3**
	Min		**7.6**	**108**	**34.4**	**4.5**	**7.51**	**96**	**29.3**	**5.9**

Abbreviations FRES, Flesch Reading Ease Score; FKGL, Flesch-Kincaid Grade Level; HPV, Human papillomavirus

comprehensive response, while both methods had the lowest scores. The Fleiss Kappa analysis showed statistically significant inter-expert agreement for the scores (All $P<0.001$), as detailed in Supplementary material 2, Table S4. However, in some cases, a few points were missed among multiple answers due to the presence of more points. For further details, please refer to Supplementary material 1, Table S4.

ChatGPT provided 30 reference links for 20 questions on CC and BC. Most of the links (except one) were from American web pages, and 90% ($n=27$) were from official websites, with 10% ($n=3$) being official guidelines. The sources for six repeated questions included papers, forum webpages, or needed more detailed information. For accessibility, 50% ($n=15$) were accessible, with 86.7% ($n=13$) being relevant and 13.3% ($n=2$) irrelevant. The other 50% ($n=15$) of the links could not be obtained. For specific results, refer to Supplementary material 1, Table S5 and Table S6 and Supplementary material 2, Table S7.

For context 1, the relevance scores of CC, BC, and both were 6.9 (range: 6.2-7.0), 5.9 (range: 5.0-7.0), and 6.6 (range: 5.0-7.0), respectively. A comparison between the results with and without context revealed that the main points of the answers remained the same. However, the answers without context provided more explanation and expansion. Furthermore, the time spent and the total number of words were significantly reduced ($P<0.001$ Cohen's $D>0.900$). No statistical difference was observed in readability. Detailed results refer to Table 4. For context 2, experts identified contextual differences in questions 4, 6, and 10 regarding CC and questions 7 and 8 regarding BC. The relevance score for both CC and BC in these identified questions was 6.6 (range: 6.0-7.0), while no differences were found in the remaining questions. The total score for this section was 6.5 (range: 4.8-7.0), and the results are shown in Supplementary material 2, Table S8. The results for the two different context-based questions are shown in Supplementary material 1, Tables S6 and S7.

The radar chart in Fig. 2 illustrates ChatGPT's comprehensive assessment of CC and BC recommendations, with an accuracy score of 100 (range: 94.3-100), readability score of 44.9 (range: 34.4-73.7), consistency score of 74.3 (range: 48.6-95.7), reliability score of 73.3 (range:

Table 3 Results of consistency test, difference comparison of ChatGPT's answers with six repetitions

Consistency test with six repetitions		**A1 to A3**		**A4 to A6**		**A1 to A6**	
	Differentiation of answering styles	Count	Median [(P_{25}, P_{75}), (range)]	Count	Median [(P_{25}, P_{75}), (range)]	Count	Median [(P_{25}, P_{75}), (range)]
Cervical Cancer	Point-by-point response	6	7.0 [(7.0, 7.0), (5.0–7.0)]	4	6.0 [(5.0, 7.0), (3.0–7.0)]	4	6.0 [(5.8, 6.0), (3.0–7.0)]
	Comprehensive response	3	7.0 [(6.0, 7.0), (5.0–7.0)]	4	6.0 [(5.8, 7.0), (5.0–7.0)]	3	5.0 [(5.0, 6.0), (4.0–7.0)]
	Both of them	1	4.0 [(5.0, 5.0), (3.0–5.0)]	2	5.0 [(5.0, 6.0), (3.0–6.0)]	3	5.0 [(3.0, 5.0), (3.0–6.0)]
	Overall consistency score*	**10**	**7.0 [(5.8, 7.0), (4.4-7.0)]**	**10**	**6.0 [(5.2, 6.8), (4.1-7.0)]**	**10**	**5.1[(4.6, 6.1), (4.4–6.7)]**
Breast cancer	Point-by-point response	6	7.0 [(7.0, 7.0), (5.0–7.0)]	3	7.0 [(6.0, 7.0), (5.0–7.0)]	3	6.0 [(6.0, 6.0), (5.0–7.0)]
	Comprehensive response	4	4.0 [(3.0, 6.0), (3.0–7.0)]	4	5.0 [(3.8, 5.3), (3.0–6.0)]	4	4.0 [(3.0, 5.0), (2.0–5.0)]
	Both of them	0	-	3	5.0 [(5.0, 5.0), (3.0–6.0)]	3	5.0 [(5.0, 5.0), (3.0–7.0)]
	Overall consistency score*	**10**	**6.6 [(4.7, 7.0), (3.3-7.0)]**	**10**	**5.4 [(4.4, 6.1), (4.1-7.0)]**	**10**	**5.3 [(4.1, 5.7), (3.4, 6.4)]**
CC and BC	Point-by-point response	12	7.0 [(7.0, 7.0), (5.0–7.0)]	7	6.0 [(6.0, 7.0), (3.0–7.0)]	7	6.0 [(6.0, 6.0), (3.0–7.0)]
	Comprehensive response	7	6.0 [(4.0, 7.0), (3.0–7.0)]	8	5.0 [(5.0, 6.0), (3.0–7.0)]	7	5.0 [(4.0, 6.0), (3.0–7.0)]
	Both of them	1	4.0 [(5.0, 5.0), (3.0–5.0)]	5	5.5[(5.0, 6.0), (3.0–6.0)]	6	5.0 [(5.0, 5.0), (3.0–6.0)]
	Overall consistency score*	**20**	**7.0 [(5.6, 7.0), (3.3-7.0)]**	**20**	**5.6 [(4.5, 6.5), (4.1-7.0)]**	**20**	**5.2 [(4.4, 5.9), (3.4–6.7)]**
Difference comparison							
		Median (P_{25}, P_{75})	Difference (P_{25}, P_{75})	*Z*	*P*		
Cervical Cancer	A1 to A3	7.0 (5.8, 7.0)	0.6 (0.0, 1.5)	1.693	0.090		
	A4 to A6	6.0 (5.2, 6.7)					
Breast cancer	A1 to A3	6.6 (4.7, 7.0)	0.0 (-0.6, 1.8)	1.352	0.180		
	A4 to A6	5.4 (4.4. 6.1)					
CC and BC	A1 to A3	7.0 (5.6, 7.0)	0.2 (0, 1.5)	2.014	0.040		
	A4 to A6	5.6 (4.5, 6.5)					

* Inconsistency test with six repetitions, the overall consistency score is calculated from the average scores of all experts, and the rest scores are calculated from the original scores of all experts

Abbreviations CC, cervical cancer; BC, breast cancer; A1, answer 1; A3, answer 3; A6, answer 6

50.0-100.0), efficiency score of 13.6 s (range: 7.6-24.0), and relevance score of 92.8 (range: 68.6-100).

Discussion

Cervical and breast cancers pose significant risks to women's well-being, influenced by health-related stigma and a lack of reliable information, leading to delayed diagnosis and potentially fatal consequences. Therefore, technologies like ChatGPT are poised to become critical sources of health information, enabling earlier identification of these cancers. This study focused on guidance generated by ChatGPT and compared outputs with physicians' repsonses. Accuracy, readability, consistency, reliability, efficiency, relevance, and the content of answers were analyzed and contrasted. The results showed that ChatGPT answers are generated efficiently, with accuracy and relevance. However, concerns arose regarding AI-generated guidance's readability, consistency, and reliability.

ChatGPT has 100% accuracy (range: 94.3-100) when answering popular science questions about CC and BC, surpassing physicians and aligning with expert consensus. Similar studies have shown high accuracy rates for ChatGPT, such as 96.9% for 'cancer myths' [17] and 88% for assessing breast cancer recommendations [18], and ChatGPT outperformed newly qualified obstetricians and gynaecologists in a virtual examination [19]. However, it is worth noting that these studies had limited question samples and contingent results, which means the findings cannot be generalized. In other studies, ChatGPT achieved a median accuracy score of 5.5 out of 6.0 in 284 medical questions [20] while achieving only 54.9% accuracy in a neurosurgery self-assessment with 477 non-image-based questions [21]. This means knowledge depends upon specific fields, and ChatGPT's accuracy in answering medical questions is influenced by question difficulty and the availability of professional knowledge in population medical science media. Overall,

Table 4 Comparison of the differences in ChatGPT's answer time, word count, and readability between the CC & BC with and without context 1

		Mean (SD)	T	*P*	Cohen's D (95% CI)
Time	Question	11.9 (2.9)	3.149	0.003	0.996 (0.331, 1.649)
	Question + Context	8.9 (3.0)			
Word	Question	167 (45)	3.462	0.001	1.095 (0.422, 1.755)
	Question + Context	118 (46)			
FREE	Question	42.1 (10.9)	0.202	0.840	0.064 (-0.556, 0.684)
	Question + Context	41.4 (10.3)			
FKGL	Question	12.5 (1.9)	0.098	0.920	0.031 (-0.589, 0.65)
	Question + Context	12.5 (1.9)			

Abbreviations FRES, Flesch Reading Ease Score; FKGL, Flesch-Kincaid Grade Level; SD, standard deviation;

CI, confidence interval

ChatGPT is highly accurate in answering general cancer science questions, making it a valuable resource for those who can craft them.

ChatGPT provides comprehensive answers to popular cancer science questions; however, they tend to be low in readability and lengthy, requiring at least a high school level of education to understand. The results of experts and physicians are similar, which aligns with the findings of Johnson et al. [17]. However, ChatGPT's answers exceeded this recommended grade level of 6th to 8th grade for popular science articles [17, 22], making them challenging for a significant portion of the population with limited literacy skills [23, 24]. The Federal Plain Language guidelines emphasize that conveying complex information is more effective with shorter sentences [25]. Therefore, ChatGPT's more complex answers result in low overall readability and may pose difficulties for some individuals with limited education.

ChatGPT's answers to popular cancer science questions are slightly inconsistent. While there is good consistency in its responses to three consecutive questions, its consistency across six questions could be improved. Though there was a greater difference in consistency score between the two sets of questions from different times and places, the overall consistency of the six answers was acceptable. However, significant inconsistencies were observed in individual question responses, conflicting information regarding the age range for administering the HPV vaccine and the recommended age for breast cancer screening. This may in part be due to reduced consistency in such guidelines across lead organizations. Potentially inaccurate information could pose inherent risks to the health and safety of people who consult. And consistency may need to be clarified for users, undermining their trust in ChatGPT or raising doubts about its effectiveness. These inconsistencies may have arisen due to the model's development, reinforcement learning strategies, and the variation in the retrieved and synthesized corpus by ChatGPT [26]. Furthermore, they may be attributed to imprecise patient queries, where ambiguous questions fail to yield nuanced responses. Moving forward, there is a necessity to intensify the model's specialized training in medical knowledge modules and equip ChatGPT with enhanced contextual understanding during patient interactions to ensure precise and consistent answers.

Reliability analysis revealed that half of the reference sources used by ChatGPT needed to be updated or inaccessible. It is important to note that these outputs were generated based on static data collected before September 2021, which may not reflect the most up-to-date medical knowledge. Furthermore, the reliance on American standards as primary reference sources for English questions may limit the accuracy and relevance of ChatGPT's responses in other regions. However, when it comes to health-related popular science inquiries that are non-specialized and not subject to significant regional variations, ChatGPT's responses are quite reliable. Despite the unavailability of many references, the accuracy of the answers remains acceptable. Therefore, future improvements to ChatGPT include its integration with the Internet for real-time consultations, ensuring the provision of even more reliable information.

As for relevance, ChatGPT demonstrates improved response results when questions are presented in different contexts compared to without added context. While providing additional context 1 reduced the word count, but readability did not significantly improve. This may be attributed to professional vocabulary and terminology, suggesting that despite more streamlined answers, readability still needs to be improved. Furthermore, when context 2 was added, ChatGPT frequently referred to "China" and used different reference sources for questions with regional differences. The fact that ChatGPT adapts its responses to different prompts indicates its sensitivity to prompts and its ability to capture relevant information [9, 27, 28]. Nevertheless, further research is necessary to understand how prompts influence medical recommendations and their response to different questioning styles. Therefore, women should utilize ChatGPT with specific contexts or prompts when seeking health science consultations to enhance the accuracy of its responses.

Based on our analysis using the radar chart, we found that the overall performance of ChatGPT is superior, and

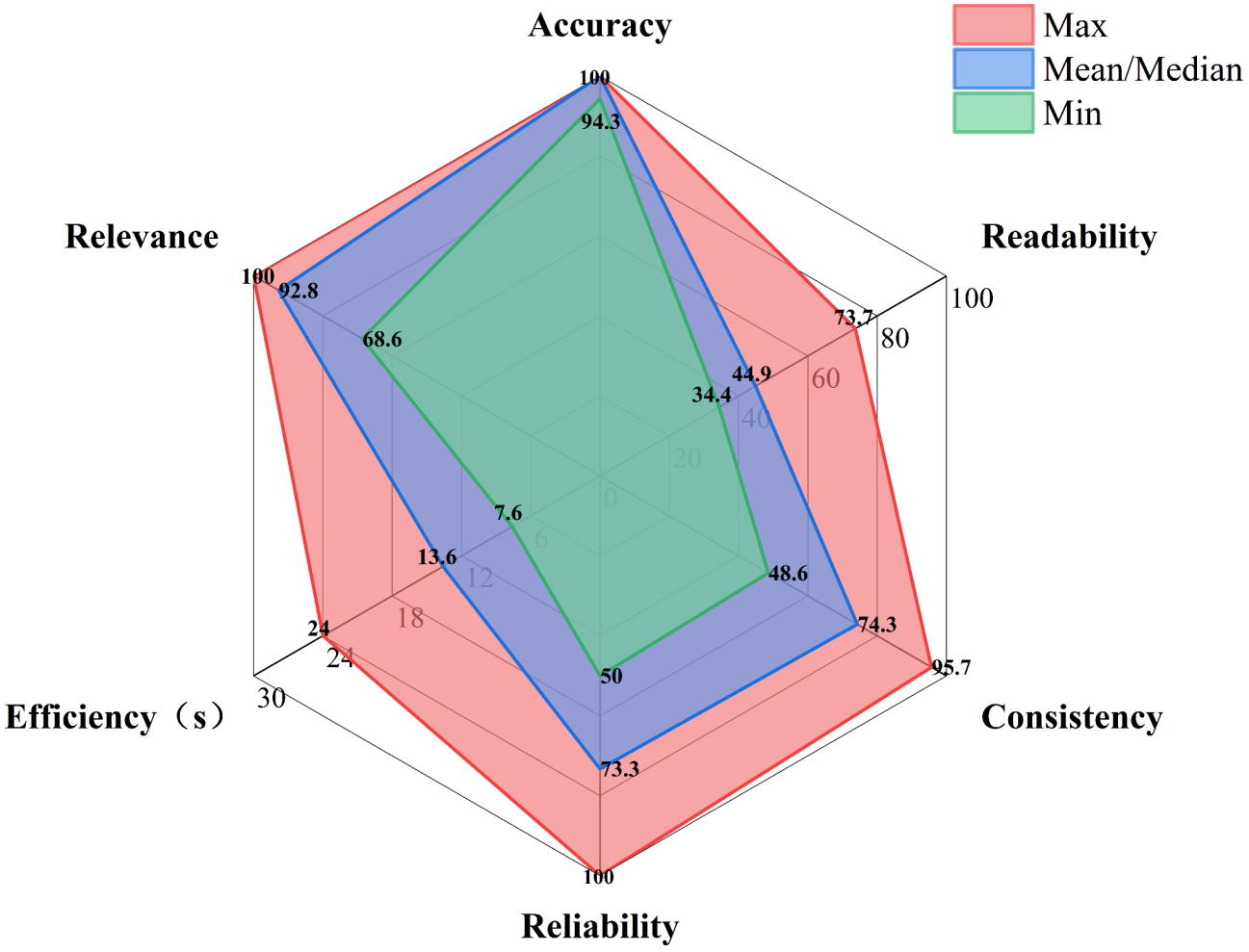

Fig. 2 Radar chart assessing ChatGPT's responses to cervical and breast cancer questions. The performance of ChatGPT was visually assessed with the accuracy, readability, consistency, reliability, efficiency, and relevance through radar images, to reflect the feasibility of applying ChatGPT to the comprehensive prevention and treatment of CC and BC. CC, cervical cancer; BC, breast cancer

it can initially be used for health consultation of CC and BC to women lacking specialized knowledge. ChatGPT can be a valuable resource for patients seeking information and support related to female cancers like CC, BC and others. It can function as a reliable and available resource for individuals seeking information about cancer risk and screening, particularly for those who may feel scared or uncomfortable discussing their concerns with a healthcare professional or those in areas with limited access to healthcare resources. ChatGPT enables patients to play an active role in their healthcare decisions by equipping them with knowledge. Besides, ChatGPT enhances efficiency in healthcare settings by quickly generating informative responses, saving time for healthcare professionals. However, it has limitations and occasional errors due to its lack of proper understanding of human language [9, 29]. Therefore, caution should be exercised when using it. It should be viewed as an auxiliary tool for health consultation rather than a substitute for professional advice. Therefore, it should always be used with human expertise to ensure the delivery of accurate information to individuals seeking cancer-related knowledge and support.

In clinical applications, the positioning of ChatGPT is of significant interest. It can assist in early disease identification and answer common patient questions [30], but ultimately, diagnosis and treatment require trained medical professionals. To support all patients, especially those with low literacy, clinicians can use the following strategies: (1) Simplified language: Use straightforward terms, e.g., ask "How severe is your headache?" instead of "Please describe the intensity of your migraine." (2) Specificity: Ask precise questions, like "Did you take any medication today?" instead of "Have you taken any medication recently?" (3) Step-by-step questioning: Break complex questions into simpler ones, e.g., ask "Are you feeling tired?" followed by "Did you sleep well last night?" and "Do you have any stomach pain?" (4) Use of

visuals: Incorporate images and charts to aid understanding. These methods will help clinicians effectively use Chat GPT to meet diverse health needs. Future research should focus on optimizing ChatGPT algorithms and expanding its medical knowledge to ensure it incorporates the latest information, enhancing its effectiveness in clinical practice and fostering patient understanding and trust. Through these efforts, ChatGPT is expected to play a larger role in clinical settings, offering new support for healthcare services.

Before providing recommendations, we need to discuss the limitations of this study. Firstly, the evaluation criterion relied primarily on expert ratings, which, despite good inter-expert agreement, still retains some degree of subjectivity and potential for a paradigm shift. Secondly, the study only included a limited number of questions on CC and BC with national/regional differences, necessitating further research to explore variations across different regions. Additionally, the accuracy of answers was evaluated only once, raising the possibility of incidental errors. Furthermore, this study focused solely on English-language questions, limiting the assessment of ChatGPT's performance in other languages. And future research may encompass a broader range of science popularization questions, evaluate ChatGPT's responses in different languages, and demonstrate its potential in promoting health knowledge dissemination and addressing health information inequality and inaccessibility, particularly for underserved communities, thus safeguarding women's health.

Conclusions

This study suggests the potential of ChatGPT to provide science-based information on CC and BC, making it a promising tool for initial exploration in women's health consultation, particularly for women lacking expertise in the field. Nevertheless, further research is necessary to delve into the underlying mechanisms of ChatGPT and address its limitations. In the future, it is imperative to develop more specialized ChatGPT tailored for health counselling and enhance its performance across various aspects.

Abbreviations

AI	Artificial intelligence
CC	cervical cancer
BC	breast cancer
FRES	Flesch Reading Ease Score
FKGL	Flesch-Kincaid Grade Level
HPV	Human papillomavirus

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12905-024-03320-8.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

I would like to express my sincere gratitude to our colleague, Yue Chen, for helping me obtain the ChatGPT dataset during her studies in Germany. And the authors thank all researchers for their support.

Author contributions

ZCY, BZ, and PX contributed to the conception and design of the study. ZCY, BZ, PX, and YJ administrated the project. ZCY, BZ, HJY, and MJGM complete data collection. ZCY, BZ and HJY contributed to data management. MYC, YMQ and KZ implemented the data analysis. ZCY, PX and JY contributed to the methodology. PX, YLQ, and YJ contributed to supervision and control. ZCY and PX wrote the manuscript and PX, YLQ, KZ, TW and YJ critically revised drafts. All authors approved the manuscript, and PX, YLQ and YJ guarantee the integrity of the work.

Funding

This study was supported by CAMS Innovation Fund for Medical Sciences (CIFMS 2021-I2M-1-004), China Postdoctoral Science Foundation (2023M740323 and 2024T170072), Postdoctoral Fellowship Program of CPSF (GZB20230076), and research on the Construction of a New Public Health Science System and Talent Training Model (201920102401).

Data availability

All data generated or analyzed during this study are included in this article and its supplementary information files. The original data can be obtained by the reasonable application of the corresponding author.

Declarations

Ethical approval

This study adhered to the Declaration of Helsinki and received approval from the Institutional Review Board of the Chinese Academy of Medical Sciences and Peking Union Medical College (No. CAMS and PUMC-IEC-2022-022).

Consent to participate

All participants in both phases signed an informed consent form before the start of the study.

Consent for publication

Not Applicable.

Competing interests

The authors declare no competing interests.

Received: 28 February 2024 / Accepted: 16 August 2024
Published online: 02 September 2024

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and Mortality Worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
3. Organization WH. Global strategy to accelerate the elimination of cervical cancer as a public health problem; 2020.
4. Organization WH. Global Breast Cancer Initiative Implementation Framework Assessing, strengthening and scaling up services for the early detection and management of breast cancer; 2023.
5. Xiao N, Sharman R, Rao HR, Upadhyaya S. Factors influencing online health information search: an empirical analysis of a national cancer-related survey. Decis Support Syst. 2014;57:417–27.
6. Borges U Jr., Riese C, Baumann W. [Internet use by Oncology outpatients: results of a survey in Germany]. Gesundheitswesen. 2018;80(12):1088–94.

7. Johnson SB, Parsons M, Dorff T, Moran MS, Ward JH, Cohen SA, Akerley W, Bauman J, Hubbard J, Spratt DE, et al. Cancer Misinformation and Harmful Information on Facebook and other Social Media: a brief report. J Natl Cancer Inst. 2022;114(7):1036–9.
8. OpenAI. GPT-4 Technical Report, vol. 2023; March 27, 2023.
9. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388(13):1233–9.
10. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med. 2023;388(13):1201–8.
11. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol. 2023;228(6):696–705.
12. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82.
13. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360–3.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
15. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221–33.
16. Badarudeen S, Sabharwal S. Readability of patient education materials from the American Academy of Orthopaedic Surgeons and Pediatric Orthopaedic Society of North America web sites. J Bone Joint Surg Am. 2008;90(1):199–204.
17. 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 Spectr 2023, 7(2).
18. 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.
19. Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, Dashraath P, Ramlal H, Mahyuddin AP, Kanayan S et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023.
20. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, Chang S, Berkowitz S, Finn A, Jahangir E et al. Assessing the accuracy and reliability of AI-Generated medical responses: an evaluation of the Chat-GPT Model. Res Sq 2023.
21. Hopkins BS, Nguyen VN, Dallas J, Texakalidis P, Yang M, Renn A, Guerra G, Kashif Z, Cheok S, Zada G et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. J Neurosurg 2023:1–8.
22. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, Jagsi R, Golden DW. Readability of Patient Education materials from High-Impact Medical journals: a 20-Year analysis. J Patient Exp. 2021;8:2374373521998847.
23. Demarco J, Nystrom R. The Importance of Health Literacy in Patient Education. J Consumer Health Internet. 2010;14(3):294–301.
24. Health Literacy Online. About Health Literacy Online: 2nd Edition [https://health.gov/healthliteracyonline/about/
25. Plain Language Action and Information Network. Federal Plain Language guidelines. US Department of Health and Human Services; March; 2011.
26. Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EHC, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol 2023.
27. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires Human Judgment. Radiology. 2023;307(2):e230171.
28. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and other large Language models are double-edged swords. Radiology. 2023;307(2):e230163.
29. Ian Bogost. ChatGPT Is Dumber Than You Think [https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/
30. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Evaluations. 2023;3(1):100105.

Publisher's note