JMIR Publications
The Leading eHealth Publisher

ChatGPT vs. DeepSeek: A Comparative Analysis of AI Models for Breast Cancer Information Retrieval

Rima Hajjo, Dima A. Sabbah, Sanaa K. Bardaweel

Submitted to: JMIR Cancer
on: February 19, 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 it's 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 expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

ChatGPT vs. DeepSeek: A Comparative Analysis of AI Models for Breast Cancer Information Retrieval

Rima Hajjo[1]; Dima A. Sabbah[1]; Sanaa K. Bardaweel[2]

[1]Department of Pharmacy Faculty of Pharmacy Al-Zaytoonah University of Jordan Amman JO
[2] University of Jordan Amman JO

Corresponding Author:
Rima Hajjo
Department of Pharmacy
Faculty of Pharmacy
Al-Zaytoonah University of Jordan
Airport Street, P.O. Box 130, Amman 11733, Jordan.
Amman
JO

Abstract

Artificial intelligence (AI) has revolutionized biomedicine, driving advancements in diagnostics, treatment, and medical data management. Breast cancer, a significant global health challenge, highlights the need for accessible and reliable medical information. AI platforms like ChatGPT-4.0 and DeepSeek-V3 have become essential tools for delivering curated medical insights. This study compared ChatGPT-4.0 and DeepSeek-V3 in retrieving and presenting medical information, focusing on readability, content quality, and reliability of information sources. Using Flesch-Kincaid Grade Level (FKGL) and a 7-point Likert scale, the analysis revealed that AI models often produce simpler responses than expert references, improving accessibility but risking oversimplification. ChatGPT demonstrated greater consistency and improved readability in multi-response scenarios, excelling in clarity and depth, while DeepSeek aligned more closely with expert reference readability in single-instance analysis but showed higher variability. DeepSeek excelled in citation efficiency and global reference diversity, but faced challenges like untagged links, corrupted references, and occasional downtime. Despite these differences, statistical analysis showed no significant differences between the models, particularly in larger datasets. Both models provided reliable information, but no single model consistently matched expert content across all questions. The findings reveal that no single AI model consistently matches expert content across all questions, emphasizing the need for careful evaluation to ensure AI-generated information meets diverse user needs. Future improvements should focus on enhancing link accessibility, platform stability, and response consistency to optimize AI-generated medical content for healthcare applications.

(JMIR Preprints 19/02/2025:72839)
DOI: https://doi.org/10.2196/preprints.72839

Preprint Settings

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

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

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

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

ChatGPT vs. DeepSeek: A Comparative Analysis of AI Models for Breast Cancer Information Retrieval

Rima Hajjo[1,2*], Dima A. Sabbah[1], Sanaa K. Bardaweel[3]

[1]*Department of Pharmacy, Faculty of Pharmacy, Al-Zaytoonah University of Jordan, P.O. Box 130, Amman 11733, Jordan.*

[2]*Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.*

[3]*Department of Pharmaceutical Sciences, School of Pharmacy, University of Jordan, Amman 11942, Jordan.*

*Correspondence

Rima Hajjo

E-mail: rhajjo@gmail.com; r.hajjo@zuj.edu.jo

ORCID: https://orcid.org/0000-0002-7090-5425

Abstract

Artificial intelligence (AI) has revolutionized biomedicine, driving advancements in diagnostics, treatment, and medical data management. Breast cancer, a significant global health challenge, highlights the need for accessible and reliable medical information. AI platforms like ChatGPT-4.0 and DeepSeek-V3 have become essential tools for delivering curated medical insights. This study compared ChatGPT-4.0 and DeepSeek-V3 in retrieving and presenting medical information, focusing on readability, content quality, and reliability of information sources. Using Flesch-Kincaid Grade Level (FKGL) and a 7-point Likert scale, the analysis revealed that AI models often produce simpler responses than expert references, improving accessibility but risking oversimplification. ChatGPT demonstrated greater consistency and improved readability in multi-response scenarios, excelling in clarity and depth, while DeepSeek aligned more closely with expert reference readability in single-instance analysis but showed higher variability. DeepSeek excelled in citation efficiency and global reference diversity, but faced challenges like untagged links, corrupted references, and occasional downtime. Despite these differences, statistical analysis showed no significant differences between the models, particularly in larger datasets. Both models provided reliable information, but no single model consistently matched expert content across all questions. The findings reveal that no single AI model consistently matches expert content across all questions, emphasizing the need for careful evaluation to ensure AI-generated information meets diverse user needs. Future improvements should focus on enhancing link accessibility, platform stability, and response consistency to optimize AI-generated medical content for healthcare applications.

Keywords: Artificial intelligence (AI); breast cancer; ChatGPT; DeepSeek; large language models (LLM).

Introduction

Breast cancer remains a major health concern for women, with early detection posing significant challenges as many individuals seek medical attention only after symptoms worsen [1-4]. The World

Health Organization (WHO) highlights the importance of enhancing screening, diagnostic, and treatment approaches to tackle breast cancer [5]. Socioeconomic factors, including education and income, influence healthcare-seeking behaviors, with women often relying on the internet for information about early signs and symptoms [6, 7]. However, the quality of online health information varies significantly, with research showing that much of the cancer-related content on social media can be misleading or inaccurate [8, 9].

To address concerns over misinformation in online health content, Large Language Models (LLMs) like ChatGPT and DeepSeek are increasingly being used to provide accessible, evidence-based medical information [10-15]. LLMs are AI models designed to understand, generate, and manipulate human language. Such models can analyze extensive medical literature to deliver quick insights into disease symptoms, risk factors, treatments, and prevention strategies [16, 17]. In fact, AI's applications in biomedicine extend beyond information retrieval, aiming to revolutionize diagnostics, improve and personalize treatments, and support healthcare data management [18]. It enhances medical imaging, disease detection, and drug discovery, often exceeding human capabilities and accuracy [19-21]. In oncology, AI-driven algorithms aid in early cancer detection, refine targeted drug delivery, and personalize treatment plans by integrating genomic data, clinical records, and real-time patient monitoring, ultimately improving patient outcomes and treatment efficacy [2, 5-8, 22, 23].

Recently, DeepSeek has emerged as a strong competitor to ChatGPT, featuring advanced natural language processing tailored for clinical domains by enabling advanced natural language processing for scientific content to deliver precise medical and clinical information. DeepSeek's ability to handle complex medical data, provide accurate diagnostic suggestions, and support evidence-based decision-making makes it an invaluable tool for healthcare professionals. By focusing on precision and reliability, DeepSeek addresses critical challenges in clinical medicine, including minimizing diagnostic errors and enhancing patient outcomes. Furthermore, its integration with electronic health records and medical literature improves clinical workflows, positioning it as a key player in medical AI innovation [24, 25].

This study evaluates the performance of ChatGPT-4.0 and DeepSeek-V3 in providing medical information on breast cancer. Ten frequently asked questions (FAQs) were sourced from Ye et al. [26], and each AI model was prompted to generate responses, followed by a request for sources and references. Responses were assessed based on readability, accuracy, completeness, clarity, depth & insight, and alignment with expert reference answers using the Flesch-Kincaid Grade Level (FKGL) [14] and Likert scoring [15]. The study also examined the reliability of sources cited by ChatGPT and DeepSeek. This analysis could help answer the question which AI model provides more reliable and comprehensive health information? This study aims to compare ChatGPT and DeepSeek in delivering medical information on breast cancer. The evaluation will focus on text readability, clarity, completeness, accuracy, depth & insight, and alignment with reference answers to assess which AI model provides more reliable and comprehensive health information.

2. Materials and Methods

2.1. Study design and workflow

Ten FAQs and reference standard answers about breast cancer were obtained from Ye et al [26]. In the second step, a researcher asked AI platforms ChatGPT version 4.0 and DeepSeek version 3.0 to answer the 10 validated questions in addition to a follow up question about the sources of each answer by explicitly asking for the basis of the response along with a link or reference. Additionally, two researchers independently accessed both AI platforms at different times, locations, and internet speeds. Each of them asked the same question two times consecutively, resulting in four recordings for each question. In the third step, an evaluation of the responses was performed. This evaluation focused on readability, accuracy, completeness, clarity, depth and insight, alignment with reference answers, reliability, and efficiency of the answers. Two tests were applied, specifically the Flesch-Kincaid Grade Level (FKGL) and Likert analysis. Next, the performance of both ChatGPT and DeepSeek was visualized using bar plots, line, plots, box plots, and radar plots. Finally, the information resource reliability was compared for both platforms.

2.2. Data Collection and Processing

Reference answers for each question (Q1–Q10) were derived from a consensus among five mammalogists, as established by Ye et al [26]. All questions and corresponding consensus answers are provided in Supplementary Table 1 and served as the baseline for comparison. The first dataset consisted of responses generated by ChatGPT-4.0 and DeepSeek-V3, where each question was queried once against both models. This resulted in a total of 20 responses (10 from ChatGPT and 10 from DeepSeek), which were compared to the expert consensus answers.

The second dataset included six responses per question from each model, obtained by three researchers who queried ChatGPT-4.0 and DeepSeek-V3 twice per question. The final analysis was conducted on a total of 50 responses from ChatGPT and 60 responses from DeepSeek, aggregating all responses across researchers.

2.3. Statistical Analysis

2.3.1. Flesch-Kincaid Grade Level (FKGL) Analysis

The textstat library in Python was used to compute the Flesch-Kincaid Grade Level (FKGL) scores for reference answers and responses generated by ChatGPT-4.0 and DeepSeek-V3. FKGL estimates the U.S. school grade level required to understand a text, with higher scores indicating greater complexity. The FKGL score were calculated using the following formula:

$$\mathrm{FKGL} = 0.39\left(\frac{\textit{Total words}}{\textit{Total sentences}}\right) +¿\ 11.8\left(\frac{\textit{Total syllables}}{\textit{Total words}}\right) - 15.59$$

This analysis utilized two datasets: Dataset 1, where expert reference answers were compared to the initial responses from ChatGPT or DeepSeek for all questions (Q1–Q10), and Dataset 2, where three researchers generated two responses each from ChatGPT and DeepSeek for Q1–Q2, resulting in 60 responses per model. To evaluate statistical significance, a paired t-test (using Python's scipy.stats library) was conducted, as FKGL scores were normally distributed. This test assessed mean differences between ChatGPT and DeepSeek while ensuring paired comparisons across identical question sets.

2.3.2. Likert Analysis

The Likert analysis, using the Python's likert library, was performed on responses from experts (Reference) and AI platforms (ChatGPT and DeepSeek). The Likert scoring is commonly used to evaluate the response quality in survey research, AI response assessment, and human-computer interaction studies. It evaluates accuracy, completeness and insightfulness of text by using 1-7 scaling system. Thus, a 7-point Likert scaling has been applied for evaluating the response quality of both ChatGPT and DeepSeek against expert consensus answers (Reference). To determine statistical significance, normality was first assessed using the Shapiro-Wilk test using (via Python's scipy.stats library. Since the Likert scores were found to be normally distributed, a paired t-test (via Python's scipy.stats library) was conducted. This allowed for paired comparisons between ChatGPT and DeepSeek responses across identical question sets. This approach ensured a robust evaluation of mean differences in response quality, including accuracy, completeness, clarity, depth & insight, and alignment with reference answers.

3. Results

A comparison of ChatGPT and DeepSeek was conducted using 10 FAQs about breast cancer, with responses evaluated against expert consensus answers (Reference) as outlined in the Methods section. The workflow, illustrated in Figure 1, consisted of three key steps: data collection, model querying, and visual/statistical analysis. In the data collection phase, 10 common breast cancer questions were identified, and expert consensus answers were established. Three independent researchers then queried ChatGPT and DeepSeek, generating responses. Two datasets were created for each model: Dataset 1, consisting of single-instance responses (one response per question), and Dataset 2, containing multi-response outputs (multiple responses per question), resulting in 10 responses from each model for Dataset 1 and 60 responses from each model for Dataset 2, as shown in Figure 1.

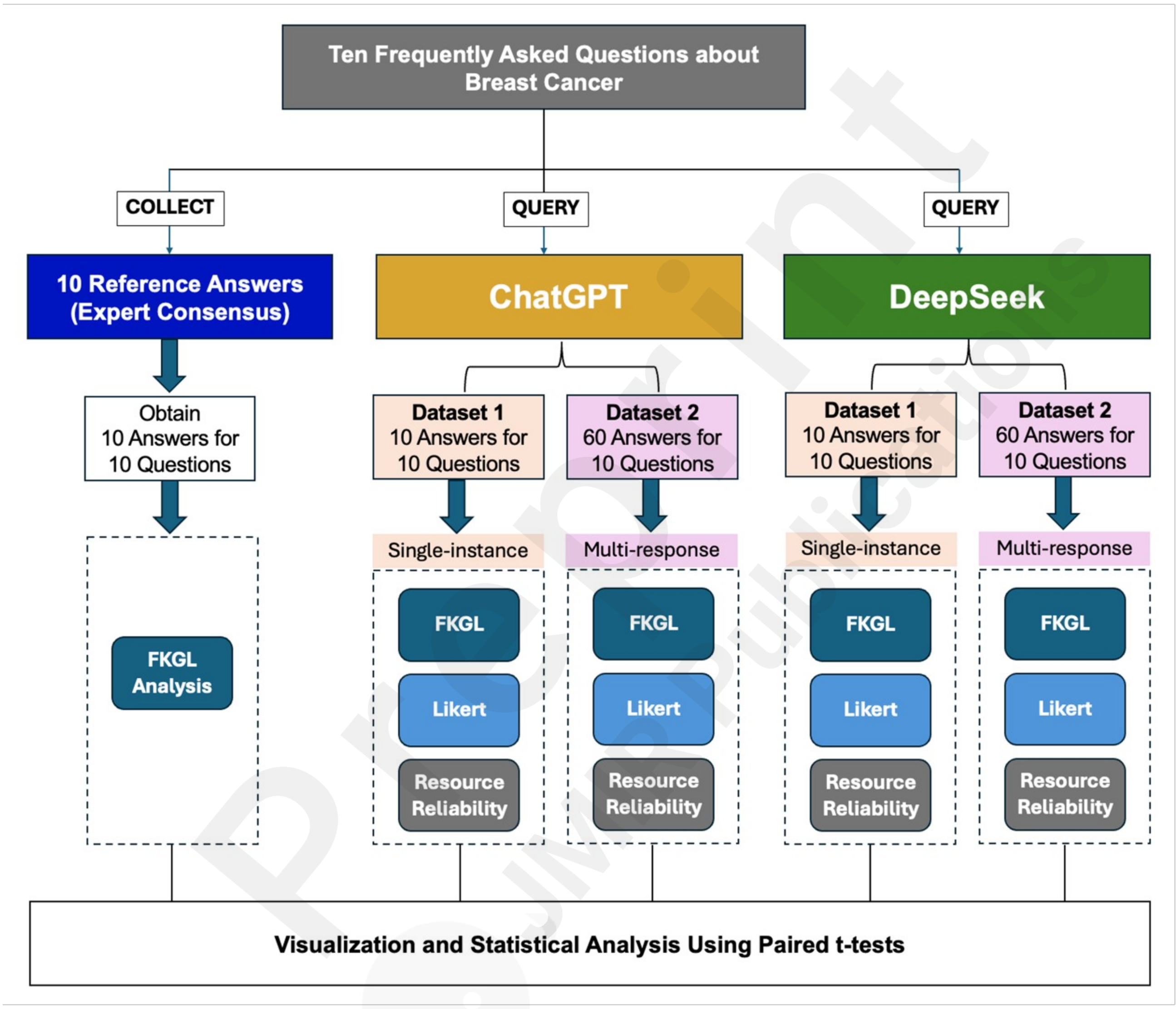
Ten Frequently Asked Questions about Breast Cancer
COLLECT
QUERY
QUERY
10 Reference Answers (Expert Consensus)
ChatGPT
DeepSeek
Obtain 10 Answers for 10 Questions
Dataset 1 10 Answers for 10 Questions
Dataset 2 60 Answers for 10 Questions
Dataset 1 10 Answers for 10 Questions
Dataset 2 60 Answers for 10 Questions
Single-instance
Multi-response
Single-instance
Multi-response
FKGL Analysis
FKGL
Likert
Resource Reliability
FKGL
Likert
Resource Reliability
FKGL
Likert
Resource Reliability
FKGL
Likert
Resource Reliability
Visualization and Statistical Analysis Using Paired t-tests

Figure 1. Workflow for evaluating the performance of ChatGPT and DeepSeek in retrieving and presenting medical information.

3.1. Readability Assessment Using FKGL Scores

The Flesch-Kincaid Grade Level (FKGL) readability scores for responses generated by ChatGPT and DeepSeek, compared against expert reference answers are presented in Table 1 and Figure 2. Two types of analyses were performed: single-instance analysis where each AI model provided a single response per question, and multi-response analysis where multiple responses per question were collected by three researchers independently from both models.

In the single-instance analysis, ChatGPT's mean FKGL score (10.11) was slightly lower than the expert reference (10.53), while DeepSeek's mean score (10.53) matched the reference exactly. In the multi-response analysis, ChatGPT's mean FKGL increased to 10.36, whereas DeepSeek's mean FKGL slightly decreased to 10.27. ChatGPT demonstrated reduced variability in the multi-response case (SD = 0.84) compared to the single-instance case (SD = 1.17), indicating greater consistency when averaging multiple responses. In contrast, DeepSeek's variability remained relatively stable across both analyses (SD = 1.02 in single-instance, 0.98 in multi-response).

In both analyses, DeepSeek's median FKGL scores were slightly lower than ChatGPT's, suggesting that DeepSeek's responses may be more concise or readable. DeepSeek also exhibited less fluctuation in readability scores, as indicated by its smaller interquartile range (IQR = 0.77 in single-instance, 1.41 in multi-response). In contrast, ChatGPT's variability decreased in the multi-response case, with its IQR narrowing from 2.05 to 1.01.

A detailed evaluation of FKGL readability scores reveals variations in how ChatGPT and DeepSeek generate responses compared to expert references. For some questions (e.g., Q1, Q2, Q3, Q4), both models produced more complex responses (higher FKGL scores), potentially reducing accessibility for general readers but could be more suitable for professionals and specialists seeking in-depth information. In contrast, for other questions (e.g., Q5, Q6, Q7, Q9), AI-models generated simpler responses (lower FKGL scores), possibly oversimplifying medical details. This is worrisome because when an AI platform oversimplifies medical details, it risks omitting critical information, leading to misinterpretation, reduced trust, inadequate decision-making, ethical concerns, and potential regulatory or legal issues. Balancing readability with accuracy is essential to ensure reliable and safe medical information. Accuracy, along with other critical aspects, will be evaluated in the next section.

DeepSeek generally remained closer to reference readability (Reference) but showed greater variability, while ChatGPT exhibited larger deviations, sometimes producing more complex or simpler responses (Figure 2A, B, D, E). The multi-response analysis improved consistency (Figure 2F), smoothing out extreme deviations (e.g., outliers in evident in Figure 2C) and leading to more balanced readability.

Table 1. Comparison of Flesch-Kincaid grade level (FKGL) scores for Reference, ChatGPT, and DeepSeek using dataset.

Single-Instance Analysis

Question	Reference	ChatGPT		DeepSeek	
	FKGL Score	FKGL Score	Difference*	FKGL Score	Difference*
Q1	7.30	10.80	3.50	9.90	2.60
Q2	9.90	11.30	1.40	10.30	0.40
Q3	9.90	11.20	1.30	9.70	-0.20
Q4	9.50	10.90	1.40	12.40	2.90
Q5	11.30	9.30	-2.00	10.60	-0.70
Q6	11.00	9.00	-2.00	10.60	-0.40
Q7	12.80	8.40	-4.40	8.40	-4.40
Q8	8.10	11.30	3.20	9.90	1.80
Q9	14.80	10.30	-4.50	10.70	-4.10
Q10	10.70	8.60	-2.10	10.90	0.20
Stats across all responses for 10 questions	Mean: 10.53 SD: 2.17 Median: 10.30 IQR: 1.63	Mean: 10.11 SD: 1.17 Median: 10.55 IQR: 2.05		Mean: 10.53 SD: 1.02 Median: 10.45 IQR: 0.77	

Multi-Response Analysis

Question	Reference	ChatGPT			DeepSeek		
	FKGL Score	FKGL Score	Difference*	SD	FKGL Score	Difference*	SD
Q1	7.30	8.63	1.33	1.32	8.73	1.43	1.43
Q2	9.90	11.20	1.30	0.77	11.47	1.57	0.78
Q3	9.90	11.17	1.27	0.76	9.73	-0.17	0.32
Q4	9.50	10.93	1.43	0.67	11.67	2.17	0.79
Q5	11.30	10.42	-0.88	1.06	10.68	-0.62	0.67
Q6	11.00	10.17	-0.83	1.09	11.20	0.20	0.96
Q7	12.80	9.47	-3.33	1.71	9.63	-3.17	1.00
Q8	8.10	10.92	2.82	0.38	9.20	1.10	0.71
Q9	14.80	10.90	-3.90	1.12	10.37	-4.43	0.66
Q10	10.70	9.83	-0.87	1.70	9.97	-0.73	0.66
Stats across all responses for 10 questions	Mean: 10.53 SD: 2.17 Median: 10.30 IQR: 1.63	Mean: 10.36 SD: 0.84 Median: 10.66 IQR: 1.01			Mean: 10.27 SD: 0.98 Median: 10.17 IQR: 1.41		

*Difference in comparison to Reference. Positive difference indicates that the FKGL score was higher than that of Reference, while negative difference means that the KFGL scores were lower than that of Reference.

A paired t-test was conducted to assess the statistical significance of mean differences in FKGL scores among the reference, ChatGPT, and DeepSeek groups. The Shapiro-Wilk test confirmed normality (p > 0.05 for all groups), validating the use of the t-test. Results indicated no significant differences in readability scores between the reference and either ChatGPT or DeepSeek, nor between ChatGPT and DeepSeek (*P*-values = .61 for single-instance, .73 for multi-response). However, high variability in scores across questions was observed. Multi-response analyses showed smaller differences between AI models and the reference, suggesting that averaging multiple responses yields more balanced readability. Normality was further confirmed by Shapiro-Wilk *P*-values of .16 (ChatGPT) and .82 (DeepSeek).

Table 2. Statistical comparison of FKGL scores for Reference, ChatGPT, and DeepSeek using a paired t-test.

Comparison	**Single-Instance**				**Multi-Response**			
	t-statistic	***P*-value**	**95% CI (Lower)**	**95% CI (Upper)**	**t-statistic**	***P*-value**	**95% CI (Lower)**	**95% CI (Upper)**
ChatGPT vs. Reference	-0.45	.66	-1.69	2.53	-0.24	.82	-1.40	1.74
DeepSeek vs. Reference	-0.24	.81	-1.58	1.96	-0.40	.70	-1.25	1.78
ChatGPT vs DeepSeek	0.52	.61	-1.23	0.77	0.35	.73	-0.53	0.72

t-statistic: Measures the magnitude and direction of the difference between paired samples. A negative t-value indicates that ChatGPT had lower scores than DeepSeek, while a positive t-value indicates the opposite; *P*-value: Represents the probability of observing the test results under the null hypothesis. A *P*-value ≤ 0.05 is considered statistically significant; 95% CI is the 95% confidence interval for the mean difference.

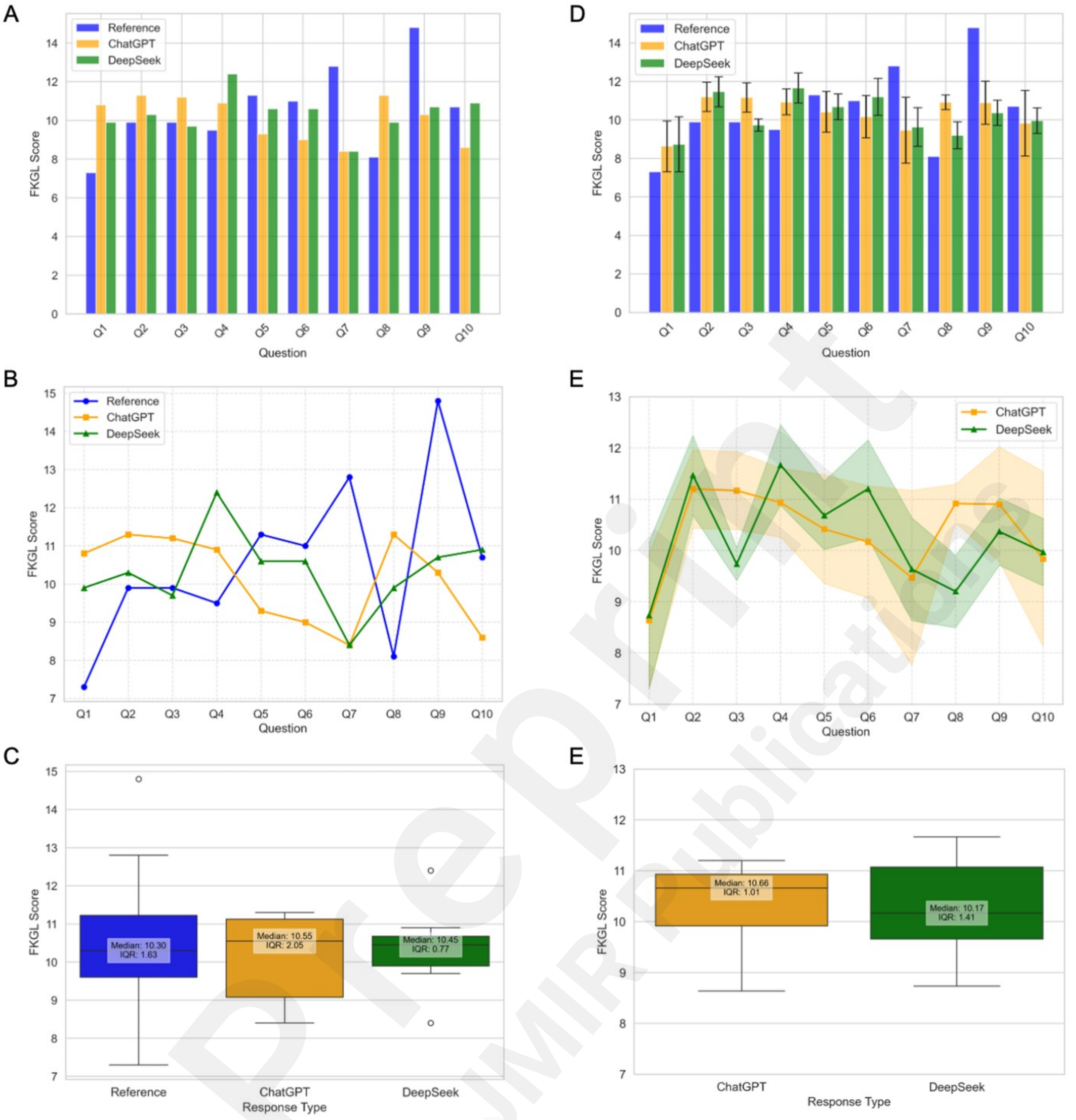

Figure 2. Comparison of Flesch-Kincaid Grade Level (FKGL) scores for Reference, ChatGPT, and DeepSeek responses. A) Bar plot comparing FKGL scores for 10 questions using one response per question from each platform. The bars represent the FKGL scores for Reference, ChatGPT, and DeepSeek. B) Line plot showing the overall trends of FKGL scores across all questions using one response per question from each platform. The lines represent the FKGL scores for Reference, ChatGPT, and DeepSeek. C) Box plot showing the distribution of FKGL scores using one response per question from each platform. The boxes represent the interquartile range (IQR), with the median annotated inside each box. Reference, ChatGPT, and DeepSeek are represented in blue, orange, and green, respectively. D) Bar plot comparing FKGL scores for 10 questions using six responses per question per platform. The bars represent the average FKGL scores, with error bars indicating the standard deviation. ChatGPT responses are shown in orange, and DeepSeek responses are shown in green. E) Line plot showing the overall trends of FKGL scores for ChatGPT and DeepSeek responses across 10 questions using six responses per question per platform. The solid lines represent the mean FKGL scores, and the shaded regions indicate the standard deviations. F) Box plot showing the

distribution of FKGL scores for ChatGPT and DeepSeek responses across 10 questions using six responses per question per platform. The boxes represent the interquartile range (IQR), with the median annotated inside each box. The y-axis for all line and box plots ranges from 7 to 15 to ensure a clear comparison of FKGL variations.

3.2. Assessing Response Quality for AI-Platforms ChatGPT and DeepSeek

The response quality of ChatGPT and DeepSeek was assessed using a 7-point Likert scale across five predefined criteria: accuracy, clarity & readability, completeness, depth & insight, and alignment with reference answer. The scale ranged from 1 (strongly disagree) to 7 (strongly agree), enabling systematic quantification of subjective judgments to compare answer quality between platforms.

In the single-instance analysis, both platforms had identical total scores (5.64), but deeper examination revealed variability in performance across criteria (Figure 3A-C). ChatGPT excelled in alignment with reference answer and accuracy, while DeepSeek achieved higher scores in other criteria, though its alignment and accuracy scores were slightly lower than ChatGPT's.

To robustly assess variability, the Likert analysis was extended to multi-response evaluations (Dataset 2, described in Methods). Averaging multiple responses tightened score distributions for both platforms (Figure 3D&E). DeepSeek's total score (5.61) was slightly higher than ChatGPT's (5.56), driven by marginally better performance in accuracy, completeness, and clarity. Figure 3F illustrates that both platforms converge when multiple responses are considered.

In both single-instance and multi-response analyses, ChatGPT and DeepSeek showed minimal differences, with average scores ranging from 5.00 to 5.90 for ChatGPT and 5.20 to 6.10 for DeepSeek. No statistically significant differences were observed (*P*-values > .05), and the 95% confidence intervals consistently overlapped zero, indicating no meaningful distinction between the models. Overall, the results suggest that ChatGPT and DeepSeek are similarly effective in generating responses across the evaluated criteria, with no significant performance differences. Paired t-test statistics are presented in Table 3.

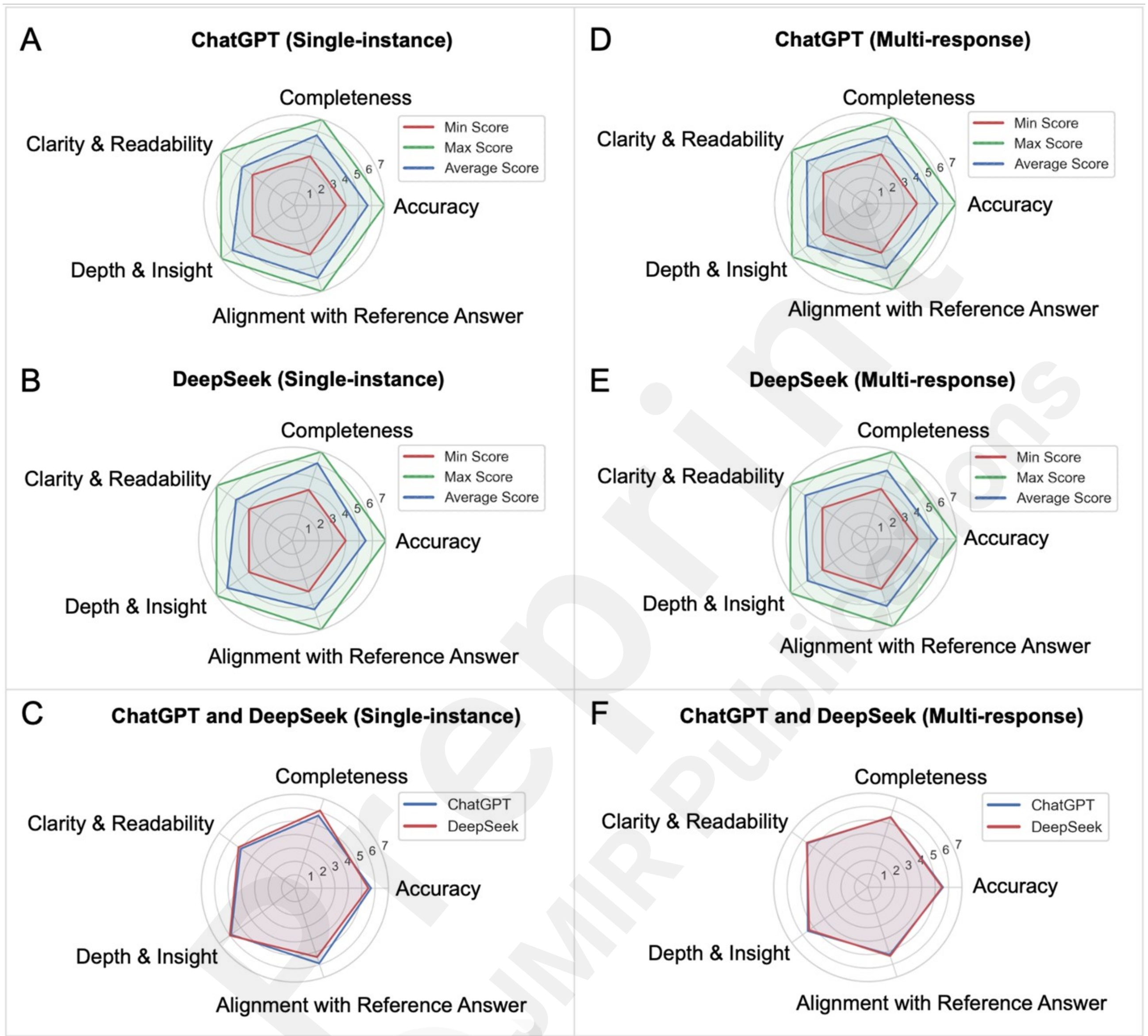

Figure 3. Radar plots of the Likert scores across five evaluation criteria for ChatGPT and DeepSeek. A) ChatGPT response evaluation based on average scores of 10 answers. B) DeepSeek response evaluation based on average score of 10 answers. C) ChatGPT vs. DeepSeek comparison based on average scores of 10 answers. D) ChatGPT response evaluation based on 60 answers. B) DeepSeek response evaluation based on 60 answers. F) ChatGPT vs. DeepSeek comparison based on average scores of 10 answers.

Table 3. Comparison of ChatGPT vs. DeepSeek across five evaluation criteria included in the Likert analysis using paired t-tests for the single-instance and multi-response analyses.

Single-Instance Results						
Evaluation Criterion	**ChatGPT Av. Score**	**DeepSeek Av. Score**	**t-Statistic**	***P*-Value**	**95% CI (Lower)**	**95% CI (Upper)**

Accuracy	5.70	5.50	0.56	.59	-0.61	1.01
Completeness	5.70	6.10	-0.63	.54	-1.84	1.04
Readability	5.00	5.20	-0.36	.73	-1.45	1.05
Depth & Insight	5.90	6.00	-0.23	.82	-1.08	0.88
Alignment with Reference Answer	5.90	5.40	1.05	.32	-0.58	1.58
Multi-Response Results						
Evaluation Criterion	**ChatGPT Av. Score**	**DeepSeek Av. Score**	**t-Statistic**	***P*-Value**	**95% CI (Lower)**	**95% CI (Upper)**
Accuracy	5.23	5.45	0.83	.41	-0.24	0.57
Completeness	5.67	5.89	-0.73	.47	-0.60	0.30
Readability	6.12	5.98	0.89	.43	-0.26	0.59
Depth & Insight	5.45	5.67	-0.88	.38	-0.59	0.22
Alignment with Reference Answer	5.78	5.56	0.17	.87	-0.32	0.39

t-statistic: Measures the magnitude and direction of the difference between paired samples. A negative t-value indicates that ChatGPT had lower scores than DeepSeek, while a positive t-value indicates the opposite; *P*-value: Represents the probability of observing the test results under the null hypothesis. A *P*-value ≤ 0.05 is considered statistically significant; 95% CI is the 95% confidence interval of the mean difference.

3.3. Assessing the Reliability of Information

This section evaluates the reliability and relevance of information sources provided by AI models in answering disease-specific questions related to breast cancer. ChatGPT primarily draws its references from U.S.-based medical institutions, cancer research centers, health organizations, and news outlets, with each question supported by 10–16 references. Key sources include the Komen Breast Cancer Foundation, Mayo Clinic, American Cancer Society, CDC, and WebMD. All references are web-based, with links tagged for retrieval; however, users must manually copy and paste URLs to access the information. Further details are provided in Table 4.

In contrast, DeepSeek offers more comprehensive answers, sourcing references globally from research articles, conference abstracts, and websites affiliated with cancer research centers, universities, pharmaceutical companies, and news outlets. Notable sources include the American Cancer Society, National Cancer Institute, and Cancer Research UK. While DeepSeek's links are generally accessible, some are untagged, and a few corrupted links were observed. The platform excels in citation efficiency, with references cited in-text like journal articles. However, it occasionally experiences downtime due to high user traffic, highlighting the need for developers to upgrade the platform to meet global demand.

Table 4. Comparison between ChatGPT and DeepSeek based on the sources of information.

Aspect	ChatGPT	DeepSeek
Total References	10-16 references per question.	More comprehensive references per question.
Source Types	• American medical institutions • Cancer research centers • Health organizations • Medical news outlets	• Cancer research centers • Universities • Pharmaceutical companies • News • Blogs • Reputable publishers (185 research/review articles) • Conferences (2 abstracts)
Major Reference Sources	Komen Breast Cancer Foundation (13), Mayo Clinic (13), American Cancer Society (12), CDC (8), National Cancer Institute (6), WebMD (5)	American Cancer Society (18), National Cancer Institute (8), Medical Xpress (7), Mayo Clinic (6), Cancer Research UK (5), Cancer Australia (5)
Additional Sources	MD Anderson Cancer Center, BCRF, City of Hope, Verywell Health, Medical Universities, Medical News Today, Wikipedia, News Media	Breast Cancer Research Foundation, AstraZeneca, CDC, Healthline, OncoLive, India Cancer Center, Novartis, Prevent Cancer Foundation, Universities
Web-Based Accessibility	All sources are web-based; URLs must be copied and pasted manually for access	All links are accessible, but some URLs are untagged; 4 corrupted links observed from ScienceDirect
Link Accessibility	• All links are web-based • Links tagged by ChatGPT (some untagged) • Links are accessible, but users must copy and paste URLs into a browser	• Links are not tagged at the terminus • All links are accessible and working (except 4 corrupted links from ScienceDirect due to internet script incompatibilities) • Users must copy and paste HTML addresses to access information
Geographical Diversity	Primarily United States	Sources from the United States, Australia, Canada, China, and India
Reference Types	Primarily institutional and nonprofit organizations	185 research and review articles; 2 conference abstracts
Citations in Text	Not explicitly cited in corresponding text	References cited in corresponding text like journal articles
Platform Performance	• Provides comprehensive answers • Links are functional but require manual copy-paste	• Surpasses ChatGPT in citation efficiency • References are cited in-text,

	• Reliable, but sources require manual verification.	similar to academic articles • Competitive and informative compared to web-based search engines • Platform occasionally out of service due to high user traffic • Occasionally out of service probably due to high worldwide demand.

Discussion

This study systematically compared the performance of ChatGPT-4.0 and DeepSeek-V3 in retrieving and presenting medical information, focusing on readability, content quality, and reliability of information sources. Therefore, a workflow that takes into account all of these aspects was devised (Figure 1) to identify the strengths and weaknesses of each AI model in handling medical FAQs. This workflow ensured a structured and comprehensive evaluation of AI models in the context of medical information retrieval, providing insights into their suitability for healthcare applications. Using quantitative (FKGL) and qualitative (7-point Likert scale) metrics, the analysis compared their performance in delivering disease-specific information regarding breast cancer. The findings highlighted each platform's strengths and limitations, offering valuable insights for users and developers.

The FKGL readability scores revealed that AI models often produce simpler responses than expert references, improving accessibility, but they risk oversimplifying complex medical information. For instance, ChatGPT's readability, assessed using FKGL scores, improved when averaging multiple responses, suggesting greater consistency and slightly enhanced readability in multi-response scenarios. On the other hand, DeepSeek's responses were closer to expert reference readability in single-instance analysis but showed increased variability in the multi-response case (IQR = 1.41), while ChatGPT's variability decreased, indicating greater stability. Notably, expert reference answers had higher FKGL scores (Mean = 10.53, SD = 2.17), reflecting more complex language, whereas AI models generally produced simpler responses, potentially improving accessibility for general audiences but risking the omission of key medical details. However, while AI-generated responses may be easier to read than expert content, readability alone does not ensure response quality, emphasizing the need for further assessment tests to balance clarity with medical precision.

Quality assessment of responses using the Likert scores, which evaluated accuracy, completeness, clarity, depth & insight, and alignment with reference answers, highlighted the strengths and weaknesses of AI models. ChatGPT demonstrated more balanced and consistent performance, making it a stronger choice for tasks requiring clarity and depth. DeepSeek, while competitive in accuracy and completeness, needs improvements in clarity and insight to match ChatGPT's overall performance. Furthermore, the response quality results indicated that single-instance analysis is more variable, and that the performance gap between ChatGPT and DeepSeek is narrower in the multi-response case.

Furthermore, the evaluation of ChatGPT and DeepSeek revealed key insights into their reliability based on the sources of information for answering breast cancer-related questions. ChatGPT relied heavily on U.S.-based medical institutions, health organizations, and news outlets, providing 10–16 references per question, all web-based but requiring manual URL access. In contrast, DeepSeek offered more comprehensive and globally diverse sources, including 185 research/review articles and 2 conference abstracts, with references cited in-text like academic articles. While DeepSeek excelled in citation efficiency and depth, it faced challenges such as untagged or corrupted links and occasional downtime due to high user traffic. Both AI models provided reliable information, but DeepSeek’s broader sourcing and academic-style citations made it more competitive, though improvements in link accessibility and platform stability were needed to enhance user experience.

Overall, these findings highlighted important considerations in AI-generated medical content. While ChatGPT may at times produce more complex responses, potentially making comprehension challenging for some users, DeepSeek’s closer alignment with reference answers may make it a better choice for delivering medical information at an appropriate readability level. However, the variability in response readability, quality, and reliability observed within and across AI models indicates that no single model consistently matches expert content across all questions. This highlights the importance of carefully evaluating AI-generated medical information to ensure it meets the needs of different user audiences.

It is also essential to recognize the limitations of the methodologies used for response evaluation. While the FKGL test offers valuable insights into text readability, it primarily measures structural complexity without considering context or meaning. Readability is also inherently subjective, influenced by factors such as the reader's familiarity with the topic and language proficiency. Moreover, FKGL is designed for English text, making its applicability to other languages uncertain; repeating this analysis across different languages would help assess result consistency. Beyond readability, AI's integration in medicine raises critical ethical, regulatory, and privacy concerns, requiring responsible implementation [17, 27]. Additionally, AI models face challenges such as biases in training data and the risk of misinformation, highlighting the need for continuous refinement and collaboration with verified medical sources to ensure reliability [28].

Conclusions

This comparative analysis of ChatGPT-4.0 and DeepSeek-V3 demonstrated that both AI platforms effectively retrieve and present medical information on breast cancer, each excelling in different areas. ChatGPT produced more polished, detailed, and readable responses, scoring higher in clarity and completeness, while DeepSeek provided more comprehensive, globally diverse references with in-text citations resembling academic articles. However, DeepSeek faced challenges such as untagged links, occasional downtime, and corrupted references, which impacted the user experience. Despite these limitations, DeepSeek demonstrated superior citation efficiency and closely aligned with expert consensus answers.

Statistical analysis revealed no significant differences between the models, particularly in larger datasets, underscoring the importance of rigorous evaluation of AI-generated medical information to ensure accuracy, accessibility, and reliability. To optimize their effectiveness in healthcare applications, future improvements should focus on enhancing platform stability, response consistency, and overall user accessibility.

Supporting Information

Supporting information consists of the following 11 supplementary tables.

Supplementary Table 1. FAQs about breast cancer and answers generated by AI models (ChatGPT and DeepSeek) by Researcher 1.

Supplementary Table 2. FAQs about breast cancer and answers generated by AI models (ChatGPT and DeepSeek) by Researcher 2.

Supplementary Table 3. FAQs about breast cancer and answers generated by AI models (ChatGPT and DeepSeek) by Researcher 3.

Supplementary Table 4. FAQs about breast cancer and expert consensus answers.

Supplementary Table 5. Average KFGL scores for the multi-response analysis and results of the paired t-test

Supplementary Table 6. Detailed FKGL scores for the multi-response analysis.

Supplementary Table 7. Average KFGL scores for the multi-response analysis and results of the paired t-test

Supplementary Table 8. Detailed FKGL scores for the multi-response analysis.

Supplementary Table 9. Likert scores for the single-instance analysis.

Supplementary Table S10. Likert scores for the multi-response analysis.

Supplementary Table 11. Information sources for ChatGPT and DeepSeek.

Acknowledgments

R.H. and D.A.S. acknowledge funding from the Deanship of Scientific Research at Al-Zaytoonah University of Jordan (Grant numbers: 2023-2022/17/50 and 2025-2024/06/29).

Conflict of Interest

The authors have no conflicts of interest to disclose.

Data Availability

Supplementary Tables 1–11 are available online on the journal's website. Additionally, all raw data, including collected responses and data resources, can be accessed on GitHub (https://github.com/rhajjo/AI-Models).

Abbreviations:

AI: Artificial Intelligence

FAQs: Frequently Asked Questions

FKGL: Flesch-Kincaid Grade Level

SD: Standard Deviation

IQR: Interquartile Range

CI: Confidence Interval

CDC: Centers for Disease Control and Prevention

URL: Uniform Resource Locator

HTML: Hypertext Markup Language

WHO: World Health Organization

Funding

Funding provided by the Deanship of Scientific Research at Al-Zaytoonah University of Jordan (Grant numbers: 2023-2022/17/50 and 2025-2024/06/29).

Author Contributions

Idea and conceptualization, (R.H. and S.K.B), data collection (R.H., D.A.S. S.K.B), methodology and analysis (R.H.), Writing and editing (R.H., D.A.S., S.K.B).

References:

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-49. PMID: 33538338. doi: 10.3322/caac.21660.
2. Sinha A, Naskar MNBJ, Pandey M, Rautaray SS. Challenges to the Early Diagnosis of Breast Cancer: Current Scenario and the Challenges Ahead. SN Comput Sci. 2024 2024/01/08;5(1):170. doi: 10.1007/s42979-023-02534-1.
3. Sunoqrot S, Abusulieh S, Sabbah D. Polymeric Nanoparticles Potentiate the Anticancer Activity of Novel PI3Kα Inhibitors Against Triple-Negative Breast Cancer Cells. Biomedicines. 2024;12(12):2676. PMID: 10.3390/biomedicines12122676. doi: 10.3390/biomedicines12122676.
4. Sweidan K, Elfadel H, Sabbah DA, Bardaweel SK, Hajjo R, Anjum S, et al. Novel Derivatives of 4,6-Dihydroxy-2-Quinolone-3-Carboxamides as Potential PI3Kα Inhibitors. ChemistrySelect. 2022;7(32):e202202263. PMID: 10.1002/slct.202202263. doi: https://doi.org/10.1002/slct.202202263.
5. World Health Organization. Global breast cancer initiative implementation framework: assessing, strengthening and scaling-up of services for the early detection and management of breast cancer: World Health Organization; 2023. ISBN: 9240065989.
6. Kasper G, Momen M, Sorice KA, Mayhand KN, Handorf EA, Gonzalez ET, et al. Effect of neighborhood and individual-level socioeconomic factors on breast cancer screening adherence in a multi-ethnic study. BMC public health. 2024 Jan 2;24(1):63. PMID: 38166942. doi: 10.1186/s12889-023-17252-9.
7. Chen J, Duan Y, Xia H, Xiao R, Cai T, Yuan C. Online health information seeking behavior among breast cancer patients and survivors: a scoping review. BMC women's health. 2025 Jan 3;25(1):1. PMID: 39754199. doi: 10.1186/s12905-024-03509-x.
8. Loeb S, Langford AT, Bragg MA, Sherman R, Chan JM. Cancer misinformation on social media. CA Cancer J Clin. 2024 Sep-Oct;74(5):453-64. PMID: 38896503. doi: 10.3322/caac.21857.
9. Johnson SB, Parsons M, Dorff T, Moran MS, Ward JH, Cohen SA, et al. Cancer Misinformation and Harmful Information on Facebook and Other Social Media: A Brief Report. J Natl Cancer Inst. 2022 Jul 11;114(7):1036-9. PMID: 34291289. doi: 10.1093/jnci/djab141.
10. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res. 2023 Jun 30;25:e47479. PMID: 37389908. doi: 10.2196/47479.
11. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. PMID: 37215063. doi: 10.3389/frai.2023.1169595.
12. Choi J, Kim JW, Lee YS, Tae JH, Choi SY, Chang IH, et al. Availability of ChatGPT to provide medical information for patients with kidney cancer. Sci Rep.

2024 2024/01/17;14(1):1542. PMID: 38233511. doi: 10.1038/s41598-024-51531-8.
13. Peng Y, Malin BA, Rousseau JF, Wang Y, Xu Z, Xu X, et al. From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare. J Biomed Inform. 2025 2025/02/10/:104791. PMID: 39938624. doi: https://doi.org/10.1016/j.jbi.2025.104791.
14. Kincaid JP, Braby R, Mears JE. Electronic authoring and delivery of technical information. J Instr Dev. 1988;11(2):8-13. PMID: 10.1007/BF02904998. doi: 10.1007/BF02904998.
15. Sullivan GM, Artino AR, Jr. Analyzing and interpreting data from likert-type scales. J Grad Med Educ. 2013 Dec;5(4):541-2. PMID: 24454995. doi: 10.4300/jgme-5-4-18.
16. Jarab AS, Al-Qerem W, Al-Hajjeh DaM, Abu Heshmeh S, Mukattash TL, Naser AY, et al. Artificial intelligence utilization in the healthcare setting: perceptions of the public in the UAE. Int J Environ Health Res. 2024:1-9. PMID: 38832887. doi: 10.1080/09603123.2024.2363472.
17. Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract. 2018 Mar;68(668):143-4. PMID: 29472224. doi: 10.3399/bjgp18X695213.
18. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31-8. PMID: 35058619. doi: 10.1038/s41591-021-01614-0.
19. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep. 2014 Jan;16(1):441. PMID: 24338557. doi: 10.1007/s11886-013-0441-8.
20. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care. 2019 Jul;8(7):2328-31. PMID: 31463251. doi: 10.4103/jfmpc.jfmpc_440_19.
21. Sahu M, Gupta R, Ambasta RK, Kumar P. Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. Prog Mol Biol Transl Sci. 2022;190(1):57-100. PMID: 36008002. doi: 10.1016/bs.pmbts.2022.03.002.
22. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. Academic medicine : journal of the Association of American Medical Colleges. 2024 Jan 1;99(1):22-7. PMID: 37651677. doi: 10.1097/acm.0000000000005439.
23. Hajjo R, Sabbah DA, Bardaweel SK, Tropsha A. Identification of tumor-specific MRI biomarkers using machine learning (ML). Diagnostics. 2021;11(5):742. PMID: 33919342. doi: 10.3390/diagnostics11050742.
24. Korea Biomedical Review. The DeepSeek dilemma: navigating innovation and security concerns in Korea's healthcare industry; available from: https://www.koreabiomed.com/news/articleView.html?idxno=26581 (accessed on 13 February 2025).
25. CNN Medicine. A shocking Chinese AI advancement called DeepSeek is sending US stocks plunging; available from: https://edition.cnn.com/2025/01/27/tech/deepseek-stocks-ai-china/index.html (accessed on 13 February 2025).
26. Ye Z, Zhang B, Zhang K, Méndez MJG, Yan H, Wu T, et al. An assessment of ChatGPT's responses to frequently asked questions about cervical and breast cancer. BMC women's health. 2024 Sep 2;24(1):482. PMID: 39223612. doi:

10.1186/s12905-024-03320-8.
27. Kitsios F, Kamariotou M, Syngelakis AI, Talias MA. Recent Advances of Artificial Intelligence in Healthcare: A Systematic Literature Review. Appl Sci. 2023;13(13):7479. PMID: 10.3390/app13137479. doi: https://doi.org/10.3390/app13137479.
28. 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. PMID: 10.1056/NEJMsr2214184. doi: 10.1056/NEJMsr2214184.

Supplementary Files

Multimedia Appendixes

Supporting Information.
URL: http://asset.jmir.pub/assets/96894d63501a669e0d3aa8e84067939e.zip