Cross-sectional Study

INTERNATIONAL JOURNAL OF SURGERY

OPEN

Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2

Linfang Deng, MD[a], Tianyi Wang, MD[b], Yangzhang, MD[g], Zhenhua Zhai, PhD[d], Wei Tao, MD[c], Jincheng Li, MD[c], Yi Zhao, MD[c], Shaoting Luo, MD[e,*], Jinjiang Xu, MD[f,*]

Background Large language models (LLMs) have garnered significant attention in the AI domain owing to their exemplary context recognition and response capabilities. However, the potential of LLMs in specific clinical scenarios, particularly in breast cancer diagnosis, treatment, and care, has not been fully explored. This study aimed to compare the performances of three major LLMs in the clinical context of breast cancer.

Methods In this study, clinical scenarios designed specifically for breast cancer were segmented into five pivotal domains (nine cases): assessment and diagnosis, treatment decision-making, postoperative care, psychosocial support, and prognosis and rehabilitation. The LLMs were used to generate feedback for various queries related to these domains. For each scenario, a panel of five breast cancer specialists, each with over a decade of experience, evaluated the feedback from LLMs. They assessed feedback concerning LLMs in terms of their quality, relevance, and applicability.

Results There was a moderate level of agreement among the raters (*Fleiss' kappa* = 0.345, $P < 0.05$). Comparing the performance of different models regarding response length, GPT-4.0 and GPT-3.5 provided relatively longer feedback than Claude2. Furthermore, across the nine case analyses, GPT-4.0 significantly outperformed the other two models in average quality, relevance, and applicability. Within the five clinical areas, GPT-4.0 markedly surpassed GPT-3.5 in the quality of the other four areas and scored higher than Claude2 in tasks related to psychosocial support and treatment decision-making.

Conclusion This study revealed that in the realm of clinical applications for breast cancer, GPT-4.0 showcases not only superiority in terms of quality and relevance but also demonstrates exceptional capability in applicability, especially when compared to GPT-3.5. Relative to Claude2, GPT-4.0 holds advantages in specific domains. With the expanding use of LLMs in the clinical field, ongoing optimization and rigorous accuracy assessments are paramount.

Keywords: breast cancer, ChatGPT, Claude2, comparison, large language models

Introduction

With the rapid advancements in natural language processing technology, chatbots are increasingly emerging in the medical and health sector, demonstrating vast potential in disease prevention, diagnosis, treatment, monitoring, and patient support[1–5]. Large language models (LLMs) in the natural language processing field have undergone comprehensive optimization in terms of data, structure, and performance, enabling them to exhibit higher

[a]Department of Nursing, Jinzhou Medical University, Jinzhou, [b]Department of Clinical Trials, [c]Department of Breast Surgery, [d]Department of General Surgery, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, [e]Department of Pediatric Orthopedics, Shengjing Hospital of China Medical University, Shenyang, [f]Department of Health Management Center, The First Hospital of Jinzhou Medical University, Jinzhou, Liaoning and [g]Department of Breast Surgery, Xingtai People's Hospital of Hebei Medical University, Xingtai, Hebei, People's Republic of China

Linfang Deng and Tianyi Wang should be considered joint first authors.

Jinjiang Xu and Shaoting Luo contributed equally as co-corresponding authors.

In breast cancer clinical applications, GPT-4.0 outperforms GPT-3.5 and Claude2 in quality, relevance, and applicability, with domain-specific advantages observed. Continuous optimization and assessment of LLMs are crucial for clinical use.

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding authors. Address: Department of Pediatric Orthopedics, Shengjing Hospital of China Medical University, Shenyang 110000, People's Republic of China. Tel.: +861 356 664 1792; fax: +243 193 9638. E-mail: lst634747560@163.com (S. Luo); Department ofHealth Management Center, The First Hospital of Jinzhou Medical University, Jinzhou 121000, Liaoning, People's Republic of China. Tel.: +861 573 394 1406; fax: +416 419 7094. 9638. E-mail: 15733941406@163.com (J. Xu).

International Journal of Surgery (2024) 110:1941–1950

Received 18 October 2023; Accepted 23 December 2023

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww.com/international-journal-of-surgery.

Published online 23 January 2024

http://dx.doi.org/10.1097/JS9.0000000000001066

accuracy and application potential when addressing medical domain issues.

Breast cancer, as one of the major health threats to women worldwide, places particular emphasis on early diagnosis and treatment. It is estimated that in 2020, about 685 000 women died from breast cancer, accounting for 16% of female cancer deaths, meaning that 1 out of every six female cancer deaths was due to breast cancer[6]. Against this backdrop, the ChatGPT series of LLMs has gained attention in the medical field for its ability to process and generate vast amounts of medical information. Walker *et al*. evaluated the medical answers generated by GPT-4. They reported that 60% of these answers were consistent with health guideline recommendations, although the quality of these recommendations was mostly low to moderate[7]. Furthermore, other studies have shown that ChatGPT's performance in medical knowledge evaluation is nearing the level of professional medical personnel, further validating its potential application value in clinical medicine[8]. However, the recently emerged model, Claude2, remains unexplored and unassessed in this field. Claude2 is a new AI language model developed by the Anthropic company after years of deep learning research. The model adopts an innovative transformer architecture, which effectively captures long-distance semantic dependencies and can recognize emotions and tones in the text[9]. Notably, models like ChatGPT and Claude2 are primarily trained on vast amounts of information from the internet. Given the varying quality of information available online, these models may pose a risk of generating inaccurate or misleading information when addressing breast cancer-related issues. Therefore, a systematic evaluation of the quality, relevance, and applicability of these LLMs in the field of breast cancer becomes especially important.

This study aimed to evaluate and compare the performance of three major LLMs: GPT-3.5, GPT-4.0, and Claude2 in simulated breast cancer clinical scenarios. Through this comparison, we hope to offer clear insights into the potential value and limitations of LLMs in clinical decision support for both breast cancer patients and medical experts.

HIGHLIGHTS

- Pioneering research comparing the capabilities of ChatGPT-3.5, ChatGPT-4.0, and Claude2 in the specialized context of breast cancer clinical scenarios.
- First-ever comprehensive evaluation of Claude2's performance in the breast cancer clinical setting.
- Advanced evaluation methodology: divided clinical scenarios into five pivotal domains, providing a multifaceted assessment platform.
- GPT-4.0 consistently demonstrated superior performance across multiple domains, signifying its potential in clinical applications.
- While GPT-4.0 excelled overall, Claude2 exhibited niche strengths, especially in the domain of assessment and diagnosis.
- Unearths the need for domain-specific optimization of LLMs, pushing the frontier of AI's applicability in critical medical scenarios.

Materials and methods

Ethics

Approval from the ethics committee was not required since no patients were involved in our study.

Study design

This study combined qualitative and exploratory research methods to deeply assess the performance of GPT-4.0, GPT-3.5, and Claude2 LLMs in the clinical breast cancer scenario. We organized an expert team consisting of five breast cancer specialists with at least 10 years of experience. The comprehensive expertise of these team members was instrumental in ensuring an exhaustive and meticulous evaluation of the LLMs' capabilities in this clinical scenario. Importantly, these specialists are not only experts in the medical and surgical management of breast cancer but also possess extensive knowledge and skills in holistic patient care, including the evaluation and support of psychosocial aspects. The study spanned from 7 August 2023 to 10 September 2023, covering two versions of ChatGPT (GPT-3.5 and GPT-4.0, OpenAI) and Claude2 (a new language model AI system developed by Anthropic after years of deep learning research) to generate responses to these queries. The work has been reported in line with the strengthening the reporting of cohort, cross-sectional, and case–control studies in surgery (STROCSS) criteria[10].

Design, classification, and evaluation of simulated clinical scenarios

Based on five key clinical areas of breast cancer: 1. assessment and diagnosis; 2. treatment decision-making; 3. postoperative care; 4. psychosocial support; 5. prognosis and rehabilitation, we designed clinical simulation scenarios. These scenarios were developed with valuable input from two experienced professionals in the field of breast cancer research and treatment. The first professional, with a strong background in oncogenetics, provided in-depth knowledge in diagnostic methodologies. The second professional, skilled in the area of breast cancer rehabilitation, ensured that the scenarios are reflective of current treatment and postoperative care practices. Their collective expertise and insights ensure the authenticity and relevance of these simulations, offering a comprehensive framework for evaluating the effectiveness of the LLMs in diverse clinical contexts.

Clinical scenario generation

For each simulated case, the LLM's response was evaluated to determine its quality, relevance, and applicability.

Obtaining responses from LLMs

Figure 1 illustrates the research design framework. First, we inputted nine case questions from five domains into each LLM-Chatbots, with each question considered as a separate query. To minimize the memory bias of LLM-Chatbots, we reset the conversation after each query. To ensure the reviewers could not identify each LLM chatbot, we formatted all responses into plain text, masking all features related to specific chatbots. These responses were then randomly shuffled and submitted to five breast specialists for scoring. The scoring process was divided into three stages, each spaced

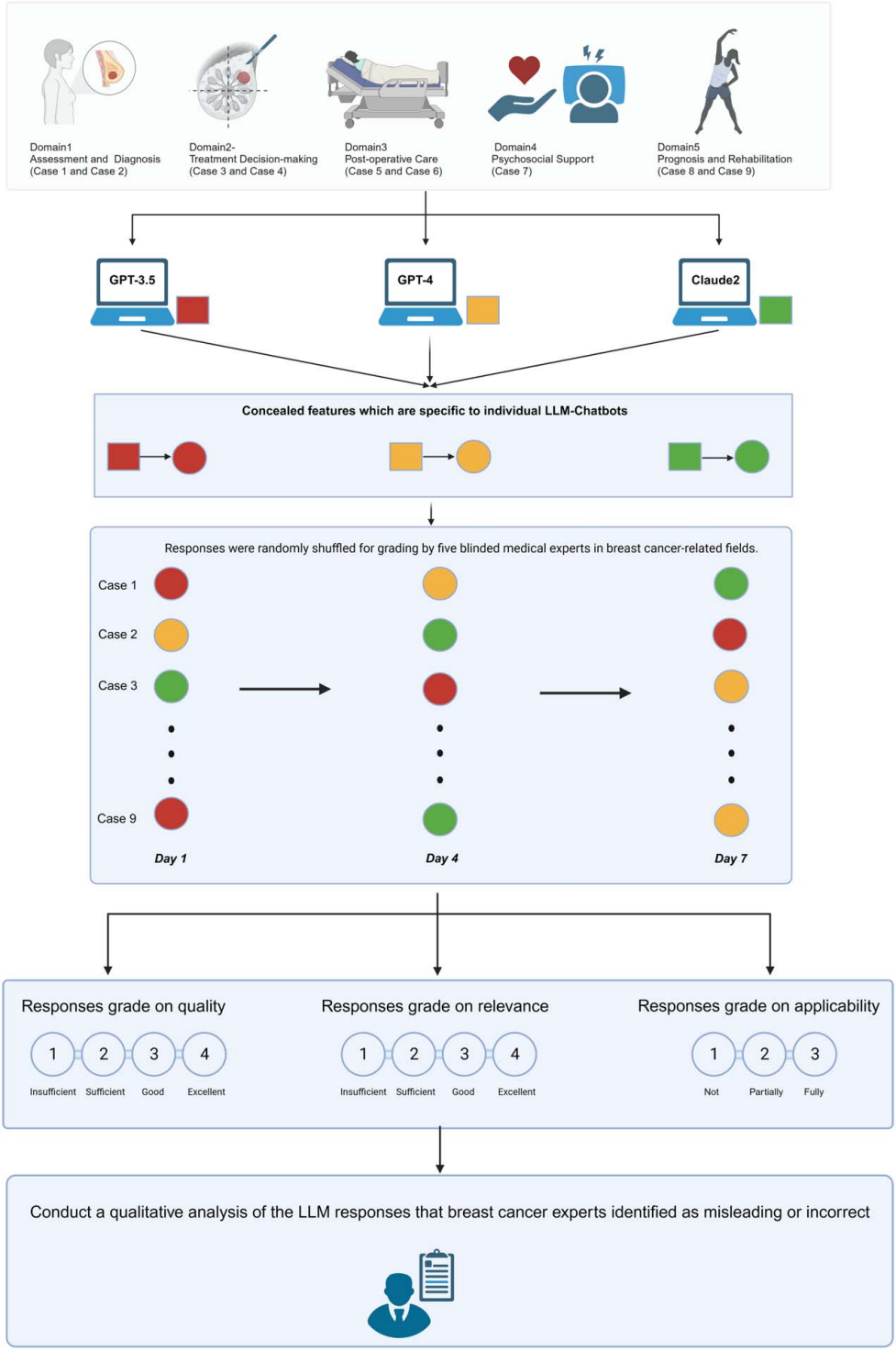

Figure 1. Flowchart of overall study design. Notes: $^{****}P$ <0.001. $^{*}P$ <0.05. Error bars represent the SD for each data set.

48 h apart, to reduce sequential effects. Each clinical case was treated as a brand-new dialog, ensuring that previous interactions did not affect the current response. This eliminated potential biases from previous interactions and ensured result consistency.

Evaluation of LLMs responses

Recruitment of breast specialists

Breast cancer specialists ($N = 5$, with at least 10 years of experience) were recruited for this study.

Evaluation method

We systematically evaluated the responses from GPT-4.0, GPT-3.5, and Claude2, utilizing pre-established evaluation criteria centered on three core dimensions: quality, relevance, and applicability. Every evaluation dimension was meticulously defined and equipped with a corresponding scoring mechanism to guarantee a scientific and consistent evaluation. A specially designed online questionnaire tool was employed to enhance the efficiency and accuracy of the evaluation by integrating all AI model feedback and their respective evaluation indicators. Experts were invited to directly access and evaluate these models by using a questionnaire. (https://www.wjx.cn/vm/Y4eYhoB.aspx#); (https://www.wjx.cn/vm/wFNfpRL.aspx#) ; (https://www.wjx.cn/vm/wFNfpRL.aspx#). Further details regarding the evaluation criteria and methodology are provided in Supplementary Texts 1, Supplemental Digital Content 3, http://links.lww.com/JS9/B747, 2, Supplemental Digital Content 4, http://links.lww.com/JS9/B748, and 3, Supplemental Digital Content 5, http://links.lww.com/JS9/B749 and Supplementary Table 1, Supplemental Digital Content 1, http://links.lww.com/JS9/B745.

Scoring method

To ensure objectivity, we employed a double-blind method. In this study, GPT-3.5, GPT-4.0, and Claude2 were designated as AI1, AI2, and AI3, respectively. Scoring experts were uninformed about which AI model corresponded to each label. All experts independently scored the answers from AI1, AI2, and AI3. Given the potential subjective biases that might emerge during the scoring process, we implemented specific strategies to mitigate these biases. We utilized the 'Majority Consensus Method' to determine the final score for each answer, that is, based on the most common score provided by experts. For instance, if three experts gave a score of 2 to a particular answer from AI1, while the other two experts gave a score of 3, then the final score for that answer was 2. When there is no identical score given by three or more out of the five scorers, it indicates a lack of consensus. In such cases, we default to the strictest evaluation standard, assigning the lowest score out of those given by the five experts.

Qualitative analysis

To systematically identify and deeply understand potentially misleading or inaccurate responses from LLMs in the breast cancer clinical scenario, we conducted a deep qualitative analysis of LLMs-generated answers that experts flagged as potentially misleading or erroneous.

Statistical methods

The study employed SPSS 21.0 software (IBM Corp.) for statistical analysis and used GraphPad Prism 8.0.1 (GraphPad Software, Inc.) for data visualization and graph plotting. Fleiss' kappa was used to quantify the consistency of scores among multiple raters. Dunn's post-hoc test was used for in-depth posterior comparisons of average overall accuracy, relevance, and applicability scores. For the responses of the three AIs in five specific areas, we used the Mann–Whitney *U* test to identify significant differences between the areas. When multiple hypothesis tests were conducted, *P*-values were adjusted using the Bonferroni correction method. A *P*-value of less than 0.05 was considered statistically significant.

Results

Inter-rater reliability

The results of the Fleiss' kappa statistic showed a statistically significant inter-rater reliability of 0.345 among the five raters ($Z = 13.573$, $P < 0.05$). This value, falling between 0 and 1, indicates a moderate level of agreement among the raters. Specifically, the strength of their agreement was within a 95% CI ranging from 0.295 to 0.394. This suggests that they mostly rated consistently across situations.

Response length analysis of GPT-3.5, GPT-4.0, and Claude2

Table 1 presents the response lengths of the LLM chatbots across nine analytical cases. The average word count and SD were 512.67 ± 195.46 for GPT-4.0, 558.78 ± 198.02 for GPT-3.5, and 255.00 ± 114.05 for Claude2. The character count averaged 3500.00 ± 1396.03 for GPT-4.0, 3861.89 ± 1400.86 for GPT-3.5, and 1725.56 ± 809.97 for Claude2.

Evaluation of the average overall quality, relevance, and applicability scores of expert responses for GPT-4.0, GPT-3.5, and Claude2

As shown in Figure 2, the average overall quality score for GPT-4.0 was 3.56 ± 0.55, which was significantly higher than that of GPT-3.5 at 2.87 ± 0.69 (Dunn's post-hoc test, $P = 0.001$) and Claude2 at 3.18 ± 0.72 (Dunn's post-hoc test, $P = 0.0196$). In terms of relevance scores, GPT-4.0 scored 3.44 ± 0.59, while GPT-3.5 scored 2.78 ± 0.56 (HSD post-hoc test, $P = 0.001$) and Claude2 scored 3.13 ± 0.59 (HSD post-hoc test, $P = 0.0315$). For applicability scores, GPT-4.0 scored 2.73 ± 0.45, significantly higher than GPT-3.5's 2.13 ± 0.34 (HSD post-hoc test, $P = 0.001$) and Claude2's 2.49 ± 0.51 (HSD post-hoc test, $P = 0.0243$).

Evaluation of GPT-4.0, GPT-3.5, and Claude2 based on consensus for nine breast cancer cases in terms of quality, relevance, and applicability

Figure 3 displays the differential consensus scores of GPT-4.0, GPT-3.5, and Claude2 in terms of accuracy, relevance, and applicability. As depicted, GPT-4.0 consistently achieves higher scores across all three metrics. Claude2 consistently outperforms GPT-3.5 in each dimension, while GPT-3.5 scores the lowest in all three areas. Regarding quality, there was a statistically significant difference between the scores of GPT-4.0 and GPT-3.5 ($z = -2.040$, $P = 0.041$). However, there was no statistical difference between GPT-4.0 and Claude2 ($z = -1.055$, $P = 0.291$) or between GPT-3.5 and Claude2 ($z = -1.060$, $P = 0.289$). In terms of relevance, scores between GPT-4.0 and GPT-3.5 showed a statistically significant difference ($z = -2.188$, $P = 0.029$). However, there was no significant difference between GPT-4.0 and Claude2 ($z = -1.352$, $P = 0.176$) or between GPT-3.5 and Claude2 ($z = -0.974$, $P = 0.330$). For applicability, there existed a statistical difference between the scores of GPT-4.0 and GPT-3.5 ($z = -2.349$, $P = 0.019$) as well as between GPT-3.5 and Claude2 ($z = -1.944$, $P = 0.049$). However, no statistical difference was observed between GPT-4.0 and Claude2 ($z = -0.470$, $P = 0.638$).

Table 1

Response length analysis of GPT-3.5, GPT-4.0, and claude2.

	Response length (words)			Response length (characters)		
LLM	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
GPT-4.0	512.67 (195.46)	337	879	3500.00 (1396.03)	2143	5875
GPT-3.5	558.78 (198.02)	336	859	3861.89 (1400.86)	2353	6011
Claude2	255.00 (114.05)	154	492	1725.56 (809.97)	1077	3401

Performance of GPT-4.0, GPT-3.5, and Claude2 in five major clinical domains

Figure 4 illustrates the performance of the three chatbots across major clinical domains. In terms of quality, Claude 2 performs slightly better than GPT-4.0 and GPT-3.5 in the domain of assessment and diagnosis. However, there was no statistically significant difference among the three ($P > 0.05$). In the other four domains, the quality score of GPT-4.0 was significantly higher than that of GPT-3.5 ($P < 0.05$), but there was no significant difference between GPT-4.0 and Claude 2 ($P > 0.05$). Regarding relevance, GPT-4.0 scored significantly better than GPT-3.5 in treatment decision-making, postoperative care, psychosocial support, and prognosis and rehabilitation ($P < 0.05$). Especially in the psychosocial support task, the score of GPT-4.0 was much higher than that of Claude 2 ($P < 0.001$). Regarding applicability, GPT-4.0 scored significantly higher than GPT-3.5 in treatment decision-making, postoperative care, and prognosis and rehabilitation ($P < 0.05$). It is particularly noteworthy that GPT-4.0 also scored higher than Claude 2 in treatment decision-making and psychosocial support ($P < 0.05$).

Qualitative assessment of responses from LLM chatbots

Supplementary Table 2 (Supplemental Digital Content 2, http://links.lww.com/JS9/B746) showcases examples of inaccurate responses from the LLM chatbots, with the erroneous portions highlighted in yellow. Additionally, these tables include detailed explanations of the identified errors, and a breast cancer specialist (ZCY) has also provided professional opinions on them.

Discussion

In this study, we comprehensively evaluated GPT-3.5, GPT-4.0, and Claude2, focusing on their application value in the clinical domain of breast cancer. By employing a robust study design with appropriate masking and randomization, we aimed to unravel the potentials and limitations of these LLMs in pertinent clinical scenarios. Including five breast cancer specialists in the evaluation process, we ensured our analysis was imbued with expert insights. Notably, this research represents, to our knowledge, the first assessment of Claude2, thereby embedding unique and valuable perspectives into our study.

While LLMs have been utilized in numerous fields[8,11–14], it becomes evident that research exploring their potential value, specifically in the clinical application for breast cancer, remains relatively limited. This study endeavors to bridge this knowledge gap, aligning itself as a critical connector by specifically targeting tasks associated with simulated clinical scenarios to evaluate the performance of each model. Our approach underscores the significance of evaluating the quality, relevance, and applicability of model responses in practical settings, thereby offering a more profound insight into the model's efficacy in real-world circumstances.

This study found that GPT-4.0 demonstrates superiority in five principal clinical domains. In parallel, Claude2 has shown considerable traits and potential. This lays a solid foundation for the application of Claude2 in future clinical practice and highlights its unique advantages in specific clinical scenarios.

However, while Claude2's performance was on par with GPT-4.0 in certain evaluative metrics, it was noticeably less proficient in the realm of psychosocial support. This indicates a superior depth and precision in GPT-4.0's responses when addressing psychosocial concerns. On the other hand, all three LLM chatbots demonstrated exceptional response capabilities within five clinical task domains, particularly in queries pertaining to breast cancer treatment strategies. They have provided comprehensive and relevant responses. This further emphasizes the potent ability of LLM chatbots to deliver relevant and precise information in medical applications, highlighting their significant potential in future medical practices.

Among the three LLM chatbots evaluated, GPT-4.0 stands out in the clinical application of breast cancer, achieving the highest average overall quality score. Arya Rao *et al*. and Jialin Liu *et al*. highlighted the superiority of ChatGPT, especially GPT-4.0, in radiologic decision-making and addressing myopia-related queries, respectively, compared to other LLMs[15,16]. One key reason for GPT-4.0's performance is its extensive parameter set, which equips it with the prowess to process medical information, especially in intricate breast cancer diagnosis and treatment scenarios. This complexity requires robust data processing capabilities to ensure the quality and relevance of medical information. Furthermore, integrating an advanced reasoning mechanism and stringent adherence to guidelines enables GPT-4.0 to address complex clinical requirements[17]. This signifies that it can provide precise answers to medical practitioners and patients. It is also imperative to note that the inclusion of a substantial volume of updated medical training data and the assimilation of lessons from real-world application experiences collectively enhance the quality and relevance of the responses provided by GPT-4.0.

With regard to erroneous information, we present a pivotal example. In evaluating the application of fine-needle aspiration in diagnosing breast lumps, GPT-3.5 provided recommendations that were somewhat general and did not adequately reflect advancements in breast lump diagnostic strategies. In particular, in certain cases, subsequent treatment plans can be directly formulated based on radiological examinations rather than solely relying on fine-needle aspiration for a definitive diagnosis[18]. Moreover, with the advancement of medical technology, the utilization rate of Fine-Needle Aspiration Cytology (FNAC) in

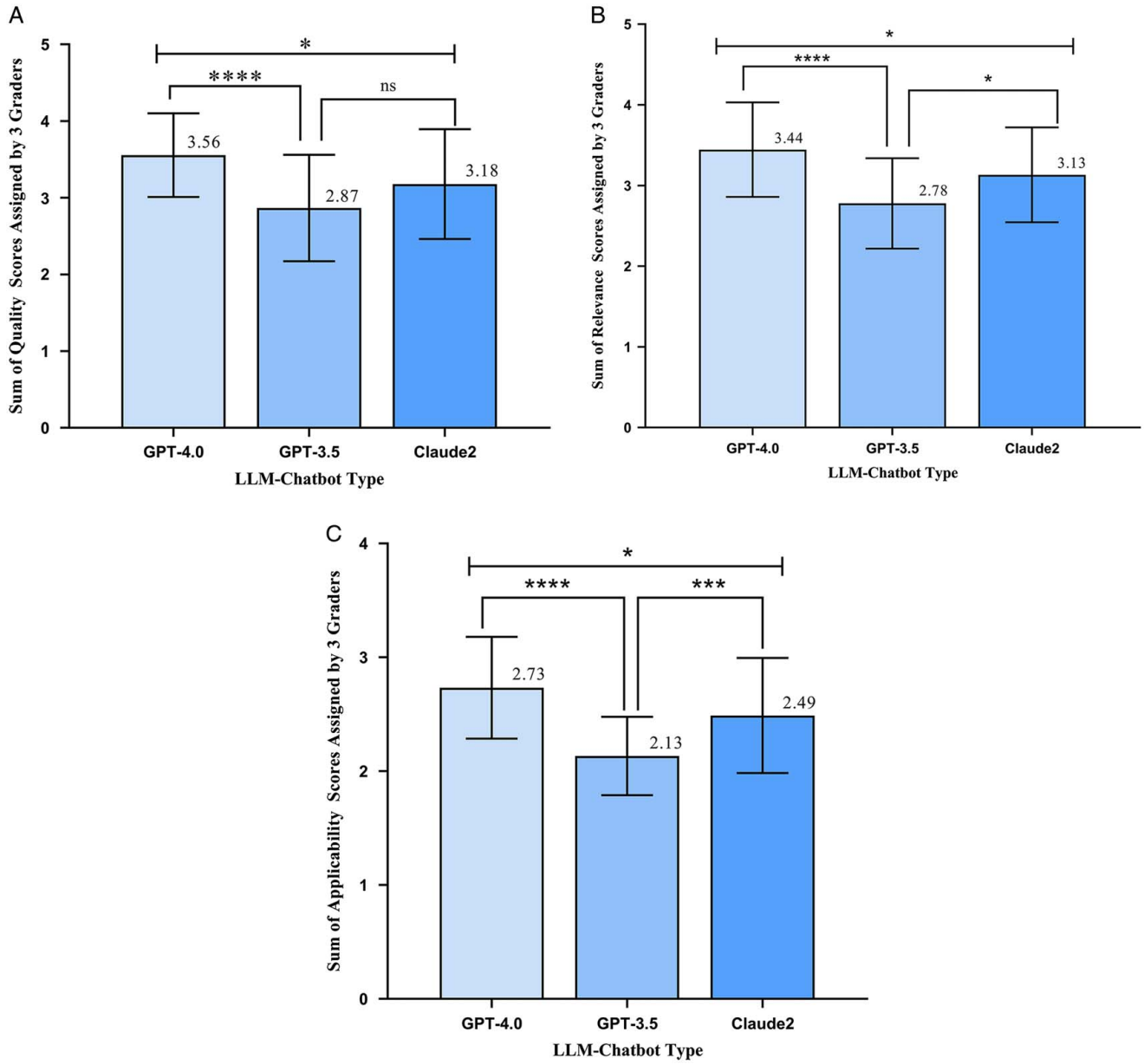

Figure 2. Average overall Quality, Relevance, and Applicability scores of responses from GPT-4.0, GPT-3.5, and Claude2 across 9 cases, as assessed by five breast specialists. (A) Average overall Quality scores of responses from GPT-4.0, GPT-3.5, and Claude2 across 9 cases. (B)Average overall Relevance scores of responses from GPT-4.0, GPT-3.5, and Claude2 across 9 cases. (C) Average overall Applicability scores of responses from GPT-4.0, GPT-3.5, and Claude2 across 9 cases.

breast cancer diagnosis has declined[19]. However, GPT-3.5 seemed to not capture this trend, particularly noticeable in medical practices in China. The cytological samples provided by FNAC have relatively weak evidential strength and are generally not used as the basis for confirming malignant tumors[20,21].

Another manifestation of this oversight is evident when considering recommendations by various AI models regarding breast cancer diagnostic methods. Distinctively, models such as GPT-3.5 and Claude2 suggested the application of either mammography or breast ultrasound for further assessment and diagnosis of breast cancer. However, GPT-4 leaned toward ultrasound as the preferred evaluation method, aligning with the 2019 'Chinese Women's Breast Cancer Screening Guidelines'[22]. While Western medical guidelines predominantly advocate for mammography as the primary screening modality for breast cancer[23–27], its efficacy is diminished for the majority of Chinese women due to their prevalent higher breast density[28]. Consequently, ultrasound emerges as a more reliable modality, outperforming mammography in both sensitivity and accuracy for Chinese women. Therefore, how LLMs adjust their medical advice according to practices and changes in different regions and consider the source and diversity of training data is an urgent issue warranting study.

A pressing concern is that potential hazards models might present when processing intricate medical data. Claude2, while

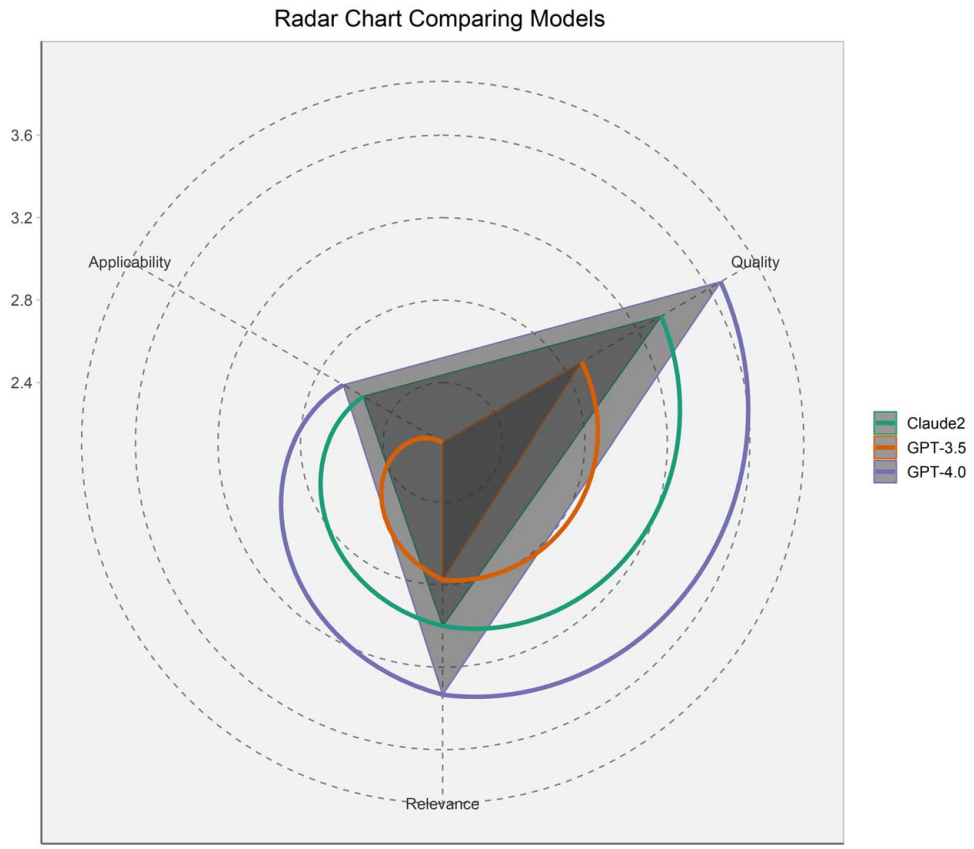

Figure 3. Consensus-based accuracy, relevance, and applicability evaluation of GPT-4.0, GPT-3.5, and Claude2 in nine breast cancer cases.

dealing with tumor data, may oversimplify the analysis and neglect vital details, as it draws conclusions based solely on partial data from mucinous carcinoma areas, overlooking the significance of the infiltrative carcinoma regions. In clinical practice, such simplification and neglect are unacceptable, as distinguishing between different types of cancer plays a pivotal role in selecting treatment strategies. Therefore, when utilizing AI models in the actual medical field, proceeding with utmost caution to ensure misinformation is not generated due to the model's limitations is imperative.

Furthermore, based on the responses from GPT-4.0, a discrepancy is noted between its description of breast self-examination and post-operative follow-up when compared to the current guidelines in China. The guidelines emphasize the significance of establishing a routine for self-examinations instead of strictly adhering to a 'once a month' directive. With regard to postoperative follow-up, GPT-4.0 recommends a check every 6 months for the first five years postsurgery. However, the guidelines propose more nuanced intervals—every 4–6 months during the initial 2 years and every 6 months thereafter for the subsequent 3 years. Crucially, such intervals should be personalized, with physicians tailoring follow-up schedules based on each patient's unique circumstances[29]. This emphasizes the necessity of avoiding a rigid, one-size-fits-all approach in medical practice.

There is an evident risk and limitation encapsulated in GPT-4.0's medical recommendations, particularly for patients shortly after surgery, and alcohol consumption may interact with post-operative medications, influencing wound healing and drug efficacy[30]. Moreover, Claude2's recommendations regarding postoperative functional exercise also deviate from actual clinical guidelines. Patients postbreast cancer surgery need to follow specific exercise and recovery plans. For instance, the guideline recommends engaging in active shoulder exercises 1–2 weeks following the operation, not immediately afterward[31]. Such details play a decisive role in postoperative recovery and should not be overlooked. This emphasizes the quality required when AI is applied in the medical domain.

In conclusion, LLM chatbots, despite their promising applications in breast cancer clinical scenarios, exhibit constraints tied to the recency and diversity of their training data, as well as potential biases, especially when it comes to providing medical advice. This is notably exemplified by even sophisticated models, such as GPT-4.0, which manifest limitations in certain contexts. Given the intricate and critical nature of breast cancer, offering treatment suggestions requires paramount precision and caution, necessitating AI models to adhere to stringent standards of accuracy and academic rigor when dispensing medical counseling on pivotal health issues.

Our research highlights the significant potential of LLMs like GPT-4, particularly within breast cancer clinical applications, thereby laying the groundwork for subsequent studies. A pivotal aspect to deliberate upon is the ongoing synchronization of the model's training and updates with the continuous advancements in medical research and specific regional practices[17]. Given the swift evolution and vast diversity of medical knowledge, periodically updating the model's training dataset is crucial.

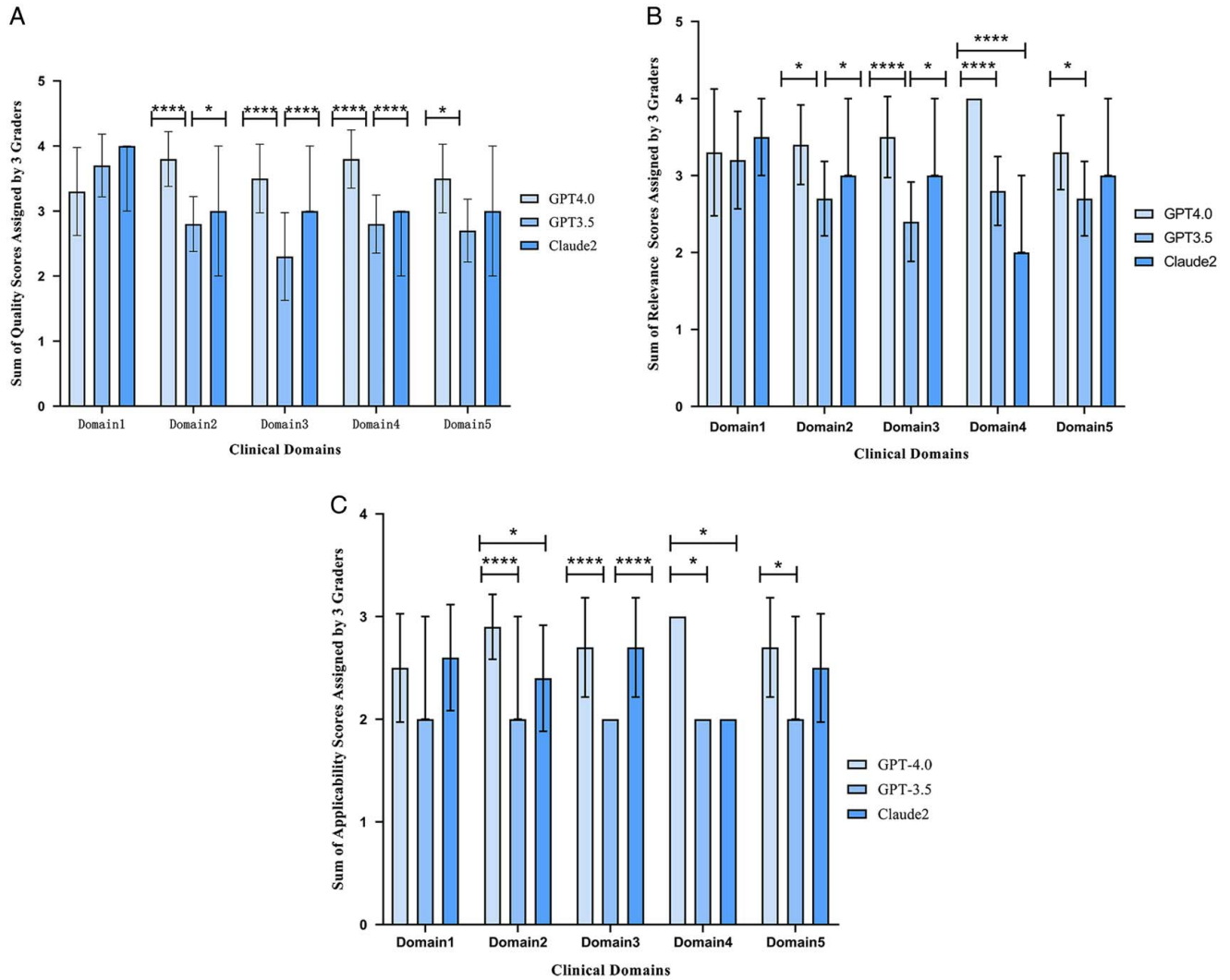

Figure 4. Performance of GPT-4.0, GPT-3.5, and Claude2 in the five major clinical areas. Notes: Domain1- Assessment and Diagnosis; Domain2- Treatment Decision-making; Domain3-Post-operative Care; Domain4- Psychosocial Support; Domain5- Prognosis and Rehabilitation (A) Quality scores of GPT-4.0, GPT-3.5, and Claude2 in the five clinical areas. Panel (B) Relevance scores of GPT-4.0, GPT-3.5, and Claude2 in the five clinical areas. (C) Applicability scores of GPT-4.0, GPT-3.5, and Claude2 in the five clinical areas.

Additionally, it is imperative to account for distinct medical practices across various global regions. Active partnerships with medical professionals and establishing robust feedback loops are essential. Input and suggestions from clinicians can fine-tune and enhance the model's outputs, solidifying its precision and applicability in practical scenarios. It is paramount to note that any medical advice provided by LLMs should be explicitly labeled as referential and cannot, under any circumstances, entirely replace the professional diagnosis and treatment recommendations of breast cancer experts[2]. This safeguard ensures that patients will not blindly follow the guidance of artificial intelligence but will engage in necessary consultations with professional doctors. Moving forward, using LLMs like GPT-4 necessitates a profound understanding of the specific circumstances across various regions. Avoiding a one-size-fits-all approach and ensuring technology adapts to the local context is crucial.

This study has several limitations. The research primarily focuses on assessing the utility of LLMs in the clinical application of breast cancer. However, the treatment and management of breast cancer is a complex and diverse field, and this study does not encompass all its multifaceted aspects. Future research, in pursuit of a more holistic evaluation of LLMs utility, may explore other breast diseases and lesions. Furthermore, while our comparative analysis among GPT-3.5, GPT-4.0, and Claude2 provides valuable insights, it is worth noting that feedback derived from a select group of five experts may not encapsulate the broader sentiment or consensus within the expansive medical fraternity. An assessment involving a wider and more varied spectrum of experts could offer a more comprehensive perspective on diverse clinical scenarios. Additionally, to enhance the accuracy and relevance of LLMs in clinical settings, integrating more clinical case studies and practice guidelines into the model training process is crucial. This approach will enable the models to provide more accurate diagnostic and treatment

recommendations, reflecting the complexities of real-world clinical practice. Finally, optimizing these models for multidisciplinary collaboration is key. Integrating insights from radiology, pathology, and oncology can lead to more comprehensive treatment recommendations. Such interdisciplinary integration is essential for supporting complex decision-making in medical practice, especially in breast cancer treatment.

In ensuring the safe and efficacious application of AI in the medical field, further in-depth research and evaluation are requisites. We harbor hope that future research will delve into a broader range of models and clinical applications, laying down a more robust theoretical and practical foundation for AI applications in the breast cancer domain and, by extension, the wider medical field, thereby advancing both the theoretical and practical aspects of AI deployment in medical scenarios.

Conclusion

The study highlights the remarkable advancements of LLMs like GPT-4.0 in clinical applications for breast cancer research. It reveals how GPT-4.0 significantly outperforms its predecessor, GPT-3.5, across four major areas and surpasses Claude2 in tasks involving psychosocial support and treatment decision-making. This underscores the growing potential of LLMs in the medical field. However, the application of these models in healthcare is not without challenges. Key issues include ensuring the recency and relevance of data, addressing inherent biases, and scrutinizing the sources of their training data. To fully leverage the benefits of LLMs in medicine, it is crucial to maintain stringent quality standards and regularly update the models. Our findings provide a solid foundation for further exploration and integration of LLMs in breast cancer management. This includes areas like prevention, diagnosis, and treatment. The continuous improvement and evaluation of these technologies are essential for their evolution into vital tools within the healthcare sector. Ensuring their effectiveness and reliability will be key in harnessing their full potential in improving patient outcomes.

Ethical approval

Approval from the ethics committee was not required since no patients were involved in our study.

Consent

This study solely focuses on the use and application of artificial intelligence and does not involve any direct human participants, patients, or volunteers. Consequently, there were no requirements for ethics committee approvals or informed consents in the context of this study. Furthermore, there are no individual details, images, or identifiers to report, ensuring full adherence to privacy standards.

Sources of funding

This study was supported by the Natural Science Foundation (Project Number: 72174183) for Research on the Construction and Mode of 5G+ 'Three Early Precautions' Health Management System.

Author contribution

L.D. and T.W.: writing – original draft preparation; Y.Z.: supervision; S.L.: data curation; Z.Z. and W.T.: conceptualization; J.L., Y.Z., J.X., and S.L.: writing – review and editing and data collection.

Conflicts of interest disclosure

The authors declare no conflicts of interest.

Research registration unique identifying number (UIN)

Our study is centered exclusively on artificial intelligence and does not involve any direct human participants. Given this nature, the requirements for registering the research study in the mentioned databases, as prescribed by the World Medical Association's Declaration of Helsinki 2013 (article 35), are not applicable to our research context. Consequently, there is no Unique Identifying Number (UIN) associated with this study.

Guarantor

Corresponding Author: Jinjiang Xu. E-mail: 15733941406@163.com.

Data availability statement

The data generated during this study are not publicly available due to privacy concerns. However, the data are available from the corresponding author upon reasonable request.

Provenance and peer review

Not commissioned, externally peer-reviewed.

References

[1] Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. Npj Digit Med 2023;6:120.
[2] Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. Nat Med 2023;29:1930–40.
[3] Mannstadt I, Mehta B. Large language models and the future of rheumatology: assessing impact and emerging opportunities. Curr Opin Rheumatol 2023;36:46–51.
[4] Ferdush J, Begum M, Hossain ST. ChatGPT and clinical decision support: scope, application, and limitations. Ann Biomed Eng 2023. doi.org/10.1007/s10439-023-03329-4
[5] Khan I, Agarwal R. Can ChatGPT help in the awareness of diabetes? Ann Biomed Eng 2023;51:2125–9.
[6] Arnold M, Morgan E, Rumgay H, *et al*. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast 2022;66:15–23.
[7] Walker HL, Ghani S, Kuemmerli C, *et al*. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023;25:e47479.
[8] Yeo YH, Samaan JS, Ng WH, *et al*. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29:721–32.
[9] Lozić Edisa, Štular Benjamin. ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing? (ver 23Q3 arXiv preprint arXiv:2309.08636, 2023) 2023.

[10] Mathew G, Agha R, Albrecht J, *et al*. STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. Int J Surg 2021;96:106165.
[11] Grünebaum A, Chervenak J, Pollet SL, *et al*. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol 2023;228: 696–705.
[12] Hristidis V, Ruggiano N, Brown EL, *et al*. ChatGPT vs google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res 2023;25:e48966.
[13] Jin JQ, Dobry AS. ChatGPT for healthcare providers and patients: practical implications within dermatology. J Am Acad Dermatol 2023;89:870–1.
[14] Marano L, Verre L, Carbone L, *et al*. Current trends in volume and surgical outcomes in gastric cancer. J Clin Med 2023;12:2708.
[15] Rao A, Kim J, Kamineni M, *et al*. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. J Am Coll Radiol 2023;20:990–7.
[16] Lim ZW, Pushpanathan K, Yew SME, *et al*. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. eBioMedicine 2023;95: 104770.
[17] Uprety D, Zhu D, West H (Jack. ChatGPT-a promising generative AI tool and its implications for cancer care. Cancer 2023;129:2284–9.
[18] George GA, Antony P. Correlation of fine needle aspiration cytology with histopathological diagnosis in assessing breast lumps at a tertiary care hospital. Int J Res Med Sci 2018;6:3738.
[19] Manfrin E, Falsirollo F, Remo A, *et al*. Cancer size, histotype, and cellular grade may limit the success of fine-needle aspiration cytology for screen-detected breast carcinoma. Cancer Cytopathol 2009;117:491–9.
[20] Nakano S, Otsuka M, Mibu A, *et al*. Significance of fine needle aspiration cytology and vacuum-assisted core needle biopsy for small breast lesions. Clin Breast Cancer 2015;15:e23–6.
[21] De Cursi JAT, Marques MEA, De Assis Cunha Castro CAC, *et al*. Fine-Needle Aspiration Cytology (FNAC) is a reliable diagnostic tool for small breast lesions ($\leq$ 1.0 cm): a 20-year retrospective study. Surg Exp Pathol 2020;3:29.
[22] Association CA-C. Breast cancer screening guideline for Chinese women. Cancer Biol Med 2019;16:822.
[23] Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2009;151:716.
[24] Gøtzsche PC, Jørgensen KJ. Screening for breast cancer with mammography. Cochrane Db Syst Rev 2013;2013:CD001877.
[25] Gradishar WJ, Moran MS, Abraham J, *et al*. NCCN Guidelines® Insights: Breast Cancer, Version 4.2023. J Natl Compr Cancer Netw 2023;21:594–608.
[26] Oeffinger KC, Fontham ETH, Etzioni R, *et al*. Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. JAMA 2015;314:1599.
[27] Shen S, Zhou Y, Xu Y, *et al*. A multi-centre randomised trial comparing ultrasound vs mammography for screening breast cancer in high-risk Chinese women. Br J Cancer 2015;112:998–1004.
[28] Hao S, Li M, Fan S, *et al*. An analysis of the status of diagnostic radiology equipment in China. Radiat Med Prot 2023;4:170–5.
[29] Breast Cancer Expert Committee of National Cancer Quality Control Center, Breast Cancer Expert Committee of China Anti-Cancer Association, Cancer Drug Clinical Research Committee of China Anti-Cancer Association. [Guidelines for clinical diagnosis and treatment of advanced breast cancer in China (2022 edition)]. Zhonghua Zhong Liu Za Zhi 2022;44:1262–87.
[30] Lavernia CJ, Villa JM, Contreras JS. Alcohol use in elective total hip arthroplasty: risk or benefit? Clin Orthop Relat Res 2013;471:504–9.
[31] Ting DXHYD. Evaluation of evidence-based resources for early post-operative functional exercise in patients with Breast Cancer. Chinese Gen Pract 2018;21:4011.