

# **Title: Clinical Application of Large Language Models for Breast Conditions: A Systematic Review**

**Authors:** Billy Ho Hung CHEUNG, Karen Gwynn POON , Cheuk Fai LAI, Ka Chun LAM , Michael CO, Ava KWONG

**Authors affiliation:** Division of Breast Surgery, Department of Surgery, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Queen Mary Hospital, Hong Kong SAR, China

## **Corresponding author:**

Professor Ava Kwong

Daniel C K Yu Professor in Breast Cancer Research

Clinical Professor

Chief of Division of Breast Surgery

Department of Surgery, Li Ka Shing Faculty of Medicine, The University of Hong Kong,

Queen Mary Hospital, Hong Kong SAR, China

(Tel: +852 2255 4773; Email address: [avakwong@hku.hk](mailto:avakwong@hku.hk))

## **Acknowledgement:**

Ethical approval: ethical approval not required for the study

Assistance with the study: none

Financial support and sponsorship: none

Conflicts of interest: none

Presentation: none

**NOTE:** This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

## Abstract:

**Background;** The application of artificial intelligence (AI) like Large Language Models (LLM) into the healthcare system has been a frequently discussed topic in recent years. **Materials and Methods;** We conducted a systemic review on primary studies about the applications of LLM in breast conditions. The studies are then categorized into their respective domains, namely diagnosis, management recommendations and communication for patients. **Results;** The diagnostic accuracy ranged from 74.3% to 99.6% across different investigation modalities. The concordance of management recommendations ranged from 50% to 70% while the prognostic evaluation of breast cancer patients of distant recurrence showed an accuracy of 75% to 88%. In regards to patient communication, it is revealed that 18-30% of the references used by the LLM were irrelevant. **Conclusion;** This study highlights the potential benefits of LLM in strengthening patient communication, diagnose and management of patients with breast conditions. With standardized protocol and guideline to minimize potential risks, LLM can be a valuable tool to support future clinicians in the field of breast management.

## 1. Introduction

Artificial intelligence (AI) has shown a promising future for enhancing various aspects of healthcare, from diagnosis to treatment planning to patient education. [1] Large language models (LLMs), a type of AI system that can interpret and generate human language, have gained particular interest for their potential in assisting with clinical tasks which involve unstructured text data. [2] LLMs, like ChatGPT, have demonstrated impressive performance on general knowledge and language understanding benchmarks. [3] However, their utility and limitations for specific clinical applications remain to be established.

Breast cancer is one of the most common malignancies worldwide with an annual incidence of 2.3 million.[4] Breast conditions-related clinical visits make up around three percent of primary health care visits among women. [5] They range from benign diseases and breast cancer with a wide variety of management. However, due to the overloading number of patients, the waiting time for patients' clinical assessment can be significantly prolonged. [6] Moreover, the clinical evaluation for breast related diseases, which involves triple assessment and personalised approach, can be strained due to time restraints and limited resources. [6]

Hence, it is our belief that the management of patients with breast conditions can be an important area where AI like LLMs could potentially aid clinicians and patients. Accurate interpretation of imaging and pathology reports is crucial for breast diagnosis,[7] while personalized management requires complex decision-making based on multiple clinical factors. Effective communication of information to patients is also key for shared decision-making and treatment compliance. [8] Prior studies have explored applications of narrow AI systems for breast imaging interpretation and clinical decision support. [9, 10] However, the use of more flexible and comprehensive LLMs for breast conditions has not been systematically examined.

The objective of this systematic review is to evaluate the current evidence on the application of LLMs in breast conditions. We further categorise them into diagnosis, management, and patient communication. Specifically, we aim to address three key questions: 1) What is the diagnostic accuracy of LLMs for interpreting breast imaging and pathology reports? 2) How well do LLM management recommendations align with those of multidisciplinary teams? 3) What is the quality and limitations of LLM-generated information for patients on breast health topics? By critically appraising and synthesizing the available studies, we seek to identify promising use cases, current gaps, and future directions for the development and validation of LLMs in breast care.

## **2. Materials and Methods**

We conducted a systematic search of PubMed, SCOPUS, and Google Scholar databases for studies published up to December 31, 2023 according to the systematic review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. [11] The search query included terms related to large language models “ChatGPT” OR “BARD” OR “LLAMA” OR “Large language model” OR “LAMDA” OR “GPT” OR “GPT” OR “natural language model” OR “natural language process\*”) in combination with “breast”.

We included only primary studies that assessed the performance or application of LLMs for any of the following in relation to breast conditions. These studies were then categorised according to the predefined domains, including diagnosis, management recommendations, and communication for patients. Commentaries, editorials, and reviews were excluded. Non-LLM

based narrow AI systems were also excluded. Two reviewers independently screened titles and abstracts, followed by full-text review of potentially eligible studies. Disagreements were resolved by consensus or consultation with a third reviewer.

### **3. Theory/Calculation**

For each included study, we extracted information on the study design, data sources, the LLM studied, comparison method, sample size, and main outcomes. For diagnostic studies, we extracted information regarding the metrics of accuracy, such as sensitivity, specificity, and area under the receiver operating curve (AUC). For management studies, we extracted concordance rates between LLM and human expert recommendations. For patient communication studies, we summarized the quality assessments and any limitations that are identified.

We synthesized results narratively and in tables, stratified by application area (diagnosis, management, patient communication). Meta-analysis was not proceeded due to the limited data.

## **4. Results**

### **4.1. Study Characteristics**

The search yielded 173 records, of which 17 studies met inclusion criteria (Figure 1). Eight studies evaluated LLMs for breast diagnosis,[12-19] three for management,[20-22] two for prognosis [23, 24] and four for patient communication. [25-28] Two studies assessed multiple applications. [26, 27] The most commonly used LLM was ChatGPT (n=6), followed by custom models based on BERT or GPT architectures. Most studies were conducted in the USA or Europe.

### **4.2. Diagnostic Accuracy**

Seven studies use developed AI models to diagnose patients with radiological free text reports, ranging between 300 to 79312 reports from each study. The accuracy of LLMs for

classifying breast conditions based on breast imaging reports (mammograms, ultrasound, MRI) and pathology reports (Table 1). Accuracies ranged from 74.3% to 99.6% across modalities, with the highest precision for MRI-based lesion detection (99.6%) and lowest for BI-RADS category assignment on mammogram-ultrasound (74.3%). [17] Two studies found that LLM accuracies were comparable to manual interpretation by radiologists. [13, 15]

Only one study evaluated a publicly-available LLM (ChatGPT) as an aid for radiologists, finding acceptable accuracy (83%) but noting limitations in currency of knowledge and potential for incorrect or fraudulent output. [18] Other studies used custom-developed LLMs based on transformer architectures and trained on site-specific imaging reports. Common challenges included extraction of granular radiographic features, assignment to proper BI-RADS categories, and integration into clinical workflows.

### **4.3. Management Recommendations and Prognosis**

Three studies compared management recommendations of LLMs to those of multidisciplinary tumor boards (MDTs) for breast cancer cases, with population sizes ranging from 10 to 20 patients (Table 2). Overall, the concordance ranged from 50% (for cases including benign/precancerous lesions) to 70% (for invasive cancers only). [21, 22] Concordance rate was highest for recommendations on radiotherapy (95%), followed by chemotherapy (94.7%), endocrine therapy (75%), genetic testing (70%). [20] Discordant recommendations were often due to LLMs not considering patient age, performance status, or specific tumor features. In addition, LLMs sometimes proposed fraudulent decisions, such as recommending extensive genetic testing for patients without positive family history of breast cancer and advising against re-excision for cases with positive surgical resection margins. Over-treatment with chemotherapy and under-treatment with surgery were also observed. The authors noted that LLMs currently have limited suitability as a decision aid for breast cancer management due to medicolegal issues and potential for harm from erroneous recommendations.

Furthermore, two studies demonstrated the application of LLM-based predictive models on real breast cancer patients for prognostic evaluation, which showed an accuracy of 75% to 88% in the prediction of distant breast cancer recurrence. [23, 24]

#### **4.4. Patient Communication**

Four studies assessed the quality of information produced by LLMs in response to common patient questions about breast health topics (Table 3). All used ChatGPT, either version 3.5 or 4. Number of questions included in each study varies, ranging from 6 to 25, each study recruited 3-5 experienced breast surgeons/ specialists to comment on the answers generated by LLMs. Across studies, clinician quality ratings of LLM-generated content averaged 4.2 on a 5-point scale, indicating high overall quality. LLMs were noted to provide comprehensive, technically accurate, and well-organized answers, generally outperforming search engines. [26, 28]

However, several important limitations were identified. Across studies, 18-30% of references cited by ChatGPT were inaccessible, irrelevant, or fabricated, raising concerns about information reliability. [25-27] ChatGPT sometimes gave outdated information, such as not reflecting the latest FDA regulations on breast implants. [28] It also tended to provide generic rather than patient-specific recommendations. [28] Additionally, ChatGPT responses were found to vary based on prompt wordings, occasionally introducing inaccurate or biased content [25]. Google searches were noted to provide more reliable and up-to-date references, especially for controversial topics, namely breast implant illness. [26, 27]

#### **5. Discussion**

This systematic review shows that LLMs such as ChatGPT have the potential in making clinical diagnosis, formulating individualised management plans, generating prognostic prediction and acting as a patient communication tool for health enquiries. However, significant limitations and potential risks were also identified, highlighting the need for further validation and refinement before clinical deployment.

In the diagnostic domain, LLMs achieved high accuracy for extracting key findings and assigning BI-RADS categories across mammography, ultrasound, MRI, and pathology reports. The performance was often comparable to expert human reviewers, supporting the potential for LLMs to enhance the efficiency and consistency of report interpretation. However, most studies used custom-developed LLMs trained on site-specific data, thus limiting generalizability. One study that evaluates a general-purpose LLM (ChatGPT), found lower

sensitivity when compared to usual care for screening mammograms. [18] This underscores the importance of training LLMs on large, diverse datasets reflective of real-world practice. Future work should also establish benchmarks and best practices for assessing LLM performance, as metrics varied widely across studies.

For management, LLMs showed moderate concordance (50-70%) with multidisciplinary tumor board decisions for breast cancer cases. Agreement was highest for recommending standard systemic therapies but lower for nuanced surgical and genetic testing decisions. Importantly, LLMs sometimes gave dangerous recommendations, such as advising against re-excision for positive margins or recommending excessive genetic testing.[20] Over- and under-treatment were also observed relative to guideline-concordant care. These findings caution against using LLMs as autonomous decision-making tools, as they lack the clinical judgment to weigh competing factors and may not reflect the latest evidence. Instead, LLMs may be best positioned as aids to prompt consideration of management options, with final decisions made by human experts. Rigorous testing in prospective studies and refinement on large, curated oncology datasets will be necessary before considering deployment.

In the realm of patient communication, LLMs like ChatGPT generated coherent, accurate, and actionable answers to common breast health questions. The information was rated highly by clinician reviewers and often outperformed search engine results for comprehensiveness and organization. This supports the potential for LLM-powered chatbots or question-answering systems to enhance patient education and engagement. However, limitations were noted regarding outdated content, fabricated references, and inconsistency across similar prompts. There were also concerns about the inability to personalize recommendations and the potential for perpetuating bias. Careful human curation and oversight will be essential to ensure reliability and transparency of LLM-generated patient materials. [25, 26]

More broadly, LLMs pose several challenges that must be addressed before widespread clinical implementation. First of all, the "black box" nature of LLMs, where the reasoning behind outputs is opaque to users. [29] This lack of explainability hinders the ability to audit decisions and identify errors. Techniques to improve LLM interpretability, such as extracting rules or decision trees, are an active area of research. [30] Another challenge is the potential for LLMs to perpetuate biases present in training data, such as over- or under-diagnosis in certain demographics. [31] Careful auditing and debiasing of datasets and models will be required to

ensure equitable performance. Data privacy and consent in the use of patient information for model development should also be considered on ethical grounds. [32]

Importantly, LLMs are not static tools, they are evolving technologies undergoing rapid development and refinement. The most recent versions, such as ChatGPT-4, have shown improved performance and safety compared to earlier iterations. Limitations identified in the studies to date may be addressed in future model updates. Additionally, ongoing efforts to build domain-specific LLMs for oncology, such as OncoGPT, [33] may yield stronger results than general-purpose LLMs. However, the fast pace of development also challenges the ability to rigorously validate models before deployment. Prospective studies with standardized evaluation frameworks and reporting will be crucial to establish the utility and safety of LLMs in real-world clinical settings.

This review has several strengths and limitations. We performed a comprehensive search across multiple databases to capture the most recent studies on a rapidly evolving technology. However, relevant studies may have been missed due to inconsistent terminology and lack of established MeSH terms for LLMs. The included studies were highly heterogeneous in LLM type, training data, clinical application, and evaluation metrics. Assessment of study quality and risk of bias was also challenging due to the lack of established tools suited for AI-focused studies. Publication bias is likely present, as studies with positive results may be more likely to be published than negative ones.

## **5. Conclusion**

In conclusion, large language models demonstrate promising potential for enhancing diagnosis, management, and patient communication for breast conditions. Across 17 studies, LLMs achieved high diagnostic accuracy, moderate management concordance, and generated high-quality patient information. However, significant risks and limitations were identified, such as biased or fraudulent recommendations, lack of transparency, and insufficient validation in prospective settings. Future work should establish standardized evaluation frameworks and reporting guidelines for LLM studies, curate large diverse training datasets reflective of practice, and incorporate human oversight to mitigate potential harms. With responsible development and validation, LLMs may become powerful tools to support clinicians and patients in the future of breast care.



## 6. Reference list

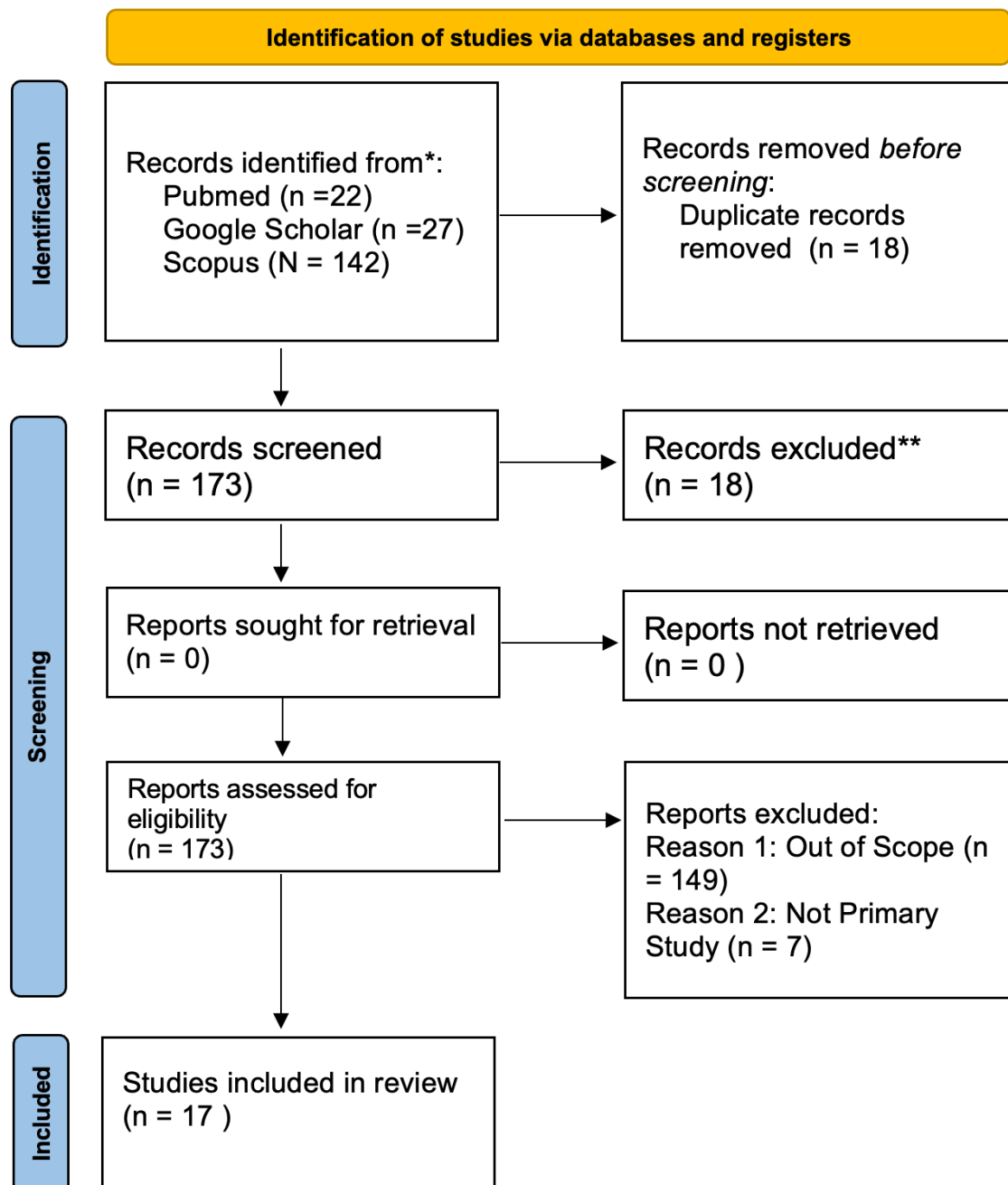
1. Topol, EJ, *High-performance medicine: the convergence of human and artificial intelligence*. Nature medicine, 2019. **25**(1): p. 44-56.
2. Katikapalli Subramanyam, K, Rajasekharan, A, and Sivanesan, S, *AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing*. arXiv.org, 2021.
3. Brown, TB, Mann, B, Ryder, N, et al., *Language Models are Few-Shot Learners*. arXiv.org, 2020.
4. Organization, WH. *Breast Cancer*. 2024 [cited 2024 20 Dec]; Available from: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
5. Eberl, MM, Phillips, RL, Jr., Lamberts, H, Okkes, I, and Mahoney, MC, *Characterizing breast symptoms in family practice*. Ann Fam Med, 2008. **6**(6): p. 528-33.
6. Mutebi, M, Anderson, BO, Duggan, C, et al., *Breast cancer treatment: A phased approach to implementation*. Cancer, 2020. **126 Suppl 10**: p. 2365-2378.
7. Ray, KMMD, Hayward, JHMD, and Joe, BNMDP, *Role of MR Imaging for the Locoregional Staging of Breast Cancer*. Magnetic resonance imaging clinics of North America, 2018. **26**(2): p. 191-205.
8. Ha, JF and Longnecker, N, *Doctor-patient communication: a review*. The Ochsner journal, 2010. **10**(1): p. 38-43.
9. McKinney, SM, Sieniek, M, Godbole, V, et al., *International evaluation of an AI system for breast cancer screening*. Nature, 2020. **577**(7788): p. 89-94.
10. Rodriguez-Ruiz, A, Lång, K, Gubern-Merida, A, et al., *Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists*. J Natl Cancer Inst, 2019. **111**(9): p. 916-922.
11. Page, MJ, McKenzie, JE, Bossuyt, PM, et al., *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. BMJ, 2021. **372**: p. n71.
12. Bozkurt, S, Gimenez, F, Burnside, ES, Gulkesen, KH, and Rubin, DL, *Using automatically extracted information from mammography reports for decision-support*. J Biomed Inform, 2016. **62**: p. 224-31.
13. Bozkurt, S, Lipson, JA, Senol, U, and Rubin, DL, *Automatic abstraction of imaging observations with their characteristics from mammography reports*. J Am Med Inform Assoc, 2015. **22**(e1): p. e81-92.
14. Choi, HS, Song, JY, Shin, KH, Chang, JH, and Jang, BS, *Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer*. Radiat Oncol J, 2023. **41**(3): p. 209-216.
15. Liu, Y, Liu, Q, Han, C, Zhang, X, and Wang, X, *The implementation of natural language processing to extract index lesions from breast magnetic resonance imaging reports*. BMC Med Inform Decis Mak, 2019. **19**(1): p. 288.
16. Liu, Y, Zhu, L-N, Liu, Q, Han, C, Zhang, X-D, and Wang, X-Y, *Automatic extraction of imaging observation and assessment categories from breast magnetic resonance imaging reports with natural language processing*. Chinese medical journal, 2019. **132**(14): p. 1673-1680.

17. López-Úbeda, P, Martín-Noguerol, T, and Luna, A, *Automatic classification and prioritisation of actionable BI-RADS categories using natural language processing models*. Clin Radiol, 2024. **79**(1): p. e1-e7.
18. Rao, A, Kim, J, Kaminen, 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**(10): p. 990-997.
19. Nandish, S, R.J, P, and N.M, N, *Natural Language Processing Approaches for Automated Multilevel and Multiclass Classification of Breast Lesions on Free-Text Cytopathology Reports*. JCO clinical cancer informatics, 2022. **6**(6): p. e2200036-e2200036.
20. Griewing, S, Gremke, N, Wagner, U, Lingenfelder, M, Kuhn, S, and Boekhoff, J, *Challenging ChatGPT 3.5 in Senology-An Assessment of Concordance with Breast Cancer Tumor Board Decision Making*. J Pers Med, 2023. **13**(10).
21. Lukac, S, Dayan, D, Fink, V, et al., *Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases*. Arch Gynecol Obstet, 2023. **308**(6): p. 1831-1844.
22. Sorin, V, Klang, E, Sklair-Levy, M, et al., *Large language model (ChatGPT) as a support tool for breast tumor board*. NPJ breast cancer, 2023. **9**(1): p. 44-44.
23. Ribelles, N, Jerez, JM, Rodriguez-Brazzarola, P, et al., *Machine learning and natural language processing (NLP) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (HR+)/HER2-negative advanced breast cancer patients*. Eur J Cancer, 2021. **144**: p. 224-231.
24. Wang, H, Li, Y, Khan, SA, and Luo, Y, *Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network*. Artif Intell Med, 2020. **110**: p. 101977.
25. Haver, HL, Ambinder, EB, Bahl, M, Oluyemi, ET, Jeudy, J, and Yi, PH, *Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT*. Radiology, 2023. **307**(4): p. e230424.
26. Liu, HY, Alessandri Bonetti, M, De Lorenzi, F, Gimbel, ML, Nguyen, VT, and Egro, FM, *Consulting the Digital Doctor: Google Versus ChatGPT as Sources of Information on Breast Implant-Associated Anaplastic Large Cell Lymphoma and Breast Implant Illness*. Aesthetic Plast Surg, 2024. **48**(4): p. 590-607.
27. Liu, HY, Alessandri Bonetti, M, Jeong, T, Pandya, S, Nguyen, VT, and Egro, FM, *Dr. ChatGPT will see you now: How do Google and ChatGPT compare in answering patient questions on breast reconstruction?* J Plast Reconstr Aesthet Surg, 2023. **85**: p. 488-497.
28. Seth, I, Cox, A, Xie, Y, et al., *Evaluating Chatbot Efficacy for Answering Frequently Asked Questions in Plastic Surgery: A ChatGPT Case Study Focused on Breast Augmentation*. Aesthet Surg J, 2023. **43**(10): p. 1126-1135.
29. Rudin, C, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. Nature machine intelligence, 2019. **1**(5): p. 206-215.
30. Lage, I, Chen, E, He, J, et al., *An Evaluation of the Human-Interpretability of Explanation*. arXiv.org, 2019.
31. Obermeyer, Z, Powers, B, Vogeli, C, and Mullainathan, S, *Dissecting racial bias in an algorithm used to manage the health of populations*. Science, 2019. **366**(6464): p. 447-453.
32. Price, WN, 2nd and Cohen, IG, *Privacy in the age of medical big data*. Nat Med, 2019. **25**(1): p. 37-43.

33. Jia, F, Liu, X, Deng, L, et al., *OncoGPT: A Medical Conversational Model Tailored with Oncology Domain Expertise on a Large Language Model Meta-AI (LLaMA)*. arXiv.org, 2024.

**Figures and Tables:**

**Figure 1:** Systematic review using the PRISMA guideline



**Table 1.** Summary table for diagnostic report usage.

<b>Imaging modalities</b>	<b>No. of papers</b>
Mammogram	2
Ultrasound and Mammogram	1
MRI	2
Pathological Report	1
Ultrasound and Pathological Report	1

**Table 2. Summary table on management and prognosis**

<b>Management</b>		
Study	Patient characteristics	Concordance rate (%)
Sorin <i>et al.</i> , 2023 [22]	10 patients (8 IDC, 1 DCIS, 1 phyllodes)	70
Lukac <i>et al.</i> , 2023 [21]	10 patients (10 IDC)	64.2
Griewing <i>et al.</i> , 2023 [20]	20 patients (17 IDC, 3 DCIS)	50 (IDC & precancerous lesions) 58 (IDC only)
<b>Prognosis</b>		
Study	Predictive model accuracy (%)	
Ribelles <i>et al.</i> , 2021 [23]	75.8	

Wang <i>et al.</i> , 2020 [24]	88.8
--------------------------------	------

IDC, invasive ductal carcinoma. DCIS, ductal carcinoma-in-situ.

**Table 3. Summary table on patient communication**

<b>Content involved</b>	<b>No. of papers</b>
Breast screening and prevention, and implication of BIRADS grading	1
Breast reconstruction and augmentation	2
Breast implants related complications	1

