Diagn Interv Radiol 2024; DOI: 10.4274/dir.2024.242876

BREAST IMAGING

ORIGINAL ARTICLE

Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition

Yasin Celal Güneş[1]
Turay Cesur[2]
Eren Çamur[3]
Leman Günbey Karabekmez[4]

[1]Kırıkkale Yüksek İhtisas Hospital, Clinic of Radiology, Kırıkkale, Türkiye

[2]Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

[3]Ankara 29 Mayıs State Hospital, Clinic of Radiology, Ankara, Türkiye

[4]Ankara Yıldırım Beyazıt University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

PURPOSE

This study aimed to evaluate the performance of large language models (LLMs) and multimodal LLMs in interpreting the Breast Imaging Reporting and Data System (BI-RADS) categories and providing clinical management recommendations for breast radiology in text-based and visual questions.

METHODS

This cross-sectional observational study involved two steps. In the first step, we compared ten LLMs (namely ChatGPT 4o, ChatGPT 4, ChatGPT 3.5, Google Gemini 1.5 Pro, Google Gemini 1.0, Microsoft Copilot, Perplexity, Claude 3.5 Sonnet, Claude 3 Opus, and Claude 3 Opus 200K), general radiologists, and a breast radiologist using 100 text-based multiple-choice questions (MCQs) related to the BI-RADS Atlas 5th edition. In the second step, we assessed the performance of five multimodal LLMs (ChatGPT 4o, ChatGPT 4V, Claude 3.5 Sonnet, Claude 3 Opus, and Google Gemini 1.5 Pro) in assigning BI-RADS categories and providing clinical management recommendations on 100 breast ultrasound images. The comparison of correct answers and accuracy by question types was analyzed using McNemar's and chi-squared tests. Management scores were analyzed using the Kruskal–Wallis and Wilcoxon tests.

RESULTS

Claude 3.5 Sonnet achieved the highest accuracy in text-based MCQs (90%), followed by ChatGPT 4o (89%), outperforming all other LLMs and general radiologists (78% and 76%) ($P < 0.05$), except for the Claude 3 Opus models and the breast radiologist (82%) ($P > 0.05$). Lower-performing LLMs included Google Gemini 1.0 (61%) and ChatGPT 3.5 (60%). Performance across different categories of showed no significant variation among LLMs or radiologists ($P > 0.05$). For breast ultrasound images, Claude 3.5 Sonnet achieved 59% accuracy, significantly higher than other multimodal LLMs ($P < 0.05$). Management recommendations were evaluated using a 3-point Likert scale, with Claude 3.5 Sonnet scoring the highest (mean: 2.12 ± 0.97) ($P < 0.05$). Accuracy varied significantly across BI-RADS categories, except Claude 3 Opus ($P < 0.05$). Gemini 1.5 Pro failed to answer any BI-RADS 5 questions correctly. Similarly, ChatGPT 4V failed to answer any BI-RADS 1 questions correctly, making them the least accurate in these categories ($P < 0.05$).

CONCLUSION

Although LLMs such as Claude 3.5 Sonnet and ChatGPT 4o show promise in text-based BI-RADS assessments, their limitations in visual diagnostics suggest they should be used cautiously and under radiologists' supervision to avoid misdiagnoses.

CLINICAL SIGNIFICANCE

This study demonstrates that while LLMs exhibit strong capabilities in text-based BI-RADS assessments, their visual diagnostic abilities are currently limited, necessitating further development and cautious application in clinical practice.

KEYWORDS

BI-RADS, breast radiology, ChatGPT 4o, Claude 3.5 Sonnet, large language models

Corresponding author: Yasin Celal Güneş

E-mail: gunesyasincelal@gmail.com

Received 27 May 2024; revision requested 23 June 2024; last revision received 07 August 2024; accepted 24 August 2024.

Epub: 09.09.2024

Publication date:

DOI: 10.4274/dir.2024.242876

You may cite this article as: Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition. *Diagn Interv Radiol.* 09 September 2024 DOI: 10.4274/dir.2024.242876 [Epub Ahead of Print].

The emergence of large language models (LLMs) marks a transformative milestone in the development of artificial intelligence (AI). These models offer unprecedented potential for understanding and generating human-like text by leveraging extensive datasets. This technological advancement holds significant promise for application in medicine.[1,2] As radiology increasingly relies on the interpretation of complex imaging data, the integration of advanced AI tools, such as LLMs, becomes crucial to enhance diagnostic accuracy and streamline workflows. LLMs have demonstrated remarkable performance in various realms of radiology, including testing radiological knowledge in different board-style examinations, simplifying radiology reports, and providing patient information.[3-7]

Recent studies have also explored the potential of LLMs specifically in breast imaging, where their capabilities show particular promise.[8-10] For instance, Rao et al.[9] evaluated the performance of two well-known LLMs, ChatGPT 3.5 and ChatGPT 4, in adhering to the American College of Radiology (ACR) eligibility criteria for breast pain and breast cancer screening, revealing impressive accuracy rates of 88.9% and 98.4%, respectively. These findings highlight the potential of LLMs as supportive tools in breast imaging, which is especially relevant given the ongoing radiologist shortages and the increasing volume of imaging studies.[11,12] Despite these advancements, it is crucial to acknowledge the limitations and challenges associated with LLMs, including their susceptibility to generating plausible-sounding but incorrect answers (hallucinations).[13]

The Breast Imaging Reporting and Data System (BI-RADS) Atlas, released in its latest edition in 2013, has provided standardized nomenclature, report organization, assessment structure, and a classification system for mammography, ultrasound, and magnetic resonance imaging (MRI) of the breast.[14] The BI-RADS Atlas is crucial for radiologists as it standardizes breast imaging terminology and reporting, ensuring clear communication and consistent, accurate patient management.[15]

While the BI-RADS Atlas offers a standardized approach to breast imaging, recent research has begun exploring how LLMs can further enhance radiological assessment and reporting accuracy. Haver et al.[16] demonstrated that ChatGPT 4 accurately predicted the BI-RADS category in 73.6% of 250 fictitious breast imaging reports. Cozzi et al.[17] evaluated the concordance between different LLMs (ChatGPT 3.5, ChatGPT 4, and Google Bard) and radiologists across 2,400 reports in three different languages, revealing a moderate agreement (Gwet's agreement coefficient: 0.52–0.42). Despite the growing emphasis on the importance of LLMs in breast imaging, there is a significant gap in the literature regarding the evaluation of multimodal LLMs' performance on breast ultrasound images. Additionally, no studies compare LLMs' knowledge of BI-RADS Atlas with that of radiologists. Hence, the first aim of this study is to evaluate the performance of nine large LLMs compared with breast and general radiologists on text-based multiple-choice questions (MCQs) related to the BI-RADS Atlas, 5th edition. The second aim is to assess the capability of five multimodal LLMs in assigning BI-RADS categories and providing clinical management recommendations for breast ultrasound images.

Main points

- This study evaluated the performance of large language models (LLMs) and multimodal LLMs in interpreting the Breast Imaging Reporting and Data System categories and providing clinical management recommendations. The evaluation involved two steps: assessing LLMs on text-based multiple-choice questions (MCQs) and evaluating multimodal LLMs on breast ultrasound images.
- Claude 3.5 Sonnet and ChatGPT 4o achieved high accuracy rates of 90% and 89%, respectively, in text-based MCQs, outperforming general radiologists, who had accuracy rates of 78% and 76%. This demonstrates the strong potential of these advanced LLMs in supporting and enhancing the diagnostic accuracy of radiologists in text-based assessments.
- Multimodal LLMs showed lower accuracy in evaluating breast ultrasound images, with Claude 3.5 Sonnet achieving only 59% accuracy. This highlights a critical limitation in their current ability to handle visual diagnostic tasks effectively compared with text-based assessments.
- The study underscores the necessity for further development of multimodal LLMs to improve their visual diagnostic capabilities. Until these improvements are realized, the use of multimodal LLMs in clinical practice should be closely supervised by experienced radiologists to prevent potential misdiagnoses and ensure patient safety.

Methods

Study design

This cross-sectional observational study had two steps. In the first step, it compared different LLMs, namely ChatGPT 4o, ChatGPT 4, ChatGPT 3.5, Google Gemini 1.5 Pro, Google Gemini 1.0, Microsoft Copilot, Perplexity, Claude 3.5 Sonnet, Claude 3 Opus, and Claude 3 Opus 200K, along with the responses of two general radiologists and a breast radiologist in answering MCQs regarding the 5th edition of the BI-RADS Atlas.

In the second step, the study compared different multimodal LLMs, namely ChatGPT 4o, ChatGPT 4V, Claude 3.5 Sonnet, Claude 3 Opus, and Google Gemini 1.5 Pro. This step focused on determining the correct BI-RADS category and clinical management by evaluating breast ultrasound images. An overview of the workflow is shown in Figure 1.

The study did not require ethics committee approval as it relied solely on fictional MCQs and a publicly available breast ultrasound dataset that had no identifiable patient information. Its design conformed to the principles articulated in the Standards for Reporting Diagnostic Accuracy Studies statement.[18]

Data collection for breast multiple-choice questions

The ACR published the 5th edition of the BI-RADS Atlas in 2013 to standardize terminology and reporting organization in breast radiology.[14] A total of 100 MCQs were prepared and categorized using the information in this atlas related to ultrasound, mammography, MRI, and general BI-RADS knowledge by general radiologist 3 (Y.C.G.). Each question had four choices, with only one correct answer and three distractors. The distractors were carefully chosen to be reasonable and related to the question. Each question was formulated to be clear and focused on a single concept to assess breast radiology knowledge. The questions were categorized according to the BI-RADS Atlas sections as follows: 16 on breast ultrasound, 39 on mammography, 22 on breast MRI, and 23 on general BI-RADS knowledge. All created MCQs are listed in Supplementary Material 1.

Design of input–output procedures and performance evaluation for large language models

The input prompt was initiated as follows: "I am working on a breast radiology quiz and will provide you MCQs. Act like a radiology professor with 30 years of expertise in breast imaging. Please indicate the correct answer. There is only one correct answer." This prompt was presented in April 2024 on eight distinct platforms with default param-

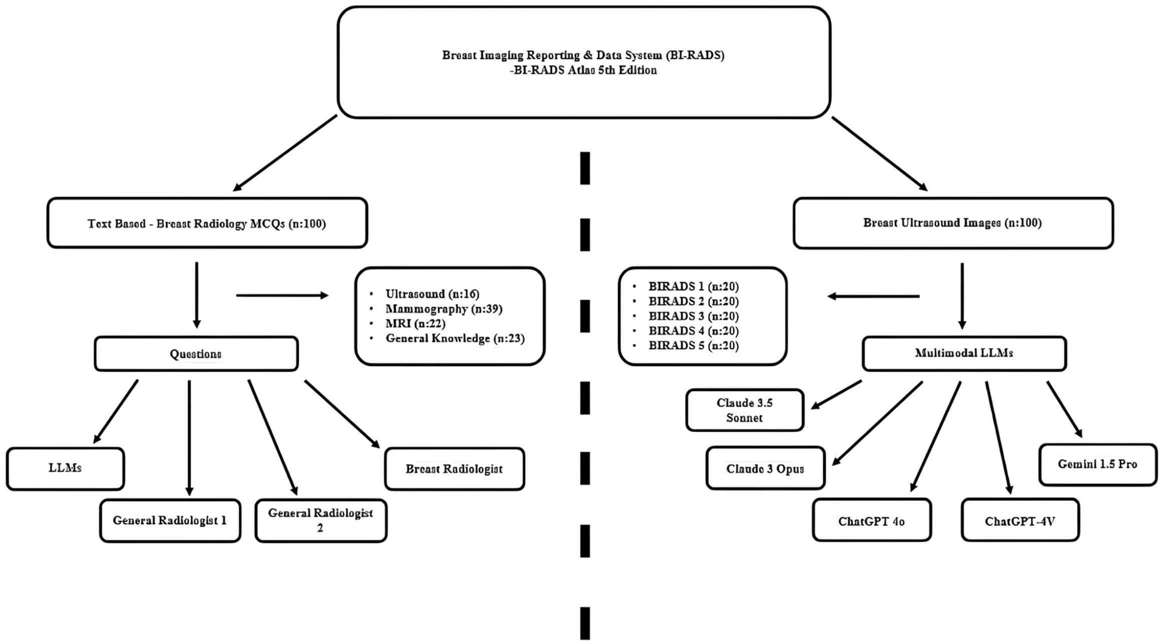

Figure 1. The workflow of the study. MCQs, multiple-choice questions; LLMs, large language models; MRI, magnetic resonance imaging.

eters: OpenAI's ChatGPT 4 and 3.5 (https://chat.openai.com), Google Gemini 1.5 Pro and 1.0 (https://gemini.google.com/), Microsoft Copilot (https://copilot.microsoft.com) (Balanced), Perplexity (https://perplexity.ai), Claude 3 Opus (https://claude.ai), and Claude 3 Opus 200K (https://poe.com). The same prompt was presented to OpenAI's ChatGPT 4o (https://chat.openai.com) in May 2024 and Claude 3.5 Sonnet (https://claude.ai) in July 2024 (Figure 2). Specific settings, such as temperature and randomness, were left at their default values unless specified otherwise by the platform.

The MCQs were sequentially added to the same chat session by copying and pasting from the MCQs list. Each LLM was presented with 100 questions by general radiologist 3, and the responses were evaluated. It is crucial to note that the employed LLMs were not pre-trained with a specific prompt or question set for this study. Each question was asked in a single chat session, without opening a new chat tab for individual inquiries.

Radiologist 3 evaluated LLMs' answers according to the correct answer list, marking them either correct (1) or incorrect (0).

Radiologists performance evaluation for breast multiple-choice questions

Two European Board of Radiology-certified junior general radiologists–radiologist 1 (T.C.) with 6 years of experience, and radiologist 2 (E.Ç.) with 6 years of experience–and a breast radiologist (L.G.K.) with 13 years of experience, independently assessed the MCQs blindly using their computers. All three answered questions in different sessions. Upon completion of all questions, radiologist 3 evaluated each other's answers according to the correct answer list, marking them either correct (1) or incorrect (0).

Multimodal large language models and visual breast ultrasound questions

The publicly available Breast Ultrasound Images dataset was utilized to assess the performance of multimodal LLMs with breast ultrasound images.[19] This dataset comprises 780 images classified as normal, benign, and malignant, sourced from 600 women aged 25–75 years. The images were acquired using the LOGIQ E9 ultrasound system [General Electric (GE) Healthcare, Wauwatosa, WI, USA] and the LOGIQ E9 Agile ultrasound system [General Electric (GE) Healthcare, Wauwatosa, WI, USA] at Baheya Hospital in Cairo, Egypt. The transducers used were 1–5 MHz on an ML6-15-D Matrix linear probe, and the images were stored in PNG format with dimensions of 500 × 500 pixels.[19]

The breast radiologist selected 20 images for each BI-RADS category from 1 to 5, resulting in a total of 100 images. These BI-RADS categories served as a reference standard. These images were presented to five different multimodal LLMs: Claude 3.5 Sonnet, Claude 3 Opus, Google Gemini 1.5 Vision Pro, ChatGPT 4o, and ChatGPT 4V.

For each image, the multimodal LLMs received the following prompt: "I am working on a breast radiology quiz and will provide you with breast ultrasound images. Please act as a radiology professor with 30 years of expertise in breast imaging. Evaluate the images and assign only one correct BI-RADS category from BI-RADS 1 to BI-RADS 5 according to the 5th edition of the BI-RADS Atlas. Lastly, provide clinical management recommendations for each category according to the same Atlas" (Figure 3).

This evaluation was conducted in July 2024, with each image presented along with the prompt, using the multimodal LLMs' default parameters.

The management recommendations provided by the multimodal LLMs, based on the BI-RADS categories, were evaluated using a 3-point Likert scale defined as the Management Score:

• 3 points: Correct management recommendations according to the BI-RADS category

• 2 points: Partially correct management recommendations according to the BI-RADS category

• 1 point: Completely incorrect management recommendations according to the BI-RADS category

Radiologist 3 provided the images and prompts to the multimodal LLMs and record-

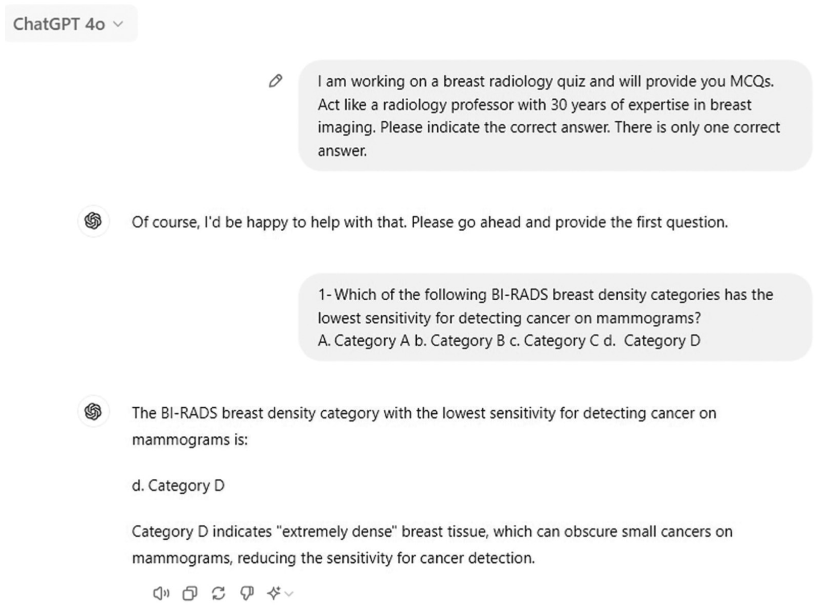

Figure 2. Demonstration of prompts provided to large language models and their responses. MCQs, multiple-choice questions; BI-RADS, Breast Imaging Reporting and Data System.

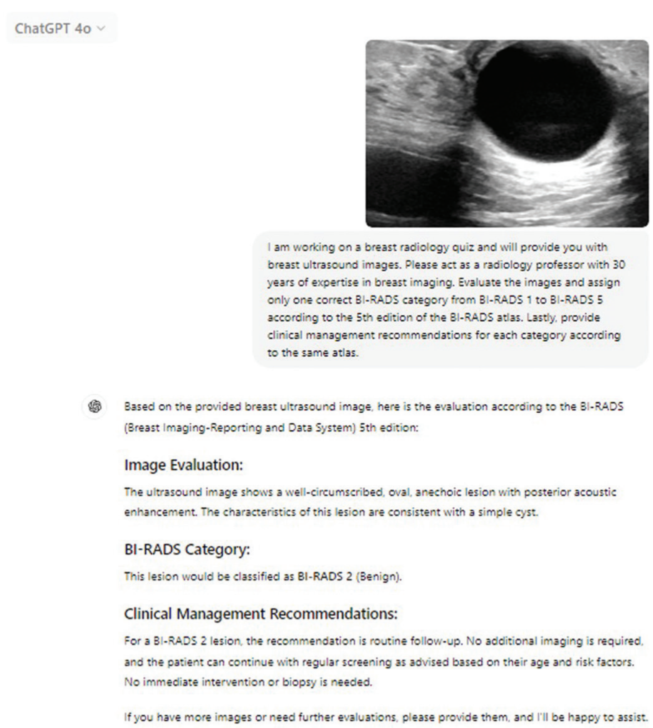

Figure 3. Illustrations of prompts and breast ultrasound images provided to multimodal large language models and their responses. MCQs, multiple-choice questions; BI-RADS, Breast Imaging Reporting and Data System.

ed their responses. The accuracy of these responses for BI-RADS categories was classified as correct (1) or incorrect (0), and clinical management recommendations were scored using the Management Score by the breast radiologist.

Statistical analysis

The distribution of variables was assessed using the Kolmogorov–Smirnov test. Descriptive statistics were represented using percentages. Non-parametric tests were employed to compare quantitative data due to the nature of the data distribution. The Kruskal–Wallis test was used to compare quantitative data, and Tamhane's T2 test was employed for multiple post-hoc comparisons following the initial Kruskal–Wallis test. McNemar's test was used to compare the proportion of correct responses between different questions. The chi-squared test was used to compare the correct answers by question types. The Wilcoxon test was used to compare the Management Scores of multimodal LLMs. The SPSS 26.0 (IBM, USA) package was used for statistical analyses, and statistical significance was set at $P < 0.05$.

Results

Accuracy of large language models on text-based breast multiple-choice questions

The highest success among the LLMs was achieved by Claude 3.5 Sonnet with an accuracy rate of 90%. ChatGPT 4o ranked second with an accuracy rate of 89%, followed by Claude 3 Opus 200K with an accuracy rate of 84%. Subsequently, Claude 3 Opus had an accuracy rate of 82%, and ChatGPT 4 had an accuracy rate of 79%. The diagnostic accuracy of the breast radiologist was 82%, radiologist 1 was 78%, radiologist 2 was 76%. Google Gemini 1.5 Pro had a 67% accuracy rate, and Microsoft Copilot with a 65% accuracy rate, while both Google Gemini 1.0 and Perplexity scored 61%, and ChatGPT 3.5 scored 60% accuracy (Figure 4).

Claude 3.5 Sonnet achieved the highest accuracy rate among the evaluated LLMs, outperforming most models with a statistically significant difference ($P < 0.05$), except when compared with ChatGPT 4o and Claude 3 Opus. Both Claude 3.5 Sonnet and ChatGPT 4o also surpassed the accuracy of the general radiologists ($P < 0.05$), although their performance was comparable with that of the breast radiologists ($P > 0.05$). Additionally, no significant differences were observed between the breast radiologist and general radiologists ($P > 0.05$).

When comparing the LLMs Claude 3 Opus 200K, Claude 3 Opus, and ChatGPT 4 with the radiologists, there were no statistically significant differences ($P > 0.05$); however, these models showed significant superiority over lower-performing LLMs, namely Google Gemini 1.5 Pro, Microsoft Copilot, and ChatGPT 3.5 ($P < 0.001$). No significant differences were found between the performances of the LLMs and radiologists across different question categories ($P > 0.05$). Detailed comparisons of the performance between radiologists and LLMs are shown in Table 1, while the performance across question categories is illustrated in Figure 5 and Table 2.

Accuracy of multimodal large language models on visual breast ultrasound questions

In a visual test consisting of 100 questions on breast ultrasound images, Claude 3.5 Sonnet achieved an accuracy rate of 59%, ChatGPT 4o 39%, Google Gemini 1.5 Pro 31%, ChatGPT 4V 20%, and Claude 3 Opus 19% (Figure 6). The performance of Claude 3.5 Sonnet was significantly higher than that of the other multimodal LLMs ($P < 0.05$). While there was no significant difference in performance between ChatGPT 4o and Google Gemini 1.5 Pro ($P = 0.067$), Claude 3 Opus and ChatGPT 4V had significantly lower performance ($P < 0.05$) (Table 3).

The accuracy rates of each model by BI-RADS categories were analyzed using the chi-squared test. The statistical analysis revealed that only Claude 3 Opus's accuracy rate did not vary by BI-RADS categories ($P = 0.992$); for other models, accuracy rates showed significant variation by category ($P < 0.05$) (Table 4).

In post-hoc tests:

• Claude 3.5 Sonnet had a higher accuracy rate for BI-RADS 5 questions (85%) compared with other categories ($P = 0.001$), while its accuracy rate for BI-RADS 1 questions (35%) was lower compared with other categories ($P = 0.001$).

• Google Gemini 1.5 Pro's accuracy rate for BI-RADS 5 questions (0%) was lower compared with other categories ($P < 0.001$).

• ChatGPT 4V had a higher accuracy rate for BI-RADS 5 questions (45%) compared with other categories ($P = 0.001$), but a lower accuracy rate for BI-RADS 1 questions (0%) ($P = 0.012$).

• ChatGPT 4o had a higher accuracy rate for BI-RADS 2 questions (65%) compared with other categories ($P = 0.007$) (Figure 7).

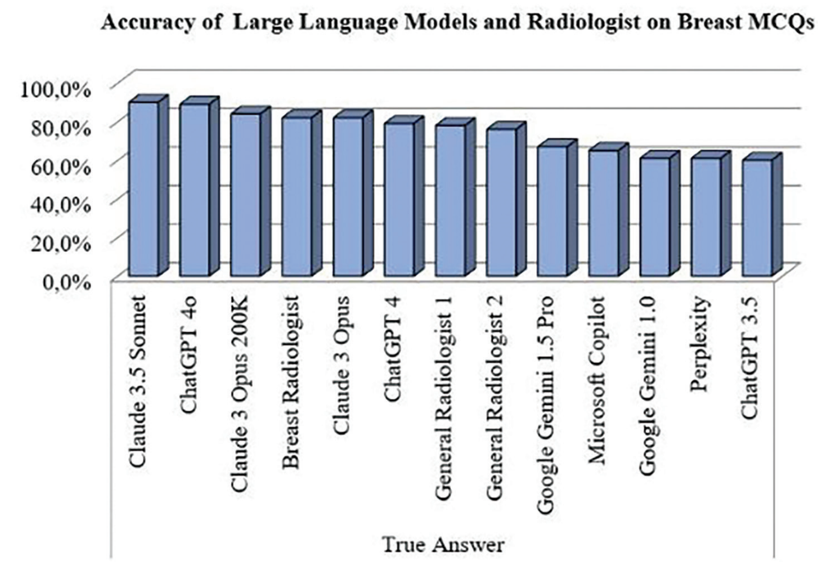

Figure 4. Accuracy of large language models and radiologists on breast multiple-choice questions. MCQs, multiple-choice questions.

Accuracy of multimodal large language models on clinical management recommendations

The mean Management Score of Claude 3.5 Sonnet (mean: 2.12 ± 0.97) was significantly superior to that of all other multimodal LLMs ($P < 0.05$). The mean Management Score of ChatGPT 4o (mean: 1.78 ± 0.98) was not significantly different from Google Gemini 1.5 Pro (mean: 1.64 ± 0.93), but it outperformed ChatGPT 4V (mean: 1.40 ± 0.80) and Claude 3 Opus (mean: 1.42 ± 0.81) ($P < 0.05$). The details of the Management Score are given in Supplementary Material 2.

Discussion

This study aimed to evaluate the performance of LLMs and multimodal LLMs in breast radiology knowledge. The most striking finding of our study is that although LLMs excel at text-based questions, their per-

Table 1. Comparison of the accuracy of LLMs and radiologists with *P* values obtained from McNemar's test

	Claude 3.5 Sonnet	Claude 3 Opus 200k	Claude 3 Opus	ChatGPT 4o	ChatGPT 4	ChatGPT 3.5	BR	R-1	R-2	Google Gemini 1.5 Pro	Google Gemini 1.0	Perplexity
Claude 3.5 Sonnet	-	0.210	0.096	1	0.019	<0.001	0.077	0.004	<0.001	<0.001	<0.001	<0.001
Claude 3 Opus 200k	0.210	-	0.774	0.302	0.359	0.001	0.832	0.327	0.152	<0.001	<0.001	<0.001
Claude 3 Opus	0.096	0.774	-	0.189	0.648	0.002	1	0.584	0.361	0.007	<0.001	<0.001
ChatGPT 4o	1	0.302	0.189	-	0.041	<0.001	0.210	0.035	0.004	<0.001	<0.001	<0.001
ChatGPT 4	0.019	0.359	0.648	0.041	-	0.004	0.710	1	0.700	0.038	0.002	0.002
ChatGPT 3.5	<0.001	0.001	0.002	<0.001	0.004	-	0.003	0.005	0.012	0.337	1	1
BR	0.077	0.832	1	0.210	0.710	0.003	-	0.208	0.327	0.029	0.002	0.002
R-1	0.017	0.327	0.584	0.035	1	0.005	0.208	-	0.805	0.091	0.005	0.005
R-2	0.004	0.152	0.361	0.004	0.700	0.012	0.327	0.805	-	0.176	0.018	0.018
Google Gemini 1.5 Pro	<0.001	<0.001	0.007	<0.001	0.038	0.337	0.029	0.091	0.176	-	0.263	0.263
Google Gemini 1.0	<0.001	<0.001	0.001	<0.001	0.002	1	0.002	0.005	0.018	0.263	-	1
Perplexity	<0.001	<0.001	0.001	<0.001	0.002	1	0.002	0.005	0.018	0.263	1	-
Microsoft Copilot	<0.001	0.002	<0.001	<0.001	0.035	0.522	0.007	0.037	0.100	0.860	0.607	0.607

LLMs, large language model; BR, breast radiologist; R-1, general radiologist 1; R-2, general radiologist.

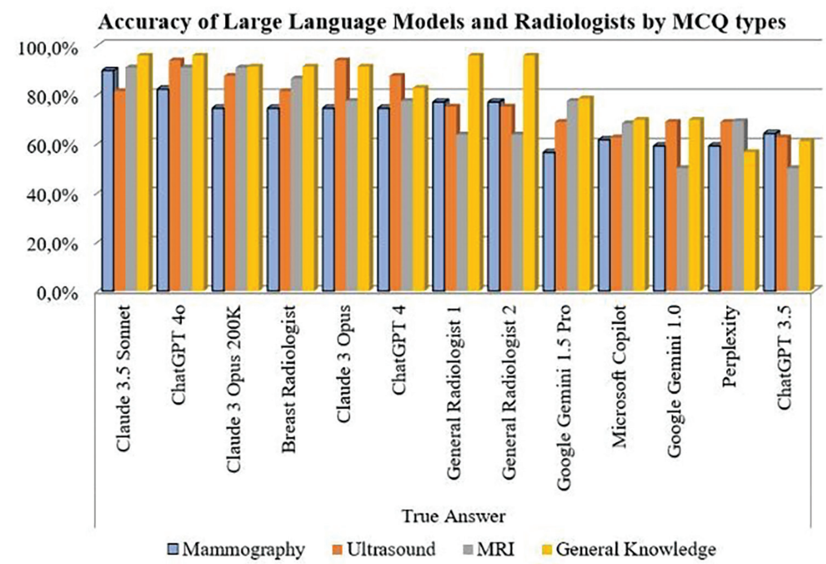

Figure 5. Accuracy of large language models and radiologists by multiple-choice question types. MCQs, multiple-choice questions; MRI, magnetic resonance imaging.

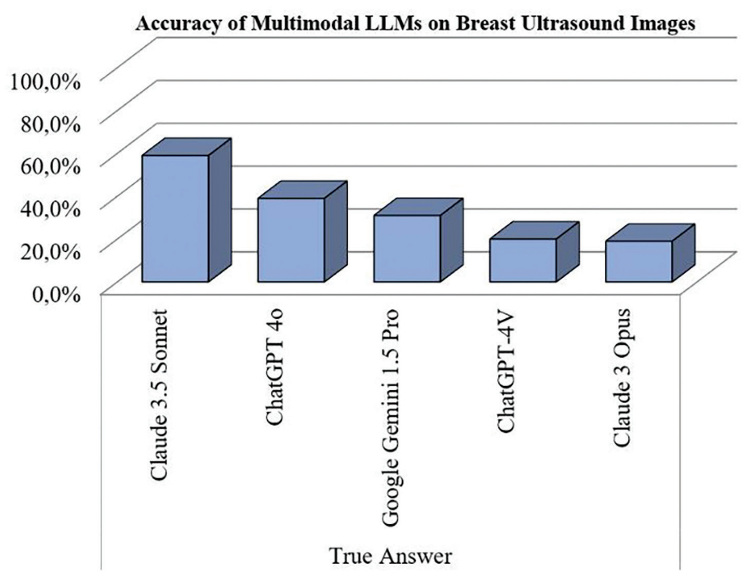

Figure 6. Accuracy of multimodal large language models on breast ultrasound images. LLMs, large language models.

formance in evaluating real-life case images is not as successful. Multimodal LLMs fall short compared with their text-based counterparts. Considering that real clinical cases are often complex and diagnoses are made through visual assessment by physicians, multimodal LLMs have not yet demonstrated sufficient performance to be used as clinical decision support systems in real-world settings.

Claude 3.5 Sonnet demonstrated the highest accuracy rate, achieving 90% in answering BI-RADS Atlas 5th edition questions. Following closely were ChatGPT 4o and Claude 3 Opus 200k with accuracy rates of 89% and 84%, respectively, while ChatGPT 4 achieved an accuracy rate of 79%. Among the radiologists, the breast radiologist exhibited the best performance with an accuracy rate of 82%, followed by general radiologist 1 with 78%, and general radiologist 2 with 76%. Claude 3.5 Sonnet demonstrated superior performance compared with all other LLMs, except for ChatGPT 4o and Claude 3 Opus models ($P < 0.05$). The performance of Claude 3.5 Sonnet and ChatGPT 4o did not show a significant difference from that of the breast radiologist ($P > 0.05$), but it notably outperformed both general radiologists ($P < 0.05$).

No statistically significant difference was found between ChatGPT 4o, Claude 3 Opus 200k, Claude 3 Opus, and ChatGPT 4 ($P > 0.05$). These LLMs, along with both the breast and general radiologists, performed significantly better than ChatGPT 3.5, Google Gemini 1.5 Pro, Google Gemini 1.0, and Perplexity ($P < 0.05$).

While interpreting real-life breast ultrasound images, Claude 3.5 Sonnet achieved an accuracy rate of 59%, ChatGPT 4o 39%, Google Gemini 1.5 Pro 31%, ChatGPT 4V 20%, and Claude 3 Opus 19%. Claude 3.5 Sonnet outperforms all the other multimodal LLMs ($P < 0.05$). The diagnostic performance of multimodal LLMs significantly differs with the BI-RADS category, except Claude 3 Opus. Claude 3.5 Sonnet (85%) and Chat GPT 4V (45%) showed superior performance in the BI-RADS 5 category ($P = 0.001$), while Google Gemini 1.5 Pro showed a higher accuracy rate (65%) for BI-RADS 2 questions ($P = 0.007$). Gemini 1.5 Pro did not correctly answer any questions in the BI-RADS 5 category, and ChatGPT 4V did not correctly answer any questions in the BI-RADS 1 category, making them the least accurate in these respective categories ($P < 0.05$).

In the Management Score, which compares the recommendations of multimodal LLMs according to BI-RADS categories, Claude 3.5 Sonnet (mean: 2.12 ± 0.97) outperformed all other multimodal LLMs ($P < 0.05$).

Notably, our study is the first to evaluate the diagnostic performance of multimodal LLMs breast radiology visual cases. Moreover, this study is the first to demonstrate the performance of the newly released Claude 3.5 Sonnet and ChatGPT 4o in breast radiology. Furthermore, there are currently no other studies that have evaluated the proficiency of different LLMs in breast radiology MCQs, both in internal comparisons and when compared with radiologists.

Multimodal LLMs, such as Claude 3.5 Sonnet and ChatGPT 4o, may perform better than a breast radiologist on text-based questions, but they can make critical errors when questions involve images that impact clinical management. For example, Gemini 1.5 Pro failed to recognize any cases in the BI-RADS 5 category, and Claude 3 Opus could not identify any normal images in the BI-RADS 0 category. This finding suggests that using multimodal LLMs without an experienced radiologist in clinical practice could lead to misdiagnoses, either missing critical conditions or misinterpreting normal findings as pathological.

On the other hand, the superior performance of LLMs on text-based questions compared with general radiologists suggests that they could serve as a supportive tool, especially for junior radiologists. They can aid in the correct use of BI-RADS nomenclature and proper classification.

Table 2. Accuracy rates of radiologists and LLMs by categories

			Mammography	MRI	General knowledge	*P*	
ChatGPT 4o	False True	n n	7 (17.9%) 32 (82.1%)	2 (9.1%) 20 (90.9%)	1 (4.3%) 22 (95.7%)	0.332	X^2
ChatGPT 4	False True	n n	10 (25.6%) 29 (74.4%)	5 (22.7%) 17 (77.3%)	4 (17.4%) 19 (82.6%)	0.700	X^2
ChatGPT 3.5	False True	n n	14 (35.9%) 25 (64.1%)	11 (50.0%) 11 (50.0%)	9 (39.1%) 14 (60.9%)	0.744	X^2
Claude 3.5 Sonnet	False True	n n	4 (10.3%) 35 (89.7%)	2 (9.1%) 20 (90.9%)	1 (4.3%) 22 (95.7%)	0.542	X^2
Claude Opus 3 200k	False True	n n	10 (25.6%) 29 (74.4%)	2 (9.1%) 20 (90.9%)	2 (8.7%) 21 (91.3%)	0.209	X^2
Claude Opus 3	False True	n n	10 (25.6%) 29 (74.4%)	5 (22.7%) 17 (77.3%)	2 (8.7%) 21 (91.3%)	0.193	X^2
Breast radiologist	False True	n n	10 (25.6%) 29 (74.4%)	3 (13.6%) 19 (86.4%)	2 (8.7%) 21 (91.3%)	0.364	X^2
General radiologist I	False True	n n	9 (23.1%) 30 (76.9%)	8 (36.4%) 14 (63.6%)	1 (4.3%) 22 (95.7%)	0.074	X^2
General radiologist II	False True	n n	9 (23.1%) 30 (76.9%)	5 (22.7%) 17 (77.3%)	5 (21.7%) 18 (78.3%)	0.905	X^2
Google Gemini Pro 1.5	False True	n n	17 (43.6%) 22 (56.4%)	4 (18.2%) 18 (81.8%)	7 (30.4%) 16 (69.6%)	0.235	X^2
Google Gemini 1.0	False True	n n	16 (41.0%) 23 (59.0%)	11 (50.0%) 11 (50.0%)	7 (30.4%) 16 (69.6%)	0.513	X^2
Microsoft Copilot	False True	n n	15 (38.5%) 24 (61.5%)	7 (31.8%) 15 (68.2%)	7 (30.4%) 16 (69.6%)	0.906	X^2
Perplexity	False True	n n	16 (41.0%) 23 (59.0%)	9 (40.9%) 13 (59.1%)	10 (43.5%) 13 (56.5%)	0.885	X^2

X^2, Chi-squared; LLM, large language model; MRI, magnetic resonance imaging.

Table 3. Comparison of accuracy of multimodal large language models with *P* values obtained from McNemar's test

	Claude 3.5 Sonnet	Claude 3 Opus	ChatGPT 4o	ChatGPT 4V	Google Gemini 1.5 Pro
Claude 3.5 Sonnet	-	<0.001	0.006	<0.001	<0.001
Claude 3 Opus	<0.001	-	0.003	1	0.067
ChatGPT 4o	0.006	0.003	-		0.302
ChatGPT 4V	<0.001	1	0.003	-	0.109
Google Gemini 1.5 Pro	<0.001	0.067	0.302	0.109	-

When multimodal LLMs correctly identify an image and assign an appropriate BI-RADS score, their management recommendations for patients closely align with the BI-RADS categories. Therefore, their success with text-based questions indicates that if they can visually determine the correct BI-RADS category, they are likely to provide accurate clinical management advice.

The variability in LLM text-based performance may be due to differences in training designs, such as different datasets, model architectures, and fine-tuning techniques.[20] LLMs such as Microsoft Copilot, Google Gemini 1.0, and Perplexity, which have internet access, sometimes provide arbitrary answers based on non-scientific information they reference.[21] This could explain their lower performance compared with other LLMs. ChatGPT and Claude 3 Opus models are trained on closed datasets, and it is unclear whether the BI-RADS Atlas was used in their training. Memorization may contribute to their high performance.

Several studies have explored the performance of LLMs on text-based radiology questions.[22,23] For instance, Almeida et al.[22] found that ChatGPT 4 achieved a 76% accuracy rate on mammography questions during the Brazilian radiology board examination, compared with 65% for ChatGPT 3.5. Our study showed higher accuracy rates, with ChatGPT 4 at 79% and ChatGPT 4o at 89%, suggesting that the difference in question difficulty may account for this variance. Furthermore, ChatGPT 4 demonstrated a general accuracy rate of 58.5%, surpassing that of 2nd-year radiology residents (52.8%) but falling short of 3rd-year residents (61.9%) in the ACR Diagnostic Radiology In-Training (DXIT) examination.[23] However, with only 10 breast radiology questions, the DXIT exam may not fully capture overall performance in this specialty. In contrast, our study's focus on a comprehensive set of BI-RADS Atlas

Table 4. Accuracy rates of multimodal large language models by categories

			BI-RADS-1	BI-RADS-2	BI-RADS-3	BI-RADS-4	BI-RADS-5	*P*	
Claude 3.5 Sonnet	False True	n n	13 (65.0%) 7 (35.0%)	5 (25.0%) 15 (75.0%)	12 (60.0%) 8 (40.0%)	8 (40.0%) 12 (60.0%)	3 (15.0%) 17 (85.0%)	0.004	X^2
ChatGPT 4o	False True	n n	13 (65.0%) 7 (35.0%)	7 (35.0%) 13 (65.0%)	17 (85.0%) 3 (15.0%)	14 (70.0%) 6 (30.0%)	10 (50.0%) 10 (50.0%)	0.015	X^2
ChatGPT 4V	False True	n n	20 (100.0%) 0 (0.0%)	14 (70%) 6 (30%)	19 (95.0%) 1 (5.0%)	16 (80.0%) 4 (20.0%)	11 (55.0%) 9 (45.0%)	0.002	X^2
Claude Opus 3	False True	n n	16 (80.0%) 4 (20.0%)	17 (85%) 3 (15%)	16 (80.0%) 4 (20.0%)	16 (80.0%) 4 (20.0%)	16 (80.0%) 4 (20.0%)	0.992	X^2
Google Gemini 1.5 Pro	False True	n n	10 (50.0%) 10 (50.0%)	13 (65.0%) 7 (35.0%)	14 (70.0%) 6 (30.0%)	12 (60.0%) 8 (40.0%)	20 (100%) 0 (0.0%)	0.010	X^2

BI-RADS, Breast Imaging Reporting and Data System.

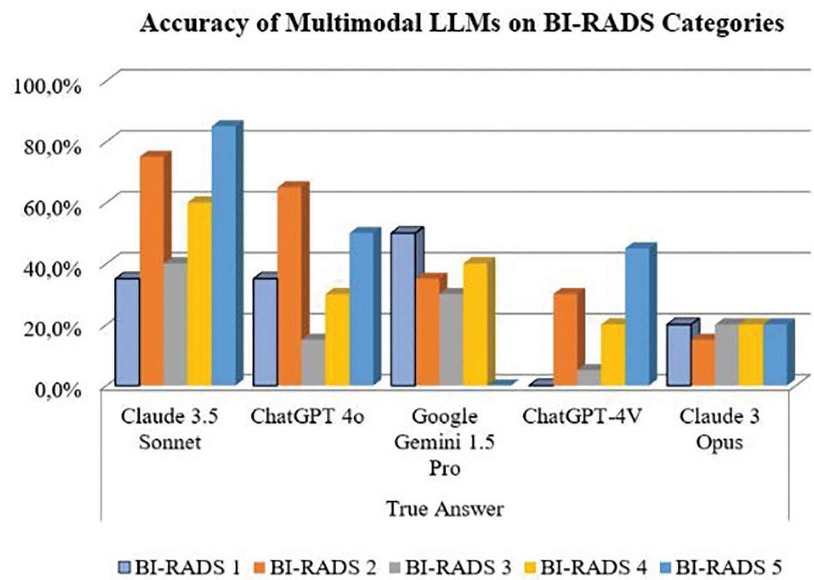

Figure 7. Accuracy of multimodal large language models on Breast Imaging Reporting and Data System (BI-RADS) categories. LLMs, large language models.

questions resulted in higher accuracy rates, underscoring that LLM performance is greatly influenced by both the specificity and quantity of the questions.

Rao et al.[9] observed that ChatGPT 4 outperformed ChatGPT 3.5 on select-all-that-apply questions related to breast pain and cancer screening, with both models performing better on these MCQs than on open-ended ones. This aligns with our findings, where the use of MCQs with a single correct answer likely contributed to the elevated success rates of LLMs. In a different context, Haver et al.[24] demonstrated ChatGPT's ability to simplify responses to frequently asked questions about breast cancer prevention and screening, achieving a 92% simplification rate. Our study, which focused on more technical and specific questions, showed that ChatGPT 4 had an accuracy rate of 79%, while ChatGPT 4o performed even better, with an accuracy rate of 89%.

When comparing the performance and readability of different LLMs, Tepe and Emekli[25] found that responses generated by Gemini 1.0 and Microsoft Copilot achieved higher readability scores ($P < 0.001$), whereas ChatGPT 4 demonstrated superior accuracy ($P < 0.001$). Our study confirmed these results, showing that ChatGPT 4 outperformed both Gemini 1.0 and Microsoft Copilot in terms of accuracy. Similarly, Griewing et al.[26] noted a 58.8% concordance between breast tumor board decisions and those generated by ChatGPT 3.5 and 4, with Sorin et al.[27] reporting a 70% agreement for ChatGPT 3.5. These findings suggest a partial alignment between LLMs and radiologists in clinical decision-making, though the variations in performance are likely due to differences in study designs and the prompts used. These studies collectively suggest that although LLMs show promise, their current performance may not yet be adequate for seamless integration into clinical decision support systems.

The challenges LLMs face in interpreting visual questions are evident in several studies.[28-30] Horiuchi et al.[30] conducted a study involving 106 musculoskeletal radiology cases, comparing the performance of ChatGPT 4 on text-based questions with ChatGPT 4V on visual questions. ChatGPT 4 correctly answered 46 out of 106 questions, significantly outperforming ChatGPT 4V, which correctly answered only 9 out of 106 ($P < 0.001$). Similarly, Dehdab et al.[28] evaluated ChatGPT 4V's performance on chest computed tomography slices across 60 different cases, including coronavirus disease-2019, non-small cell lung cancer, and control cases, finding an overall diagnostic accuracy of 56.76%, with variability depending on the case type.

In breast radiology, Haver et al.[29] compared ChatGPT 4V's performance on 151 mammography images from the ACR BI-RADS Atlas, reporting an accuracy rate of 28.5% (43/151). Although ChatGPT 4V correctly identified more than 50% of cases involving mass shape, architectural distortion, and associated features, it performed poorly on calcifications, intramammary lymph nodes, skin lesions, and solitary dilated ducts, with less than 15% correct responses.[29] In our study, ChatGPT 4V similarly showed low performance, correctly answering only 20% of breast ultrasound questions. Notably, it had an accuracy rate of 45% (9/20) for BI-RADS 5 lesions but failed to correctly identify any BI-RADS 0 lesions (0/20), indicating a tendency to misinterpret normal parenchymal tissue as pathology.

Nonetheless, as LLMs and multimodal LLMs continue to rapidly evolve and newer, more advanced models emerge, they are poised to become supportive tools for radiologists in the future. However, ethical considerations, such as ensuring patient privacy and obtaining informed consent from patients involved in the integration of LLMs into clinical decision support systems, are paramount.[31] Moreover, the lack of transparency in the decision-making mechanisms of LLMs during the diagnostic process is a significant concern.[32] Therefore, it is imperative that LLMs and multimodal LLMs are utilized under the supervision of a responsible radiologist to ensure their contribution to the diagnostic process aligns with the highest standards of patient care and safety.

An intriguing finding of our study is the notable performance of the recently introduced Claude 3.5 Sonnet, which closely rivals ChatGPT 4o. This suggests that the Claude models hold promise in the medical domain as well. Furthermore, our study contributes significantly to the existing literature by evaluating the performance of various LLMs, including both free and paid versions, alongside radiologists in the realm of breast radiology.

While our study offers valuable insights into LLMs' and multimodal LLMs' understanding of the BI-RADS Atlas, it does have limitations. First, the number of text-based questions was limited and presented in an MCQ format. Considering LLMs capacity to handle open-ended questions in real clinical scenarios, their performance may better reflect real-world situations with such questions. Further research comparing LLM performance on both open-ended and MCQs is warranted. Second, in our study evaluating multimodal LLMs' performance, we only used breast ultrasound images. Further research should include ultrasound, mammography, and MRI images to better understand the comprehensive capabilities of multimodal LLMs across different imaging modalities. Last, our study employed a single prompt to assess the performances, highlighting the need for research into the impact of different prompts and various prompt settings on LLMs' performance in breast radiology.

In conclusion, although LLMs such as Claude 3.5 Sonnet and ChatGPT 4o show potential in supporting radiologists with text-based BI-RADS assessments, their current limitations in visual diagnostics suggest that these tools should be used with caution and under the supervision of experienced radiologists to avoid misdiagnoses.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [Crossref]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. [Crossref]
3. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):e230582. [Crossref]
4. Horiuchi D, Tatekawa H, Oura T, et al. Comparison of the diagnostic performance from patient's medical history and imaging findings between GPT-4 based ChatGPT and radiologists in challenging neuroradiology cases. *medRxiv*. 2023. [Crossref]
5. Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. *Acad Radiol*. 2021;28(9):1225-1235. [Crossref]
6. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG. Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR)" case of the month". *Cureus*. 2023;15(8):e43958. [Crossref]
7. Güneş YC, Cesur T. Diagnostic Accuracy of Large Language Models in the European Board of Interventional Radiology Examination (EBIR) sample questions. *Cardiovasc Intervent Radiol*. 2024;47:836-837. [Crossref]
8. Diab KM, Deng J, Wu Y, Yesha Y, Collado-Mesa F, Nguyen P. Natural language processing for breast imaging: a systematic review. *Diagnostics (Basel)*. 2023;13(8):1420. [Crossref]
9. 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(10):990-997. [Crossref]
10. Sorin V, Glicksberg BS, Artsi Y, et al. Utilizing large language models in breast cancer management: systematic review. *J Cancer Res Clin Oncol*. 2024;150(3):140. [Crossref]
11. Nakaura T, Ito R, Ueda D, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol*. 2024;42:685-696. [Crossref]
12. Warner-Smith M, Ren K, Mistry C, et al. Protocol for evaluating the fitness for purpose of an artificial intelligence product for radiology reporting in the BreastScreen New South Wales breast cancer screening programme. *BMJ Open*. 2024;14(5):e082350. [Crossref]
13. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15(2):e35179. [Crossref]
14. Sickles EA. ACR BI-RADS® Atlas, Breast imaging reporting and data system. *American College of Radiology*. 2013:39. [Crossref]
15. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS Atlas. *AJR Am J Roentgenol*. 2016;206(5):1119-1123. [Crossref]
16. Haver HL, Yi PH, Jeudy J, Bahl M. Use of ChatGPT to assign BI-RADS assessment categories to breast imaging reports. *AJR Am J Roentgenol*. 2024. [Crossref]
17. Cozzi A, Pinker K, Hidber A, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and Google Bard: a multilanguage study. *Radiology*. 2024;311(1):e232133. [Crossref]
18. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277(3):826-832. [Crossref]
19. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data Brief*. 2020;28:104863. [Crossref]
20. Yao Y, Duan J, Xu K, et al. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confidence Computing*. 2024;4:100211. [Crossref]
21. Aydın Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*. 2023;11(3):118-134. [Crossref]
22. Almeida LC, Farina EMJM, Kuriki PEA, Abdala N, Kitamura FC. Performance of ChatGPT on the Brazilian Radiology and Diagnostic Imaging and Mammography Board Examinations. *Radiol Artif Intell*. 2023;6(1):e230103. [Crossref]
23. Payne DL, Purohit K, Borrero WM, et al. Performance of GPT-4 on the American College of Radiology in-training examination: evaluating accuracy, model drift, and fine-tuning. *Acad Radiol*. 2024;31(7):3046-3054. [Crossref]
24. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. *Radiology Imaging Cancer*. 2024;6(2):e230086. [Crossref]
25. Tepe M, Emekli E. Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. *Cureus*. 2024;16(5):e59960. [Crossref]
26. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, 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):1502. [Crossref]
27. 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):44. [Crossref]
28. Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. *Jpn J Radiol*. 2024. [Crossref]
29. Haver HL, Bahl M, Doo FX, et al. Evaluation of multimodal ChatGPT (GPT-4V) in describing mammography image features. *Can Assoc Radiol J*. 2024:8465371241247043. [Crossref]
30. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol*. 2024. [Crossref]
31. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*. 2023;90:104512. [Crossref]
32. Tokayev KJ. Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*. 2023;8(9):17-33. [Crossref]

Supplementary Material 1. MCQs

1. Which of the following BI-RADS breast density categories has the lowest sensitivity for detecting cancer on mammograms?

A) Category A

B) Category B

C) Category C

D) Category D

2. According to the BI-RADS breast density categories, in which of the following categories is mammography the most sensitive for detecting cancer?

A) Category A

B) Category B

C) Category C

D) Category D

3. Which of the following statements about BI-RADS breast density categories is true?

A) Mammography is equally sensitive for detecting cancer in all breast density categories.

B) Mammography is more sensitive for detecting cancer in women with dense breasts than in women with fatty breasts.

C) Mammography is less sensitive for detecting cancer in women with dense breasts than in women with fatty breasts.

D) There is no relationship between breast density and the sensitivity of mammography.

4. According to the BI-RADS assessment categories, which of the following categories has the highest likelihood of malignancy?

A) Category 1

B) Category 2

C) Category 3

D) Category 4

5. Which of the following BI-RADS assessment categories has a likelihood of malignancy of > 2% but <95%?

A) Category 1

B) Category 2

C) Category 3

D) Category 4

6. Which of the following management recommendations is associated with BI-RADS assessment category 3?

A) Routine mammography screening

B) Short-interval follow-up or continued surveillance mammography

C) Tissue diagnosis

D) Surgical excision

7. When is it appropriate to use BI-RADS assessment category 6?

A) When a mammographic examination is incomplete

B) When a finding is probably benign

C) When a malignancy has been biopsy-proven

D) When a finding is highly suggestive of malignancy

8. What is the management recommendation for a BI-RADS category 4 assessment?

A) Routine mammography screening

B) Short-interval follow-up or continued surveillance mammography

C) Tissue diagnosis

D) Surgical excision

9. Which of the following findings is NOT typically assessed as BI-RADS category 3?

A) Non-calcified circumscribed solid mass

B) Palpable lesion

C) Focal asymmetry

D) Solitary group of punctate calcifications

10. What is the likelihood of malignancy for a finding assessed as BI-RADS category 3?

A) Essentially 0%

B) > 0% but ≤2%

C) >2% but <95%

D) ≥ 95%

11. Which of the following is NOT a characteristically benign finding that may be assessed as BI-RADS category 2?

A) Involuting calcified fibroadenoma

B) Skin calcifications

C) Non-calcified circumscribed solid mass

D) Oil cyst

12. Which of the following findings may be described in a BI-RADS category 2 assessment?

A) Non-calcified circumscribed solid mass

B) Skin calcifications

C) Architectural distortion

D) Solitary group of punctate calcifications

13. Which of the following findings is NOT validated as being probably benign (BI-RADS category 3)?

A) Non-calcified circumscribed solid mass

B) Focal asymmetry

C) Solitary group of punctate calcifications

D) Palpable lesion

14. A screening mammogram shows unilateral axillary adenopathy with no suspicious findings in the breasts. The patient has no known infectious or inflammatory cause for the adenopathy. What should the BI-RADS® final assessment be?

A) Negative (BI-RADS® category 1)

B) Benign (BI-RADS® category 2)

C) Probably benign (BI-RADS® category 3)

D) Suspicious (BI-RADS® category 4)

15. Which of the following US descriptors for tissue composition corresponds most closely to the BI-RADS® breast density category "heterogeneously dense"?

A) Homogeneous background echotexture-fat

B) Homogeneous background echotexture-fibroglandular

C) Heterogeneous background echotexture

D) Not given in the provided text

16. Which of the following is NOT a finding that may be described in a BI-RADS® category 2 US assessment?

A) Simple cyst

B) Intramammary lymph node

C) Non-palpable solid mass

D) Postsurgical fluid collection

17. What is the recommended follow-up interval for a stable probably benign (BI-RADS® category 3) finding on US after the initial 6-month follow-up examination?

A) 3 months

B) 6 months

C) 1 year

D) 2 years

18. A US examination reveals a large axillary mass in a patient with known metastatic melanoma. The mass was previously biopsied and confirmed to be an axillary lymph node with metastatic melanoma. Except for the axillary mass, the US examination shows no abnormalities in the breast. What is the appropriate BI-RADS® assessment for this examination?

A) BIRADS 1

B) BIRADS 2

C) BIRADS 3

D) BIRADS 4

19. Which of the following is NOT a category of background parenchymal enhancement (BPE) on breast MRI?

A) Minima

B) Mild

C) Moderate

D) Severe

20. Which of the following is NOT a descriptor for the margin of a mass on breast MRI?

A) Circumscribed

B) Not circumscribed

C) Irregular

D) Rounded

21. Which of the following is NOT a modifier describing non-mass enhancement distribution?

A) Focal

B) Linear

C) Granular

D) Segmental

22. Which of the following is NOT an internal enhancement pattern for non-mass enhancement?

A) Homogeneous

B) Heterogeneous

C) Focal

D) Clumped

23. Which of the following is NOT an intracapsular silicone rupture finding on MRI?

A) Linguine sign

B) Subcapsular line

C) Keyhole sign

D) Peri-implant fluid

24. Which BI-RADS® assessment category is not recommended for screening for mammography?

A) BIRADS 1

B) BIRADS 2

C) BIRADS 3

D) BIRADS 4

25. According to BI-RADS classification, which of the following is NOT a type of asymmetry?

A) Asymmetry

B) Global asymmetry

C) Focal asymmetry

D) Diffuse asymmetry

26. Which of the following calcification morphologies should be assigned to BI-RADS® category 4C?

A) Amorphous

B) Coarse heterogeneous

C) Fine pleomorphic

D) Fine linear or fine-linear branching

27. Which of the following statements is true regarding the margin of a mass in mammography?

A) The margin must be completely well-defined for the mass to be classified as circumscribed.

B) At least 75% of the margin must be well-defined for the mass to qualify as circumscribed.

C) If any portion of the margin is indistinct, the mass should be classified as such.

D) Spiculated margins are less suspicious than microlobulated margins.

28. What is an obscured margin in mammography?

A) A margin that is completely hidden by other tissue

B) A margin that is mostly well-defined, but part of it is hidden

C) A margin that is indistinct and irregular

D) A margin that is spiculated and jagged

29. Which of the following statements about density in mammography is true?

A) Breast cancers are always lower in density than normal breast tissue.

B) Breast density is the most reliable mammographic feature of masses.

C) Breast cancers can be fat-containing.

D) The likelihood of malignancy for a high-density mass is significantly greater than that for equal- and low-density masses.

30. Which of the following is a characteristic of a fat-containing mass in mammography?

A) It is always malignant.

B) It is almost always benign.

C) It is a mixed-density mass.

D) It is associated with a high risk of breast cancer.

31. Which of the following types of calcifications is typically benign in mammography?

A) Fine and linear

B) Pleomorphic

C) Coarse or "popcorn-like"

D) Punctate

32. Which of the following statements about coarse or "popcorn-like" calcifications in mammography is true?

A) They are typically associated with breast cancer.

B) They are small and difficult to see on mammograms.

C) They are a sign of a benign breast lesion.

D) They are more common in younger women

33. Which of the following is a characteristic of large rod-like calcifications in mammography?

A) They are typically associated with breast cancer.

B) They are small and difficult to see on mammograms.

C) They are more common in younger women.

D) They follow a ductal distribution.

34. Which of the following is a characteristic of benign round calcifications in mammography?

A) They are typically clustered together.

B) They are always larger than 1 mm in size.

C) They are more common in younger women.

D) They are often formed in the acini of lobules.

35. Which of the following statements about dystrophic calcifications in mammography is true?

A) They are typically associated with breast cancer.

B) They are always smaller than 1 mm in size.

C) They are more common in younger women.

D) They are caused by radiation therapy or trauma.

36. Which of the following is a characteristic of milk of calcium calcifications in mammography?

A) They are typically associated with breast cancer.

B) They always appear as round, smudgy deposits on all mammographic projections.

C) They are more common in younger women.

D) They change shape on different mammographic projections.

37. Which of the following distributions of amorphous calcifications is suspicious and generally warrants biopsy?

A) Bilateral, diffuse

B) Grouped, linear, or segmental

C) Punctate

D) Coarse

38. What is the BI-RADS® assessment category for a single group of coarse heterogeneous calcifications, which has a positive predictive value of slightly less than 15%?

A) 4A

B) 4B

C) 4C

D) 5

39. What is the BI-RADS® assessment category for fine pleomorphic calcifications, which have a positive predictive value (PPV) of 29%?

A) 4A

B) 4B

C) 4C

D) 5

40. What is the BI-RADS® assessment category for fine linear and fine-linear branching calcifications, which have the highest PPV (70%) among suspicious calcifications?

A) 4A

B) 4B

C) 4C

D) 5

41. Which distribution of calcifications is of concern because it suggests deposits in a duct or ducts and their branches, raising the possibility of extensive or multifocal breast cancer?

A) Clustered

B) Grouped

C) Linear

D) Segmental

42. What is a possible cause of asymmetry that is visible on only one mammographic projection?

A) Summation artifacts

B) Real lesions

C) Cancer

D) Calcification

43. Where are intramammary lymph nodes frequently located?

A) Medial and lower portions of the breast

B) Lateral and usually upper portions of the breast closer to the axilla

C) Central portion of the breast

D) Posterior portion of the breast

44. Which mammographic finding is a rare finding and has been reported to be associated with non-calcified DCIS?

A) Intramammary lymph node

B) Solitary dilated duct

C) Asymmetric breast tissue

D) Architectural distortion

45. Which of the following is a sign of malignancy?

A) Bilateral nipple inversion

B) New nipple retraction

C) Stable nipple inversion for a long period of time

D) Nipple eversion

46. Which of the following is NOT a concern for skin thickening?

A) Focal or diffuse skin thickening greater than 2 mm in thickness

B) Skin thickening that represents a change from previous mammograms

C) Unilateral skin thickening after radiation therapy

D) Diffuse skin thickening with no other suspicious findings

47. Why did the fifth edition of BI-RADS® remove ranges of percentage dense tissue for the four density categories?

A) To emphasize the text descriptions of breast density

B) To indicate that percentage breast density is not associated with changes in mammographic sensitivity

C) To indicate that percentage breast density is more important than text descriptions of breast density for breast cancer risk assessment

D) To simplify the BI-RADS® reporting system

48. What is the key difference between an asymmetry and a mass on a mammogram?

A) Asymmetry has concave-outward borders, while a mass has convex-outward borders.

B) Asymmetry is unilateral, while a mass can be bilateral.

C) Asymmetry is interspersed with fat, while a mass is not.

D) Asymmetry is less conspicuous than a mass.

49. Why was the shape descriptor "lobular" eliminated in the 2013 edition of BI-RADS®?

A) Because it was redundant with the margin descriptor "microlobulated"

B) Because it was always associated with benign masses

C) Because it was difficult to distinguish from other shape descriptors

D) Because it was not a reliable indicator of malignancy

50. Which of the following is a key difference between "round" and "punctate" calcifications in the 2013 edition of BI-RADS®?

A) Round calcifications are typically benign, while punctate calcifications may be associated with malignancy

B) Punctate calcifications are defined as particles <0.5 mm, while round calcifications are defined as particles ≥0.5 mm

C) Round calcifications are typically isolated, while punctate calcifications are typically grouped

D) Punctate calcifications are more common in younger women, while round calcifications are more common in older women

51. Which of the following statements about coarse heterogeneous calcifications is true?

A) They are always associated with malignancy.

B) They are typically benign when present as multiple bilateral groupings.

C) They are larger than dystrophic calcifications.

D) They are more likely to be malignant when they occur together with fine pleomorphic calcifications.

52. What percentage of cases of developing asymmetry are found to be malignant?

A) 5%

B) 15%

C) 25%

D) 35%

53. What is the range of likelihood of malignancy for findings placed in BI-RADS® category 4A?

A) >2% TO ≤10%

B) >10% TO ≤20%

C) >20% TO ≤30%

D) >30% TO ≤50%

54. Which of the following findings is an example of a category 4A finding in BI-RADS®?

A) A circumscribed solid mass with smooth margins

B) A partially (<75%) circumscribed solid mass with US features suggestive of a fibroadenoma

C) A mass with spiculated margins and heterogeneous internal echogenicity

D) A cluster of irregular microcalcifications

55. Which of the following findings is an example of a category 4C finding in BI-RADS®?

A) A circumscribed solid mass with smooth margins

B) A partially circumscribed solid mass with US features suggestive of a fibroadenoma

C) A new indistinct, irregular solid mass

D) A cluster of punctate microcalcifications

56. When is BI-RADS® category 6 used?

A) When a tissue diagnosis of malignancy has been made but prior to complete surgical excision

B) When a biopsy is recommended for a suspicious lesion

C) When a benign lesion is found on imaging

D) When a patient has a history of breast cancer

57. What is the primary use of BI-RADS® category 0?

A) To indicate a finding that is highly suggestive of malignancy

B) To indicate the recommendation for additional imaging evaluation

C) To indicate the presence of a benign lesion

D) To indicate the need for a biopsy

58. Which of the following is a common mammographic finding associated with gynecomastia?

A) A circumscribed solid mass with smooth margins

B) A cluster of round calcifications

C) A "flame-shaped" area of increased density extending postero-laterally from the nipple

D) A spiculated mass with heterogeneous internal echogenicity

59. Which of the following is recommended by the ACR Practice Guideline for the Performance of a Breast Ultrasound Examination (2011) for optimal US image quality?

(A) Use of a low-frequency linear array transducer

B) Use of a broad bandwidth linear array transducer with a center frequency of at least 10 MHz

C) Use of a handheld, high-frequency breast US system

D) Use of a system with a low-resolution imaging capability

60. What is an important consideration when setting the field of view (FOV) on a breast ultrasound examination?

A) The FOV should be deep enough to include the pleura and lung.

B) The FOV should be set to a shallower depth when a lesion is found.

C) The FOV should be set deeply enough to include breast tissue and the pectoralis muscle posterior to it.

D) The FOV should be set to a very narrow depth to improve image resolution.

61. Which of the following is the correct method for taking measurements of a breast lesion on ultrasound?

A) Take two measurements from the same view, and take the third measurement from a view that is perpendicular to the first two.

B) Take three measurements from the same view, with each measurement representing a different plane.

C) Take two measurements from the same view, and take the third measurement from a view that is parallel to the first two.

D) Take three measurements from different views, with each measurement representing the longest axis of the lesion.

62. Which of the following is NOT a type of margin that can be used to characterize a mass on ultrasound?

A) Circumscribed

B) Indistinct

C) Spiculated

D) Irregular

63. What is the key feature of an indistinct margin on ultrasound?

A) The margin is clearly demarcated from the surrounding tissue.

B) The margin is poorly defined and blends into the surrounding tissue.

C) The margin is spiculated and irregular.

D) The margin is angular and has sharp corners.

64. What is the normal skin thickness in the periareolar area and inframammary folds on ultrasound?

A) Up to 2 mm

B) Up to 3 mm

C) Up to 4 mm

D) Up to 5 mm

65. Which of the following is a characteristic of edema on ultrasound?

A) Increased echogenicity of the surrounding tissue

B) Decreased echogenicity of the surrounding tissue

C) A mass-like appearance

D) Calcifications

66. Which of the following is a standardized descriptor for lesion stiffness on ultrasound elastography?

A) Soft

B) Intermediate

C) Hard

D) All of the above

67. What is the key difference between a "complicated cyst" and a "complex cystic and solid" mass on ultrasound?

A) The presence of internal echoes

B) The presence of septations

C) The presence of a discrete solid component

D) The size of the mass

68. Which of the following is a common benign mass that can be found in or on the skin on ultrasound?

A) Sebaceous cyst

B) Metastasis

C) Cancer

D) Abscess

69. What is the characteristic ultrasound appearance of extravasated silicone or silicone gel bleed?

A) A well-defined mass with posterior acoustic shadowing

B) A cystic mass with internal echoes

C) An echogenic mass with a "snowstorm" appearance

D) A hypoechoic mass with indistinct margins

70. Which of the following is NOT a US descriptor for tissue composition?

A) Homogeneous background echotexture-fat

B) Homogeneous background echotexture-fibroglandular

C) Heterogeneous background echotexture

D) Coarse background echotexture

71. According to the BI-RADS fifth edition, what is the correct term for a mass that contains solid and cystic components on ultrasound?

A) Complex mass

B) Complicated mass

C) Complex cystic and solid mass

D) Cystic mass

72. Which type of calcification is typically associated with an involuting fibroadenoma?

A) Fine linear and branching

B) Round

C) Coarse or "Popcorn-Like"

D) Amorphous

73. What is the recommended time point for assessing breast parenchymal enhancement (BPE) on breast MRI?

A) 2 minutes

B) 5 minutes

C) 90 seconds

D) 15 minutes

74. Which of the following is NOT a characteristic of breast parenchymal enhancement (BPE) on breast MRI?

A) Occurs regardless of menstrual cycle or menopausal status

B) Directly related to the amount of fibroglandular tissue

C) Evaluated with respect to the amount of fibroglandular tissue

D) May demonstrate progressive enhancement over time

75. Which of the following is a consideration when scheduling a breast MRI for elective examinations?

A) Scheduling the patient early in her menstrual cycle to minimize background enhancement

B) Scheduling the patient late in her menstrual cycle to maximize breast enhancement

C) Avoiding the use of contrast agents in pre-menopausal women

D) Performing the MRI regardless of the menstrual cycle or menstrual status

76. Which of the following features of a focus on breast MRI is suggestive of malignancy?

A) Not unique compared to the BPE

B) Bright on bright-fluid imaging

C) Washout kinetics

D) Persistent kinetics

77. Which of the following is a suggestive feature of a fibroadenoma on breast MRI?

A) Enhancing internal septations

B) Non-enhancing dark internal septations

C) Washout kinetics

D) Irregular shape

78. Which of the following is a cause of a false-positive interpretation of a rim-enhancing lesion on contrast-enhanced ultrasound?

A) Galactocele

B) Fat necrosis

C) Fibroadenoma

D) Malignant tumor

79. Which of the following internal enhancement patterns of non-mass enhancement (NME) is suggestive of malignancy?

A) Homogeneous

B) Heterogeneous

C) Clumped

D) Clustered ring

80. What is the primary factor used to determine the second phase of a contrast-enhanced lesion on MRI?

A) Initial-phase enhancement pattern

B) Delayed-phase enhancement pattern

C) Lesion morphology

D) Lesion size

81. Which of the following delayed-phase enhancement patterns is most commonly associated with malignant lesions?

A) Persistent

B) Plateau

C) Washout

D) Mixed

82. Which of the following is a potential cause of asymmetric breast parenchymal enhancement (BPE) on contrast-enhanced MRI?

A) Radiation therapy

B) Menstrual cycle

C) Age

D) Menopausal status

83. What is the criterion for classifying the initial phase of enhancement on contrast-enhanced MRI?

A) Percent increase in signal intensity compared to precontrast image

B) Time to peak enhancement

C) Shape of the enhancement curve

D) Type of contrast agent used

84. Which of the following is NOT a characteristic of intracapsular silicone rupture?

A) Linguine sign

B) Intraparenchymal oil cyst

C) Subcapsular line

D) Keyhole sign

85. How can you differentiate between a focal bulge in an intact breast implant and extruded silicone from an extracapsular rupture on MRI?

A) The focal bulge will have signs of intracarpsular rupture on MRI, while the extracapsular rupture will not.

B) The extracapsular rupture will have signs of intracapsular rupture inside the implant on MRI, while the focal bulge will not.

C) The focal bulge will be located on the outer edge of the implant, while the extracapsular rupture will be located in the center of the implant.

D) The extracapsular rupture will be larger than the focal bulge.

86. What is the appearance of a subcapsular line on MRI in an intracapsular silicone implant rupture?

A) A dark line paralleling the implant edge

B) A white line paralleling the implant edge

C) A dark line perpendicular to the implant edge

D) A white line perpendicular to the implant edge

87. What are the four categories used to describe the amount of background enhancement on contrast-enhanced MRI?

A) None, minimal, moderate, marked

B) Minimal, mild, moderate, marked

C) Mild, moderate, marked, severe

D) None, mild, moderate, severe

88. Which of the following statements about breast parenchymal enhancement (BPE) on contrast-enhanced MRI is true?

A) BPE is only seen in patients with dense breasts.

B) BPE can occur regardless of the menstrual cycle or menopausal status of the patient.

C) BPE is always related to the amount of fibroglandular parenchyma present.

D) Younger patients with dense breasts are less likely to demonstrate BPE than older patients with dense breasts.

89. What is the key distinguishing feature of a focus on contrast-enhanced breast MRI?

A) It is a small, punctate enhancing dot that is non-specific

B) It is a small, punctate enhancing dot that shows washout kinetics

C) It is a non-enhancing dot that corresponds to a precontrast finding

D) It is a small, punctate enhancing dot that is separated by intervening normal breast parenchyma

90. What type of enhancement pattern on MRI most closely resembles the pleomorphic pattern on mammography?

A) Punctate

B) Linear

C) Clumped

D) Regional

91. What is the appropriate BI-RADS® assessment for isolated unilateral axillary adenopathy in the absence of a known infectious or inflammatory cause?

A) Benign (category 2)

B) Probably benign (category 3)

C) Suspicious (category 4)

D) Malignant (category 5)

92. What is the standard term for a mammographic view that is angled toward the axilla?

A) Craniocaudal view

B) Mediolateral oblique view

C) Lateral view

D) Axillary view

93. What is the standard abbreviation for a tangential mammographic view?

A) TAN

B) CV

C) XCCM

D) XCCL

94. What is not the standard abbreviation for mammography views?

A) MLO75

B) SIO

C) FL

D) XCCL

95. What is not the standard abbreviation for a step-oblique view?

A) MLO15

B) MLO45

C) MLO75

D) MLO90

96. Which of the following is NOT a type of fat-containing lesion that can be seen on a mammogram?

A) Oil cyst

B) Lipoma

C) Galactocele

D) Fibroadenoma

97. What is the standard abbreviation for a superolateral-to-inferomedial oblique view?

A) IOS

B) SOI

C) SIO

D) ISO

98. What is the recomended term for ultrasound regarding special cases?

A) Clustered microcyst

B) Fibrocystic changes

C) Clustered fibrocyst

D) Microcyst

99. What is the recomended term for implants regarding location?

A) Postglandular

B) Postpectoral

C) Retroglandular

D) Glandular

100. Which year was the fifth edition of the BI-RADS® Atlas released?

A) 2013

B) 2010

C) 2008

D) 2007

Supplementary Material 2. Accuracy rates of multimodal LLMs on BI-RADS categories

			BIRADS-1	BIRADS-2	BIRADS-3	BIRADS-4	BIRADS-5	*P*	
Claude 3.5 Sonnet	False	n	13 (65.0%)	5 (25.0%)	12 (60.0%)	8 (40.0%)	3 (15.0%)	0.004	X^2
	True	n	7 (35.0%)	15 (75.0%)	8 (40.0%)	12 (60.0%)	17 (85.0%)		
ChatGPT 4o	False	n	13 (65.0%)	7 (35.0%)	17 (85.0%)	14 (70.0%)	10 (50.0%)	0.015	X^2
	True	n	7 (35.0%)	13 (65.0%)	3 (15.0%)	6 (30.0%)	10 (50.0%)		
ChatGPT 4V	False	n	20 (100.0%)	14 (70%)	19 (95.0%)	16 (80.0%)	11 (55.0%)	0.002	X^2
	True	n	0 (0.0%)	6 (30%)	1 (5.0%)	4 (20.0%)	9 (45.0%)		
Claude Opus 3	False	n	16 (80.0%)	17 (85%)	16 (80.0%)	16 (80.0%)	16 (80.0%)	0.992	X^2
	True	n	4 (20.0%)	3 (15%)	4 (20.0%)	4 (20.0%)	4 (20.0%)		
Gemini 1.5 Pro	False	n	10 (50.0%)	13 (65.0%)	14 (70.0%)	12 (60.0%)	20 (100%)	0.010	X^2
	True	n	10 (50.0%)	7 (35.0%)	6 (30.0%)	8 (40.0%)	0 (0.0%)		

X^2, chi-square; LLMs, large language models; BI-RADS, Breast Imaging Reporting and Data System.