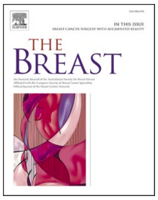
The Breast 65 (2022) 124–135

Contents lists available at ScienceDirect

The Breast

journal homepage: www.journals.elsevier.com/the-breast



Evaluation of physician performance using a concurrent-read artificial intelligence system to support breast ultrasound interpretation

Yi-Chen Lai [a,b], Hong-Hao Chen [c], Jen-Feng Hsu [c,d], Yi-Jun Hong [c], Ting-Ting Chiu [c], Hong-Jen Chiou [b,e,*]

[a] *Comprehensive Breast Health Center, Taipei Veterans General Hospital, No.201, Sec. 2, Shipai Rd., Beitou District, Taipei, 11217, Taiwan, ROC*
[b] *School of Medicine, National Yang Ming Chiao Tung University, No.155, Sec. 2, Linong St., Beitou District, Taipei, 112304, Taiwan, ROC*
[c] *TaiHao Medical Inc., 6F.-1, No.100, Sec. 2, Heping E. Rd., Da'an District, Taipei, 10663, Taiwan, ROC*
[d] *Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Da'an District, Taipei, 10617, Taiwan, ROC*
[e] *Department of Radiology, Taipei Veterans General Hospital, No.201, Sec. 2, Shipai Rd., Beitou District, Taipei, 11217, Taiwan, ROC*

ARTICLE INFO

Keywords:
Artificial intelligence (AI)
Breast cancer
Computer-assisted detection (CADe)
Computer-assisted diagnosis (CADx)
Deep learning

ABSTRACT

Purpose: The purpose of this study was to compare the diagnostic performance and the interpretation time of breast ultrasound examination between reading without and with the artificial intelligence (AI) system as a concurrent reading aid.

Material and methods: A fully crossed multi-reader and multi-case (MRMC) reader study was conducted. Sixteen participating physicians were recruited and retrospectively interpreted 172 breast ultrasound cases in two reading scenarios, once without and once with the AI system (BU-CAD™, TaiHao Medical Inc.) assistance for concurrent reading. Interpretations of any given case set with and without the AI system were separated by at least 5 weeks. These reading results were compared to the reference standard and the area under the LROC curve (AUCLROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for performance evaluations. The interpretation time was also compared between the unaided and aided scenarios.

Results: With the help of the AI system, the readers had higher diagnostic performance with an increase in the average AUCLROC from 0.7582 to 0.8294 with statistically significant. The sensitivity, specificity, PPV, and NPV were also improved from 95.77%, 24.07%, 44.18%, and 93.50%–98.17%, 30.67%, 46.91%, and 96.10%, respectively. Of these, the improvement in specificity reached statistical significance. The average interpretation time was significantly reduced by approximately 40% when the readers were assisted by the AI system.

Conclusion: The concurrent-read AI system improves the diagnostic performance in detecting and diagnosing breast lesions on breast ultrasound images. In addition, the interpretation time is effectively reduced for the interpreting physicians.

1. Introduction

The estimated number of new cases of breast cancer is the first of all cancers and breast cancer is the second leading cause of cancer death in women [1]. Mammography is the most common screening tool for breast cancer, which is an effective examination to find breast cancer at an early stage. In general, regular mammography scan is not recommended for young women who do not have a personal history or a strong family history of breast cancer [2,3]. The decision to screen with mammography in young women should be an individual one. Therefore, although breast cancer mortality tends to decline due to early detection and improved treatment, it is not reflected in young women [4].

Breast density is highly associated with the risk of breast cancer. Many researchers [5,6] observed an inverse relationship between patient age and breast density. Women with dense breasts that appear as a solid white area on mammograms are more likely to affect radiologists' interpretation [7]. Hence, breast ultrasound plays an integral part in breast cancer screening, used primarily as an adjunct to diagnostic

* Corresponding author. Department of Radiology Taipei Veterans General Hospital, No.201, Sec.2, Shipai Rd., Beitou District, Taipei City, 11217, Taiwan, ROC.
E-mail addresses: yclai3@vghtpe.gov.tw (Y.-C. Lai), honghowc@taihaomed.com (H.-H. Chen), jfhsu@taihaomed.com (J.-F. Hsu), sheenahong@taihaomed.com (Y.-J. Hong), tammychiu@taihaomed.com (T.-T. Chiu), hjchiou@vghtpe.gov.tw (H.-J. Chiou).

https://doi.org/10.1016/j.breast.2022.07.009
Received 15 April 2022; Received in revised form 11 July 2022; Accepted 13 July 2022
Available online 18 July 2022

mammographies, such as in cases of palpable lesions, focal clinical findings (pain, possible infection, nipple-areolar changes, etc.), and equivocal mammographic findings (especially in dense breasts). With the popularization of breast ultrasound use, many technologies in artificial intelligence (AI) assisted detection and diagnosis of breast cancer have been investigated.

Most AI systems focus on computer-assisted detection (CADe) or computer-assisted diagnosis (CADx). CADe has been evaluated to improve reader performance and reduce interpretation time for automated breast ultrasound (ABUS) interpretation [8,9], while CADx has been evaluated to improve accuracy for breast ultrasound assessment and to be comparable to that of experienced readers, based on the reader-identified lesion [10–12].

In this study, we evaluated an innovative computer-assisted detection and diagnosis system (BU-CAD™, TaiHao Medical Inc.) by a fully crossed multi-reader and multi-case (MRMC) reader study. The BU-CAD™ automatically identifies lesions on breast ultrasound images and generates a score of lesion characteristics (SLC) in terms of malignancy or benignity of a lesion, a corresponding breast imaging reporting and data system (BI-RADS) [13] category, and BI-RADS descriptors including shape, orientation, margin, echo pattern, and posterior features. The study compared the reader (radiologists and breast surgeons) performance between interpreting with BU-CAD™ assistance and interpreting with breast ultrasound images alone.

2. Materials and methods

2.1. Data collection

This study was approved by the Institutional Review Board of the Taipei Veterans General Hospital (Taipei, Taiwan) with IRB No. 2019-01-017CC. The requirement to obtain informed consent was waived. All B-mode breast ultrasound images were retrospectively collected at the time of diagnostic sonography or before biopsy. Lesions were identified by breast palpation, mammography, ultrasound, or magnetic resonance imaging (MRI). When the scanning physician performed an ultrasound scan, a lesion was imaged to demonstrate the subtle findings in the perpendicular planes (transverse plus longitudinal or radial plus anti-radial). The images were acquired by the linear transducer with a length of 50 mm for the Philips ultrasound device and 58 mm for the Toshiba/Canon ultrasound device with a minimum 8-MHz acquisition frequency.

Each case consisted of at least two orthogonal projections of a single lesion. In cases where multiple images were present, the two most representative images were selected for use by a senior radiologist with over 30 years of work experience in breast image interpretation. The selection priority was based on (1) the concordance between the image and the clinical report; (2) the images that demonstrate the highest proportion of each lesion; and (3) the highest image quality. A total of 172 B-mode breast ultrasound patient cases (mean age: 54.07 ± 12.01 years; range: 22–77 years) scanned by 11 operators were anonymized and retrospectively collected from the Taipei Veterans General Hospital in Taiwan. The reference standard of the region of interest (ROI) for each lesion was defined by the expert panel of five radiologists.

2.2. Demographic and Clinical Characteristics

The 172 patient cases consisted of 65 biopsy-proven malignant, 71 biopsy-proven benign, and 36 benign cases with at least a 2-year follow-up. The distribution of ultrasound scanner models and case characteristics including patient age and lesion size are shown in Table 1 and Fig. 1, respectively. Patients aged 40–70 years account for more than 77% of all patients.

Table 2 shows the distribution of cancer types. Among the 65 malignant cases, the invasive ductal carcinoma (IDC) accounts for 80%, and it far outnumbered the ductal carcinoma in situ (DCIS), invasive lobular carcinoma (ILC), and others.

Table 1
Collected breast ultrasound case distribution.

Scanner Model	Malignant	Benign (Biopsy-Proven, Follow-Up)	All Case
Philips iU22	33	67 (36, 31)	100
Toshiba Aplio 500/Canon Aplio i800	32	40 (35, 5)	72
Total	65	107 (71, 36)	172

2.3. Computer-assisted detection and diagnosis system

The computer-assisted detection and diagnosis system used in this study was a medical device software BU-CAD™ (TaiHao Medical Inc.), which had been approved by the U.S. Food and Drug Administration (FDA) in 2021. This system is indicated for breast ultrasound lesion identification and providing diagnostic recommendations. The BU-CAD™ system is an artificial intelligence software application that adopts deep learning neural network techniques and implements instance segmentation [14] for the identification and diagnosis of soft tissue lesion. The CADe functionality identifies regions of interest (automated ROIs) of a single suspicious soft tissue lesion in up to two orthogonal views of breast ultrasound images to assist users in detecting soft tissue lesions. In addition, CADe generates ROI and lesion contour on each breast ultrasound image. The lesion contour in each image will automatically be delineated by the given ROI.

The CADx functionality generates an SLC, and a corresponding BI-RADS [13] category is also presented to respond to clinical practice. The SLC ranging [0, 26), [26, 51), [51, 98), and [98, 100) correspond to BI-RADS Category 2, 3, 4, and 5 respectively by experimental fitting. In addition, lesion morphology by means of BI-RADS lexicon of descriptors including shape, orientation, margin, echo pattern (i.e., internal sonographic texture), and posterior features are also provided for the concurrent read. During the reader study, the users were able to replace the automated ROIs with re-delineated rectangular ROIs for analysis by CADx. Only the last analysis results were displayed on the user interface and could be modified by the user. After reading each case, the system automatically recorded the rating results and interpretation time for further analysis. Fig. 2 shows the results of the AI analysis of a benign case (fibrocystic change) and a malignant case (invasive ductal carcinoma).

3. Study design

A fully crossed MRMC reader study was conducted. Sixteen participating physicians were recruited and asked to retrospectively interpret each case under the following two scenarios: (1) reading without BU-CAD™ assistance, and (2) reading with BU-CAD™ assistance. Location-specific receiver operating characteristic (LROC) curves were used to evaluate the diagnostic performance of the readers between the unaided and aided scenarios. In addition, the time spent on the review and interpretation of each case was recorded for analysis.

3.1. Reader recruitment and training

A total of 16 readers were recruited including 14 U S. board-certified radiologists and 2 breast surgeons from Virginia, Maryland, Pennsylvania, North Carolina, New Hampshire, Illinois, Ohio, and California. Radiologists' years of work experience range from 1 to 24 years (average: 10.1 years). Four of the fourteen radiologists had received breast imaging fellowship training. Both board-certified breast surgeons have practiced for more than 30 years. Table 3 lists the specialty, the Mammography Quality Standards Act (MQSA) certification, training of breast imaging fellowship, and years of work experience for each reader.

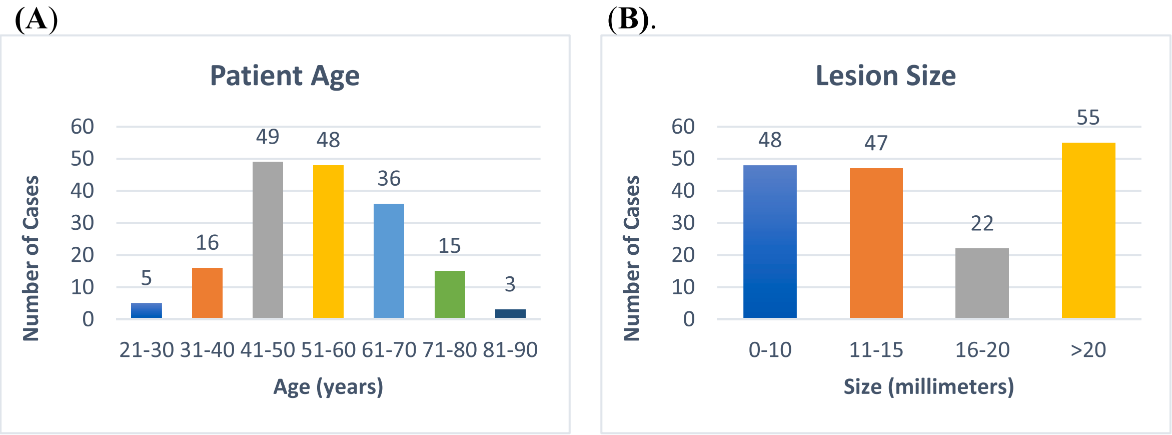


Fig. 1. Distribution of demographic and clinical characteristics. (A) Patient age; (B) Lesion size.

Table 2
Cancer type distribution of malignant lesions.

Cancer Type	Percentage	Count
DCIS	10.77%	7
IDC	80.00%	52
ILC	4.62%	3
Others[a]	4.62%	3

[a] It included two metaplastic carcinoma and one encapsulated papillary carcinoma.

Each study workstation was equipped with the BU-CAD™ system and standard image display functions including ruler, window level/window width, zoom in/out, and pan functions were provided to each reader. In addition, a function key was provided for the study reader to stop the study and shade the screen to measure real interpretation time in case the reader was interrupted.

Training for each reader consisted of two parts: a written guideline and one-on-one instruction. Twenty cases (8 malignant and 12 benign cases in random order) were used for each training session before the study. These training cases were not part of the reader study. All functionalities of the reading tool were demonstrated to the reader, and then the readers used the tools under observation to ensure their competence.

3.2. *Study scenario and workflow*

One hundred and seventy-two cases were randomized and split into dataset A and dataset B with the same number of cases in each group. Sixteen readers were randomized and split into two groups X and Y with the same number of readers in each group. For each reader, the study was conducted in 2 reading sessions. In the first session, each reader in group X interpreted dataset A in different random order first in the unaided scenario and then interpreted dataset B in different random order in the aided scenario. Each reader in group Y interpreted dataset A in different random order first in the aided scenario and then interpreted dataset B in different random order in the unaided scenario. In the second session, each reader in group X interpreted dataset A in different random order in the aided scenario and then interpreted dataset B in different random order in the unaided scenario. Each reader in group Y interpreted dataset A in different random order first in the unaided scenario and then interpreted dataset B in different random order in the aided scenario. Interpretations of any given case set with and without the AI support were separated by at least five weeks. The study workflow of the two study scenarios is depicted in Fig. 3.

The readers were blinded to any information on the study cases, including previous radiology, histopathology report, and mammographic findings. Each breast ultrasound case was presented with two orthogonal views of a single lesion. In the unaided scenario, each reader scored an SLC on a quasi-continuous scale of 0–100, a BI-RADS final assessment of BI-RADS Category (2, 3, 4A, 4B, 4C, or 5), and a set of BI-RADS descriptors including shape, orientation, margin, echo pattern, and posterior features. In the aided scenario, the parameters including ROI(s) and lesion contour(s) to indicate lesion location, SLC, BI-RADS Category, and BI-RADS descriptors were presented on the system. All these parameters were modifiable by the reader. After review by the reader, the parameters and interpretation time were recorded by the study workstation. The specific reading workflows for the unaided and aided scenarios are shown below.

3.2.1. *Unaided reading scenario*

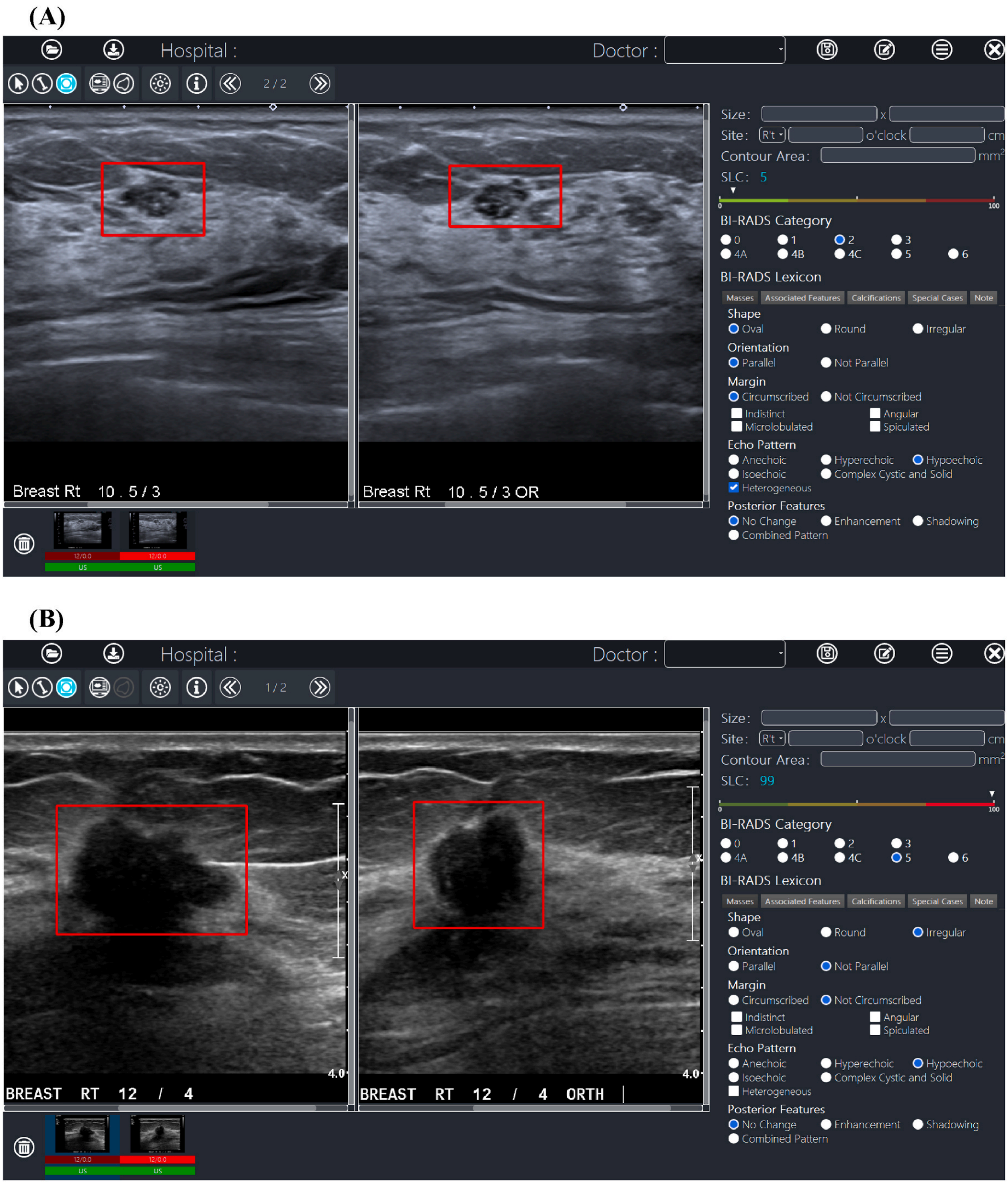
1. A breast ultrasound case with two orthogonal views showing the same lesion is presented to the reader.
2. The reader draws an ROI on each of the orthogonal views.
3. The reader assigns an SLC associated with its likelihood of malignancy/benignity (between 0 and 100).
4. The reader selects a BI-RADS Category (2, 3, 4A, 4B, 4C, and 5).
5. The reader selects the options of each BI-RADS descriptor (shape, orientation, margin, internal sonographic pattern, and posterior features).
6. The study workstation records the reader's assessments.

3.2.2. *Aided reading scenario*

1. A breast ultrasound case with two orthogonal views showing the same lesion is presented to the reader.
2. The system provides an automated ROI on each image for reference, which is completely changeable if the location of the ROI is unsatisfactory to the reader. In that case, the reader can perform the same operation as in the unaided scenario to create a rectangular ROI to encompass the lesion, including the edges, while minimizing the background.
3. The system outputs an SLC associated with its likelihood of malignancy or benignity (between 0 and 100), a BI-RADS category (i.e., 2, 3, 4A, 4B, 4C, and 5), and BI-RADS descriptors.
4. The reader modifies the SLC value, BI-RADS Category, and BI-RADS descriptors as desired.
5. The study workstation records the reader's assessments.

4. Evaluation

This study compared the reader's performance between the unaided and aided scenarios. To avoid the possibility that results were rewarded for malignant cases when the readers delineated the wrong locations of



(A)

Size: x

Site: R't o'clock cm

Contour Area: mm²

SLC: 5

BI-RADS Category: 0 1 [✓ 2] 3 4A 4B 4C 5 6

BI-RADS Lexicon

Shape: [✓ Oval] Round Irregular

Orientation: [✓ Parallel] Not Parallel

Margin: [✓ Circumscribed] Not Circumscribed
□ Indistinct □ Angular
□ Microlobulated □ Spiculated

Echo Pattern: Anechoic Hyperechoic [✓ Hypoechoic] Isoechoic Complex Cystic and Solid
☑ Heterogeneous

Posterior Features: [✓ No Change] Enhancement Shadowing Combined Pattern

(B)

Size: x

Site: R't o'clock cm

Contour Area: mm²

SLC: 99

BI-RADS Category: 0 1 2 3 4A 4B 4C [✓ 5] 6

BI-RADS Lexicon

Shape: Oval Round [✓ Irregular]

Orientation: Parallel [✓ Not Parallel]

Margin: Circumscribed [✓ Not Circumscribed]
□ Indistinct □ Angular
□ Microlobulated □ Spiculated

Echo Pattern: Anechoic Hyperechoic [✓ Hypoechoic] Isoechoic Complex Cystic and Solid
□ Heterogeneous

Posterior Features: [✓ No Change] Enhancement Shadowing Combined Pattern

Fig. 2. Illustrated Example of the Results of the BU-CAD™ System. (A) A benign case (fibrocystic change) of a 50-year-old woman with analysis results of SLC: 5; BI-RADS Category: 2; Shape: oval; Orientation: parallel; Margin: circumscribed; Echo Pattern: heterogeneously hypoechoic; Posterior Features: no change; (B) A malignant case (invasive ductal carcinoma) of a 55-year-old woman with analysis results of SLC: 99; BI-RADS Category: 5; Shape: irregular; Orientation: not parallel; Margin: not circumscribed; Echo Pattern: hypoechoic; Posterior Features: no change.

the lesions, a rating adjustment was performed [15]. For a patient case, the center of the ROI placed by the reader (or by the AI system for standalone performance evaluation) was within the ROI delineated by the expert panel and the overlapped ROI region was more than 50% of the ROI delineated by the expert panel, the rating was considered a true positive determination; otherwise, the rating was considered a false negative determination.

The primary endpoint of this MRMC reader study was to compare the reader's performance between the unaided and aided reading scenarios. Reader performance was quantified by the area under the LROC curve, which gave a continuous SLC rating score (0–100), but the wrong location for a malignant case was penalized as a false negative. Then, given an adjusted reader SLC, the trapezoidal method [16] was used to obtain the area under the LROC curve. For analysis of readers' performance, the OR-DBM MRMC-ROC (OR-DBM for short) model was applied using the MRMCaov package (version 2.51) in the R 4.1.2 software [17]. We specified both *reader* and *case* as random effects and *modality* (aided vs. unaided by the AI system) as a fixed effect in the OR-DBM model [18, 19].

Next, we conducted a linear regression analysis of the reader-specific

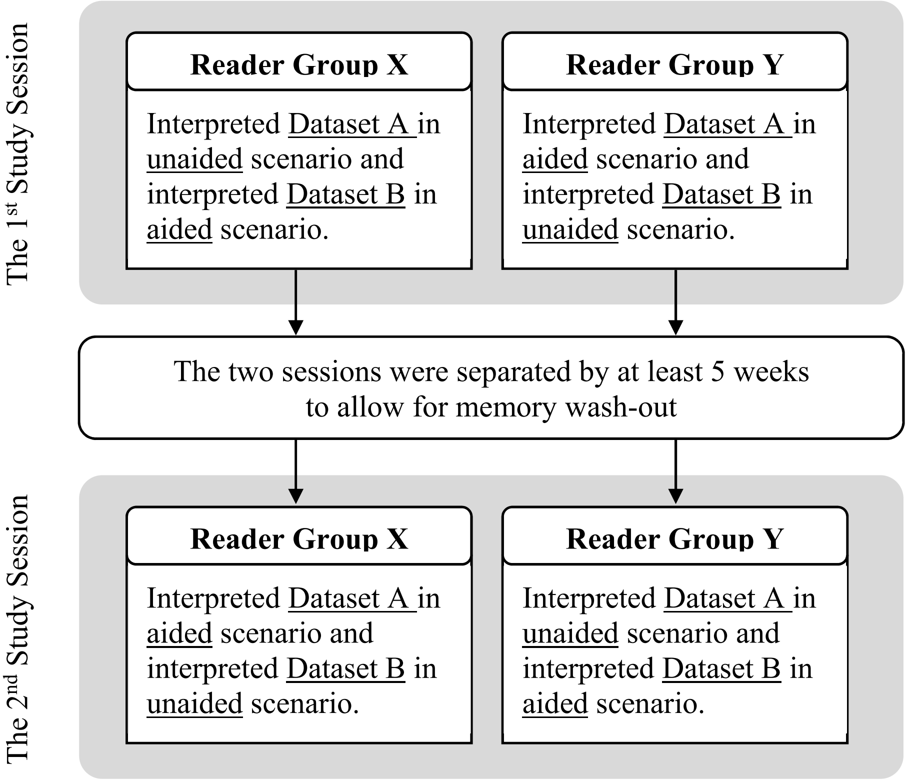
Table 3
Reader characteristics.

Reader	Specialty	MQSA	Received Breast Image Fellowship	Years of Experience as Radiologist/Breast Surgeon
Dr. X01	Radiologist	Yes	No	24
Dr. X02	Radiologist	Yes	Yes	3
Dr. X03	Radiologist	Yes	No	13
Dr. X04	Radiologist	Yes	No	14
Dr. X05	Radiologist	Yes	No	8
Dr. X06	Radiologist	Yes	Yes	5
Dr. X07	Radiologist	Yes	Yes	2
Dr. X08	Radiologist	Yes	No	10
Dr. X09	Radiologist	Yes	Yes	12
Dr. X10	Radiologist	Yes	No	11
Dr. X11	Breast Surgeon	No	No	>30 (breast surgeon)
Dr. X12	Breast Surgeon	No	No	>30 (breast surgeon)
Dr. X13	Radiologist	Yes	No	21
Dr. X14	Radiologist	Yes	No	1
Dr. X15	Radiologist	Yes	No	13
Dr. X16	Radiologist	Yes	No	5

data ($16 \times 2 = 32$) to assess the effects of reading type (unaided vs. aided) and reader characteristics (specialty, breast image fellowship, and years of work experience) on the AUC of ROC. The *logit* transformations (i.e., $\log(x/(1-x))$) were applied to the AUC of ROC for making its distribution more symmetric, where "log" was the natural logarithm. To account for the within-reader correlations in such clustered data, we fitted a linear regression model to the reader-specific data with the *generalized estimating equations* (GEE) method [20,21]. Computationally, assuming an *exchangeable* working correlation structure, we used the geeglm() function (with the default robust estimator of standard error) of the geepack package [23–25] to fit a GEE marginal linear regression model of correlated continuous responses in R. Since the sample size (i.e., the number of clusters) of 16 was small in this MRMC study, the robust estimate of standard error would preferably be obtained by computing the *fully iterated jackknife variance estimator* (specifying std.err = "fij" in the geeglm() function) in R.

To ensure the analysis quality, the model-fitting techniques for (1) variable selection, (2) goodness-of-fit (GOF) assessment, and (3) regression diagnostics and remedies were used in our GEE linear regression analysis. Specifically, the modern *stepwise variable selection procedure* (with the iterations between the *forward* and *backward* steps) was applied to obtain the best candidate final GEE marginal linear regression model. The available covariates, including the reading type (unaided vs. aided), reader characteristics (specialty, breast image fellowship, and years of work experience), and some of their interaction terms such as specialty × reading type, were placed on the variable list to be selected. The significance levels for entry (SLE) and stay (SLS) were both set to 0.25 due to the small sample size (i.e., the number of clusters) of 16 in this MRMC study. Then, with the aid of substantive knowledge, the best final GEE marginal linear regression model was identified manually by dropping the covariates with p value > 0.10 one at a time until all regression coefficients were significantly or borderline significantly different from 0.

The GOF measure, *coefficient of determination* (R^2), was examined to



Interpreted Dataset A in unaided scenario and interpreted Dataset B in aided scenario.

Interpreted Dataset A in aided scenario and interpreted Dataset B in unaided scenario.

The two sessions were separated by at least 5 weeks to allow for memory wash-out

Interpreted Dataset A in aided scenario and interpreted Dataset B in unaided scenario.

Interpreted Dataset A in unaided scenario and interpreted Dataset B in aided scenario.

Fig. 3. Study workflow of Two Study Scenarios.

assess the GOF of the fitted GEE marginal linear regression model. Technically, the R^2 statistic ($0 \leq R^2 \leq 1$) for a linear regression model is equal to the square of Pearson's correlation between the observed and predicted response values and indicates how much of the response variability is explained by the covariates included in the linear regression model.

Simple and multiple *generalized additive models* (GAMs) were fitted to draw the GAM plots for detecting nonlinear effects of continuous covariates and identifying the appropriate cut-off point(s) to discretize continuous covariates, if necessary, during the stepwise variable selection procedure. Computationally, we used the vgam() function with the default values of the smoothing parameters (e.g., s(age, df = 4, spar = 0) for the cubic smoothing splines) of the VGAM package to fit the GAMs for our continuous responses, and then used the plotvgam() function of the same package to draw the GAM plots for visualizing the linear or nonlinear effects of continuous covariates in R. Finally, the statistical tools of regression diagnostics for residual analysis, detection of influential cases, and check of multicollinearity were applied to discover any model or data problems. The value of the *variance inflating factor* (VIF) ≥ 10 in continuous covariates or ≥ 2.5 in categorical covariates indicates the occurrence of the multicollinearity problem among some of the covariates in the fitted linear regression model, but interaction terms inevitably enlarge VIF values.

Then, we performed secondary analyses of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and interpretation time. Sensitivity, specificity, PPV, and NPV were calculated according to the reference standard and the reader's rating of the BI-RADS category. Since the clinical decision threshold for cancer vs. non-cancer is BI-RADS 3 vs. BIRADS 4A and the BI-RADS fifth edition [13] concluded that patients with Category ≥ 4A lesions are recommended to undergo biopsy, the analysis of sensitivity, specificity, PPV, and NPV are still based on BI-RADS 4A as the cutoff point (i.e., a BI-RADS Category of 4A or higher defines a positive call for cancer diagnosis). To compare the sensitivity, specificity, PPV, and NPV between the aided and unaided reading scenarios, we applied the McNemar's test to compare two binomial proportions from the pair-matched data of size = 16 [26].

The interpretation time of each reading was automatically recorded by the study workstation. The research by Hupse et al. [27] appeared that the most experienced readers tend to repeat the exploration of the results of the AI analysis for prudence. Therefore, our study also calculated the interpretation time that excluded outliers. The interpretation time of a case longer than the reader's average interpretation time plus three times his/her standard deviation of the interpretation time was considered an outlier. The paired *t*-test [28] was used to examine the difference in the interpretation time between the two reading scenarios.

Finally, the area under the LROC curve, sensitivity, specificity, PPV, and NPV were also calculated to assess the standalone performance of the AI system.

5. Results

5.1. Reader LROC analysis

With the help of the AI system, 16 readers (14 radiologists and 2 breast surgeons) had higher diagnostic performance with increased AUC_{LROC} ranging from 0.0053 to 0.1223. The average AUC_{LROC} differed significantly from zero between the aided and unaided scenarios (unaided: 0.7582 vs. aided: 0.8294, $p < 0.0001$). Table 4 lists the AUC_{LROC} of each reader and the average AUC_{LROC} of the unaided and aided scenarios. Specifically, the AUC_{LROC} differed significantly from zero between the aided and unaided scenarios in the following 12 readers, Drs. X01, X03, X05, X07, X08, X10, X11, X12, X13, X14, X15, X16. The LROC curves for each reader and all readers are summarized in Fig. 4 and Fig. 5.

Table 4
The AUC_{LROC} for each reader and the average AUC_{LROC} for unaided and aided scenarios.

Reader	Unaided AUC_{LROC} (95% CI)	Aided AUC_{LROC} (95% CI)	*p*-Value
Dr. X01	0.7487 (0.6699, 0.8275)	0.8234 (0.7562, 0.8905)	0.0097*
Dr. X02	0.8230 (0.7607, 0.8852)	0.8377 (0.7784, 0.8971)	0.5401
Dr. X03	0.7333 (0.6564, 0.8102)	0.7925 (0.7250, 0.8600)	0.0452*
Dr. X04	0.7815 (0.7101, 0.8527)	0.8346 (0.7702, 0.8990)	0.1616
Dr. X05	0.7638 (0.6910, 0.8367)	0.8764 (0.8256, 0.9274)	0.0002*
Dr. X06	0.8533 (0.7938, 0.9129)	0.8777 (0.8262, 0.9293)	0.3659
Dr. X07	0.7907 (0.7199, 0.8615)	0.8440 (0.7836, 0.9043)	0.0378*
Dr. X08	0.7656 (0.6936, 0.8376)	0.8418 (0.7814, 0.9022)	0.0009*
Dr. X09	0.8185 (0.7556, 0.8815)	0.8238 (0.7626, 0.8850)	0.8449
Dr. X10	0.7421 (0.6678, 0.8164)	0.8131 (0.7509, 0.8754)	0.0239*
Dr. X11	0.7113 (0.6324, 0.7902)	0.8159 (0.7497, 0.8822)	0.0086*
Dr. X12	0.7374 (0.6602, 0.8146)	0.8306 (0.7690, 0.8922)	0.0130*
Dr. X13	0.7248 (0.6439, 0.8057)	0.8104 (0.7430, 0.8779)	0.0211*
Dr. X14	0.7017 (0.6219, 0.7815)	0.8052 (0.7404, 0.8700)	0.0242*
Dr. X15	0.7389 (0.6664, 0.8114)	0.8250 (0.7648, 0.8852)	0.0153*
Dr. X16	0.6972 (0.6174, 0.7770)	0.8195 (0.7541, 0.8848)	0.0001*
Average	0.7582 (0.7014, 0.8151)	0.8294 (0.7777, 0.8813)	<0.0001*

*The difference achieves statistical significance. That is the *p*-value calculated from the OR-DBM model is less than or equal to 0.05.

Breast density, which is a parameter of mammograms, changes with age [29]. On average, older women have lower density breast tissue than younger women, and the obvious demarcation occurs at menopause [30]. Because the density information was not available for all the cases in this study and Bissell et al. [31] recommended that observations were considered postmenopausal when 55 years or older in the absence of other information, our study conducted an alternative subgroup analysis using an age threshold of 55 years (Table 5). In the subgroup of 102 patients with age ≤55 years, the AUC_{LROC} differed significantly from zero between the aided and unaided scenarios (unaided: 0.7804, aided: 0.8454, $p = 0.0054$). By contrast, in the subgroup of 70 patients >55 years, the AUC_{LROC} also differed significantly from zero between the aided and unaided scenarios (unaided: 0.7282, aided: 0.8132, $p = 0.0021$). Additionally, subgroup analysis based on lesion size was performed. In the subgroup of 43 cases with lesion size less than 1 cm, the AUC_{LROC} differed significantly from zero between the aided and unaided scenarios (unaided: 0.7998, aided: 0.8428, $p = 0.0229$). In the subgroup of 74 cases with a lesion size between 1 and 2 cm, the AUC_{LROC} differed significantly from zero between the aided and unaided scenarios (unaided: 0.7438, aided: 0.8111, $p = 0.0207$). In the subgroup of 55 cases with lesion size larger than 2 cm, the AUC_{LROC} differed marginally significantly from zero between the aided and unaided scenarios (unaided: 0.7588, aided: 0.8092, $p = 0.0953$).

Next, to explain the observed variation in the AUC_{LROC} among the 16 readers in the unaided and aided scenarios (see Table 4), the results of the GEE marginal linear regression analysis of logit(AUC_{LROC}) are shown in Table 6. We found that after adjusting for the effects of the other covariates, the mean value of logit(AUC_{LROC}) increased by 0.7285 in the aided scenario compared to the unaided scenario ($p < 0.0001$), which was the most striking finding in this MRMC study. Specifically, after adjusting for the effects of the other covariates, we learned that.

(1) In the unaided scenario, the estimated mean value of logit (AUC_{LROC}) of the breast surgeon was 1.6741–0.7285 = 0.9456, the radiologist without the breast image fellowship was 1.6741–0.7285 + 0.1176 = 1.0632, and the radiologist with the breast image fellowship was 1.6741–0.7285 + 0.5630 = 1.5086, respectively.
(2) By contrast, in the aided scenario, the estimated mean value of logit(AUC_{LROC}) of all readers was 1.6741, but those with years of work experience ≤1.4 years or > 12.4 years were just 1.6741–0.1661 = 1.5080.

It is interesting to see that the readers in the aided scenario can perform as well as the radiologist with the breast image fellowship. The

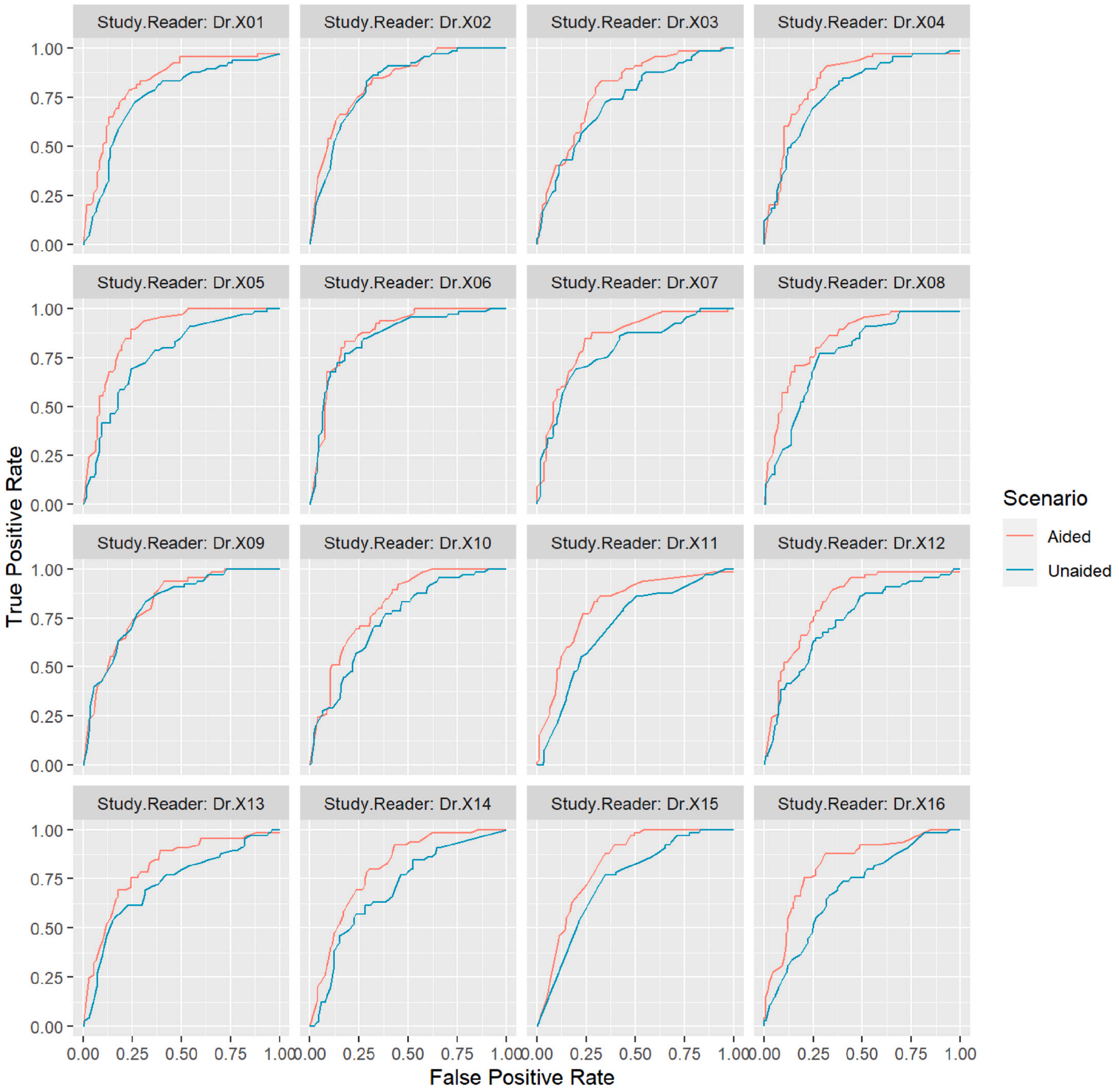


Fig. 4. LROC curves of unaided and aided scenarios for each reader.

value of the GOF measure $R^2 = 0.7975$ was quite high, indicating that Pearson's correlation between the observed and the predicted values of logit(AUCLROC) was $(0.7975)^{½} = 0.8930$.

5.2. Reader sensitivity, specificity, PPV, and NPV

In the unaided scenario, the mean ± standard error of sensitivity was 0.9577 ± 0.0250, ranging from 0.8154 to 1, with a median sensitivity of 0.9769. The mean ± standard error of specificity was 0.2407 ± 0.0413, ranging from 0.0187 to 0.6262, with a median specificity of 0.1963. The mean ± standard error of PPV was 0.4418 ± 0.0415, ranging from 0.3824 to 0.5699, with a median PPV of 0.4233. The mean ± standard error of NPV was 0.9350 ± 0.0462, ranging from 0.8462 to 1, with a median NPV of 0.9297.

In the aided scenario, the mean ± standard error of sensitivity was 0.9817 ± 0.0166, ranging from 0.9231 to 1, with a median sensitivity of 0.9846. The mean ± standard error of specificity was 0.3067 ± 0.0446, ranging from 0.0374 to 0.5327, with a median specificity of 0.3037. The mean ± standard error of PPV was 0.4691 ± 0.0425, ranging from 0.3832 to 0.5455, with a median PPV of 0.4660. The mean ± standard error of NPV was 0.9610 ± 0.0332, ranging from 0.8 to 1, with a median NPV of 0.9667. All indicators were improved in the aided scenario (Table 7) and the specificity shows significant a difference between the unaided and aided reading scenarios.

5.3. Interpretation time

A total of 192 outliers were identified in the 5504 readings. The

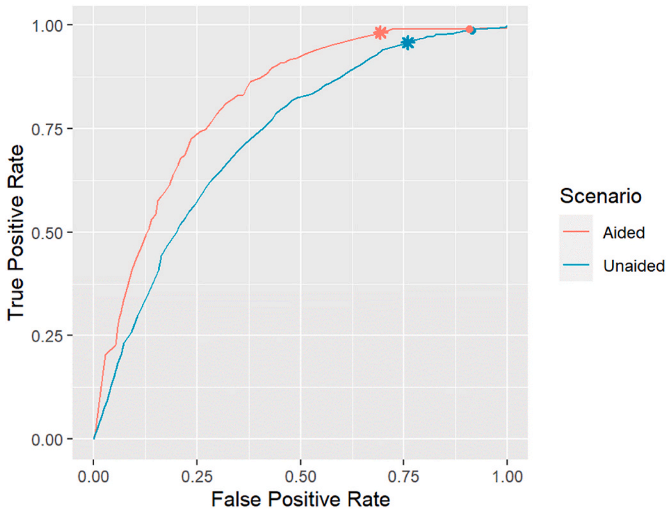


Fig. 5. Pooled LROC Curves of Unaided and Aided Scenarios for All Readers. The star denotes the performance corresponding to the cut-off threshold of BI-RADS 4A., i.e., positive if BI-RADS $\geq$ 4A; whereas the dot denotes the performance corresponding to the cut-off threshold of BI-RADS 3, i.e., positive if BI-RADS $\geq$ 3.

Table 5
AUC_{LROC} for each reader and average AUC_{LROC} for unaided and aided scenarios.

Characteristics	Unaided AUC_{LROC} (95% CI)	Aided AUC_{LROC} (95% CI)	*p*-Value
Age of Cut-off Point 55 Years			
≤55 years (102 cases)	0.7804 (0.7098, 0.8510)	0.8454 (0.7818, 0.9090)	0.0054*
>55 years (70 cases)	0.7282 (0.6362, 0.8202)	0.8132 (0.7270, 0.8993)	0.0021*
Lesion Size (cm)			
less than 1 cm (43 cases)	0.7998 (0.7005, 0.8991)	0.8428 (0.7461, 0.9394)	0.0229*
1–2 cm (74 cases)	0.7438 (0.6471, 0.8405)	0.8111 (0.7253, 0.8969)	0.0207*
larger than 2 cm (55 cases)	0.7588 (0.6607, 0.8569)	0.8092 (0.7058, 0.9125)	0.0953

*The difference achieves statistical significance. That is the *p*-value calculated from the OR-DBM model is less than or equal to 0.05.

average interpretation time of each reader and the other sample statistics for all cases are shown in Table 8 and Fig. 6. The average interpretation times of the 16 readers with and without outliers were compared between the aided and unaided scenarios using the paired *t*-test [28].

In the unaided scenario, the mean (with the 95% CI) ± standard error of interpretation time was 30.15 (9.92, 50.37) ± 10.31 s, ranging from 18.57 to 57.76 s, with a median interpretation time of 30.0 s. In the aided scenario, the mean (with the 95% confidence interval) ± standard error of the interpretation time was 18.11 (2.49, 33.72) ± 7.97 s, ranging from 9.16 to 40.94 s, with a median interpretation time of 16.11 s. The difference in interpretation time between the aided and unaided scenarios was 12.04 s per reading (39.9% interpretation time reduction) and was statistically significant ($p < 0.0001$).

When the outlier interpretations were excluded, the mean (with the 95% CI) ± standard error of the interpretation time in the unaided scenario was 28.54 (8.62, 48.46) ± 10.16 s, ranging from 17.73 to 56.39 s, with a median interpretation time of 28.95 s. In the aided scenario, the mean (with the 95% confidence interval) ± standard error of the interpretation time was 16.87 (2.29, 31.45) ± 7.44 s, ranging from 8.34 to 38.35 s, with a median of interpretation time of 15.09 s. The difference in interpretation time between the aided and unaided scenarios was 11.67 s per reading (40.9% interpretation time reduction) and was statistically significant ($p < 0.0001$).

Table 6
Marginal linear regression analysis of the logit-transformed AUCLROC over the 16 readers using the generalized estimating equations (GEE) method.[a]

Covariate[b]	Regression Coefficient Estimate	Robust Standard Error[c]	Wald's χ^b Test	*p*-Value[d]
Intercept	1.6741	0.0521	1034.3511	<0.0001
Scenario: Unaided vs. Aided	−0.7285	0.0660	121.8694	<0.0001
Radiologist × Fellowship: No × Scenario: Unaided	0.1176	0.0442	7.0891	0.0078
Radiologist × Fellowship: Yes × Scenario: Unaided	0.5630	0.0783	51.6320	<0.0001
Years of Experience ≤1.38 or >12.42 × Scenario: Aided	−0.1661	0.0554	8.9871	0.0027

[a] The clustered data of this MRMC study were analyzed by fitting a multiple marginal linear regression model with the *generalized estimating equations* (GEE) method (assuming an *exchangeable* correlation structure) to assess the effects of reading type (unaided vs. aided), reader characteristics (specialty, breast image fellowship, and years of work experience), and some of their interaction terms on the mean value of logit(AUCLROC). The number of clusters (n) = 16 with cluster size of 2, and thus the total number of observations (m) = 32. The coefficient of determination (R^b) = 0.7975 was quite high, indicating that Pearson's correlation between the observed and the predicted values of logit (AUCLROC) was $(0.7975)^{½}$ = 0.8930.

[b] The symbol " × " between two covariates was used to indicate an "interaction" and it can literally be interpreted as "and" for categorical variables in this table.

[c] Since the sample size (i.e., the number of clusters) of 16 was small in this MRMC study, the robust estimate of standard error would preferably be obtained by computing the *fully iterated jackknife variance estimator* (specifying std.err = "fij" in the geeglm() function) in R.

[d] The p values ≤ 0.05 indicate statistical significance.

Table 7
Reader sensitivity, specificity, PPV, and NPV.

Performance Measure	Unaided (95% CI)	Aided (95% CI)	*p*-Value
Sensitivity	95.77% (0.9088, 1.0066)	98.17% (0.9492, 1.0143)	0.2991
Specificity	24.07% (0.1597, 0.3217)	30.67% (0.2193, 0.3940)	0.0448*
PPV	44.18% (0.3606, 0.5231)	46.91% (0.3858, 0.5523)	0.0580
NPV	93.50% (0.8445, 1.0255)	96.10% (0.8959, 1.0261)	0.0580

*The difference achieves statistical significance. That is the *p*-value calculated from McNemar's test is less than or equal to 0.05.

5.4. *Influence of AI system on readers*

The readers decided to accept or modify the AI-derived results based on comprehensive consideration of their own medical expertise and the recommendation of the AI system. To investigate the influence of the AI system on the readers, the number of readings that the reader's BI-RADS category changed after the artificial intelligence system was recommended is listed in Table 9. In this study, the readers increased a total of 37 true positive readings, of which the readers accepted 34 recommendations from the AI system. The net increase in true positive readings was 29. For the 217 increased true negative readings, the readers accepted the 187 readings suggested by the AI system. The net increase in true negative readings was 113.

Table 8
The interpretation time for unaided and aided scenarios.

Sample Statistic	Interpretation Time (Seconds)		
	Unaided Scenario	Aided Scenario	Difference, *p*-Value [*]
All Readings			
Mean	30.15	18.11	12.04 (39.9%), $p <$ 0.0001
Standard Error	10.31	7.97	–
Median	30.0	16.11	–
Min	18.57	9.16	–
Max	57.76	40.94	–
95% Lower CI	9.92	2.49	8.87
95% Upper CI	50.37	33.72	15.21
All Readings Exclude Outlier			
Mean	28.54	16.87	11.67 (40.9%), $p <$ 0.0001
Standard Error	10.16	7.44	–
Median	28.95	15.09	–
Min	17.73	8.34	–
Max	56.39	38.35	–
95% Lower CI	8.62	2.29	8.74
95% Upper CI	48.46	31.45	14.60

*The p value was obtained from the paired t-test ($df = 15$).

5.5. Standalone performance of AI system

Of the 172 study cases, six benign cases (including 4 cases not identified by the AI system) and three malignant cases (including 1 case not identified by the AI system) failed to pass the location correctness determination. Therefore, the accuracy of the lesion identification algorithm was 94.77%. Further exploration of these fail-identified cases revealed that they were non-mass lesions that do not have clear boundaries on ultrasound images. Fig. 7 shows one example of benign lesion with a non-mass-like pattern and architectural distortion that the AI system did not identify. Fig. 8 shows an example of malignant lesion with a no-mass-like pattern and microcalcifications. The AI system falsely identified the fatty tissue.

Table 10 lists the system's standalone performances with and without adjustment for the wrong-location penalty. With adjustment for the wrong location penalty, the AUC under the LROC curve (AUC_{LROC}) was 0.8384 with 95% CI = (0.7726, 0.9041). The sensitivity, specificity, PPV, and NPV were 93.85%, 55.14%, 55.96%, and 93.65%, respectively. Without adjustment for wrong-location penalty, the AUC under the ROC curve (AUC_{ROC}) was 0.8591 with 95% CI = (0.8028, 0.9155).

Table 9
Reader determination changes after the AI system recommendation.

AI System Recommendation	Changes in Reader Determination			
	Malignant Lesion		Benign Lesion	
	Incorrect to Correct	Correct to Incorrect	Incorrect to Correct	Correct to Incorrect
No biopsy[a]	1	3	187	42
Biopsy[b]	34	3	21	54
Fail-identified	2	2	9	8
Total	37	8	217	104

[a] The AI system recommended BI-RADS Category 2 or 3.
[b] The AI system recommended BI-RADS Category 4A, 4B, 4C, or 5.

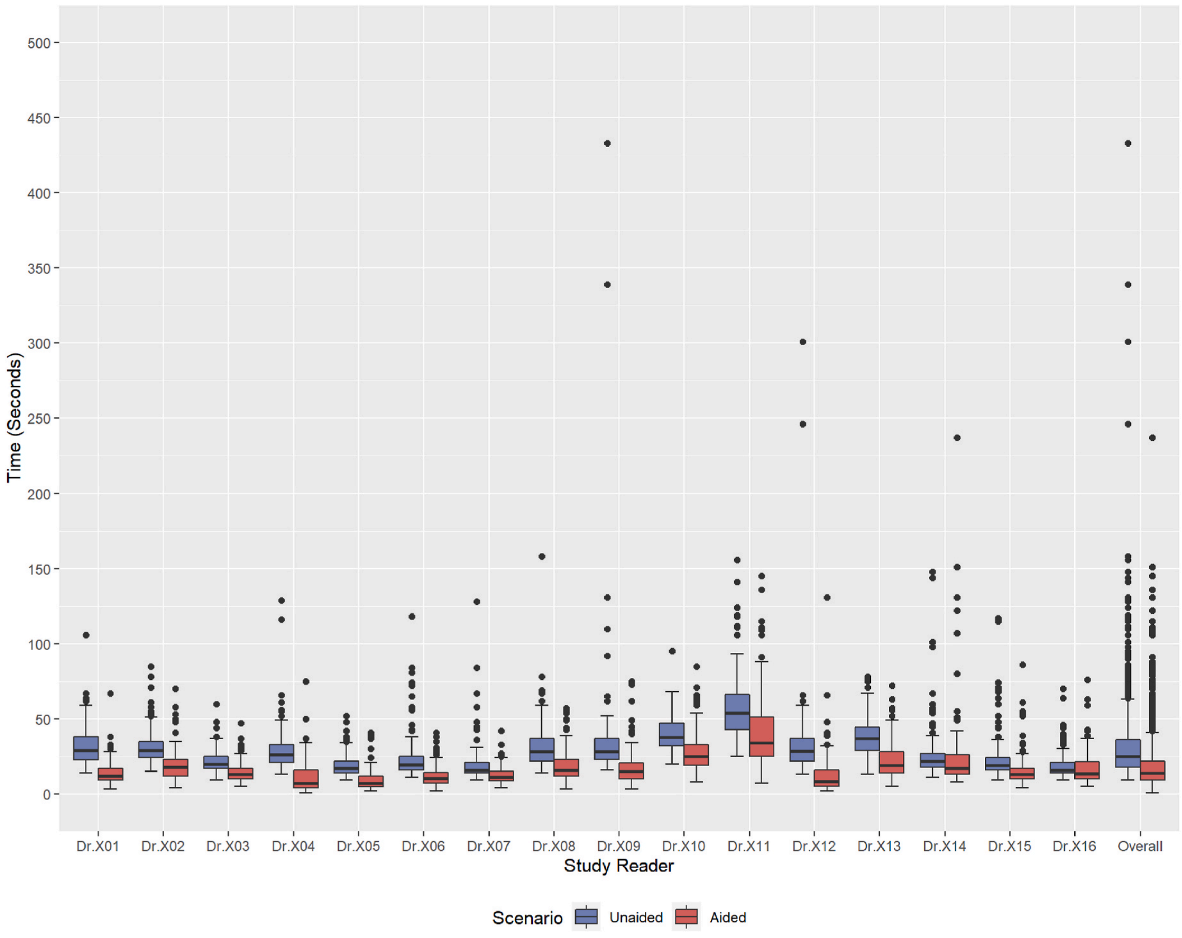


Fig. 6. Boxplots of the interpretation time for each reader.

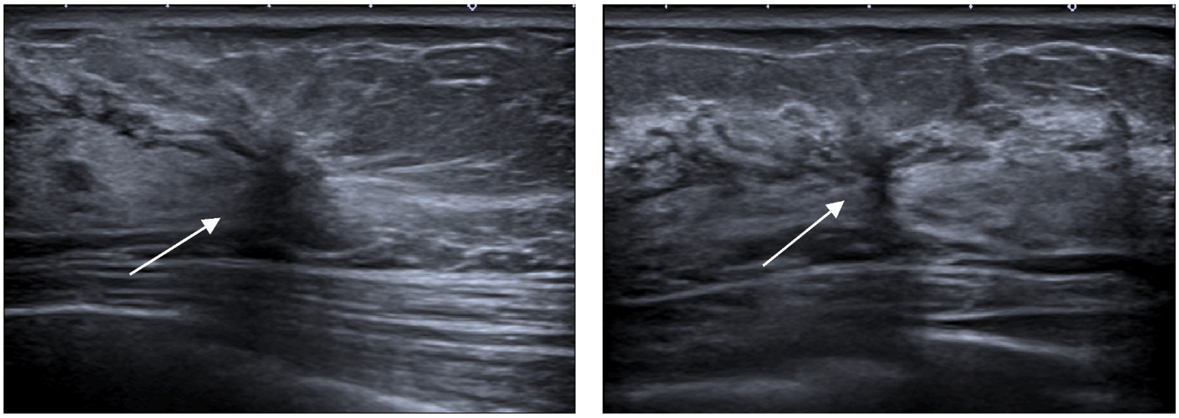


Fig. 7. Two Orthogonal Views: A fifty-one-year-old woman had a non-mass benign lesion (white arrow) with architectural distortion in her right breast, which was not identified by the AI system.

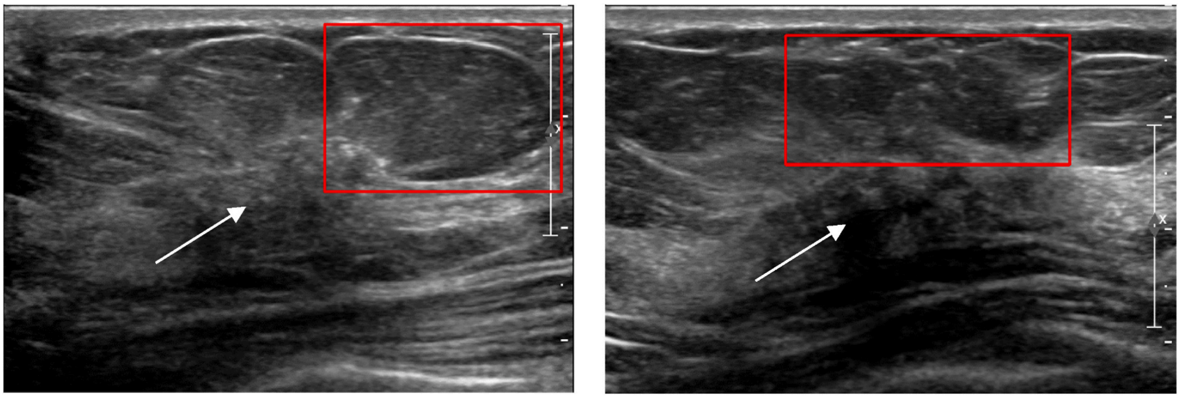


Fig. 8. Two Orthogonal Views: A fifty-six-year-old woman had a non-mass malignant lesion (white arrow) with microcalcifications in her left breast. The fatty tissue was falsely identified by the AI system.

Table 10
System standalone performance.

Statistical Parameter	Without Adjustment for Wrong-Location Penalty	With Adjustment for Wrong-Location Penalty
AUC	0.8591 (AUC_{ROC})	0.8384 (AUC_{LROC})
Sensitivity	96.92%	93.85%
Specificity	55.14%	55.14%
PPV	56.76%	55.96%
NPV	96.72%	93.65%

The sensitivity, specificity, PPV, and NPV were 96.92%, 55.14%, 56.76%, and 96.72%, respectively.

6. Discussion

In our study, when assisted by the AI system, readers effectively improved their diagnostic performance with increased average AUC_{LROC} that provides an aggregate measure of performance across all possible thresholds. For individual readers, the AUC_{LROC} value of each reader is improved, but four readers do not achieve statistical significance. Further analysis revealed that three of the four non-significant improved readers had completed breast imaging fellowship training receiving in-depth training in mammography, tomosynthesis, breast ultrasound, breast MRI for screening and diagnosis, as well as image-guided interventions. Therefore, their unaided AUC_{LROC} of 0.8230 (Dr. X02), 0.8533 (Dr. X06), and 0.8185 (Dr. X09) are among the top three of all readers. It is expected that AI support has less improvement for experienced readers, since their diagnostic performances are already relatively high in the unaided scenario.

In the subgroup analyses, the readers improved their diagnostic performance in the age subgroups (age ≤55 years and age >55 years) and the lesion size subgroups (less than 1 cm, between 1 and 2 cm, and larger than 2 cm). Except for the subgroup with lesion sizes larger than 2 cm, the improvements were statistically significant in all subgroups. The readers' sensitivity, specificity, PPV, and NPV were also improved, especially the specificity, reached a statistically significant improvement. Additionally, the average interpretation time was significantly reduced by approximately 40% when readers were aided by the AI system.

The system's standalone performance presents the performance evaluations based on the automated ROIs. Due to the performance of automated ROIs that achieved an accuracy of 94.77%, the system standalone performance adjusted for the wrong-location penalty was similar to the unadjusted performances. Both achieved AUC above 0.8 (AUC_{ROC}: 0.8591 vs. AUC_{LROC}: 0.8384). By comparing the system standalone performance and the unaided reading scenario, it is noted that the AUC_{LROC} (0.8384) of the AI system was higher than every reader's AUC_{LROC} value for those without breast imaging fellowship training, with an unaided average AUC_{LROC} of 0.7582, ranging from 0.6972 to 0.7815. Their average performance in AUC_{LROC} was lifted to 0.8240 by the assistance of the AI system, which was close to the performance of readers with breast imaging fellowship training (AUC_{LROC}: 0.8214) in the unaided reading scenario.

Clinically, cancer cell growth and division are cell proliferation activities that are usually accompanied by angiogenesis and infiltration of surrounding tissues [32,33], and thus the variation of the ROIs range delineated by the readers can affect the system analysis results. Most studies [11,12] adopted the original ROIs placed by the original interpreting physician from the data source to reduce the intra-reader variability and focused on the diagnostic results. In our study, although the AI system provided automated ROIs for assisting users in detecting the location of breast soft tissue lesions, the users were still allowed to replace the automated ROIs with re-delineated ROIs for further analysis by the AI system. To evaluate the robustness of the AI system when different rectangular ROIs were drawn around the same lesion, two reproducibility experiments of the same lesion cropped by different rectangular ROIs were conducted. In the first reproducibility experiment, each corner point of an ROI was shifted by randomly changing the horizontal and vertical dimensions up to 20% respectively from the reference standard ROI defined by the expert panel. The experiment was repeated 20 times and the results show that AUCs remained stable between 0.8525 and 0.8682 (Fig. 9). In the second reproducibility experiment, each corner point of the reference standard ROI was altered by systematically shrinking the horizontal and vertical dimensions respectively from 1% to 30%. The results showed that as long as the shrinking percentage of the width and height of the ROIs is within 15%, the AUC remained above 0.8 (Fig. 10).

There are some limitations of this study. First, the fifth edition of BI-RADS concluded that patients with BI-RADS Category ≥ 4A lesions are recommended to undergo biopsy, while most BI-RADS Category 3 lesions that have a probability of malignancy less than 2% are recommended for follow-up instead of biopsy to reduce the number of false-positive biopsies. In this study, not all benign lesions were pathologically proven that exists a slight likelihood of false negatives. Secondly, the retrospectively collected cases were scanned by 11 operators. Although different operators are more compliant with the real clinical scenario, ultrasound is still an operator-dependent imaging technique that would be a variable in imaging quality. Lastly, in clinical practice, the physicians report a breast cancer diagnosis decision by referencing relevant clinical information such as patient age, family medical history, mammography report, and context of patient symptoms. This retrospective reader study only focused on assessing the assistance of the AI system to readers in breast ultrasound image interpretation.

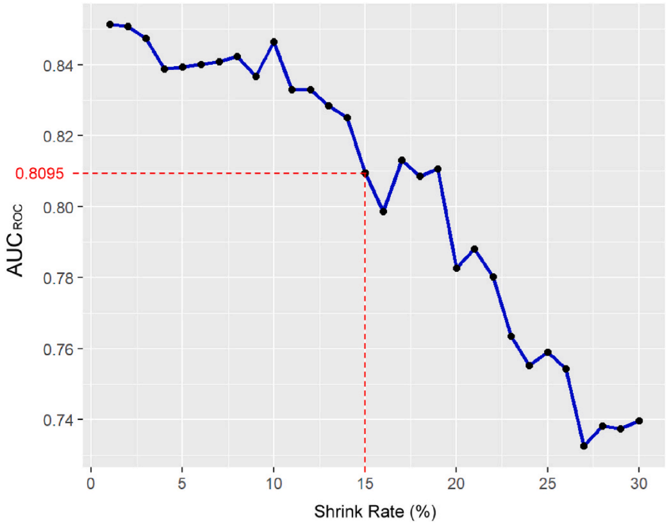


Fig. 10. Variations of AUC in the reproducibility (shrinking) experiment.

7. Conclusions

In conclusion, radiologists and breast surgeons improved their diagnostic performance in detecting and diagnosing breast lesions on breast ultrasound images with the assistance of the AI system. In particular, the AI system was able to help breast surgeons and radiologists who have no breast imaging fellowship training to improve their diagnostic performance to the level of radiologists specializing in breast imaging those who have received breast imaging fellowship training and read breast images daily. Additionally, the interpretation time was reduced by approximately 40% by the AI system support. However, even if the AI system is helpful in this study, the practical application in a real clinical environment should be further explored.

Declaration of competing interest

This work was supported in part by a research agreement between

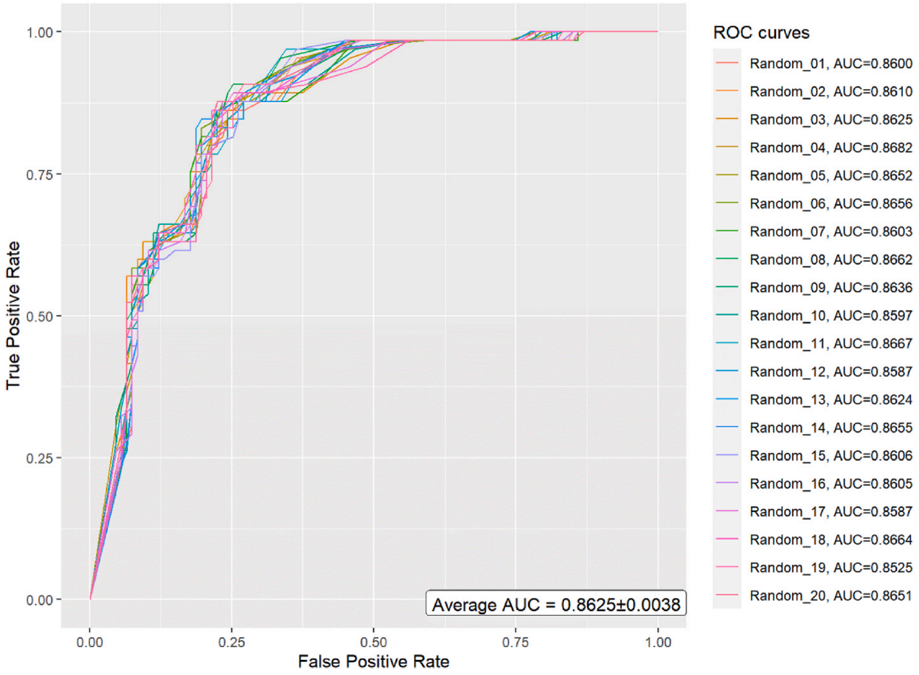


Fig. 9. ROC curves and AUC values of the reproducibility (enlargement) experiment.

TaiHao Medical and Taipei Veterans General Hospital. H. H. Chen, J. F. Hsu, Y. J. Hong, and T. T. Chiu are employed by TaiHao Medical. All authors have no other financial interest that could have appeared to influence the work reported in this paper.

Acknowledgements

The clinical study was sponsored by the TaiHao Medical Inc. (Taipei, Taiwan) providing the AI system BU-CAD™ for evaluation. We would like to thank Dr. Fu-Chang Hu and Ms. Fang-Yu Wen in the International-Harvard Statistical Consulting Company (Taipei, Taiwan) for the GEE linear regression analysis of the $\mathrm{AUC}_{\mathrm{LROC}}$.

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA A Cancer J Clin Jan. 2022;72(1):7–33. https://doi.org/10.3322/CAAC.21708.

[2] Jones BA, Patterson EA, Calvocoressi L. Mammography screening in African American women. Cancer Jan. 2003;97(S1):258–72. https://doi.org/10.1002/CNCR.11022.

[3] Gilbert FJ, Pinker-Domenig K. Diagnosis and staging of breast cancer: when and how to use mammography, tomosynthesis, ultrasound, contrast-enhanced mammography, and magnetic resonance imaging. 2019. p. 155–66. https://doi.org/10.1007/978-3-030-11149-6_13.

[4] Hendrick RE, Helvie MA, Monticciolo DL. Breast cancer mortality rates have stopped declining in U.S. Women younger than 40 years. Radiology Apr. 2021;299 (1):143–9. https://doi.org/10.1148/RADIOL.2021203476/ASSET/IMAGES/LARGE/RADIOL.2021203476.FIG4C.JPEG.

[5] Checka CM, Chun JE, Schnabel FR, Lee J, Toth H. The relationship of mammographic density and age: implications for breast cancer screening. AJR Am J Roentgenol Mar. 2012;198(3). https://doi.org/10.2214/AJR.10.6049.

[6] Ji Y, Li B, Zhao R, Zhang Y, Liu J, Lu H. The relationship between breast density, age, and mammographic lesion type among Chinese breast cancer patients from a large clinical dataset. BMC Med Imag Dec. 2021;21(1). https://doi.org/10.1186/S12880-021-00565-9.

[7] Thigpen D, Kappler A, Brem R. The role of ultrasound in screening dense breasts—a review of the literature and practical solutions for implementation. Diagnostics Mar. 2018;8(1). https://doi.org/10.3390/DIAGNOSTICS8010020.

[8] Yang S, et al. Performance and reading time of automated breast us with or without computer-aided detection. Radiology Jun. 2019;292(3):540–9. https://doi.org/10.1148/RADIOL.2019181816/ASSET/IMAGES/LARGE/RADIOL.2019181816.FIG6.JPEG.

[9] Jiang Y, Inciardi MF, Edwards Av, Papaioannou J. Interpretation time using a concurrent-read computer-aided detection system for automated breast ultrasound in breast cancer screening of women with dense breast tissue. AJR Am J Roentgenol Aug. 2018;211(2):452–61. https://doi.org/10.2214/AJR.18.19516.

[10] O'Connell AM, Bartolotta Tv, Orlando A, Jung SH, Baek J, Parker KJ. Diagnostic performance of an artificial intelligence system in breast ultrasound. J Ultrasound Med Jan. 2022;41(1):97–105. https://doi.org/10.1002/JUM.15684.

[11] Barinov L, et al. Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems. J Digit Imag Jun. 2019;32(3):408–16. https://doi.org/10.1007/S10278-018-0132-5.

[12] Mango VL, Sun M, Wynn RT, Ha R. Should we ignore, follow, or biopsy? Impact of artificial intelligence decision support on breast ultrasound lesion assessment. AJR Am J Roentgenol Jun. 2020;214(6):1445–52. https://doi.org/10.2214/AJR.19.21872.

[13] Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS® fifth edition: a summary of changes. Diagnostic and Interventional Imaging Mar. 2017;98(3): 179–90. https://doi.org/10.1016/J.DIII.2017.01.001.

[14] Tian Z, Shen C, Chen H. Conditional convolutions for instance segmentation. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). vol. 12346. LNCS; Aug. 2020. p. 282–98. https://doi.org/10.1007/978-3-030-58452-8_17.

[15] Rodríguez-Ruiz A, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology Mar. 2019;290(3):305–14. https://doi.org/10.1148/RADIOL.2018181371/ASSET/IMAGES/LARGE/RADIOL.2018181371.FIG6B.JPEG.

[16] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics Sep. 1988;44(3):837. https://doi.org/10.2307/2531595.

[17] Smith BJ, Hillis SL. Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities. Proc SPIE-Int Soc Opt Eng Mar. 2020;11316:18. https://doi.org/10.1117/12.2549075.

[18] Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations Jan. 2007;24(2):285–308. https://doi.org/10.1080/03610919508813243. doi: 10.1080/03610919508813243.

[19] Hillis SL, Schartz KM. Multireader sample size program for diagnostic studies: demonstration and methodology. J Med Imaging Nov. 2018;5(4):1. https://doi.org/10.1117/1.JMI.5.4.045503.

[20] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Accessed: Jul. 11, 2022. [Online]. Available: https://academic.oup.com/biomet/article/73/1/13/246001; 1986. 13-22.

[21] Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics Sep. 1991;47(3):825. https://doi.org/10.2307/2532642.

[23] Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations. J Stat Software 2006;15(2):1–11. https://doi.org/10.18637/JSS.V015.I02.

[24] Yan J, Fine J. Estimating equations for association structures. Stat Med Mar. 2004; 23(6):859–74. https://doi.org/10.1002/SIM.1650.

[25] Yan J. Geepack: yet another package for generalized estimating equations, 2/3. Request PDF," *R-News*; 2002. p. 12–4.

[26] Le CT, Eberly LE. Introductory biostatistics. 2016.

[27] Hupse R, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology Jan. 2013;266(1):123–9. https://doi.org/10.1148/RADIOL.12120218/ASSET/IMAGES/LARGE/120218T03.JPEG.

[28] Hsu H, Lachenbruch PA. Paired t test. Wiley StatsRef: Statistics Reference Online; Sep. 2014. https://doi.org/10.1002/9781118445112.STAT05929.

[29] Mokhtary A, Karakatsanis A, Valachis A. Mammographic density changes over time and breast cancer risk: a systematic review and meta-analysis. Cancers Oct. 2021; 13(19). https://doi.org/10.3390/CANCERS13194805/S1.

[30] Burton A, et al. Mammographic density and ageing: a collaborative pooled analysis of cross-sectional data from 22 countries worldwide. PLoS Med Jun. 2017;14(6): e1002335. https://doi.org/10.1371/JOURNAL.PMED.1002335.

[31] Bissell MCS, et al. Breast cancer population attributable risk proportions associated with body mass index and breast density by race/ethnicity and menopausal status. Cancer Epidemiol Biomark Prev Oct. 2020;29(10):2048–56. https://doi.org/10.1158/1055-9965.EPI-20-0358/70488/AM/BREAST-CANCER-POPULATION-ATTRIBUTABLE-RISK.

[32] Moon WK, Chen HH, Shin SU, Han W, Chang RF. Evaluation of TP53/PIK3CA mutations using texture and morphology analysis on breast MRI. Magn Reson Imaging Nov. 2019;63:60–9. https://doi.org/10.1016/J.MRI.2019.08.026.

[33] Chang RF, Chen HH, Chang YC, Huang CS, Chen JH, Lo CM. Quantification of breast tumor heterogeneity for ER status, HER2 status, and TN molecular subtype evaluation on DCE-MRI. Magn Reson Imaging Jul. 2016;34(6):809–19. https://doi.org/10.1016/J.MRI.2016.03.001.