Radiology

ORIGINAL RESEARCH • BREAST IMAGING

Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System

Alejandro Rodríguez-Ruiz, MSc • Elizabeth Krupinski, PhD • Jan-Jurre Mordang, MSc • Kathy Schilling, MD • Sylvia H. Heywang-Köbrunner, MD, PhD • Ioannis Sechopoulos, PhD • Ritse M. Mann, MD, PhD

From the Department of Radiology and Nuclear Medicine, Radboud University Medical Center, PO Box 9101, 6500 HB Nijmegen, Geert Grooteplein 10, 6525 GA, Post 766, Nijmegen, the Netherlands (A.R.R., I.S., R.M.M.); Department of Radiology & Imaging Sciences, Emory University, Atlanta, Ga (E.K.); ScreenPoint Medical BV, Nijmegen, the Netherlands (J.J.M.); Lynn Women's Health & Wellness Institute, Boca Raton Regional Hospital, Boca Raton, Fla (K.S.); Referenzzentrum Mammographie Munich, Brustdiagnostik München and FFB, Munich, Germany (S.H.H.); and Dutch Expert Centre for Screening, Nijmegen, the Netherlands (I.S.). Received June 10, 2018; revision requested July 30; final revision received September 21; accepted September 28. **Address correspondence to** R.M.M. (e-mail: *Ritse.Mann@radboudumc.nl*).

Conflicts of interest are listed at the end of this article.

See also the editorial by Bahl in this issue.

Radiology 2019; 00:1–10 • https://doi.org/10.1148/radiol.2018181371 • Content code: BR

Purpose: To compare breast cancer detection performance of radiologists reading mammographic examinations unaided versus supported by an artificial intelligence (AI) system.

Materials and Methods: An enriched retrospective, fully crossed, multireader, multicase, HIPAA-compliant study was performed. Screening digital mammographic examinations from 240 women (median age, 62 years; range, 39–89 years) performed between 2013 and 2017 were included. The 240 examinations (100 showing cancers, 40 leading to false-positive recalls, 100 normal) were interpreted by 14 Mammography Quality Standards Act–qualified radiologists, once with and once without AI support. The readers provided a Breast Imaging Reporting and Data System score and probability of malignancy. AI support provided radiologists with interactive decision support (clicking on a breast region yields a local cancer likelihood score), traditional lesion markers for computer-detected abnormalities, and an examination-based cancer likelihood score. The area under the receiver operating characteristic curve (AUC), specificity and sensitivity, and reading time were compared between conditions by using mixed-models analysis dof variance and generalized linear models for multiple repeated measurements.

Results: On average, the AUC was higher with AI support than with unaided reading (0.89 vs 0.87, respectively; $P = .002$). Sensitivity increased with AI support (86% [86 of 100] vs 83% [83 of 100]; $P = .046$), whereas specificity trended toward improvement (79% [111 of 140]) vs 77% [108 of 140]; $P = .06$). Reading time per case was similar (unaided, 146 seconds; supported by AI, 149 seconds; $P = .15$). The AUC with the AI system alone was similar to the average AUC of the radiologists (0.89 vs 0.87).

Conclusion: Radiologists improved their cancer detection at mammography when using an artificial intelligence system for support, without requiring additional reading time.

Breast cancer screening with mammography is considered effective in reducing breast cancer–related mortality (1,2). However, the large number of women screened and the use of double reading of examinations in some countries creates a high workload that poses a threat to efficiency, especially considering the increasing scarcity of screening radiologists (3). Moreover, it is important to minimize misses and interpretation errors of visible lesions at digital mammography, which contribute to at least 25% of detectable cancers being missed (4–7).

Computer-aided detection (CAD) systems were introduced as an aid for radiologists trying to improve human detection performance. Although some studies indicated that single reading plus CAD could be an alternative to double reading (8–11), few, if any, have identified the actual benefit of using single reading plus CAD versus single reading alone (ie, the actual benefit on radiologists' performance in screening) (12). In general, the benefit of using CAD in screening is still unclear. Most evidence shows no clear improvement in the cost-effectiveness of screening, mainly because of the low specificity of most traditional CAD systems (12–14).

However, substantial improvements in artificial intelligence (AI) with deep convolutional neural networks (commonly known as deep learning algorithms) are reducing the difference in performance between humans and computers in many medical imaging applications (15), including breast cancer detection (16). Therefore, this new generation of deep learning–based CAD systems may finally allow for an improvement in the performance of breast cancer screening programs (17). Apart from the evolution of AI algorithms, the aid that the AI system provides can also help improve screening. Previous studies have shown that using CAD concurrently as a decision support tool helps radiologists more than does the traditional approach with prompts for assessing soft-tissue lesions (18,19).

The benefit, if any, of interactive AI-based systems on radiologists' performance remains to be assessed in terms of overall diagnostic performance and efficiency. The purpose of this study was to compare breast cancer detection performance of radiologists reading mammographic images unaided versus supported by a commercially available AI system.

Abbreviations

AI = artificial intelligence, AUC = area under the ROC curve, BI-RADS = Breast Imaging Reporting and Data System, CAD = computer-aided detection, POM = probability of malignancy, ROC = receiver operating characteristic

Summary

Radiologists had improved diagnostic performance for detection of breast cancer at mammography when using an artificial intelligence computer system for support, with no additional reading time required.

Implications for Patient Care

- An artificial intelligence support system for mammography improved radiologists' breast cancer detection, without lengthening their reading time.
- Improvement was observed for all breast density categories and was independent of lesion type and vendor image quality.

Materials and Methods

This retrospective study was compliant with the Health Insurance Portability and Accountability Act. Our study was performed with anonymized, retrospectively collected digital mammographic images obtained from screening examinations. Women were included from two institutions: one in the United States (collection center A) and one in Europe (collection center B). The requirement to obtain informed consent and ethical approval to use anonymized data was waived after review of the institutional review board at collection center A and under national law at collection center B. The study was financially supported by ScreenPoint Medical (Nijmegen, the Netherlands). The authors who were not employees of or consultants for ScreenPoint Medical had control of the data and information submitted for publication at all times.

Study Population

The flowchart of collection and final selection of digital mammographic examinations is detailed in Figure 1. First, the sample size and examination type distribution for our observer evaluation study population were estimated on the basis of the results of a similar previous study (18), by using the unified method proposed by Hillis et al (20), to yield a study power greater than 0.8. This resulted in a target data set of 240 digital mammographic examinations (100 showing cancer, 40 with false-positive results, and 100 with normal results).

Mammogram collection.—To ensure there were enough digital mammographic examinations from which to select the final sample, at least 55 examinations showing cancer, 30 examinations with false-positive results, and 60 examinations with normal results were set to be collected by each collection center. For collection, the single inclusion criterion was women presenting for screening with no symptoms or concerns. Women with implants and/or a history of breast cancer were excluded. Digital mammographic examinations were consecutively collected: From collection center A (performed by K.S.), examinations were retrieved over several blocks of samples between June 2013 and March 2017; from collection center B (performed by S.H.H.) a single block of examinations acquired between January 2014 and February 2015 was retrieved. Examinations from each type (cancer, false-positive, normal) were consecutively collected until the numbers defined above were met. A total of 546 digital mammographic examinations were collected (110 showing cancer, 76 with false-positive recalls, and 360 with normal results). For each collected examination, a case report form was obtained, detailing patient demographic characteristics, lesion characteristics, and histopathologic features (Fig 1).

Mammogram selection.—To ensure appropriate image quality, all 546 collected examinations were reviewed by one radiologist (R.M.M., with 13 years of experience with digital mammography) who did not participate in the observer study. Nine cancer examinations were excluded during this revision (three because of poor image quality, three because it was not possible to link the case report form findings to the digital mammography examination, and three because the examinations showed extremely obvious signs of breast cancer). From the remaining data, a randomized selection was performed to meet the predefined distribution of examinations. The same number of examinations was included from each collection center.

Population characteristics.—The characteristics of the populations and the digital mammographic examinations included for the observer study are shown in Table 1. All digital mammographic examinations were bilateral and contained two views (craniocaudal and mediolateral oblique). The digital mammographic examinations were performed with two different systems: a Lorad Selenia unit (Hologic, Bedford, Mass) at collection center A and a Mammomat Inspiration unit (Siemens Healthineers, Erlangen, Germany) at collection center B. Previous digital mammographic examinations were included for evaluation if available (192 women underwent previous examinations: 76 of those with cancer, 37 with false-positive results, and 79 with normal results).

Cancers were verified by means of histopathologic evaluation (Table 2), and false-positive findings were verified with histopathologic evaluation (n = 11) or with negative follow-up findings for at least 1 year (n = 29). All normal examinations had at least 1 year of negative follow-up findings. Seventy cancers manifested as soft-tissue lesions (including mass lesions, architectural distortions, and asymmetries, which were grouped together because of the relatively low number of the latter two categories) and 35 as calcifications (five lesions presented both soft-tissue lesions and calcifications).

The reference standard for each digital mammographic examination was established by an experienced breast radiologist (R.M.M.) with access to the case report form. Each examination was defined as showing cancer, a false-positive result, or a normal result. The location in all views (lesions were delineated) and characterization (morphologic appearance and histologic features) of cancers and of findings that led to false-positive recalls was recorded. According to the reference standard, the median size of the cancers at mammography was 13 mm^2 (interquartile range, 4–22 mm^2).

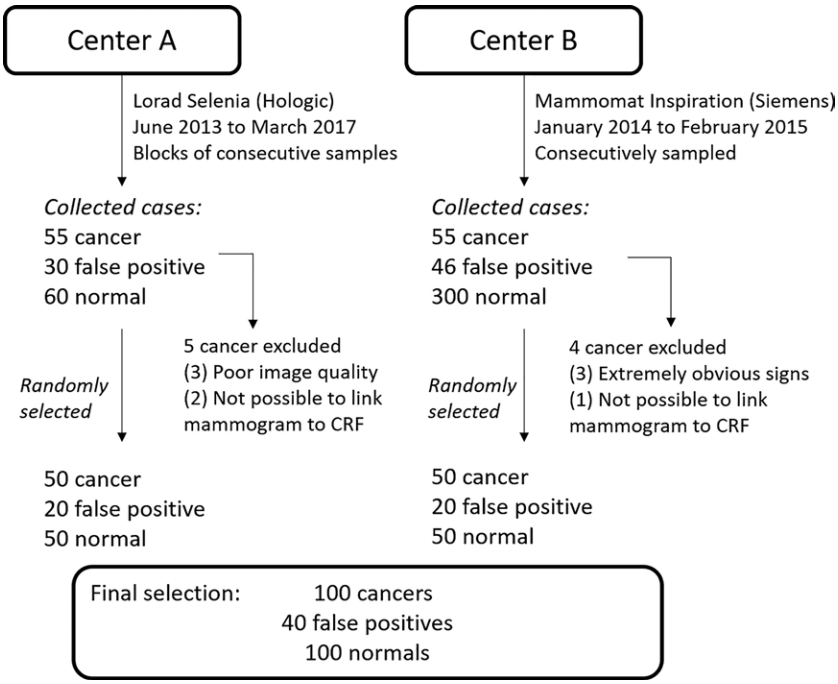


Figure 1: Flowchart of examination selection. *CRF* = case report form.

Table 1: Characteristics of the Population and Digital Mammographic Examinations Selected for the Study

Variable	Collection Center A (n = 120)	Collection Center B (n = 120)	Total (n = 240)
Patient age (y)			
Mean	61	60	61
Median	62	61	62
Range	39–89	50–70	39–89
Interquartile range	47–72	56–64	53–66
Median breast thickness (mm)*	58 (48–66)	55 (48–64)	57 (48–65)
BI-RADS breast density†			
a	6 (5)	22 (18)	28 (12)
b	78 (65)	55 (46)	133 (55)
c	30 (25)	34 (28)	64 (27)
d	6 (5)	9 (8)	15 (6)
Mean glandular dose (mGy)*	1.60 (1.39–1.93)	1.18 (0.97–1.37)	1.38 (1.14–1.64)

Note.—BI-RADS = Breast Imaging Reporting and Data System.
* Numbers in parentheses are the interquartile range.
† Data are numbers of examinations, with percentages in parentheses.

AI Support System

The AI computer system used by the radiologists for support was Transpara (version 1.3.0, ScreenPoint Medical). This system is designed for automated breast cancer detection in mammography and breast tomosynthesis. The system works on processed mammograms, is compatible with examinations performed with digital mammography and/or breast tomosynthesis systems from different vendors, and analyzes information across the four standard views of a digital mammographic examination (craniocaudal and mediolateral oblique views of both breasts).

The system uses deep learning convolutional neural networks and features classifiers and image analysis algorithms to depict calcifications (21,22) and soft-tissue lesions (16,23–25) in two different modules. Soft-tissue and calcification findings are later combined to determine suspicious region findings. A value between 1 and 100 is assigned to each region, representing the level of suspicion that cancer is present (with 100 indicating the highest suspicion). Finally, proprietary algorithms are used to combine the scores of all detected regions in craniocaudal and/or mediolateral oblique right and/or left breast images into the examination-based score (the Transpara score), which ranges from 1 to 10 (with 10 indicating the highest likelihood that cancer is present on the mammogram). The Transpara score is calibrated such that the number of mammograms in each category is roughly equal in a screening setting (eg, 10% of screening mammograms fall into category 1, 10% in category 2).

The AI system is trained, validated, and tested by using a database containing more than 9000 mammograms with cancer (one-third of which are presented as lesions with calcifications) and the same number of mammograms without abnormalities. The mammograms originate from devices from four different vendors (Hologic, Siemens, GE Healthcare [Waukesha, Wis], and Philips Healthcare [Sölna, Sweden]). The AI system is validated on an independent internal multivendor data set that has not been used for training or validation of the algorithms. The mammograms used in this study have never been used to train, validate, or test the algorithms.

In practice, when using this system for support, radiologists can use an interactive decision support mode as well as traditional CAD. Interactive decision support can be activated for any specific breast region by clicking on it. The system then displays its level of suspicion (on a scale of 1 to 100) if something in that area has been detected (otherwise nothing is displayed except for a small cross indicating the clicked location). Traditional CAD is available to display calcification and soft-tissue lesion markers, with the false-positive rate of the prompts set lower for soft-tissue lesions than for calcifications (0.02 and 0.2 per image, respectively). In addition, on a whole-examination basis, the system always displays a proprietary examination score (Transpara score) between 1 and 10.

Table 2: Characteristics of the 100 Malignant Cancers

Characteristic	No. of Examinations
Histologic type	
Invasive ductal carcinoma	64
Ductal carcinoma in situ	13
Invasive lobular carcinoma	18
Invasive tubular carcinoma	6
Other	3
Lesion type	
Mass	54
Calcifications	35
Asymmetry	10
Architectural distortion	6

Note.—Five examinations showed both calcifications and mass lesions, and four examinations showed two histologic cancer types (eg, invasive ductal carcinoma and invasive lobular carcinoma).

Observer Evaluation

A fully crossed, multireader, multicase evaluation with two sessions (separated by at least 4 weeks) was performed to test both reading conditions: unaided or with AI support. The evaluation was performed at two different centers (evaluation centers A and B, both in the United States).

Fourteen Mammography Quality Standard Act–qualified radiologists performed the evaluation. Three were general radiologists and 11 were dedicated breast radiologists. The median experience with Mammography Quality Standard Act qualification was 9.5 years (range, 3–25 years), and the approximate mean number of mammograms read per year during the past 2 years was 5900 (range, 1200–10 000).

During each session, radiologists read half the examinations with AI support and half unaided. Radiologists were blinded to any information about the patient, including previous radiology and histopathology reports. Before the first session, each radiologist was individually trained in a session with 45 examinations not included in the final evaluation. The training was intended to familiarize radiologists with the evaluation workstation, the evaluation criteria, and the AI support system (eg, to understand how to use all its functionalities). Readers were also informed that the study data set was enriched with cancer mammograms with respect to the standard prevalence seen in screening.

For each examination, radiologists provided a forced Breast Imaging Reporting and Data System (BI-RADS) score (range, 1–5) and assigned a probability of malignancy (POM) between 1 and 100 (with 100 indicating highly suspicious for malignancy). During training, radiologists were instructed to use the full extent of the POM scale with anchor points as a guide. For instance, transition from BI-RADS category 2 to BI-RADS category 3 was recommended at a POM of 40, transition from BI-RADS category 3 to BI-RADS category 4 was recommended at a POM of 60, and transition from BI-RADS category 4 to BI-RADS category 5 was recommended at a POM of 80.

The evaluation was performed with an in-house–developed workstation by using a 12-MP mammographic display (Coronis Uniti; Barco, Kortrijk, Belgium) calibrated to the Digital Imaging and Communications in Medicine Grayscale Standard Display Function. Readers could adjust window and level settings and could zoom and pan. Ambient lights were set to approximately 45 lux. Half the readers used the AI system integrated in the reading workstation (at evaluation center A), and the other half used the AI system on a separate screen from the workstation (Microsoft Surface Pro [Redmond, Wash], at evaluation center B).

Statistical Analysis

The main end points of the study were to compare the area under the receiver operating characteristic (ROC) curve, sensitivity and specificity, and reading time between reading unaided or reading with AI support. Secondary analyses (explained in the next section) were also performed to obtain detailed knowledge of the effect of using AI system for support in reading mammograms. The reported *P* values of the secondary end points were not adjusted for testing multiple hypotheses, and therefore we refrain from claims about the significance of secondary end points. Instead, the secondary analyses are meant only to be supportive of our main hypotheses. Areas under the ROC curve (AUCs), specificity and sensitivity, and radiologists' reading time were compared between reading conditions by using mixed-model analysis of variance and generalized linear models for multiple repeated measurements.

Statistical analysis was performed with SPSS software (version 24; IBM, Armonk, NY), and open-access Obuchowski-Rockette and Dorfman-Berbaum-Metz software (version 2.5; Medical Image Perception Laboratory–University of Iowa, Iowa City, Iowa; available from *http://perception.radiology.uiowa.edu/*).

ROC performance.—ROC curves and their AUCs were computed by using the POM score. The Obuchowski-Rockette and Dorfman-Berbaum-Metz mixed-model analysis of variance yielded a *P* value for rejecting the null hypothesis that readings performed unaided or with AI support have equal performance (26–29). $P < .05$ was indicative of a statistically significant difference between both reading conditions. Secondarily, to identify possible strengths and weaknesses of the study as a function of the different types of mammograms and readers used, five subgroup subanalyses were also performed: *(a)* subgroups of examinations according to lesion type (soft tissue or calcifications); *(b)* subgroups of examinations according to digital mammography system used (Hologic or Siemens); *(c)* subgroups of examinations according to breast density (lower density [BI-RADS categories a and b] or higher density [BI-RADS categories c and d]); *(d)* equal subgroups of radiologists based on years of experience (lower 50% vs higher 50%); and *(e)* subgroups of radiologists based on the use of the AI system integrated in the workstation or on a separate viewer.

Similarly, in a separate secondary subanalysis, the location of the reader's findings was considered to avoid the possibility that readers were rewarded for detecting a cancer when they marked the wrong location. In this analysis, if a reader did not annotate a malignant lesion within 1.5 cm from the center of the ground

truth of the lesion, this reading was modified to the lowest POM used by that reader across our whole study.

Sensitivity and specificity.—Sensitivity and specificity for each reading condition (ie, with or without AI support) were computed by using the BI-RADS scores. The reader-averaged sensitivity and specificity for each modality was computed by using a generalized linear model (30); thus, repeated measures by multiple readers were taken into account. This binary logistic generalized linear model was built with consideration of reading condition, reader, and the interaction term as factors. Parameters were bootstrapped (n = 1000). χ^2 statistics and confidence intervals were based on the Wald test. $P < .05$ was indicative of a statistically significant difference between reading conditions.

Reading time.—The reading time per case was automatically measured by the workstation software used for the observer evaluation. Average reading times per case were compared between reading conditions with a generalized linear model similar to the one described for sensitivity and specificity but with use of reading time per case as the dependent variable. For this analysis, outliers, defined as values extending beyond 1.5 times the standard deviation of the data, were removed. These were considered unreliable because readers might have been interrupted. $P < .05$ was indicative of a statistically significant difference between reading conditions.

A learning curve for the AI system in relation to reading time was evaluated. The generalized linear model analysis for reading times was repeated for two subsets of the data, representing data from the first reading session (first-time use of AI system) and the data from the second reading session, after the washout period (second-time use of AI system).

Secondarily, reading time subanalysis was also performed in two scenarios: *(a)* differentiating between the radiologists who used the AI system on the workstation and those who used the AI system on the separate viewer and *(b)* as a function of the Transpara score (score, 1–10), creating a subgroup of low-suspicion examinations (score, 1–5) and another one of high-suspicion examinations (score, 6–10).

Stand-alone computer system performance.—The AUC of the stand-alone computer system was compared with the radiologists' AUC when radiologists read mammograms in the unaided mode as a secondary study outcome. The ROC of the AI system was computed by using a continuous version of the Transpara score. This analysis was done by using the single-modality multiple-reader Obuchowski-Rockette model described in an article by Hillis (28). This test yields a P value for rejecting the null hypothesis that computer and the radiologists have equal performance.

Results

ROC Performance

Radiologists improved their detection performance when using AI support, with the average AUC increasing from 0.87 to 0.89 (difference, 0.02; P = .002) (Fig 2, Table 3). Per reader, the changes in AUC ranged from 0.0 to 0.05 and were higher with AI support for 12 of the 14 radiologists (there was no change in AUC for the other two readers). The AUC was higher with AI support in all subgroup scenarios, with a similar effect of 0.02 (Table 3).

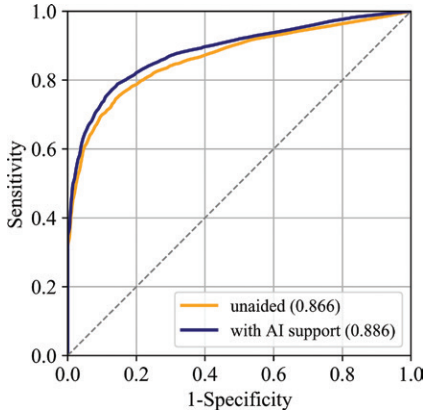


Figure 2: Average receiver operating characteristic (ROC) curves under two reading conditions: unaided and with artificial intelligence (*AI*) support. Average is computed across 14 radiologists participating in this evaluation. Numbers in parentheses are areas under ROC curve.

Sensitivity and Specificity

On average, sensitivity was 3 percentage points higher with AI support (P = .046); specificity also trended toward improvement (2 percentage points higher with AI support; P = .06) (Table 4). Examples of examinations in which the total number of correct recall assessments across readers changed between reading conditions are shown in Figures 3 and 4. In total, there was a disagreement (ie, at least three radiologists changed their assessment between reading conditions) in 32 examinations: In 72% of examinations (23 of 32) there was an increase in the number of readers making the right interpretation when using AI support, whereas the opposite occurred in the other 28% of examinations (nine of 32).

Reading Time

On average, reading time per case was similar in the unaided sessions (146 seconds; 95% confidence interval: 143 seconds, 149 seconds) and the sessions with AI support (149 seconds; 95% confidence interval: 146 seconds, 152 seconds); the difference was not significant (P = .15). Reading time increased for nine of 14 radiologists (range, 0.5%–10%) and decreased for five (range, 0.3%–22%) (Fig 5a). Of all reading times, 2.7% (181 of 6720) were defined as outliers and were excluded from this analysis.

The reading times were closer to each other between reading unaided and reading with AI support during the second block of sessions (2 seconds; not significant at P = .70) than in the first (5 seconds; not significant at P = .09).

Reading unaided and with AI support differed as a function of the computer Transpara score ($P < .001$) (Fig 5b). For the

Table 3: AUC for Each Radiologist and Reader-averaged AUCs for Reading Mammograms Unaided and with AI Support

Variable	Unaided	With AI Support	Difference	*P* Value
Radiologist No.				
1	0.87	0.90	0.04	
2	0.82	0.84	0.02	
3	0.91	0.92	0.01	
4	0.85	0.85	0.01	
5	0.79	0.85	0.05	
6	0.84	0.86	0.02	
7	0.93	0.95	0.01	
8	0.87	0.90	0.04	
9	0.87	0.87	0.0	
10	0.90	0.92	0.02	
11	0.86	0.90	0.04	
12	0.86	0.86	0.0	
13	0.87	0.90	0.03	
14	0.87	0.88	0.01	
Average	0.87 (0.83,0.90)	0.89 (0.85,0.92)	0.02 (0.01, 0.03)	.002
Subgroup secondary analyses*				
Soft-tissue lesions	0.89	0.90	0.02 (0.0, 0.03)	.03
Calcifications	0.88	0.90	0.02 (0.0, 0.05)	.10
Hologic examinations	0.85	0.86	0.02 (0.0, 0.04)	.09
Siemens examinations	0.89	0.91	0.02 (0.0, 0.04)	.03
Low breast density	0.88	0.90	0.02 (0.01, 0.03)	.003
High breast density	0.83	0.85	0.02 (−0.01, 0.05)	.15
Least experienced	0.87	0.89	0.03 (0.01, 0.04)	.003
Most experienced	0.87	0.88	0.01 (0.0, 0.03)	.08
AI workstation	0.87	0.89	0.02 (0.0, 0.03)	.04
AI separate viewer	0.86	0.88	0.02 (0.01, 0.04)	.01
Location specific	0.84	0.87	0.02 (0.01, 0.04)	.003

Note.—AI = artificial intelligence, AUC = area under the receiver operating characteristic curve. Numbers in parentheses are 95% confidence intervals.

* Data are averages.

Table 4: Mean Sensitivity and Specificity across Radiologists

Variable	Unaided	With AI Support	Difference (Percentage Points)	*P* Value
Sensitivity (%)	83 (83/100) [81, 85]	86 (86/100) [84, 88]	3	.046
Specificity (%)	77 (108/140) [75, 79]	79 (111/140) [77, 81]	2	.06

Note.—Breast Imaging Reporting and Data System category 3 or higher was used the recall threshold. Numbers in parentheses are raw data, and numbers in brackets are 95% confidence intervals.

Stand-Alone Computer System

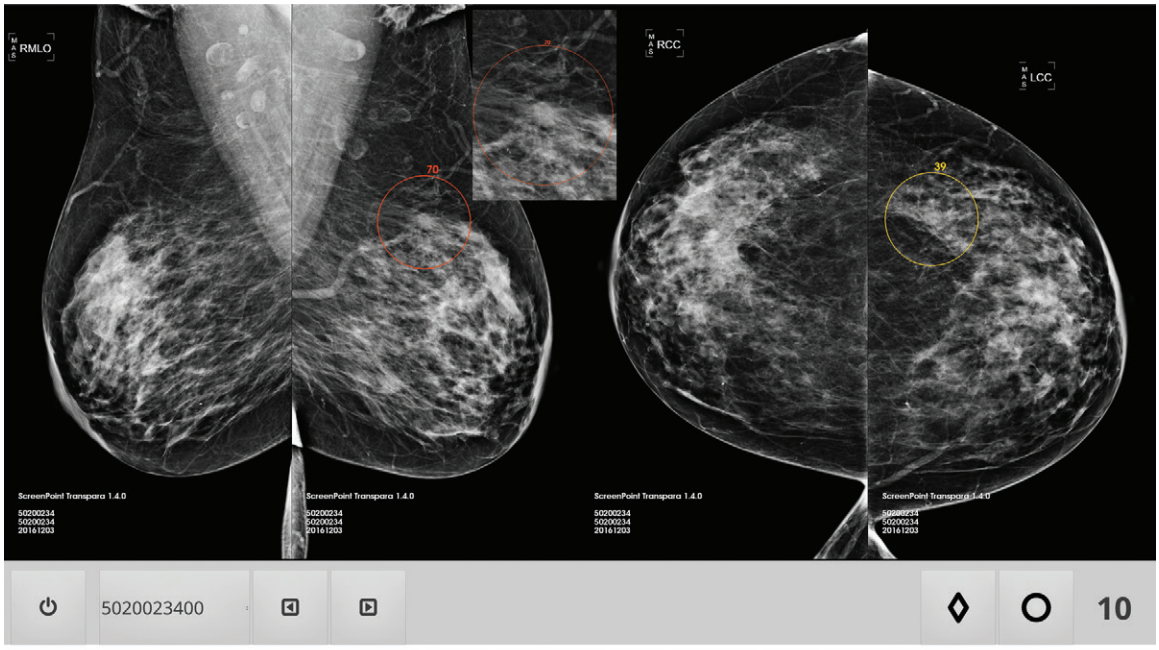
The breast cancer detection performance of the stand-alone AI computer system was similar to the radiologists' average performance (radiologists' average AUC, 0.87; computer AUC, 0.89; difference, 0.02; P = .33) (Fig 6).

Discussion

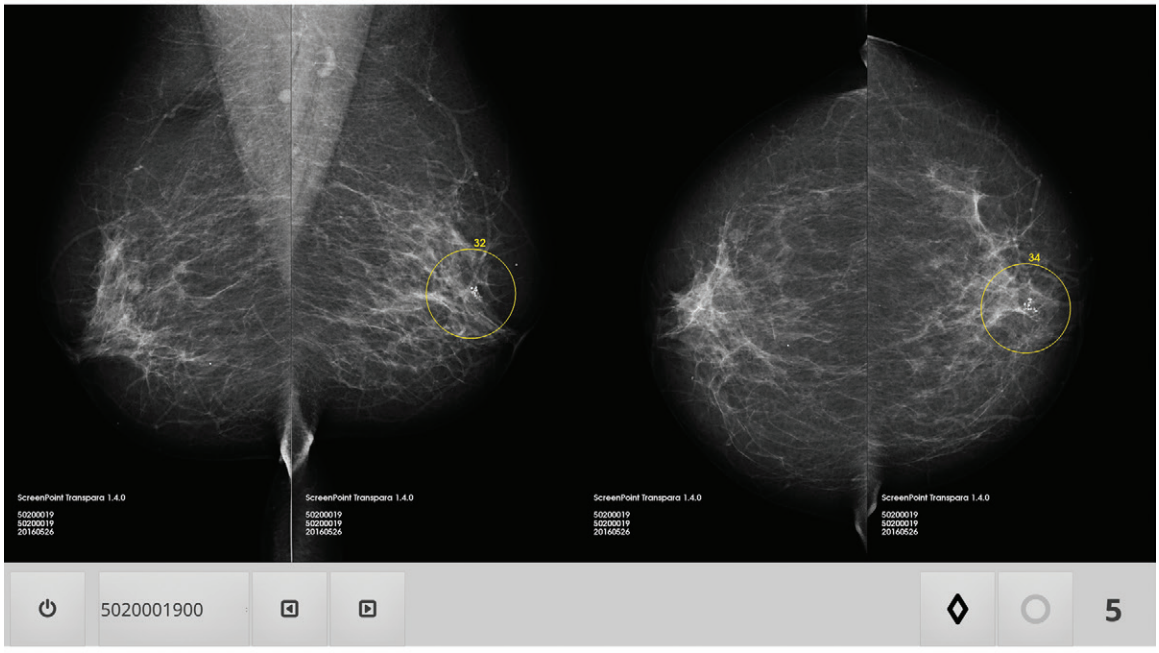
In our study, we found that breast radiologists had a higher diagnostic performance (as measured with the AUC) with support from an AI system compared with reading unaided. The average reading times per case were similar under both conditions. This improvement in diagnostic performance was observed in a cancer-enriched data set of digital mammographic examinations with a representative sample of abnormalities that may be observed in asymptomatic women undergoing mammographic screening. The improvement in diagnostic performance with the AI system was due to an increase in the middle part of the ROC curve. This suggests that the AI system improves the evaluation of equivocal cases, suggesting the clinical relevance of this tool.

low-suspicion examinations (score, 1–5), radiologists decreased their average reading time per case by 11% when using the AI system. Conversely, reading time per case was 2% higher with use of AI support for the high-suspicion examinations (score, 6–10). Assuming that in a screening population each Transpara score category includes the same number of examinations (and, therefore, that examinations with a score of 1–5 are 50% of the total and that those with a score of 6–10 make up the remaining 50%), averaging the above-mentioned results is expected to lead to an overall 4.5% reduction in reading time per case with use of the AI system in screening.

As expected, the improvement in AUC with AI support was higher for the radiologists who were least experienced with mammography. We did not observe a difference in unaided performance on the basis of experience. As suggested by Hupse et al (18), some experienced radiologists might tend to query the decision support more times, obtaining most of the available prompt marks, which might reduce their performance. The finding is remarkably similar to the reported benefits in performance with the addition of digital breast tomosynthesis to mammography (31), which is higher for the least-experienced radiologists. This might imply that



a.



b.

Figure 3: **(a)** Mammograms in 71-year-old woman with invasive ductal carcinoma (outlined and with level of suspicion score assigned by computer system). Patient was recalled (Breast Imaging Reporting and Data System [BI-RADS] score, ≥3) by four of 14 radiologists when reading unaided and by 11 of 14 radiologists using artificial intelligence (AI) system for support. Outlined areas and scores are shown as in viewer of AI system. **(b)** Mammograms in 62-year-old woman without cancer, who was recalled (BI-RADS score, ≥3) by 12 of 14 radiologists when reading unaided and by seven of 14 readers when using AI system for support. Outlined areas and scores are shown as in viewer of AI system.

more-experienced radiologists are less likely, or slower, to adopt new techniques to improve their performance.

Given the high workload of screening programs, from a cost-effectiveness point of view the performance benefit of using AI support is further enhanced by the fact that radiologists do not lengthen their reading time when using this system. In fact, in a real screening scenario, the average reading time per case would actually decrease by approximately 4.5%. This means that the examination-based score provided by the system has the potential to make radiologists' readings more efficient, increasing their attention in the most suspicious examinations while reassuring them in faster readings of the least suspicious examinations. Moreover, the observed learning curve implies that more practice with the system might yield even shorter reading times. In

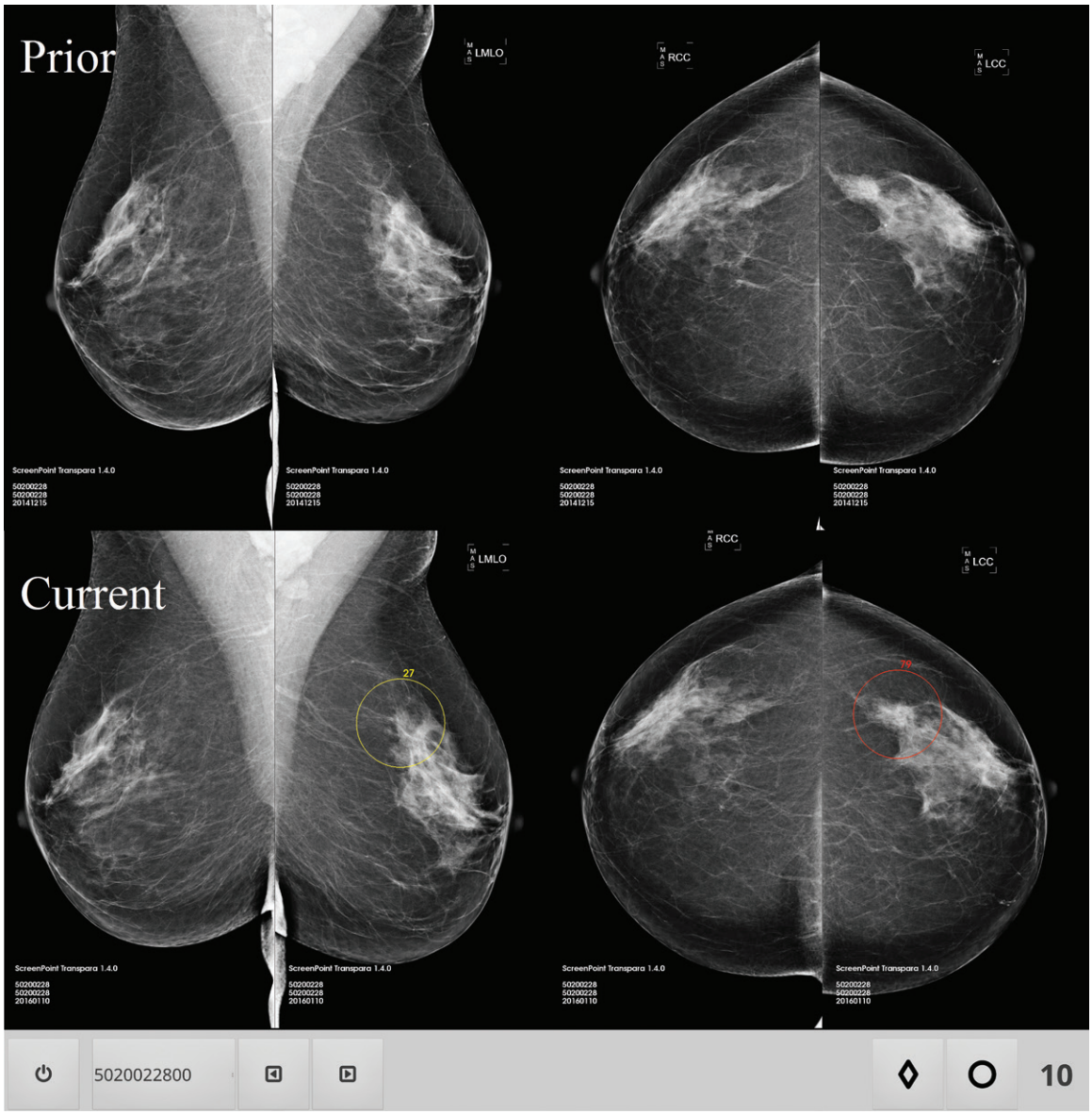


Figure 4: Mammograms in 62-year-old woman without cancer who was incorrectly recalled (Breast Imaging Reporting and Data System score, ≥3) by one of 14 radiologists when reading unaided but by five of 14 radiologists when using artificial intelligence (AI) system for support. Outlined areas and scores are shown as in viewer of AI system.

the secondary analysis, the stand-alone performance of the computer system was similar to the average performance of the radiologists. Even though larger studies are needed to validate these findings, our results suggest that using computer systems as a stand-alone first or second reader in screening programs might be feasible. Given the increasing lack of (experienced) breast radiologists (3), this might even allow the development or continuation of screening programs.

There is a paucity of literature about the clinical performance of AI systems or deep learning–based traditional CAD systems to support reading of mammograms. So far, published studies have mainly evaluated the stand-alone performance of AI. Kooi et al (16) and Becker et al (32) found that AI algorithms developed in-house could achieve a performance similar to that of the radiologist with the lowest performance in enriched and selected data sets, but only in very limited scenarios (eg, only soft-tissue lesions). A study by Kim et al (33) found that in-house–developed AI algorithms achieved a sensitivity of 76% and a specificity of 89% in a screening data set. Despite the differences in data sets, our results support the observed trend that AI algorithms are reaching a performance similar to that of radiologists for breast cancer detection in mammography.

Our study had some limitations. The main limitation is that the study was performed with a highly enriched data set with screening-detected cancers instead of using a prospective assessment in screening practice. Although the readers trended to improve their recall when using the AI system, in some examinations the computer might have misled radiologists into

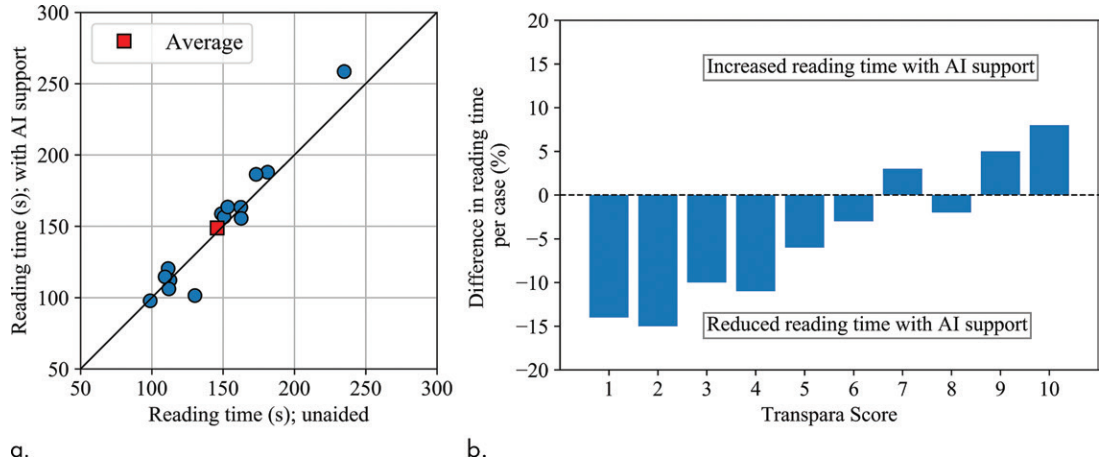


Figure 5: **(a)** Graph shows differences in reading time per case for each radiologist (circles) and on average (square). **(b)** Bar chart shows differences in reading times as function of examination-based Transpara score assigned by system. *AI* = artificial intelligence.

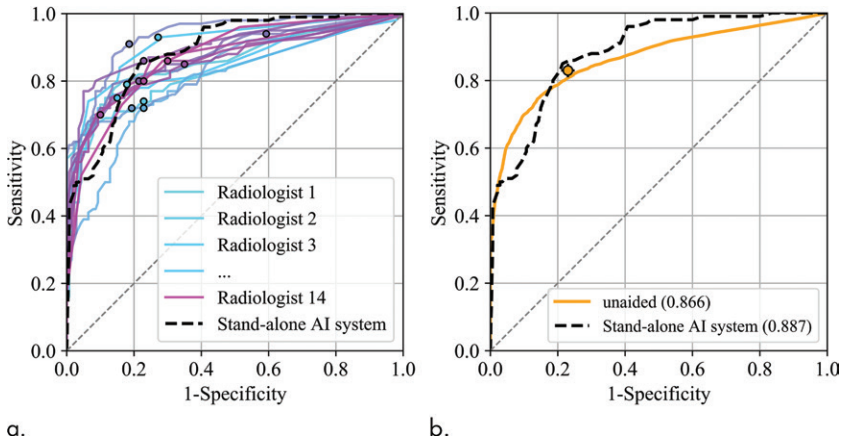


Figure 6: Receiver operating characteristic (ROC) curves for **(a)** individual radiologists reading mammograms unaided and stand-alone artificial intelligence (AI) computer system and **(b)** average of radiologists and stand-alone AI computer system. Radiologists' operating points at Breast Imaging Reporting and Data System category 3 thresholds are indicated with circles. Areas under ROC curve are shown in parentheses in **b**.

making false-positive assessments. Future improvements of the algorithms, especially those using temporal information, are likely to improve the benefit of AI support. Moreover, readers were aware of the high rate of malignancies in the case set, which may have resulted in a "laboratory effect" (34,35). Ideally, future studies should assess the benefit of AI support in an actual screening setting. Furthermore, our study was performed with radiologists from the United States only, whereas screening practice and recall rates vary substantially around the world. Consequently, the net effect of the AI system might also vary on the basis of geographic regions and local policies (36–38).

In conclusion, radiologists improved their diagnostic performance in the detection of breast cancer at mammography by using an AI computer system for support without the need for additional reading time. However, as promising as these findings may be, studies within a screening scenario should be performed to validate them and seize the real effect of AI support in screening.

Acknowledgments: The authors thank Barco (Kortrijk, Belgium) for providing the displays for the study. ScreenPoint Medical is a spinoff company from the Radboudumc. While there is no financial relationship between ScreenPoint Medical and I.S. and R.M.M., we do work closely together with its CEO, who is also a professor at our department. There is a master research agreement (MRA) between the Radboudumc (department of radiology) and ScreenPoint Medical that describes terms of cooperation. For this project, an addendum to the MRA was signed that details the roles of the investigators and ScreenPoint Medical in this specific study. ScreenPoint Medical was responsible for data generation and paid all external costs of the study. I.S. and R.M.M work as independent investigators for the Radboudumc and did not receive any financial compensation from ScreenPoint Medical for this work. We guarantee the quality of the data and are responsible for the statistical analysis. We were free to publish the results, with the only precondition that we first reported the results to ScreenPoint Medical.

Author contributions: Guarantors of integrity of entire study, E.K., R.M.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, A.R.R., E.K., J.J.M., R.M.M.; clinical studies, A.R.R., J.J.M., K.S., S.H.H.; statistical analysis, A.R.R., E.K., J.J.M., R.M.M.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: **A.R.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is employed by ScreenPoint Medical. Other relationships: disclosed no relevant relationships. **E.K.** disclosed no relevant relationships. **J.J.M.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is employed by ScreenPoint Medical. Other relationships: disclosed no relevant relationships. **K.S.** disclosed no relevant relationships. **S.H.H.** Activities related to the present article: software for testing was provided by ScreenPoint. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **I.S.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: see acknowledgments. **R.M.M.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives money for board membership from Siemens Healthineers; institution has grants/grants pending from Siemens Healthineers, Bayer Healthcare, Medtronic, Elswood, Identification Solutions, Micrima, Screenpoint Medical, MR Coils, and Sigma Screening; institution receives payment for lectures including service on speakers bureaus from Siemens Healthineers; has stock/stock options in Transonic Imaging. Other relationships: see acknowledgments.

References

1. Smith RA, Cokkinides V, Brooks D, Saslow D, Brawley OW. Cancer screening in the United States, 2010: a review of current American Cancer Society guidelines and issues in cancer screening. CA Cancer J Clin 2010;60(2):99–119.
2. Broeders M, Moss S, Nyström L, et al. The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies. J Med Screen 2012;19(Suppl 1):14–25.
3. Rimmer A. Radiologist shortage leaves patient care at risk, warns Royal College. BMJ 2017;359:j4683.
4. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992;184(3):613–617.
5. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. RadioGraphics 2003;23(4):881–895.
6. Weber RJ, van Bommel RM, Louwman MW, et al. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. Breast Cancer Res Treat 2016;158(3):471–483.
7. Broeders MJ, Onland-Moret NC, Rijken HJ, Hendriks JH, Verbeek AL, Holland R. Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. Eur J Cancer 2003;39(12):1770–1775.
8. Gilbert FJ, Astley SM, Gillan MG, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med 2008;359(16):1675–1684.
9. Bargalló X, Santamaría G, Del Amo M, et al. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. Eur J Radiol 2014;83(11):2019–2023.
10. Fenton JJ, Xing G, Elmore JG, et al. Short-term outcomes of screening mammography using computer-aided detection: a population-based study of Medicare enrollees. Ann Intern Med 2013;158(8):580–587.
11. Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR Am J Roentgenol 2008;190(4):854–859.
12. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356(14):1399–1409.
13. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175(11):1828–1837.
14. Azavedo E, Zackrisson S, Mejàre I, Heibert Arnlind M. Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. BMC Med Imaging 2012;12(1):22.
15. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
16. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–312.
17. Trister AD, Buist DSM, Lee CI. Will machine learning tip the balance in breast cancer screening? JAMA Oncol 2017;3(11):1463–1464.
18. Hupse R, Samulski M, Lobbes MB, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology 2013;266(1):123–129.
19. Samulski M, Hupse R, Boetes C, Mus RD, den Heeten GJ, Karssemeijer N. Using computer-aided detection in mammography as a decision support. Eur Radiol 2010;20(10):2323–2330.
20. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. Acad Radiol 2011;18(2):129–142.
21. Bria A, Karssemeijer N, Tortorella F. Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications. Med Image Anal 2014;18(2):241–252.
22. Mordang JJ, Janssen T, Bria A, Kooi T, Gubern-Mérida A, Karssemeijer N, eds. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. International Workshop on Digital Mammography. Cham, Switzerland: Springer, 2016.
23. Hupse R, Karssemeijer N. Use of normal tissue context in computer-aided detection of masses in mammograms. IEEE Trans Med Imaging 2009;28(12):2033–2041.
24. Karssemeijer N, Te Brake GM. Detection of stellate distortions in mammograms. IEEE Trans Med Imaging 1996;15(5):611–619.
25. Karssemeijer N. Automated classification of parenchymal patterns in mammograms. Phys Med Biol 1998;43(2):365–378.
26. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992;27(9):723–731.
27. Obuchowski NA. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. Acad Radiol 1995;2(Suppl 1):S22–S29; discussion S57–S64, S70–S71 pas.
28. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Stat Med 2007;26(3):596–619.
29. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Acad Radiol 2008;15(5):647–661.
30. McCullagh P, Nelder JA. Generalized linear models. Boca Raton, Fla: CRC, 1989.
31. Tucker L, Gilbert FJ, Astley SM, et al. Does reader performance with digital breast tomosynthesis vary according to experience with two-dimensional mammography? Radiology 2017;283(2):371–380.
32. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52(7):434–440.
33. Kim EK, Kim HE, Han K, et al. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. Sci Rep 2018;8(1):2762.
34. Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. PLoS One 2013;8(5):e64366.
35. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology 2008;249(1):47–53.
36. Gennaro G, Hendrick RE, Ruppel P, et al. Performance comparison of single-view digital breast tomosynthesis plus single-view digital mammography with two-view digital mammography. Eur Radiol 2013;23(3):664–672.
37. Warren RM, Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. Br J Radiol 1995;68(813):958–962.
38. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191(1):241–244.