Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis

Emily F. Conant, MD • Alicia Y. Toledano, ScD • Senthil Periaswamy, PhD • Sergei V. Fotin, PhD • Jonathan Go, MASc • Justin E. Boatsman, MD • Jeffrey W. Hoffmeister, MD, MSEE

From the Department of Radiology, Perelman School of Medicine at the University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104 (E.F.C.); Biostatistics Consulting, Kensington, Md (A.Y.T.); iCAD, Nashua, NH (S.P., S.V.F., J.G., J.W.H.); and Intrinsic Imaging, Bolton, Mass (J.E.B.). Received January 14, 2019; revision requested March 25; final revision received May 27; accepted June 20. **Address correspondence to** E.F.C. (e-mail: *emily.conant@pennmedicine.upenn.edu*).

E.F.C. supported by grants from the National Cancer Institute (U54CA163313), iCAD, and Hologic.

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

See also the commentary by Hsu and Hoyt in this issue.

Radiology: Artificial Intelligence 2019; 1(4):e180096 • https://doi.org/10.1148/ryai.2019180096 • Content codes: BR IN

Purpose: To evaluate the use of artificial intelligence (AI) to shorten digital breast tomosynthesis (DBT) reading time while maintaining or improving accuracy.

Materials and Methods: A deep learning AI system was developed to identify suspicious soft-tissue and calcified lesions in DBT images. A reader study compared the performance of 24 radiologists (13 of whom were breast subspecialists) reading 260 DBT examinations (including 65 cancer cases) both with and without AI. Readings occurred in two sessions separated by at least 4 weeks. Area under the receiver operating characteristic curve (AUC), reading time, sensitivity, specificity, and recall rate were evaluated with statistical methods for multireader, multicase studies.

Results: Radiologist performance for the detection of malignant lesions, measured by mean AUC, increased 0.057 with the use of AI (95% confidence interval [CI]: 0.028, 0.087; $P < .01$), from 0.795 without AI to 0.852 with AI. Reading time decreased 52.7% (95% CI: 41.8%, 61.5%; $P < .01$), from 64.1 seconds without to 30.4 seconds with AI. Sensitivity increased from 77.0% without AI to 85.0% with AI (8.0%; 95% CI: 2.6%, 13.4%; $P < .01$), specificity increased from 62.7% without to 69.6% with AI (6.9%; 95% CI: 3.0%, 10.8%; noninferiority $P < .01$), and recall rate for noncancers decreased from 38.0% without to 30.9% with AI (7.2%; 95% CI: 3.1%, 11.2%; noninferiority $P < .01$).

Conclusion: The concurrent use of an accurate DBT AI system was found to improve cancer detection efficacy in a reader study that demonstrated increases in AUC, sensitivity, and specificity and a reduction in recall rate and reading time.

Screening with digital breast tomosynthesis (DBT) has been shown to improve cancer detection (1–4) and reduce false-positive recalls (2–7) compared with screening with digital mammography (DM) alone. Whether DBT is combined with DM or with reconstructed synthetic mammography (SM) images to reduce x-ray dose, the time to interpret a DBT examination is almost twice that of interpreting a DM-alone study (1,8,9). The increased reading time is due to the added time for the radiologist to "scroll" through the in-focus planes or "images" of the reconstructed DBT stack, with the number of DBT images being proportional to the thickness of the breast in compression. The ability of the radiologist to assess the DBT reconstructed images in a quasi–three-dimensional (3D) format reduces the impact of confounding or superimposed breast tissue, which may "mask" or obscure lesions in two-dimensional (2D) planar mammography. The DBT reconstructed dataset also adds information on 3D localization of lesions within the breast.

As DBT increasingly becomes the standard of care for mammographic imaging, there is a need for algorithms to optimize reading efficiency while maintaining or improving the accuracy achieved with DBT compared with that achieved with DM-alone imaging. Applications that "flag" and "bookmark" lesion location in the reconstructed DBT image stack could help with detection and localization of clinically significant breast lesions while also decreasing reading time.

In the United States, computer-aided detection (CAD) is used in approximately 83% of all screening DM (10). While some studies have shown that when a single reader interprets DM with CAD the accuracy for cancer detection is often increased to a level similar to that of double reading (11–13), another large study evaluating the performance of radiologists both with and without CAD from 43 facilities over a 4-year period found that CAD use was associated with reduced accuracy, as well as an increase in biopsy recommendation (14). Additional studies have also shown that interpretation times for DM with CAD increase by approximately 19%, compared with reading DM without CAD (15). In contrast, recent advances in CAD algorithms applied to DBT datasets have resulted in faster reading times, with maintenance of DBT reader performance (16–18). Newer algorithms based on deep learning artificial intelligence (AI) trained on large DBT datasets have the potential to further improve reader accuracy while also improving

Abbreviations

AI = artificial intelligence, AUC = area under the ROC curve, BI-RADS = Breast Imaging Reporting and Data System, CAD = computer-aided detection, CI = confidence interval, DBT = digital breast tomosynthesis, DM = digital mammography, ROC = receiver operating characteristic, SM = synthetic mammography, 2D = two-dimensional

Summary

Reading times were significantly reduced, and sensitivity, specificity, and recall rate improved in a nonclinical reader study when an artificial intelligence system was utilized concurrently with image interpretation for digital breast tomosynthesis.

Key Points

- The results of this reader study suggest that the combination of a highly accurate artificial intelligence (AI) system and concurrent rather than second-reader workflow may reduce reading times.
- The study also suggests the potential to increase sensitivity, specificity, and recall rates in comparison with interpretation without an AI system.
- The results of this study suggest that both improved efficiency and accuracy could be achieved in clinical practice by using an effective AI system.

reading efficiency. The purpose of this study was to evaluate the concurrent use of a DBT AI system to reduce reading time and maintain or improve area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, and specificity.

Materials and Methods

Case data were retrospectively collected in compliance with the Health Insurance Portability and Accountability Act, with institutional review board approval and waiver of informed consent. The study was financially supported by iCAD (Nashua, NH) and was performed by Intrinsic Imaging (Bolton, Mass). The truthing radiologist (J.E.B.) was Intrinsic Imaging's medical director and had control of the study data and information submitted for publication that might present a conflict of interest for authors who are employees of iCAD (S.P., S.V.F., J.G., and J.W.H.) or consultants for iCAD (E.F.C. and A.Y.T.).

Study Design

A multireader, multicase study of a DBT AI system was performed with 24 readers and 260 cases, including 65 cancer cases with 66 malignant lesions and 65 biopsy-proven benign cases. Readers reviewed 130 cases without AI and another 130 cases with AI during one session and complementary cases during a second session such that each case was read by each reader both with and without AI. Sessions were separated by a memory washout period of at least 4 weeks to minimize recall bias (19).

Cases

Screening and diagnostic DBT examinations with 2D DM or SM (Selenia Dimensions; Hologic, Marlborough, Mass) were collected from blocks of sequential series, including 500 female patients across seven U.S. acquisition sites. A pool of 474 cases met the inclusion and exclusion criteria (Fig 1). The 260 reader study cases were randomly selected by a statistician (A.Y.T.) from the case pool to meet predefined targets on the basis of characteristics of a screening population within each case type (negative, recalled, benign, cancer) and breast density distribution across case types (Tables 1, 2).

The 260 cases were from women aged 26–85 years (median, 55 years) imaged between June 2012 and October 2017 and included 65 cases with biopsy-proven malignancies (Tables 1, 2). Cancer cases included 64 cases with one malignant lesion and one case with two malignant lesions (invasive ductal carcinoma with ductal carcinoma in situ in the left breast and invasive ductal carcinoma in the right breast). The 66 malignant lesions included 14 ductal carcinomas in situ and 52 invasive carcinomas. Invasive lesions were predominantly invasive ductal carcinomas (46 of 52 lesions, 88%), along with five invasive lobular carcinomas only (10%) and one invasive papillary carcinoma with ductal carcinoma in situ (2%). The 46 invasive ductal carcinoma lesions included 24 invasive ductal carcinomas only, 17 invasive ductal carcinomas with ductal carcinoma in situ, four invasive ductal and lobular carcinomas, and one invasive ductal and lobular carcinoma with ductal carcinoma in situ. Maximum lesion size within each cancer case ranged from 0.1 to 6.0 cm (median, 1.4 cm). Although invasive carcinomas larger than 2.5 cm were excluded, ductal carcinomas in situ were included regardless of size. Maximum lesion size in cancer cases with soft-tissue lesions ranged from 0.1 to 4 cm (median, 0.9 cm), and maximum lesion size in cancer cases with calcifications only ranged from 0.3 to 6 cm (median, 1.4 cm).

AI System

An AI system (PowerLook Tomo Detection 2.0; iCAD, Nashua, NH) based on deep convolutional neural networks processed DBT images in all 260 study cases to detect soft-tissue and calcific lesions. The system was trained offline in a data-driven manner by using an expertly annotated tomosynthesis image dataset collected independently, meaning that none of the study cases were used to develop or train the algorithm. Unlike conventional CAD systems, the AI system acquired knowledge necessary for lesion detection directly from the provided training data and did not rely on explicit encoding or replication of human expert decision processes. An operating point controlling the trade-off between the detection sensitivity and specificity of the algorithm was chosen in favor of higher sensitivity. The development and configuration of the algorithm were completed prior to the study. A workstation (WorkstationOne; Three Palm Software, Carmel, Calif) displayed outlines of lesions detected by AI with faint outlines on every image and a bold outline on the image in which AI detected the lesion. The outlines could be toggled on and off. Calibrated scores (0–100) at the lesion and case levels were also provided to indicate the algorithm's confidence that a finding or case showed malignancy. The scores were calibrated with the training dataset, which was weighted to have 50% cancers and 50% noncancers such that, for example, a 70% lesion score meant that, of all the detections in

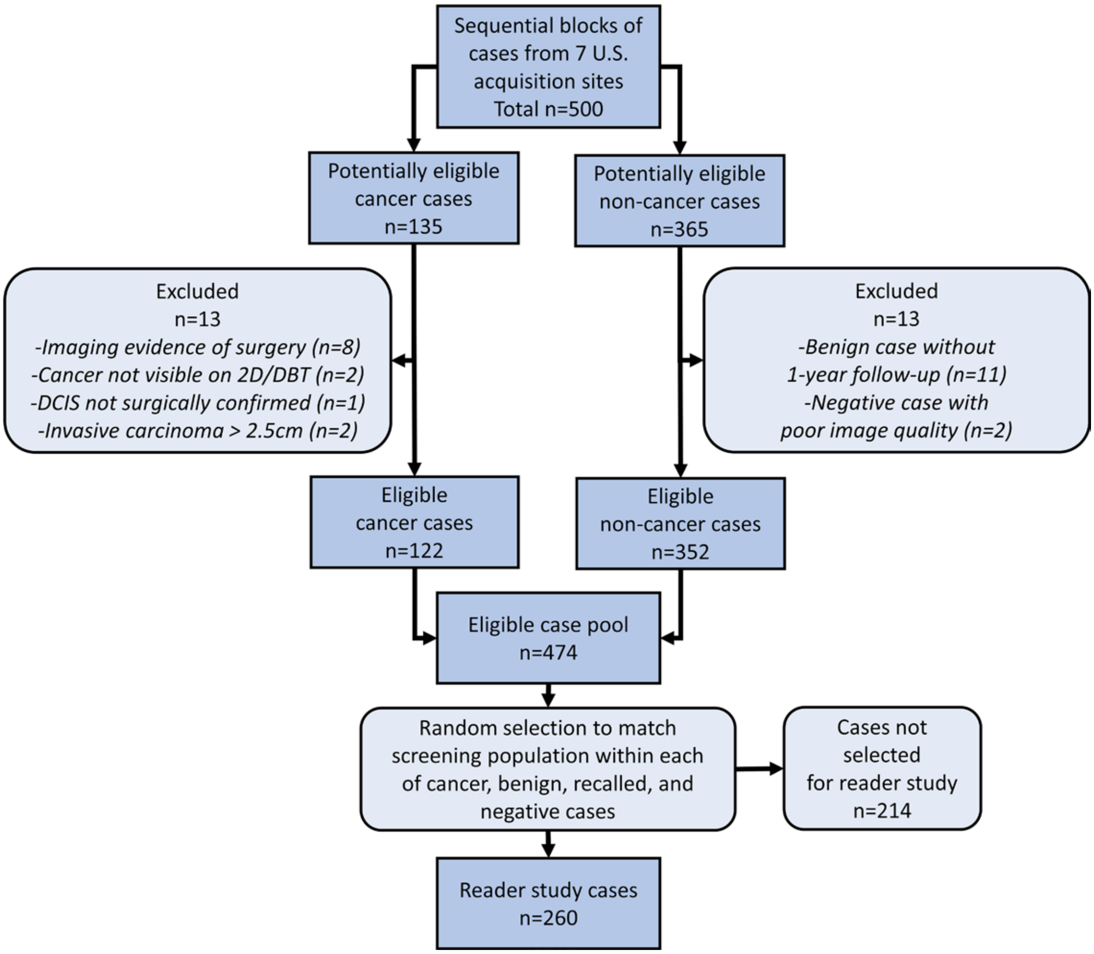
Sequential blocks of cases from 7 U.S. acquisition sites Total n=500
Potentially eligible cancer cases n=135
Potentially eligible non-cancer cases n=365
Excluded n=13 -Imaging evidence of surgery (n=8) -Cancer not visible on 2D/DBT (n=2) -DCIS not surgically confirmed (n=1) -Invasive carcinoma > 2.5cm (n=2)
Excluded n=13 -Benign case without 1-year follow-up (n=11) -Negative case with poor image quality (n=2)
Eligible cancer cases n=122
Eligible non-cancer cases n=352
Eligible case pool n=474
Random selection to match screening population within each of cancer, benign, recalled, and negative cases
Cases not selected for reader study n=214
Reader study cases n=260

Figure 1: Case selection flowchart. Cases with imaging evidence of prior breast surgery (*n* = 8) were excluded because readers were not provided history or prior examinations. The following cancer cases were also excluded: Cases with primary breast cancers that were not visible mammographically (*n* = 1 detected with US; *n* = 1 detected because palpable), cases with biopsy results of ductal carcinoma in situ (DCIS) that were not surgically confirmed (*n* = 1), and invasive carcinomas larger than 2.5 cm (*n* = 2). Lesion size was based on surgical pathologic findings when available or longest linear dimension on study images. A breast subspecialist truthing radiologist (J.E.B) who was not a study reader annotated the location and extent of malignant lesions on two-dimensional (2D) images and digital breast tomosynthesis (DBT) images. The reference standard was biopsy proof for all cancer cases and excision of any benign histopathologic findings or concordant biopsy of fibroadenoma or fibrocystic changes for benign cases. Benign cases in which the patient had undergone aspiration and those with discordant biopsy or concordant biopsy of histopathologic findings other than fibroadenoma or fibrocystic changes also required normal imaging at least 1 year (320 days) after the study DBT examination (*n* = 11 excluded for lack of 1-year follow-up). Normal 1-year follow-up imaging findings was the reference standard for recalled (Breast Imaging Reporting and Data System [BI-RADS] category 0) and negative (BI-RADS category 1 or 2) cases. Two negative cases were excluded because of poor image quality. Cases with implants that had implant-displaced views were included (*n* = 4 in 474 case pool, *n* = 3 in 260 reader study cases).

the weighted dataset that were similar to the detected lesion on the basis of the algorithm classifier score, 70% were malignant and 30% were benign. The AI system did not process 2D images. The study cases were processed offsite by the AI system. The output of the AI system and the case images were loaded onto workstations for use during the reader study.

Readers

All 24 radiologists were Mammography Quality Standards Act qualified to read DBT studies and had read more than 500 DBT examinations in the prior 2 years; none practiced at acquisition sites. Readers had been in practice from 1 year to 34 years (median, 8 years); 13 (54%) were breast subspecialists who had devoted 75% or more of their time to breast imaging in the prior 3 years, and 11 (46%) were general radiologists who had devoted less than 75% of their time to breast imaging. Readers were trained with 30 cases separate from the 260 study cases. Readers were trained to review standard-view (left and right craniocaudal and mediolateral oblique) 2D images followed by standard-view DBT images with concurrent use of AI outlines and scores to assist in the identification of soft-tissue and calcific lesions.

Readings with and Those without AI

Cases were evaluated independently, with an individually randomized reading order, at Intrinsic Imaging's reading facility from January through March 2018. Readers were told that the sample of cases was enriched but were blinded to specific proportions. Readers were also blinded to types of cases and were not provided patient history, prior images, or acquisition site results. When detecting suspicious lesions, readers provided location, mammographic appearance (soft tissue, calcifications, or mixed), "forced" Breast Imaging Reporting and Data System (BI-RADS) (20) assessment category (1, 2, 3, 4A, 4B, 4C, or 5), and level of suspicion on a 0–100-point scale (with a score of 100 indicating the highest suspicion of malignancy). If no lesions were detected, readers provided case-level BI-RADS category and level of suspicion. Readers knew they were being timed but were blinded to the measurement from first viewing case images until determining whether the case had suspicious lesions. Assessments and reading time were missing for seven readings (three with AI, four without AI) because of reader and software errors; thus, 12 473 of 12 480 planned readings had complete data.

Table 1: Selection of 260 Study Cases from 474-Case Pool to Match a Screening Population within each of Negative, Recalled, Benign, and Cancer Cases

Case and Lesion Mammographic Characteristics	Case Pool (*n* = 474)	Randomly Selected Cases (*n* = 260)
Breast density in all cases (*n* = 474)		
Almost entirely fatty or scattered areas of fibroglandular density	219 (46.2)	133 (51.2)
Heterogeneously dense or extremely dense	255 (53.8)	127 (48.8)
Negative cases: BI-RADS category 1 or 2; not suspicious, no recall, no biopsy	212 (44.7)	109 (41.9)
BI-RADS category 1	173/212 (81.6)	89/109 (81.7)
BI-RADS category 2	39/212 (18.4)	20/109 (18.3)
Recalled cases: BI-RADS category 0; suspicious, recall, but no biopsy warranted	37 (7.8)	21 (8.1)
Soft-tissue densities (with or without calcifications)	26/37 (70.3)	16/21 (76.2)
Soft-tissue densities and calcifications	1/26 (3.8)	1/16 (6.3)
Calcifications only	11/37 (29.7)	5/21 (23.8)
Benign cases: BI-RADS category 3, 4, or 5; suspicious, recall, biopsy-proven benign	103 (21.7)	65 (25.0)
Soft-tissue densities (with or without calcifications)	63/103 (61.2)	46/65 (70.8)
Soft-tissue densities and calcifications	6/63 (9.5)	6/46 (13.0)
Calcifications only	40/103 (38.8)	19/65 (29.2)
Cancer cases: BI-RADS category 3, 4, or 5; suspicious, recall, biopsy-proven cancer	122 (25.7)	65 (25.0)
Soft-tissue densities (with or without calcifications)	100/122 (82.0)	50/65 (76.9)
Soft-tissue densities and calcifications	12/100 (12.0)	7/50 (14.0)
Calcifications only	22/122 (18.0)	15/65 (23.1)

Note.—Data are numbers of cases, with percentages in parentheses. BI-RADS = Breast Imaging Reporting and Data System.

Table 2: Histopathologic Findings in Cancer Cases

Histopathologic Findings of Malignant Lesions in Cancer Cases	Pool of Cancer Cases (*n* = 122)	Randomly Selected Cancer Cases (*n* = 65)
Invasive cancer (all ≤ 2.5 cm in size) with or without ductal carcinoma in situ	99/122 (81.1)	51/65 (78.5)
≤1.4 cm in size	49/99 (49.5)	26/51 (51.0)
Invasive lobular cancer	10/99 (10.1)	5/51 (9.8)
Ductal carcinoma in situ only (no size restriction)	23/122 (18.9)	14/65 (21.5)

Note.—Data are numbers of cases, with percentages in parentheses.

Statistical Analysis

Sample size was calculated (21,22) from a previous pilot reader study with the DBT AI system. Co-primary end points were noninferiority of case-level AUC and superiority (reduction) of reading time with versus without AI. Secondary end points were superiority of case-level AUC, noninferiority and superiority of case-level and lesion-level sensitivity, and noninferiority of specificity and recall rate in noncancers. End points were evaluated hierarchically in this prespecified, fixed sequence to protect the study type I error rate (α = .05) from inflation associated with multiple comparisons; hypothesis tests for secondary end points were performed only once the co-primary end points were met. When testing noninferiority, we placed limits on the amount by which performance with AI could be inferior to performance without AI (noninferiority margins) of 0.05. Nonparametric AUCs were based on case-level level of suspicion scores requiring correct localization of malignant le-

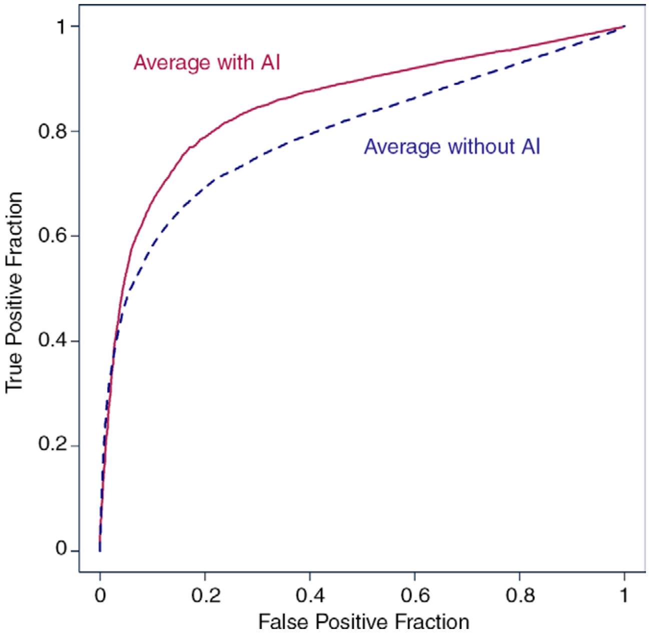

Figure 2: Average of empirical receiver operating characteristic plots with and without artificial intelligence (AI). True-positive fraction = case-level sensitivity, false-positive fraction = 1 − specificity.

sions. A normalizing transformation (23) was used to assess reading times, which were not normally distributed. Sensitivity and specificity were based on BI-RADS categories provided by readers, with category 3 or higher considered to be positive, while also requiring correct localization. Recall rate was based on whether the reader detected a suspicious lesion. End points were assessed (using $P < .05$ to indicate significance) with multireader, multicase analysis methods (24–26) programmed by a statistician (A.Y.T.) and allowing for missing reading data (27) and accounting for correlation between lesions in the same case for lesion-level sensitivity (28,29). The stand-alone performance of AI was evaluated through case-level sensitivity and specificity at the operating point with Wilson (30) 95% confidence intervals (CIs). A nonparametric ROC curve plotted the case-level sensitivity of AI and specificity based on case-level AI scores requiring correct lesion localization.

Results

Co-Primary End Point: Noninferior AUC

The average empirical nonparametric case-level ROC across readers with AI was higher than that without AI (Fig 2). Each reader's nonparametric trapezoidal AUC without AI, that with AI, and the difference between them are shown in Table 3. Twenty-two (92%) of 24 readers had higher AUCs with AI than without AI. The average AUC across readers without AI was 0.795, and the average AUC across readers with AI was 0.852. The average difference in AUC was 0.057 (two-sided 95% CI: 0.028, 0.087). The study successfully demonstrated noninferior AUC for noninferiority margin = 0.05 ($P < .01$).

Table 3: Estimated AUCs without and with AI

Reader Number	AUC without AI	AUC with AI	AUC Difference
Reader 1*	0.631 (0.035)	0.746 (0.035)	0.115 (0.045)
Reader 2	0.812 (0.033)	0.868 (0.028)	0.056 (0.028)
Reader 3*	0.790 (0.035)	0.852 (0.028)	0.062 (0.033)
Reader 4	0.844 (0.033)	0.868 (0.029)	0.024 (0.028)
Reader 5	0.747 (0.037)	0.825 (0.031)	0.078 (0.026)
Reader 6	0.824 (0.032)	0.894 (0.023)	0.070 (0.030)
Reader 7	0.850 (0.032)	0.853 (0.029)	0.003 (0.023)
Reader 8	0.905 (0.023)	0.891 (0.026)	−0.014 (0.025)
Reader 9	0.885 (0.026)	0.882 (0.026)	−0.004 (0.024)
Reader 10*	0.681 (0.035)	0.831 (0.029)	0.150 (0.037)
Reader 11	0.870 (0.028)	0.874 (0.028)	0.003 (0.023)
Reader 12*	0.784 (0.039)	0.849 (0.033)	0.066 (0.039)
Reader 13*	0.707 (0.038)	0.806 (0.032)	0.099 (0.032)
Reader 14	0.786 (0.038)	0.814 (0.033)	0.028 (0.040)
Reader 15*	0.828 (0.032)	0.878 (0.026)	0.049 (0.025)
Reader 16*	0.801 (0.033)	0.886 (0.025)	0.085 (0.034)
Reader 17	0.828 (0.032)	0.842 (0.030)	0.014 (0.028)
Reader 18	0.727 (0.037)	0.833 (0.031)	0.106 (0.036)
Reader 19*	0.820 (0.034)	0.878 (0.028)	0.058 (0.025)
Reader 20*	0.801 (0.035)	0.820 (0.034)	0.020 (0.033)
Reader 21	0.802 (0.043)	0.888 (0.027)	0.086 (0.043)
Reader 22	0.860 (0.030)	0.882 (0.027)	0.022 (0.027)
Reader 23*	0.756 (0.042)	0.836 (0.033)	0.079 (0.038)
Reader 24*	0.740 (0.041)	0.854 (0.030)	0.114 (0.034)
Average	0.795 (0.026)	0.852 (0.023)	0.057 (0.015)†
95% CI for average‡	0.743, 0.847	0.807, 0.897	0.028, 0.087

Note.—Data in parentheses are standard errors of the estimate. There were 65 cancer cases except for reader 4 (*n* = 64) and reader 17 (*n* = 64) without artificial intelligence (AI) and reader 1 (*n* = 64) with AI. There were 195 noncancer cases except for reader 2 (*n* = 194) and reader 7 (*n* = 194) without AI and reader 10 (*n* = 193) with AI. AUC = area under the receiver operating characteristic, CI = confidence interval.

* General radiologists who had devoted less than 75% of their time to breast imaging in the prior 3 years.

† Test of noninferiority for noninferiority margin = 0.05: T* (T statistic adjusted for correlation) = 7.17, $P < .01$. Test for difference: T* = 3.82, $P < .01$.

‡ CIs were obtained by using the Student *t* distribution, with Hillis (23) degrees of freedom limited to one less than the number of contributing observations: 340.4 for without AI, 2543.3 for with AI, and (1, 186.7) for the difference.

Co-Primary End Point: Superior Reading Time

Reading time improved with AI, as shown by the difference of reading time in seconds without AI minus that with AI and the normalizing-transformed percentage difference: natural log of (percentage difference + 100) minus natural log of 100 (Table 4). For difference in seconds, the average decrease in reading time with AI was 34.7 seconds (95% CI: 23.4, 46.0 seconds; $P < .01$). For percentage difference, the average on the untransformed scale was likely to underestimate the center of the distribution of improvement because that center was more heavily influenced by reading times that are longer without AI. Using the transformation, we obtained an average of −0.75 (95% CI: −0.95, −0.54; $P < .01$). Transforming back to the percentage difference scale provided an average improvement of 52.7% with AI (95% CI: 41.8%, 61.5%). Average reading time, if calculated on the untransformed scale, is likely to overestimate the center of the distribution because that center is influenced by longer reading times. Therefore, average reading times and associated 95% CIs were obtained by using a natural log transformation. Transforming back to seconds gives an average reading time without AI of 64.1 seconds (95% CI: 53.0, 77.5) and a reading time with AI of 30.4 seconds (95% CI: 24.8, 37.2). Figure 3 shows average reading times for each reader without and those with AI.

Table 4: Analysis of Differences in Reading Times without AI minus Those with AI

Reader Number	Reading Time Difference (sec)	Percentage Difference in Reading Time	log(Percentage Difference in Reading Time + 100) − log(100)
Reader 1*	−2.5 (1.6)	2.92 (5.67)	−0.38 (0.06)
Reader 2	−13.6 (2.4)	−5.51 (5.03)	−0.30 (0.04)
Reader 3*	−11.1 (1.4)	−10.49 (3.07)	−0.24 (0.03)
Reader 4	−26.0 (2.8)	−23.01 (5.05)	−0.76 (0.07)
Reader 5	−16.1 (3.1)	−2.99 (5.46)	−0.46 (0.06)
Reader 6	−34.6 (2.2)	−34.84 (2.52)	−0.61 (0.04)
Reader 7	−58.3 (3.2)	−44.80 (3.97)	−0.93 (0.05)
Reader 8	−65.7 (3.6)	−52.47 (2.93)	−1.11 (0.05)
Reader 9	−90.4 (2.8)	−71.67 (1.43)	−1.58 (0.05)
Reader 10*	−13.6 (0.9)	−36.57 (3.59)	−0.67 (0.04)
Reader 11	−18.8 (3.1)	−12.26 (4.22)	−0.46 (0.06)
Reader 12*	−84.3 (3.8)	−55.26 (7.02)	−1.53 (0.07)
Reader 13*	−8.0 (1.9)	−9.96 (4.28)	−0.35 (0.04)
Reader 14	−58.6 (2.9)	−74.61 (2.29)	−1.94 (0.06)
Reader 15*	−45.1 (4.8)	−18.26 (3.97)	−0.51 (0.05)
Reader 16*	−26.4 (1.0)	−45.59 (1.52)	−0.69 (0.02)
Reader 17	−48.6 (1.7)	−64.93 (1.38)	−1.21 (0.03)
Reader 18	1.4 (1.4)	24.63 (6.40)	−0.04 (0.05)
Reader 19*	−3.8 (3.6)	3.19 (4.12)	−0.14 (0.04)
Reader 20*	−52.3 (3.4)	−46.14 (3.27)	−1.07 (0.06)
Reader 21	−21.5 (1.6)	−27.01 (3.95)	−0.58 (0.05)
Reader 22	−26.8 (2.2)	−34.32 (3.08)	−0.67 (0.05)
Reader 23*	−72.5 (4.0)	−52.80 (3.29)	−1.13 (0.05)
Reader 24*	−35.4 (2.3)	−35.32 (2.10)	−0.58 (0.03)
Average	−34.7 (5.5)†	…	−0.75 (0.10)†
95% CI for average	−46.0, −23.4‡	…	−0.95, −0.54‡
Back-transformed average	…	−52.7%	…
95% CI for back-transformed average	…	−61.5%, −41.8%	…

Note.—Data in parentheses are standard errors of the estimate. There were 260 paired reading times for 65 cancer cases and 195 noncancer cases except for reader 1 (n = 259), reader 2 (n = 259), reader 4 (n = 259), reader 7 (n = 259), reader 10 (n = 258), and reader 17 (n = 259). AI = artificial intelligence, CI = confidence interval.

* General radiologists who had devoted less than 75% of their time to breast imaging in the prior 3 years.

† P value for test of null hypothesis: no difference versus alternate hypothesis. $P < .01$ for difference (seconds) and $P < .01$ for percentage difference using natural log of (percentage difference + 100) minus natural log of 100 to normalize.

‡ CIs were obtained by using the Student t distribution, with Hillis (23) degrees of freedom 24.1 for difference (normalizing transformation not required) and 25.3 for percentage difference using natural log of (percentage difference + 100) − natural log of 100 to normalize.

Secondary End Point: Superior AUC

The superior case-level average AUC with AI, 0.852, versus that without AI, 0.795 (Table 3) was statistically significant ($P < .01$ for test of difference in the hypothesis-testing sequence).

Secondary End Point: Case-level Sensitivity

Radiologists had superior sensitivity at the case level with AI (Table 5). Average sensitivity increased by 0.080 (95% CI: 0.026, 0.134), from 0.770 without AI to 0.850 with AI (in the hypothesis-testing sequence, $P < .01$ for noninferiority margin = 0.05, and $P < .01$ for test of difference). Eighteen (75%) of 24 readers had higher case-level sensitivity with AI, three (13%) readers had lower sensitivity, and sensitivity for three (13%) readers did not change. Average sensitivity increased by 0.120 with AI in the subgroup of 15 cancer cases with only calcifications (standard error of the estimate, 0.040) and by 0.068 in the subgroup of 50 cancer cases with at least one soft-tissue or mixed lesion (standard error of the estimate, 0.031). Figures 4 and 5 provide case examples where 10–12 more radiologists detected small invasive cancers with AI while reducing reading time.

Secondary End Point: Lesion-level Sensitivity

At the lesion level, radiologists also had superior sensitivity with AI. Average per-lesion sensitivity across readers increased

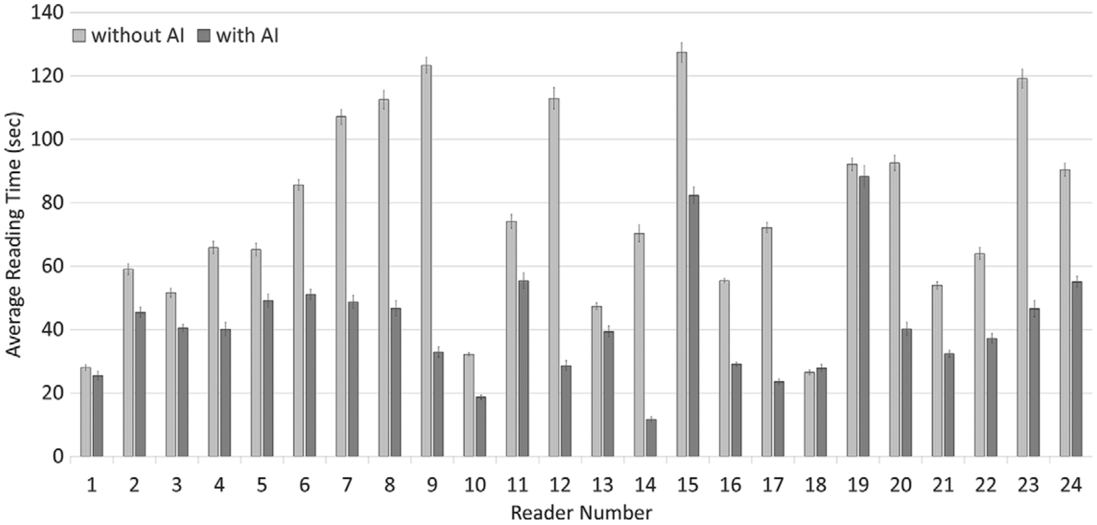

Figure 3: Bar graph shows average reading times for each reader without and with artificial intelligence (AI).

by 0.084 (95% CI: 0.029, 0.139), from 0.769 without AI to 0.853 with AI (in the hypothesis-testing sequence, $P < .01$ for noninferiority margin = 0.05, and $P < .01$ for test of difference). Nineteen (79%) of 24 readers had higher per-lesion sensitivity with AI, three (13%) readers had lower sensitivity, and sensitivity for two (8%) readers did not change.

Secondary End Point: Specificity

Radiologists had noninferior specificity with AI (Table 5). Twenty-one (88%) of the 24 readers had higher specificity with AI than without AI, and three (13%) readers had lower specificity. These averaged out to an increase of 0.069 in specificity (95% CI: 0.030, 0.108), from 0.627 without AI to 0.696 with AI (in the hierarchical, prespecified, fixed hypothesis-testing sequence, $P < .01$ for noninferiority margin = 0.05). Superiority of specificity was not included in the prespecified testing sequence, so a hypothesis test for superiority was not performed.

Secondary End Point: Recall Rate in Noncancers

Radiologists had a noninferior recall rate in noncancer cases with AI. In noncancer cases, lower recall rates are better than higher recall rates. Average recall rate in noncancer cases was 0.380 without AI and 0.309 with AI, with an average reduction of 0.072 (95% CI: 0.031, 0.112) (in the hierarchical, prespecified, fixed hypothesis-testing sequence, $P < .01$ for noninferiority margin = 0.05). Superiority of recall rate in noncancers was not included in the prespecified testing sequence, so a hypothesis test for superiority was not performed.

AI Stand-Alone Performance and Summary of Reader Study Results

Average case-level sensitivity, specificity, and reading time are summarized for each reader without and with AI, along with AI's stand-alone performance (no human reader), in Figure 6. The AI operating point case-level sensitivity was 91% (59 of 65; 95% CI: 81%, 96%), and its specificity was 41% (79 of 195; 95% CI: 34%, 48%). Sensitivity in cases with only calcifications was 100% (15 of 15; 95% CI: 80%, 100%) and was 88% (44 of 50; 95% CI: 76%, 94%) in cases with soft-tissue densities with or without calcifications.

Discussion

In our enriched multireader, multicase study, we demonstrated that when AI is concurrently incorporated into the interpretation of DBT studies, reading times can be significantly reduced while accuracy is improved. The AI system, which is based on a deep convolutional neural network algorithm that processes individual reconstructed DBT images, presents the readers with outlines and locations of soft-tissue and calcific breast lesions. In addition, the AI system assigns a "certainty of finding" score for each lesion and for the entire case. When DBT readers are concurrently presented with this information during interpretation, reading times, on average, are cut in half. This significant reduction in reading time mitigates the approximately twofold increase in reading times reported for DBT compared with reading DM-alone studies (1,8,9). This improvement in reading time thus has the potential to significantly improve the efficiency of DBT workflow in busy breast imaging clinics.

Reading times varied by reader; the average time for DBT reading without AI was 64.1 seconds, compared with 30.4 seconds with AI. Note that in our study, when readers interpreted DBT with AI, the AI lesion and case-based data appeared concurrently, at the opening of the DBT study. This reading scheme is different from the typical presentation of CAD results at the end of reading, which adds to interpretation time (15). The high stand-alone performance of this AI system (91% sensitivity, 41% specificity) and the availability of confidence scores for each lesion along with concurrent use may have been important contributing factors that resulted in better outcomes than those with conventional CAD systems (14). The concurrent use of AI from the start of DBT reading will be a change and will possibly

Table 5: Estimated Case-Level Sensitivity and Specificity without and Those with AI

Reader Number	Sensitivity without AI	Sensitivity with AI	Sensitivity Difference	Specificity without AI	Specificity with AI	Specificity Difference
Reader 1*	0.385 (0.060)	0.625 (0.061)	0.240 (0.078)	0.846 (0.026)	0.821 (0.027)	−0.026 (0.034)
Reader 2	0.754 (0.053)	0.831 (0.047)	0.077 (0.050)	0.758 (0.031)	0.821 (0.027)	0.063 (0.028)
Reader 3*	0.769 (0.052)	0.815 (0.048)	0.046 (0.063)	0.667 (0.034)	0.723 (0.032)	0.056 (0.032)
Reader 4	0.859 (0.043)	0.877 (0.041)	0.018 (0.041)	0.518 (0.036)	0.641 (0.034)	0.123 (0.034)
Reader 5	0.677 (0.058)	0.815 (0.048)	0.138 (0.043)	0.718 (0.032)	0.733 (0.032)	0.015 (0.038)
Reader 6	0.846 (0.045)	0.954 (0.026)	0.108 (0.049)	0.595 (0.035)	0.626 (0.035)	0.031 (0.031)
Reader 7	0.831 (0.047)	0.846 (0.045)	0.015 (0.041)	0.655 (0.034)	0.754 (0.031)	0.099 (0.030)
Reader 8	0.923 (0.033)	0.877 (0.041)	−0.046 (0.046)	0.728 (0.032)	0.790 (0.029)	0.062 (0.025)
Reader 9	0.862 (0.043)	0.862 (0.043)	0.000 (0.044)	0.805 (0.028)	0.815 (0.028)	0.010 (0.027)
Reader 10*	0.523 (0.062)	0.846 (0.045)	0.323 (0.066)	0.821 (0.027)	0.736 (0.032)	−0.085 (0.036)
Reader 11	0.938 (0.030)	0.892 (0.038)	−0.046 (0.040)	0.400 (0.035)	0.631 (0.035)	0.231 (0.039)
Reader 12*	0.877 (0.041)	0.877 (0.041)	0.000 (0.058)	0.221 (0.030)	0.477 (0.036)	0.256 (0.042)
Reader 13*	0.585 (0.061)	0.769 (0.052)	0.185 (0.057)	0.738 (0.031)	0.759 (0.031)	0.021 (0.034)
Reader 14	0.877 (0.041)	0.877 (0.041)	0.000 (0.049)	0.292 (0.033)	0.487 (0.036)	0.195 (0.039)
Reader 15*	0.831 (0.047)	0.908 (0.036)	0.077 (0.040)	0.646 (0.034)	0.682 (0.033)	0.036 (0.030)
Reader 16*	0.723 (0.056)	0.908 (0.036)	0.185 (0.065)	0.790 (0.029)	0.713 (0.032)	−0.077 (0.033)
Reader 17	0.781 (0.052)	0.769 (0.052)	−0.012 (0.050)	0.774 (0.030)	0.872 (0.024)	0.097 (0.027)
Reader 18	0.585 (0.061)	0.785 (0.051)	0.200 (0.062)	0.785 (0.029)	0.795 (0.029)	0.010 (0.032)
Reader 19*	0.800 (0.050)	0.892 (0.038)	0.092 (0.042)	0.600 (0.035)	0.651 (0.034)	0.051 (0.033)
Reader 20*	0.754 (0.053)	0.800 (0.050)	0.046 (0.051)	0.677 (0.033)	0.738 (0.031)	0.062 (0.036)
Reader 21	0.846 (0.045)	0.908 (0.036)	0.062 (0.053)	0.585 (0.035)	0.662 (0.034)	0.077 (0.033)
Reader 22	0.846 (0.045)	0.877 (0.041)	0.031 (0.049)	0.656 (0.034)	0.733 (0.032)	0.077 (0.032)
Reader 23*	0.892 (0.038)	0.938 (0.030)	0.046 (0.046)	0.256 (0.031)	0.374 (0.035)	0.118 (0.043)
Reader 24*	0.723 (0.056)	0.862 (0.043)	0.138 (0.053)	0.523 (0.036)	0.672 (0.034)	0.149 (0.038)
Average	0.770 (0.039)	0.850 (0.032)	0.080 (0.027)†	0.627 (0.041)	0.696 (0.033)	0.069 (0.019)‡
95% CI for average§	0.692, 0.848	0.788, 0.913	0.026, 0.134	0.544, 0.710	0.631, 0.761	0.030, 0.108

Note.—Data in parentheses are standard errors of the estimate. There were 65 cancer cases except for reader 4 (*n* = 64) and reader 17 (*n* = 64) without artificial intelligence (AI) and reader 1 (*n* = 64) with AI. There were 195 noncancer cases except for reader 2 (*n* = 194) and reader 7 (*n* = 194) without AI and reader 10 (*n* = 193) with AI. CI = confidence interval.

* General radiologists who had devoted less than 75% of their time to breast imaging in the prior 3 years.

† Test of case-level sensitivity noninferiority for noninferiority margin = 0.05: T^* (T statistic adjusted for correlation) = 4.76, $P < .01$. Test for case-level sensitivity difference: $T^* = 2.93$, $P < .01$.

‡ Test of specificity noninferiority for noninferiority margin = 0.05: $T^* = 6.12$, $P < .01$.

§ CIs were obtained by using the Student *t* distribution with Hillis (23) degrees of freedom limited to one less than the number of contributing observations: 92.7 for case-level sensitivity without AI, 600.2 for case-level sensitivity with AI, and (1, 95.5) for case-level sensitivity difference; 35.9 for specificity without AI, 77.6 for specificity with AI, and (1, 38.9) for specificity difference.

require a learning curve for most breast imagers. However, because concurrent use of AI yields significant improvements in both efficiency and accuracy, such a reading protocol with AI should be considered by all breast imagers.

AUC is an overall measure of accuracy that combines case-level sensitivity and specificity into a single metric. Although improvements in AUC have been difficult to show with the use of AI or CAD with DBT (16–18), our study showed a statistically significant 0.057 average improvement in AUC, even while significantly cutting reading time approximately in half on average. In terms of case-level sensitivity, specificity, and reading time for each reader, all readers benefitted from using AI: Most of the readers (14 of 24) had improvements in sensitivity, specificity, and reading time; the three readers who had reductions in sensitivity and the three readers who had no change in sensitivity all had increases in specificity and faster reading times; the three readers who had reductions in specificity had increases in sensitivity and faster reading times; and the one reader who had a slower reading time had increases in sensitivity and specificity. Of note is the overall decrease in the variability of the group of readers' performances and reading times when reading with AI, demonstrated by the general shift of readers' paired (sensitivity, specificity) locations to the upper left corner of the ROC plane (increases in AUC) and the overall decreases in the reading times and variability of the paired (sensitivity, specificity) locations.

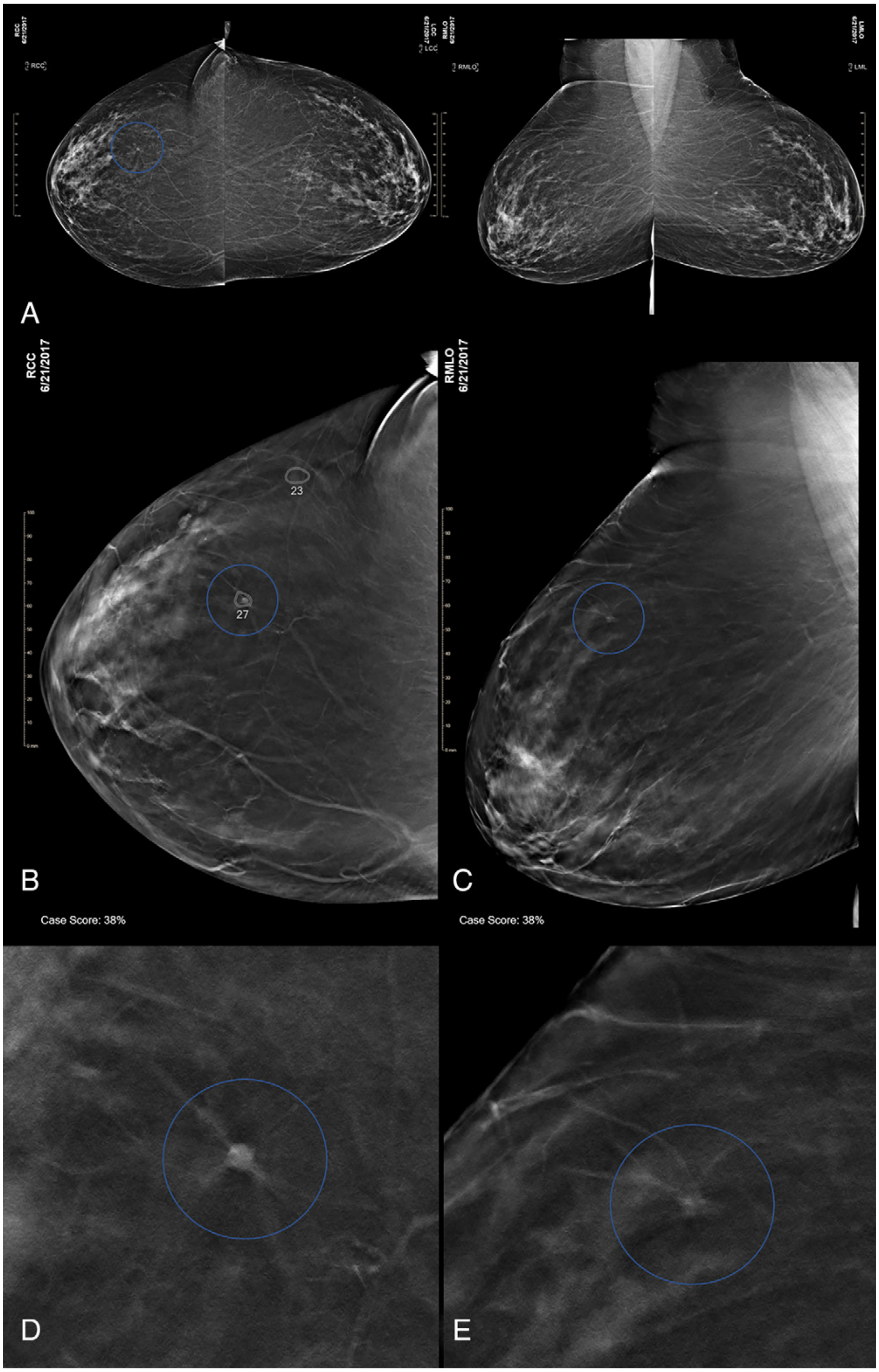

Figure 4: Images in a 74-year-old woman at screening with combination digital mammography (DM) and digital breast tomosynthesis (DBT). *A,* DM views show small, focal asymmetry that is seen only in the right craniocaudal (RCC) view. Right DBT views show small spiculated mass in upper outer quadrant, better seen on, *B,* RCC than, *C,* right mediolateral oblique (RMLO) view. The artificial intelligence (AI) case score of 38% was displayed at the bottom of the DBT views, and two AI outlines and lesion scores (27 for small spiculated mass; 23 for false-positive finding) were displayed on the RCC DBT view. Readers could click on the outlines on any DBT image and automatically advance to the DBT image where the lesion was detected by AI. *D, E,* Zoomed, *D,* RCC, and, *E,* RMLO DBT views show small spiculated mass that proved to be an 8-mm invasive ductal carcinoma (estrogen receptor positive, progesterone receptor positive, human epidermal growth factor receptor 2 negative, low Ki67 level). Twelve more readers detected the malignant mass with AI, while reducing average reading time across all 24 readers: Six (25%) of the 24 readers detected the mass without AI (reading time, 77.6 seconds), and 18 (75%) readers detected it with AI (reading time, 57.0 seconds).

At the case level, 75% of readers had higher sensitivities with AI, and at the lesion level, 79% had higher sensitivities. Improvements in overall sensitivity were greater for cancer cases with calcifications only than for those that were soft tissue only or mixed soft tissue and calcifications. This may be related to the size of the lesions and the overall improved conspicuity of soft-tissue lesions compared with calcifications-only lesions at DBT. In addition, some of the cases did not have DM images but instead had only synthetic 2D images combined with DBT (SM/DBT), for which there have been reports of lower rates of detection of ductal carcinomas in situ (8,31,32). Perhaps AI provides an even larger improvement in sensitivity for calcifications-only malignant lesions when SM/DBT is performed (without DM); subgroup analysis and larger sample sizes are needed to investigate the role of AI in SM/DBT.

There were limitations to our study. First, it is well known that radiologists may behave differently in reader studies than they do in clinical practice, and therefore, some of the findings from our enriched reader study may not translate to the clinical workplace (33). Although the study included 13 breast subspecialists and 11 general radiologists, other

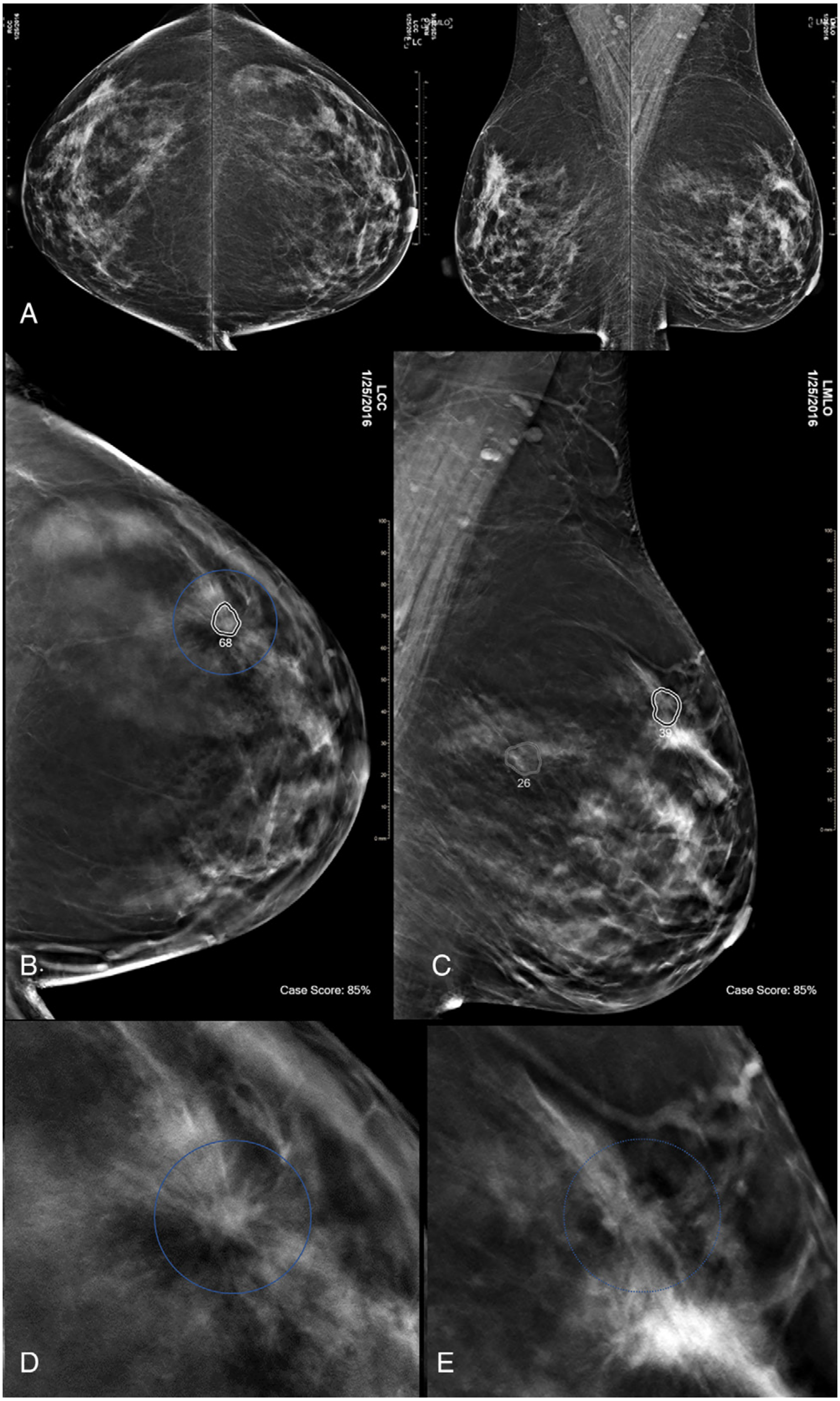

Figure 5: Images in a 47-year-old woman at screening with combination digital mammography (DM) and digital breast tomosynthesis (DBT). *A,* DM views show no suspicious findings. Left DBT views show 7-mm spiculated mass in outer breast seen only in, *B,* left craniocaudal (LCC) view and not seen in, *C,* left mediolateral oblique (LMLO) view. The artificial intelligence (AI) case score of 85% was displayed at the bottom of DBT views, with one AI outline and lesion score (68 for spiculated mass) displayed on the LCC DBT view and two AI outlines and lesion scores (39 for potential correlate of spiculated mass; 26 for false-positive finding) displayed on the LMLO DBT view. Readers could click on the outlines on any DBT image and automatically advance to the DBT image where the lesion was detected by AI. *D,E,* Zoomed, *D,* LCC, and, *E,* LMLO DBT views show spiculated mass (dotted circle for potential correlate on LMLO), which proved to be a 5-mm invasive ductal carcinoma with associated ductal carcinoma in situ (estrogen receptor positive, progesterone receptor positive, human epidermal growth factor receptor 2 negative, low Ki67 level). Ten more readers detected the malignant mass with AI, while reducing average reading time across all 24 readers: Twelve (50%) of 24 readers detected the mass without AI (reading time, 110.3 seconds) and 22 (92%) readers detected it with AI (reading time, 62.3 seconds).

limitations to our study may have been the number of readers and variability of their experience. In addition, outcome analyses by patient, cancer subtypes, and detection method (ie, cancers detected vs those that had prior false-negative screening results) are needed to fully understand the clinical impact of this AI system. These limitations and others might be addressed with a larger prospective study.

Finally, any strategy to achieve gains in clinical efficiency must not come at the risk of decreasing patient outcomes (here, the accuracy of DBT imaging interpretations). We have shown in our reader study that the concurrent use of AI in DBT interpretation has the potential to improve both efficiency and accuracy. As machine learning methods advance with exposure to larger and larger datasets and the adoption of AI expands, we expect the impact on patient outcomes at the individual level to only improve.

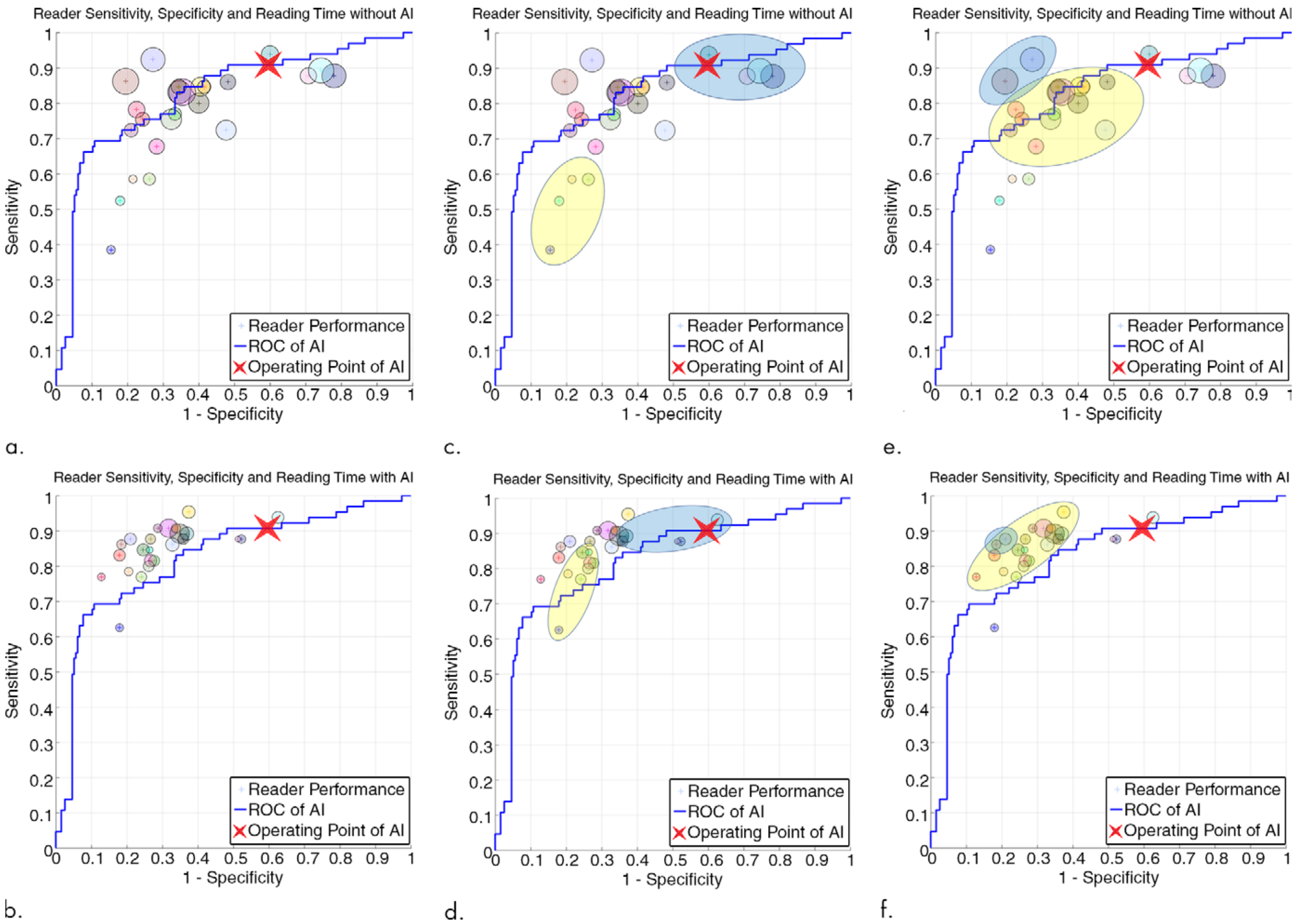

Figure 6: Graphs show **(a, c, e)** average case-level performance for each reader without artificial intelligence (AI) and **(b, d, f)** with AI. Locations of small circles = sensitivity and specificity. Diameters of small circles are proportional to reading time; a decrease in circle size from readings without AI to readings with AI reflects the relative decrease in reading time for the individual reader. **(c–f)** Graphs highlight specific groups of readers and their changes in sensitivity, specificity, and reading time. The large yellow circle in **c** and **d** shows a group of four readers who without AI have high specificity, low sensitivity, and short reading times. With AI, these readers maintain their relatively short reading times and high specificity but improve their sensitivity. As demonstrated by the large blue circle in **c** and **d**, four readers who have low specificity but high sensitivity without AI improve their reading times and specificity and maintain their relatively high sensitivity. The large yellow circle in **e** and **f** shows a group of readers with generally high sensitivities and specificities without AI who with AI generally improve all three parameters (reading time, sensitivity, and specificity). The large blue circle in **e** and **f** indicates the only two readers in our study who had slight decreases in area under the receiver operating characteristic (ROC) curve (AUC) when reading with AI compared with reading without AI. Their reductions in AUC were quite small at −0.014 (right small circle) and −0.004 (left small circle); however, they both experienced significant reductions in reading time. The right reader had the fourth largest reduction of all readers of −65.7 seconds, and the left reader had the largest reduction of all, −90.4 seconds. **(a–f)** Graphs also show the AI stand-alone performance ROC curve (no human reader, blue line) and operating point (red "X") with the 260 enriched reader study cases. The AI operating point case-level sensitivity was 91% (59 of 65; 95% confidence interval [CI]: 81%, 96%), and its specificity was 41% (79 of 195; 95% CI: 34%, 48%).

Acknowledgments: The authors thank the 24 radiologist readers for their dedication and participation in the study, including Ronald J. Bick, MD, David J. Bindman, MD, Sandra B. Brennan, MD, Carolynn M. DeBenedectis, MD, Erica T. Ghosh, MD, Neil A. Grossman, MD, Lauren A. Hollowell, MD, Rebecca C. Hultman, DO, Gregg A. Miller, MD, Adrienne R. Newburg, MD, Brittany L. Ritchie, MD, Hamid Salamipour, MD, Hadeer H. Shaikhly, MD, Anatoly M. Sukharsky, MD, Michelle N. Swire, MD, Monique M. Tyminski, DO, Jade E. Watkins, MD, and Kira A. Wendorf, MD. We also appreciate Meridith Blevins Peratikos, MS, for providing statistical expertise and support.

Author contributions: Guarantors of integrity of entire study, S.P., J.E.B.; 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, E.F.C., J.E.B., J.W.H.; clinical studies, S.P., J.G., J.E.B.; experimental studies, S.V.F., J.G., J.E.B.; statistical analysis, E.F.C., A.Y.T., J.E.B.; and manuscript editing, all authors.

Disclosures of Conflicts of Interest: **E.F.C.** Activities related to the present article: institution received grant from iCAD to support E.F.C.'s time for oversight of the reader study and the resulting data analysis. Activities not related to the present article: is a consultant for Hologic; institution has received a grant from Hologic for data review and analysis; institution has received a PO1 grant (multi-institutional) from the National Cancer Institute to allow retrospective review of data and analysis; has received payment from ICPME for lectures on digital breast tomosynthesis. Other relationships: disclosed no relevant relationships. **A.Y.T.** Activities related to the present article: is an employee of Biostatistics Consulting and served as the statistical team lead for the study from design through analysis, in the role of consultant to the device manufacturer; was paid for writing or reviewing the manuscript as part of the consulting services provided by institution to the device manufacturer; is a consultant for iCAD. Activities not related to the present article: Biostatistics Consulting provides similar statistical services for entities other than this device manufacturer. None of those projects are CAD/AI for breast imaging, the device studied in the current project. Other relationships: disclosed no relevant relationships. **S.P.** Activities related to the present article: is employed by iCAD as Vice President of Research. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **S.V.F.** Activities related to the present article: is an employee of and owns stock or stock options in iCAD. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **J.G.** Activities related to the present article: is employed by iCAD as Chief Technical Officer; owns stock or stock options in iCAD as a term of employment. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **J.E.B.** Activities related to the present article: is employed by and is a shareholder in Intrinsic Imaging; Intrinsic Imaging performed the reader study, which was financially supported by iCAD. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **J.W.H.** Activities related to the present article: is employed by iCAD as Vice President and Medical Director. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

References

1. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. Radiology 2013;267(1):47–56.
2. Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA 2014;311(24):2499–2507.
3. Sharpe RE Jr, Venkataraman S, Phillips J, et al. Increased cancer detection rate and variations in the recall rate resulting from implementation of 3D digital breast tomosynthesis into a population-based screening program. Radiology 2016;278(3):698–706.
4. Hooley RJ, Durand MA, Philpotts LE. Advances in digital breast tomosynthesis. AJR Am J Roentgenol 2017;208(2):256–266.
5. ACR Statement on Breast Tomosynthesis. American College of Radiology Web site. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Breast-Tomosynthesis. Published November 24, 2014. Accessed October 24, 2018.
6. McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. JAMA Oncol 2016;2(6):737–743.
7. 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.
8. Gilbert FJ, Tucker L, Gillan MGC, et al. Accuracy of digital breast tomosynthesis for depicting breast cancer subgroups in a UK retrospective reading study (TOMMY Trial). Radiology 2015;277(3):697–706.
9. Bernardi D, Ciatto S, Pellegrini M, et al. Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time. Br J Radiol 2012;85(1020):e1174–e1178.
10. 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.
11. 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.
12. James JJ, Gilbert FJ, Wallis MG, et al. Mammographic features of breast cancers at single reading with computer-aided detection and at double reading in a large multicenter prospective trial of computer-aided detection: CADET II. Radiology 2010;256(2):379–386.
13. 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.
14. 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.
15. Tchou PM, Haygood TM, Atkinson EN, et al. Interpretation time of computer-aided detection at screening mammography. Radiology 2010;257(1):40–46.
16. Benedikt RA, Boatsman JE, Swann CA, Kirkpatrick AD, Toledano AY. Concurrent computer-aided detection improves reading time of digital breast tomosynthesis and maintains interpretation performance in a multireader multicase study. AJR Am J Roentgenol 2018;210(3):685–694.
17. Balleyguier C, Arfi-Rouche J, Levy L, et al. Improving digital breast tomosynthesis reading time: A pilot multi-reader, multi-case study using concurrent Computer-Aided Detection (CAD). Eur J Radiol 2017;97:83–89.
18. Chae EY, Kim HH, Jeong JW, Chae SH, Lee S, Choi YW. Decrease in interpretation time for both novice and experienced readers using a concurrent computer-aided detection system for digital breast tomosynthesis. Eur Radiol 2019;29(5):2518–2525.
19. Clinical Performance Assessment. Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions - Guidance for Industry and FDA Staff. U.S. Food & Drug Administration Web site. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology. Published July 3, 2012. Accessed October 24, 2018.
20. Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS Mammography. In: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. 5th ed. Reston, Va: American College of Radiology, 2013; 46–74.
21. 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.
22. Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR Am J Roentgenol 2000;175(3):603–608.
23. Bartlett MS. The use of transformations. Biometrics 1947;3(1):39–52.
24. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. Commun Stat Simul Comput 1995;24(2):285–308.
25. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Stat Med 2007;26(3):596–619.
26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
27. Zhou XH, Gatsonis CA. A simple method for comparing correlated ROC curves using incomplete data. Stat Med 1996;15(15):1687–1693.
28. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. Biometrics 1992;48(2):577–585.
29. Obuchowski NA. On the comparison of correlated proportions for clustered data. Stat Med 1998;17(13):1495–1507.
30. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927;22(158):209–212.
31. Zuckerman SP, Conant EF, Keller BM, et al. Implementation of Synthesized Two-dimensional Mammography in a Population-based Digital Breast Tomosynthesis Screening Program. Radiology 2016;281(3):730–736.
32. Aujero MP, Gavenonis SC, Benjamin R, Zhang Z, Holt JS. Clinical Performance of Synthesized Two-dimensional Mammography Combined with Tomosynthesis in a Large Screening Population. Radiology 2017;283(1):70–76.
33. 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.