Radiology

ORIGINAL RESEARCH • BREAST IMAGING

Artificial Intelligence for Reducing Workload in Breast Cancer Screening with Digital Breast Tomosynthesis

Yoel Shoshan, BSc • Ran Bakalo, MSc* • Flora Gilboa-Solomon, MSc • Vadim Ratner, PhD • Ella Barkan, MA • Michal Ozery-Flato, PhD • Mika Amit, PhD • Daniel Khapun, BSc • Emily B. Ambinder, MD • Eniola T. Oluyemi, MD, MPH • Babita Panigrahi, MD • Philip A. DiCarlo, MD • Michal Rosen-Zvi, PhD • Lisa A. Mullen, MD*

From the Department of Healthcare Informatics, IBM Research, IBM R&D Laboratories, University of Haifa Campus, Mount Carmel, Haifa 3498825, Israel (Y.S., R.B., F.G.S., V.R., E.B., M.O.F., M.A., D.K., M.R.Z.); and The Russell H. Morgan Department of Radiology and Radiological Science, Breast Imaging Division, Johns Hopkins Medicine, Baltimore, Md (E.B.A., E.T.O., B.P., P.A.D., L.A.M.). Received May 6, 2021; revision requested June 23; revision received October 27; accepted November 8. **Address correspondence to** F.G.S. (e-mail: *flora@il.ibm.com*).

* Y.S. and R.B. contributed equally to this work.

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

See also the editorial by Philpotts in this issue.

Radiology 2022; 303:69–77 • https://doi.org/10.1148/radiol.211105 • Content codes: BR AI

Background: Digital breast tomosynthesis (DBT) has higher diagnostic accuracy than digital mammography, but interpretation time is substantially longer. Artificial intelligence (AI) could improve reading efficiency.

Purpose: To evaluate the use of AI to reduce workload by filtering out normal DBT screens.

Materials and Methods: The retrospective study included 13 306 DBT examinations from 9919 women performed between June 2013 and November 2018 from two health care networks. The cohort was split into training, validation, and test sets (3948, 1661, and 4310 women, respectively). A workflow was simulated in which the AI model classified cancer-free examinations that could be dismissed from the screening worklist and used the original radiologists' interpretations on the rest of the worklist examinations. The AI system was also evaluated with a reader study of five breast radiologists reading the DBT mammograms of 205 women. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and recall rate were evaluated in both studies. Statistics were computed across 10 000 bootstrap samples to assess 95% CIs, noninferiority, and superiority tests.

Results: The model was tested on 4310 screened women (mean age, 60 years ± 11 [standard deviation]; 5182 DBT examinations). Compared with the radiologists' performance (417 of 459 detected cancers [90.8%], 477 recalls in 5182 examinations [9.2%]), the use of AI to automatically filter out cases would result in 39.6% less workload, noninferior sensitivity (413 of 459 detected cancers; 90.0%; $P = .002$), and 25% lower recall rate (358 recalls in 5182 examinations; 6.9%; $P = .002$). In the reader study, AUC was higher in the standalone AI compared with the mean reader (0.84 vs 0.81; $P = .002$).

Conclusion: The artificial intelligence model was able to identify normal digital breast tomosynthesis screening examinations, which decreased the number of examinations that required radiologist interpretation in a simulated clinical workflow.

Online supplemental material is available for this article.

Breast cancer is the second-leading cause of cancer-related death among women in developed countries (1). Although digital mammography is the most common examination used for breast cancer screening, digital breast tomosynthesis (DBT) improves cancer detection (2) and lowers recall rate (3). Although DBT interpretation time is almost twice that required for reading digital mammograms (4), its use is expected to show progressive growth worldwide (5). This results in an increased burden for the radiologist and higher cost for screening programs.

The use of artificial intelligence (AI) models could help to save time in the assessment of breast screening examinations. Several studies have introduced successful AI technologies for two-dimensional mammographic interpretation (6–10). Yala et al (10) reported a 19.3% worklist reduction, and McKinney et al (8) reported a 34.8% reduction at AI testing on digital mammograms. Conant et al (11) presented a reader study showing that reading time was reduced by highlighting suspicious areas in each image by computer-aided detection software. However, with computer-aided detection the radiologist still needs to read all screening examinations, although most of them are cancer free (12). Raya-Povedano et al (13) recently showed impressive work reduction (up to approximately 70%) when comparing a simulation of an AI-assisted system to radiologist interpretation in digital mammography and DBT screening. However, the test set included examinations from only one screening site, and the number of cancer cases was relatively small (113 cancers).

In our study, we propose an AI model to detect cancer-free screening examinations that could be dismissed without consulting a radiologist to reduce workloads. Our study included a large DBT screening data set with a substantial number of biopsy-proven examinations (1472 malignant cases and 2232 benign cases) collected from 22 clinical sites. In addition, our AI model examined both the DBT images and the clinical information with each DBT examination. The purpose of our study was to develop an

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, BI-RADS = Breast Imaging Reporting and Data System, DBT = digital breast tomosynthesis

Summary

An artificial intelligence system filtered out cancer-free digital breast tomosynthesis examinations, led to lower recall rates, and reduced the number of examinations in the simulated workflow.

Key Results

- In a retrospective study of 5182 digital breast tomosynthesis screening examinations, the artificial intelligence (AI) model reduced screening workload by 39.6% while maintaining noninferior sensitivity (90.0% vs 90.8%; $P < .001$).
- In a simulation, the AI model filtered out cancer-free examinations, which also led to a 25% decrease in the number of women who would have been recalled (6.9% vs 9.2%; $P < .001$).

AI model that could filter out normal DBT studies to reduce screening workloads while improving diagnostic accuracy. We also performed a reader study to assess the effect of the use of an AI model in a simulated clinical workflow.

Materials and Methods

Our study included the following two health care networks: Johns Hopkins Medicine institutional review boards approved the use of their data, with a waiver of the need to obtain written informed consent for this study, which was compliant with the Health Insurance Portability and Accountability Act; and a U.S. health care network, which provided institutional review board–exempt, retrospective, deidentified data that was approved for secondary use by IBM. The study was not financially supported by a grant or external company.

Data Collection and Ground Truth Definition

We randomly sampled DBT examinations of women examined between June 2013 and November 2018 in two large health care networks in the United States, spanning 22 imaging sites. In total, we gathered examinations of 13 043 individuals (9938 from health care network 1 and 3105 from health care network 2). Images were acquired with a Selenia Dimensions device (Hologic) and with combined digital mammography and DBT or synthesized mammography and DBT. We excluded men, as well as women with pacemakers, implants, and prior breast surgery (Fig 1). We extracted the women's age, ethnicity, hormone therapy, gynecologic history, family history, and prior Breast Imaging Reporting and Data System (BI-RADS) assessments from the medical records. A DBT examination was labeled as positive for cancer if there was a biopsy with a positive finding within 12 months of the examination date. An examination was considered cancer free if there was no positive biopsy finding within 12 months of the examination date and a follow-up study was performed 11–36 months from the examination date (Fig 2B). We conducted two studies: a retrospective study in which we developed the AI system and tested worklist reduction in screening population prevalence, and a reader study.

Development of the AI System

The cohort of 9919 women (13 306 DBT examinations) was split into a training set of 3948 women (804 cancers [20.3%]), a validation set of 1661 women (182 cancers [11.0%]), and a test set of 4310 women (453 cancers [10.5%]) (main test). Examinations in the same woman appear within only one of these data sets. The training data set included examinations from 18 imaging facilities. The validation set was used for model selection and calibration of the AI model (Appendix E1 [online]). A subset of the main test (Fig 2C, Table E1 [online]), which we named sites test (2375 women), included examinations from four sites and was used to test the generalizability of AI to unseen sites. The AI model is an ensemble of 50 different classifiers; 45 of them are deep learning classifiers that processed all four DBT views, and five are machine learning classifiers that processed the clinical information (eg, age, ethnicity, body mass index, hormone therapy, gynecologic history, family history, breast density) and information from the Digital Imaging and Communications in Medicine tags, such as compression force. Figure E1 (online) shows the analysis of the most important clinical features to affect the classification. The AI model outputs, per DBT case, a single malignancy score between 0 and 1.0, where higher scores are more indicative for cancer. Additional details are provided in Appendix E2 (online), and code is available at *https://github.com/IBM/work-reduction-dbt*.

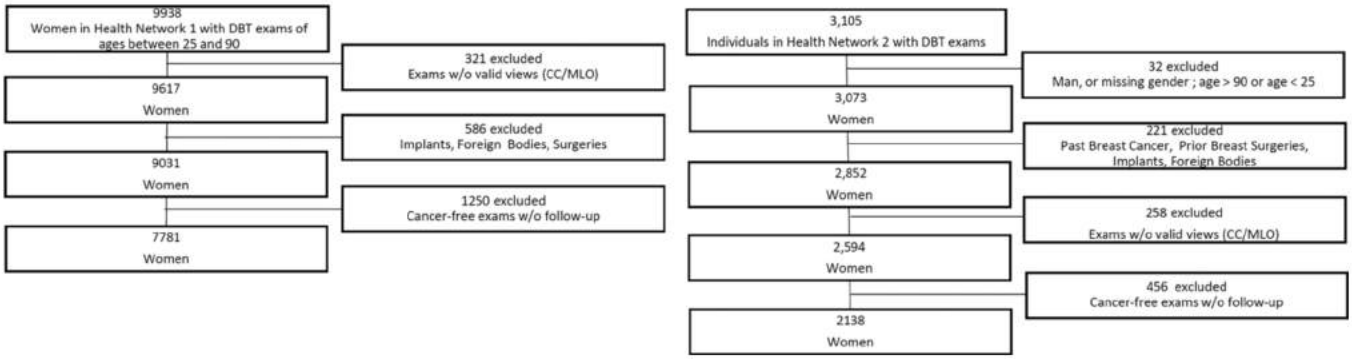

Figure 1: Flowchart of study inclusion and exclusion for both health care networks. Women with implants were excluded because they would have more than the four standard views; women with breast surgeries were excluded because the breast distortion and scars might be learned incorrectly by the artificial intelligence model as breast cancer. However, a radiologist interpreting the images would have access to the clinical reports and would be aware of the previous surgery, such that the possibility of breast cancer in that location would be dismissed. CC = craniocaudal view, DBT = digital breast tomosynthesis, MLO = mediolateral oblique view.

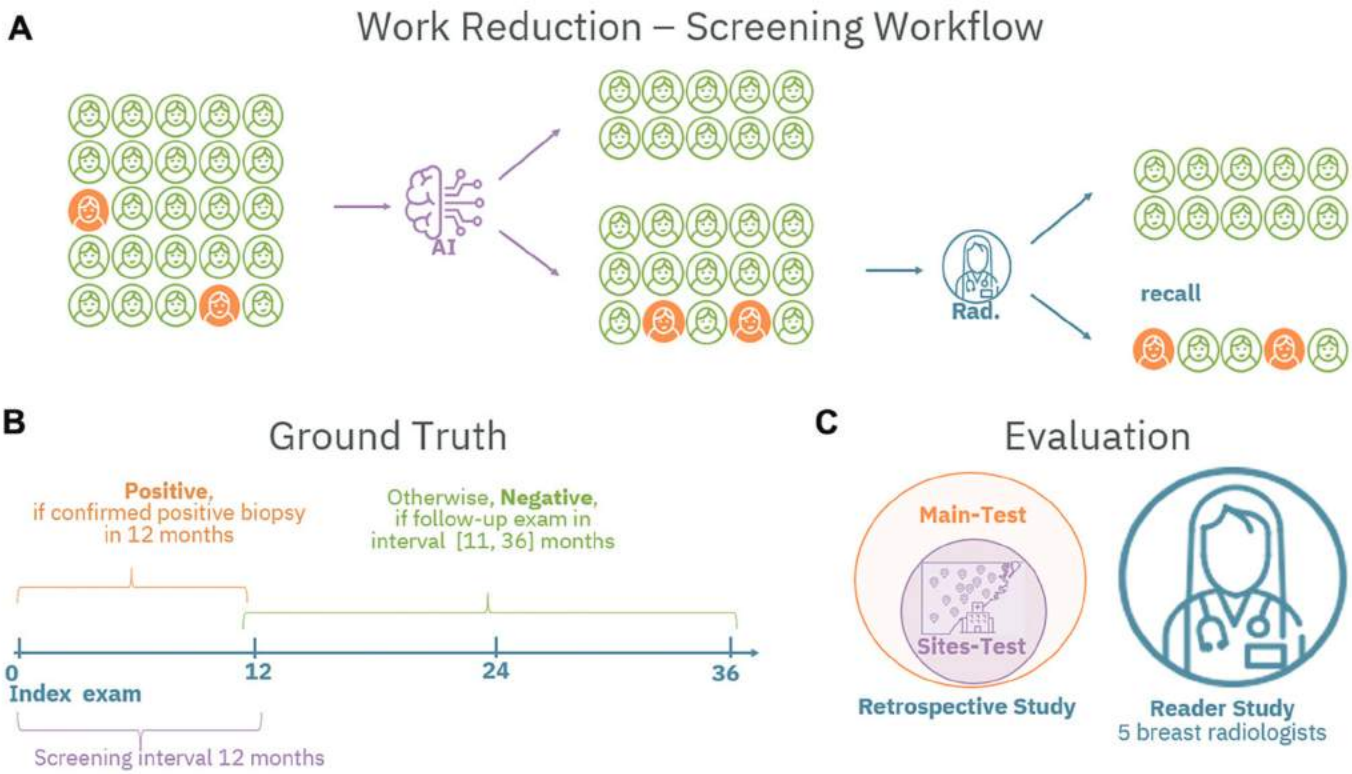

Figure 2: Artificial intelligence (AI) model to assist digital breast tomosynthesis screening. **(A)** Study goal was to demonstrate the ability of AI to reduce radiologists' workload by filtering out a portion of the cancer-free examinations. **(B)** Outcomes were derived from the biopsy results and longitudinal follow-up. **(C)** Evaluation included testing on enriched data set; simulation of AI with reader in worklist-reduction flow (main test); generalization, examined at unseen sites (sites test); and assessment of the potential assistance of AI to readers via a reader study. Rad = radiologist.

Worklist Reduction Simulation

Performance of radiologists combined with the AI model was simulated as follows: the AI model analyzed all DBT examinations; examinations that were classified as cancer free with high confidence were removed from the radiologists' simulated worklist as no recall. For the remaining examinations, the original radiologist recall or no-recall assessments were used (Fig 2A). Performance was measured by comparing recall or no-recall decisions to biopsy-based ground truth.

Reader Study Design

The reader study included five U.S. board-certified breast imaging–specialized radiologists; years of breast imaging and DBT experience, respectively, are given in parentheses: B.P. (1 year for both), E.B.A. (1 year for both), E.T.O. (6 years for both), P.A.D. (8 and 7 years), and L.A.M. (23 and 7 years). The 205 screening examinations were randomly sampled from the main test to fit a distribution of 83 cancer examinations (40.5%) and 122 (59.5%) noncancer examinations (13.6% benign biopsy, 2.5% BI-RADS 0, 43.4% BI-RADS 1 and 2). The readers were blinded to the enrichment levels in this data set, and the reading was conducted in a double-blind manner. All examinations had one or two previous studies available for review. The readers were also provided with the information available in the clinical setting, including age, family history of breast cancer, hormone therapy, and gene mutation.

Readers reported a BI-RADS score indicating that they would recall the case (BI-RADS 0) or would not recall the case (BI-RADS 1 or 2), as if they were interpreting the screening examination in a routine practice. They then provided a forced diagnostic BI-RADS score (designated as "forced" because radiologists do not give a final BI-RADS assessment at screening) by using the values 1, 2, 3, 4A, 4B, 4C, or 5, and probability of malignancy score. Forced BI-RADS and probability of malignancy were used to compare between AI and the readers' area under the receiver operating characteristic curve (AUC; Fig E2 [online]).

Statistical Analysis

The 95% CIs for sensitivity, specificity, negative predictive value, positive predictive value, and recall rate were on the basis of 10 000 bootstrap replications and a one-sided *t* test (14). We assessed noninferiority in sensitivity and negative predictive value at a relative margin of 5% and assessed superiority in specificity, recall rate, and positive predictive value at an absolute margin of 1%. A *P* value less than .05 was considered to indicate a statistically significant difference. Our 10 primary comparisons were sensitivity noninferiority, specificity superiority, and recall rate superiority, tested on the main test, sites test, and reader study; and AUC noninferiority of mean reader compared with standalone AI. Bonferroni correction of an α value of .05/10 was applied to account for multiple hypothesis testing on the primary comparisons (15,16). To simulate prevalence of

Table 1: Patient Characteristics in Training, Validation, and Test Data Sets

Characteristic	Training Set		Validation Set		Test Set	
	Data Set	Cancer	Data Set	Cancer	Data Set	Cancer
All women	3948	804	1661	182	4310	453
All DBT examinations	5595	831	2529	182	5182	459
Age (y)	60 ± 10	63 ± 11	60 ± 11	63 ± 10	60 ± 11	62 ± 11
Age by groups (y)						
<50	1314 (23.5)	142 (17.1)	487 (19.3)	17 (9.3)	1091 (21.1)	63 (13.7)
50–59	1602 (28.6)	215 (25.8)	764 (30.2)	54 (29.7)	1330 (25.7)	116 (25.2)
60–69	1582 (28.2)	236 (28.4)	752 (29.7)	56 (30.7)	1539 (29.7)	142 (30.1)
70–79	887 (15.8)	168 (20.2)	406 (16.1)	41 (22.5)	970 (18.7)	111 (24.2)
>80	210 (3.75)	70 (8.4)	120 (4.7)	14 (7.7)	252 (4.9)	27 (5.9)
Ethnicity						
White	3748 (67.0)	544 (65.4)	1880 (74.3)	140 (76.9)	3864 (74.6)	351 (76.5)
African American	1490 (29.6)	221 (26.6)	497 (19.6)	35 (19.2)	775 (15.0)	71 (15.5)
Asian	134 (2.4)	36 (4.3)	63 (2.5)	3 (1.6)	190 (3.7)	21 (4.6)
Other	9 (0.2)	2 (0.2)	3 (0.1)	1 (0.6)	9 (0.8)	0 (0)
Unknown	214 (3.8)	28 (3.3)	86 (3.4)	3 (1.7)	344 (6.6)	16 (3.5)
Body mass index (kg/m^2)						
<25	1142 (20.4)	168 (20.2)	498 (19.7)	39 (21.4)	1235 (23.8)	132 (28.8)
25–29.9	1127 (20.1)	196 (23.6)	461 (18.2)	45 (24.7)	1056 (20.4)	125 (27.2)
30–34.9	920 (16.4)	183 (22.0)	348 (13.8)	36 (19.8)	672 (13.0)	82 (17.9)
>35	873 (15.6)	157 (18.9)	354 (14.0)	33 (18.1)	531 (10.3)	70 (15.3)
Unknown	1533 (27.4)	127 (15.3)	868 (34.3)	29 (15.9)	1688 (32.6)	50 (10.9)
Breast density						
Fatty	307 (5.5)	21 (2.5)	173 (6.8)	8 (4.4)	299 (5.8)	21 (4.6)
Scattered	2302 (41.1)	350 (42.1)	1060 (41.9)	88 (48.4)	1756 (33.9)	169 (36.8)
Heterogeneously dense	2322 (41.5)	422 (50.8)	1014 (40.1)	81 (44.5)	2714 (52.4)	242 (52.7)
Extremely dense	261 (4.7)	34 (4.1)	123 (4.9)	5 (2.8)	370 (7.1)	27 (5.88)
Unknown	403 (7.2)	4 (0.5)	159 (6.3)	0 (0)	43 (0.8)	0 (0)

Note.—Data are number of women; data in parentheses are percentages. Mean data are ± standard deviation. Multiple examinations per woman were allowed. DBT = digital breast tomosynthesis.

screening population, we used inverse probability weighting (17) on the enriched retrospective test set to match it with screening population statistics (18) (Appendix E3 [online]). Confidence bands on receiver operating characteristic curves were computed by using the Kolmogorov-Smirnov method (19,20).

Results

Patient Characteristics

A total of 9919 women were included in the study (mean age, 60 years ± 11). Of those women, 3589 had biopsy results (2159 benign results and 1439 cancers). The women's clinical characteristics are described in Table 1. Cancer characteristics are described in Table 2. Clinical data analysis is described in Tables E2 and E3 (online) and Figure E1 (online).

Table 2: Cancer Examination Characteristics

Characteristic	Training Set	Validation Set	Test Set
No. of cancer examinations	831	182	459
Dominant imaging feature			
Mass	406 (48.8)	60 (33.0)	123 (26.8)
Calcification	142 (17.1)	43 (23.6)	132 (28.8)
Asymmetry	101 (12.2)	27 (14.8)	85 (18.5)
Architectural distortion	85 (10.2)	30 (16.5)	89 (19.4)
Other	8 (0.9)	1 (0.6)	6 (1.3)
Unknown (not visible finding)	89 (10.7)	21 (11.5)	24 (5.2)
Cancer type			
Invasive	675 (81.2)	136 (74.7)	333 (72.5)
Noninvasive	149 (17.9)	45 (24.7)	120 (26.1)
Unknown	7 (0.8)	1 (0.6)	6 (1.3)

Note.—Unless otherwise indicated, data are number of individuals in each group; data in parentheses are percentages.

Retrospective Study Results

Worklist reduction.—In the simulated workflow, the addition of AI reduced 39.6% (95% CI: 38.1, 41.0) of the worklist while improving radiologist specificity, recall rate, and positive predictive value and maintaining noninferior sensitivity (reference to radiologist performance in Appendix E4 [online]) and noninferior negative predictive value.

Table 3: Evaluation of Worklist Reduction Simulation in Retrospective Study

Variable	Radiologist (%)	AI (%)	Radiologist with AI (%)	*P* Value*
Specificity	91.3 (4312/4723)	39.7 (1875/4723) [38.3, 41.2]	93.6 (4421/4723) [93.3, 93.9]	.002
Sensitivity	90.8 (417/459)	96.2 (442/459) [94.9, 97.5]	90.0 (413/459) [89.0, 90.7)]	.002
Recall rate	9.2 (477/5182)	60.4 (3130/5182) [59.1, 61.9)]	6.9 (358/5182) [6.6, 7.2]	.002
DBT read	100 (5182/5182)	Not available	60.4 (3130/5182)	

Note.—Data in parentheses are numerator/denominator; data in brackets are 95% CIs. AI enabled a 39.6% worklist reduction in screening while maintaining noninferior sensitivity, superior specificity, and superior recall rate. Radiologists' performance is given for Conant et al (3). Radiologist with AI is the worklist reduction simulation. Stand-alone AI specificity of 39.7% (95% CI: 38.3, 41.2) results in overall worklist reduction of 39.6% (95% CI: 38.1, 41.0) for radiologists with AI because specificity is calculated only on the negative examinations, but overall work reduction is calculated on both positive and negative examinations. The 477 stand-alone radiologist recalls were reduced to 358 in the radiologist-with-AI model. Focusing on the reduced recalls, 3.4% (four of 119) were originally radiologist true-positive findings, and 96.6% (115 of 119) were originally radiologist false-positive findings. AI = artificial intelligence, DBT = digital breast tomosynthesis.

* The *P* values were computed for comparisons between radiologist and radiologist-with-AI models. Specificity and recall rate superiority were calculated with a 1% absolute margin. Sensitivity noninferiority is on the basis of a 5% relative noninferiority margin. All *P* values were Bonferroni multihypothesis corrected.

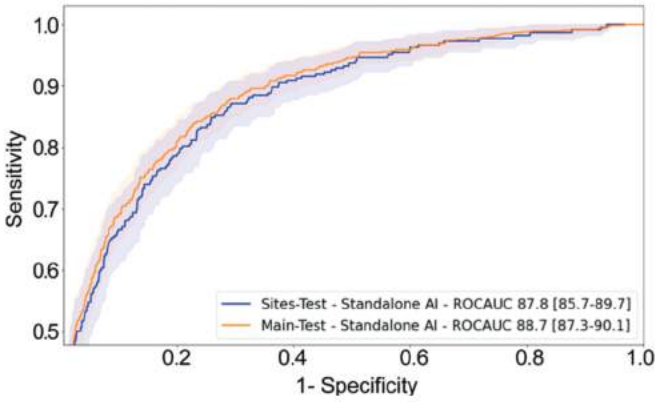

Figure 3: Receiver operating characteristic (ROC) curves and operation points of main test and sites test with Kolmogorov-Smirnov confidence bands (19,20). AI = artificial intelligence, AUC = area under the curve.

As shown in Table 3, recall rate decreased by 25%: from 9.2% (477 recalls in 5182 examinations) to 6.9% (358 recalls in 5182 examinations) (95% CI: 6.6, 7.2; $P = .002$). Specificity improved from 91.3% (4312 of 4723) to 93.6% (4421 of 4723) (95% CI: 93.3, 93.9; $P = .002$), and noninferior sensitivity of 90.0% (413 of 459) (95% CI: 89.0, 90.7; 5% relative margin; $P < .001$) was maintained (compared with 90.8%; 417 of 459). Positive predictive value improved from 5.9% (95% CI: 5.9, 6.0) to 7.7% (95% CI: 7.3, 8.0; $P = .02$), and negative predictive value was noninferior: 99.9% (95% CI: 99.9, 99.9) versus 99.9% (95% CI: 99.9, 99.9; $P = .02$). Among the examinations dismissed from the simulated screening worklist, 0.2% (four of 2052) were originally radiologists' true-positive findings and 5.6% (115 of 2052) were false-positive findings.

AI error analysis.—Analysis of the AI false-negative findings by two experienced breast radiologists (20 and 23 years) found that the majority (18 of 26) were mammographically occult (full details in Appendix E5 [online]). For further analysis of AI errors, see Figures E3 and E4 (online).

Generalization.—In addition to testing on unseen women's examinations, we examined how well our AI system generalizes across different screening settings by using examinations from four imaging facilities (ie, sites test) that were not used for training or validation. The AUC of the standalone AI model (Fig 3) was statistically equivalent (5% relative margin, $P = .004$), as follows: 0.89 (95% CI: 0.87, 0.90) for the main test and 0.88 (95% CI: 0.86, 0.90) for the sites test. Worklist-reduction simulation resulted in 39.9% (1264 of 3168 examinations) (95% CI: 38.0, 41.7), similar to results obtained on the main test. The sensitivity of 89.6% (265 of 296 detected cancers; 95% CI: 88.3, 90.6) was noninferior to radiologists' sensitivity of 90.9% (269 of 296 detected cancers; noninferiority margin of 5%, $P = .002$). The 93.3% specificity (2680 of 2872; 95% CI: 92.9, 93.7) was higher than the radiologists' specificity of 91.3% (2622 of 2872; $P = .002$) (3). Recall rate was reduced from 9.0% (285 recalls in 3168 examinations; 95% CI: 9.0, 9.1) to 7.2% (228 recalls in 3168 examinations; 95% CI: 6.8, 7.6; $P = .002$). Radiologists' positive predictive value of 5.9 (95% CI: 5.9, 6.0) improved to 7.3 (95% CI: 6.9, 7.8; $P = .03$). Negative predictive value was noninferior at 99.9% (95% CI: 99.9, 99.9) compared with the radiologists' negative predictive value of 99.9% (95% CI: 99.9, 99.9; $P = .002$).

In addition, the AUC of AI across ages, ethnicities, and breast density categories showed similar AI performance (Table E4 [online]).

Reader Study Results

Performance of readers and standalone AI.—The AUC of the mean reader was 0.81 (95% CI: 0.76, 0.85), and the AUC for AI was 0.84 (95% CI: 07.78, 0.89). Comparing reader's performance with standalone AI performance showed noninferiority of AI versus the mean reader (5% relative noninferiority margin, $P = .002$). The receiver operating characteristic curves of each reader in comparison to AI are presented in Figures 4A, E2, and E5 (online).

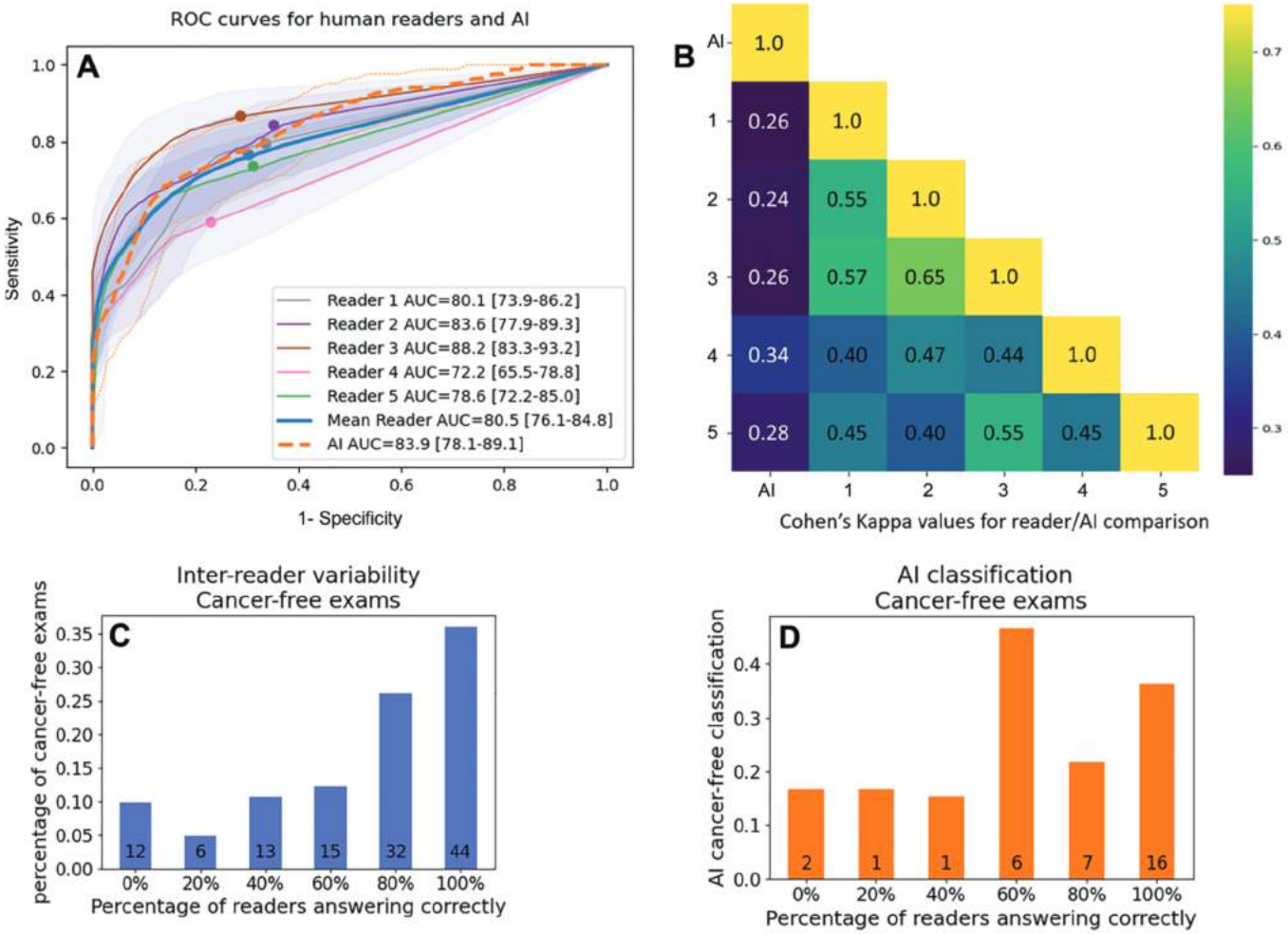

Figure 4: Reader study results. **(A)** Readers and artificial intelligence (AI) model receiver operating characteristic (ROC) curves. All ROC curves include Kolmogorov-Smirnov confidence bands (19,20), marked by a blue area around the readers' curve, mean reader, and dotted lines for AI. Dots on the curves mark the sensitivity and specificity achieved by the readers. In the high sensitivity range, AI exceeds readers performance. **(B)** Each cell depicts agreement, as measured by Cohen κ (21) between pairs of readers or between AI classification and each one of the five readers. For this comparison, an AI operation point of 0.79 sensitivity and 0.66 specificity was chosen because it was the closest point to the reader mean of 0.79 sensitivity and 0.67 specificity. Although most readers are in moderate agreement, with κ values between 0.4 and 0.65 (warmer colors), AI differs from a human reader, with κ values between 0.24 and 0.34 (darker, colder colors). **(C)** Interreader variability per cancer-free decision. Each bar shows the percentage of readers who provided identical interpretation (number on bar represents number of examinations). For example, all five readers agreed correctly on 35% of examinations, whereas none of them answered correctly on 10% of the examinations. **(D)** The AI answers on each one of the interreader variability bars (number on bar represents number of examinations). AUC = area under the receiver operating characteristic curve.

Interpretation variability.—The degree of agreement of all combination pairs of readers and for each reader with AI is presented by using Cohen κ scores (21,22). There was moderate agreement between readers ($\kappa > 0.4$) (Fig 4B). The agreement between readers for the cancer-free examinations is shown in Figure 4C. All five readers agreed on only 35% of the cancer-free examinations. Figure 4D shows how the standalone AI system assessed examinations in groups showing different levels of agreement between readers. In the group where 0% of the readers classified correctly as cancer free, AI would reduce recall rate by 16%.

Simulation of worklist reduction in the reader study.—We repeated the worklist reduction simulation on the reader study data set, which had 40.5% cancer cases (83 of 205), a rate much higher than in the typical screening population (5.9 of 1000) (12). The mean reader specificity increased from 70% (85 of 122; 95% CI: 64, 75) to 76% (93 of 122; 95% CI: 71, 82; superiority with 1% absolute margin, $P = .001$), whereas sensitivity was noninferior at 77% (64 of 83 detected cancers; 95% CI: 70, 83) compared with 76% (63 of 83 detected cancers; 95% CI: 70, 83; noninferiority margin of 5%, $P = .001$). Mean reader recall rate improved from 49% (101 of 205; 95% CI: 44, 54) to 45% (92 of 205; 95% CI: 40, 50; superiority with 1% absolute margin, $P = .001$) (Table 4). Applying AI on the 205 examinations yielded an 18.5% worklist reduction (38 of 205). This is lower than the worklist reduction in the retrospective study because of the enriched data set. This also explains the high recall rates of all readers. The mean radiologist positive predictive value of 63% (95% CI: 55, 71) was improved to 69% (95% CI: 61, 77; $P = .001$. Worklist-reduction simulation maintained a noninferior negative predictive value of 82% (95% CI: 76, 87) compared with the radiologists' negative predictive value of 83% (95% CI: 77, 88; $P = .01$). Reader study statistics with and without the use of AI are shown in

Table 4: Evaluation of Worklist Reduction Simulation in Reader Study

Variable	Mean Reader (%)	AI (%)	Mean Reader with AI (%)	*P* Value*
Specificity	70 (85/122) [64, 75]	30 (36/122) [21–38]	76 (93/122) [71, 82]	.01
Sensitivity	77 (64/83) [70, 83]	96 (80/83) [92, 100]	76 (63/83) [70, 83]	.01
Recall rate	49 (101/205) [44, 54]	82 (167/205) [76, 87]	45 (92/205) [40, 50]	.01
DBT read	100 (205/205)	Not available	82 (167/205)	

Note.—Data in parentheses are numerator/denominator; data in brackets are 95% CIs. Radiologist with AI is the worklist-reduction simulation. The cause for the lower performance of the readers compared with the known mean radiologist screening performance is the biopsied cases–enriched data set, which differs from screening population prevalence. AI = artificial intelligence, DBT = digital breast tomosynthesis.

* The *P* values are for comparisons of radiologist with AI with radiologist models. Specificity and recall rate superiority were calculated with a 1% absolute margin. Sensitivity noninferiority is on the basis of a 5% relative noninferiority margin. All *P* values were Bonferroni multihypothesis corrected.

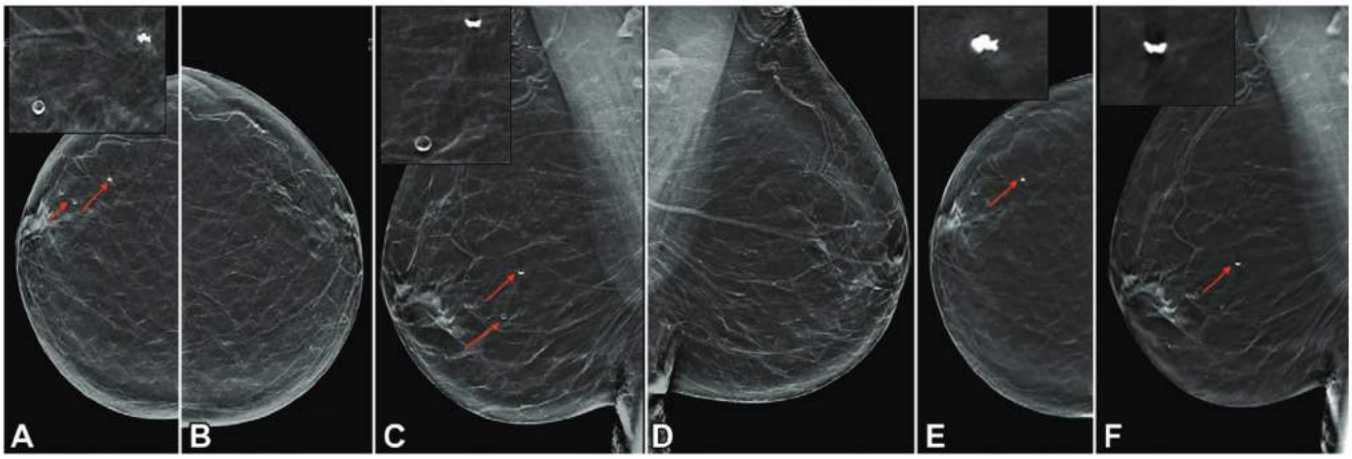

Figure 5: Screening mammography images in a 77-year-old woman. Bilateral reconstructed screening mammogram (C-view) with **(A, B)** craniocaudal and **(C, D)** mediolateral oblique views and right digital breast tomosynthesis image with **(E)** craniocaudal and **(F)** mediolateral oblique views, show coarse benign-appearing calcifications in the right breast (arrows). Inset images show magnified coarse calcifications in the right craniocaudal (**A**) and mediolateral oblique (**C**) views, and in the right digital breast tomosynthesis craniocaudal (**E**) and mediolateral (**F**) views.The study was categorized as Breast Imaging Reporting and Data System 2 by the radiologist. The artificial intelligence (AI) score was 17. AI would have correctly categorized this study as normal.

Figure 6: Screening mammography images in a 50-year-old woman. Bilateral reconstructed screening mammogram (C-view) with **(A, B)** craniocaudal and **(C, D)** mediolateral oblique views and right digital breast tomosynthesis image with **(E)** craniocaudal and **(F)** mediolateral oblique views, show architectural distortion in the upper right breast (red circles). Subsequent US showed a 12 × 8 × 12 mm irregular mass, and biopsy yielded a diagnosis of invasive lobular carcinoma. The artificial intelligence (AI) score was 66. AI would have correctly triaged this case as abnormal.

Table 5: Reader Study Statistics

Variable	Reader without AI (%)	Reader with AI (%)	*P* Value
Specificity			
R1	66 (81/122) [58, 75]	73 (89/122) [66, 81]	.003
R2	65 (79/122) [56, 73]	74 (90/122) [66 82]	<.001
R3	71 (87/122) [63, 79]	75 (91/122) [68, 82]	.09
R4	77 (94/122) [70, 84]	84 (103/122) [77, 90]	.001
R5	69 (84/122) [61, 77]	76 (93/122) [69, 84]	.001
Sensitivity			
R1	80 (66/83) [71, 88]	80 (66/83) [70, 88]	<.001
R2	84 (70/83) [76, 92]	83 (69/83) [75, 91]	.02
R3	87 (72/83) [79, 94]	87 (72/83) [80, 94]	<.001
R4	59 (49/83) [49, 70]	59 (49/83) [48, 70]	<.001
R5	74 (61/83) [64, 83]	74 (61/83) [64, 83]	<.001
Recall rate			
R1	52 (107/205) [45, 59]	48 (99/205) [42, 55]	.01
R2	55 (113/205) [48, 62]	49 (101/205) [42, 56]	.001
R3	52 (107/205) [45, 59]	50 (103/205) [43, 57]	.24
R4	38 (77/205) [31, 44]	33 (68/205) [26, 41]	.005
R5	48 (99/205) [42, 55]	44 (90/205) [37, 51]	.004
PPV			
R1	62 (66/107) [53, 71]	67 (66/99) [58, 76]	.003
R2	62 (70/133) [53, 70]	68 (69/101) [59, 77]	.001
R3	67 (72/107) [59, 76]	70 (72/103) [62, 78]	.09
R4	64 (49/77) [53, 74]	72 (49/68) [61, 82]	<.001
R5	62 (61/99) [52, 72]	68 (61/90) [58, 78]	.001
NPV			
R1	83 (81/98) [75, 90]	84 (89/106) [77, 90]	<.001
R2	86 (79/92) [79, 93]	87 (90/104) [80, 93]	<.001
R3	89 (87/98) [82, 95]	89 (91/102) [83, 95]	<.001
R4	73 (94/128) [65, 81]	75 (103/137) [68, 82]	<.001
R5	79 (84/106) [71, 87]	81 (93/115) [73, 88]	<.001

Note.— Data in parentheses are numerator/denominator; data in brackets are 95% CIs. Recall rate of the radiologists was high because of the enrichment with cancer and benign biopsies (in the screening population, recall rate is approximately 10%). AI = artificial intelligence, NPV = negative predictive value, PPV = positive predictive value, R = reader.

Table 5. Figures 5 and 6 present two examples of AI model results on DBT studies.

Discussion

We developed an artificial intelligence (AI) system that analyzed imaging and clinical information and classified digital breast tomosynthesis (DBT) screening examinations as cancer-free, allowing these examinations to be dismissed from the worklist without consultation with a radiologist. The purpose was to address the long reading times of DBT compared with those of digital mammography (4) because of increased use of DBT worldwide (5). Because 99.5% of screening examinations are cancer free (18), deploying such an AI system to optimize screening reads could be of substantial value.

In our retrospective study, AI demonstrated the potential to reduce radiologists' worklist by 39.6%, with improved specificity (from 91.3% to 93.6%; $P = .002$) and noninferior sensitivity (from 90.8% to 90.0%; $P = .002$). In a simulated workflow, the recall rate was reduced by 25% (from 9.2% to 6.9%; $P = .002$). When we analyzed the AI false-negative findings, we found that almost 70% were occult at mammography. We presented evidence of generalizability of the AI model, both to unseen patients and to unseen sites. AI performance was stable across all age groups, ethnicities, and body mass indexes, suggesting that AI may be widely applicable to diverse patient populations.

In a reader study, the readers had access to all information typically available during screening (eg, previous studies and clinical information). The AI standalone performance was noninferior to that of the mean reader (AUC, 0.81 vs 0.84; $P = .002$). When worklist reduction for the mean reader was simulated, the specificity increased (from 70% to 76.4%; $P < .01$) and recall rate decreased (from 49% to 45%; $P < .01$), with maintenance of noninferior sensitivity (from 77% to 76%; $P < .01$); these findings strengthen the potential contribution of AI. Our analysis also showed that although AI performance was better in some metrics and noninferior in others, its method of analysis is different from that of the human readers. This diversity provides additional support for AI's potential to augment human decision making.

Several studies introduced successful AI technologies for interpretation of digital mammography (6–10). Conant et al (11) and Raya-Povedano (13) reported AI-based computer-aided detection assistance on limited DBT data sets. Our study focused on DBT by using a large and diverse DBT screening data set with high number of biopsy-proven examinations (1472 malignant and 2232 benign) collected from 22 clinical sites.

We theorize that trusting AI to perform radiologist's work requires substantial evidence. We believe that AI should be introduced into clinical practice gradually. Before AI is allowed to automatically interpret complex cases, it will first be used for tasks that are considered repetitive work, which was the approach we took in this study. We believe that with time and with enough accumulated evidence, AI will be trusted in the same way we trust results of automated blood tests.

Our study had several limitations. All DBT data were acquired with Hologic devices. Future research should assess the performance of the AI system across a variety of manufacturers. Our simulation of the potential benefit of worklist reduction assumed that radiologists would have read the remaining examinations the same way, regardless of whether AI reduced their worklist. This assumption should be further tested in a prospective study. Our study did not include women with foreign bodies (eg, implants, pacemakers) or women with a history of breast cancer. In the reader study, although the readers were in their regular environment, they had access to one or two previous examinations, whereas in routine practice they would have had access to all previous examinations. The readers were unaware that the data set was enriched with 40% cancer cases, which may have affected their performance.

To conclude, we developed an artificial intelligence (AI) system to filter out normal digital breast tomosynthesis (DBT) examinations. We envision that implementation of this type of model within the clinic could affect three different levels: for radiologists, by reducing both workload and fatigue arising from

routine clinical tasks; for health systems, by improving workflow and facilitating further introduction of DBT, especially where there is a shortage of breast radiologists; and for women, by reducing unnecessary recalls, stress, and exposure to radiation. Future research should include prospective evaluation of our AI model, to assess the percentage of DBT examinations that would be removed from a prospective reading worklist, and to assess how readers perform when interpreting the remaining cases (knowing that some of the "normal" cases have already been removed). Future research should also evaluate generalizability to multiple DBT manufacturers.

Acknowledgments: We acknowledge multiple contributors to this project: Susan Harvey, MD, for the original concept and design of the project. Without Dr Harvey's vision, dedication, and enthusiasm, the collaboration between Johns Hopkins and IBM Research would not have been possible. The Johns Hopkins Medical Information Technology team, including Charlene Tomaselli, MBA, RT (R)(M), CIIP, Maisy Steirhoff, BA, MBA, Daniele Bananto, BA, Boris Feldman, BSc, and Dushyant Gupta, MSc, for their help with gathering, deidentification, and transmission of DBT images and clinical reports; Epic analyst Jenn Zuk, BSc, for her help with clinical data gathering; research coordinator Mary Kate Jones, MA, for her help with IRB protocol creation, submission, and maintenance, as well as transmission of data; Aviad Zlotnick, PhD, for his algorithmic advice and infrastructure support in the underlying AI system; Oren Kagan, MSc, for curating the clinical data and ingesting it into the database; Yoni Keren, BSc, for his database access layer development,; the IBM Haifa IT team for their support in development needs, including transmission and storage of data, maintenance of many GPU machines, and software installations; Paula Simovitz, MD, for her help with graphical annotation of DBT images; and Oksana Greg, BA, for her practical advice and ground truthing from pathology and radiology reports.

Author contributions: Guarantors of integrity of entire study, **Y.S., F.G.S., V.R.**; 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, **Y.S., R.B., F.G.S., V.R., D.K., E.B.A., M.R.Z., L.A.M.**; clinical studies, **F.G.S., E.T.O., B.P., P.A.D., L.A.M.**; experimental studies, **Y.S., R.B., F.G.S., E.B., M.A., D.K., L.A.M.**; statistical analysis, **Y.S., R.B., F.G.S., V.R., E.B., M.O.F., M.A., D.K., E.B.A., L.A.M.**; and manuscript editing, **Y.S., R.B., F.G.S., V.R., E.B., M.O.F., M.A., D.K., E.B.A., E.T.O., M.R.Z., L.A.M.**

Disclosures of conflicts of interest: Y.S. Employed by IBM Research. **R.B.** No relevant relationships. **F.G.S.** Employed by IBM Research. **V.R.** Employed by IBM Research; patent application filed for method used by the artificial intelligence system. **E.B.** Employed by IBM Research. **M.O.F.** Employed by IBM and worked on this study as part of IBM-Research Haifa; patents submitted with colleagues at IBM; owns IBM stocks. **M.A.** No relevant relationships. **D.K.** No relevant relationships. **E.B.A.** No relevant relationships. **E.T.O.** No relevant relationships. **B.P.** No relevant relationships. **P.A.D.** No relevant relationships. **M.R.Z.** Employed by IBM; stock in IBM. **L.A.M.** Payment to institution for salary support from IBM Research; grants for salary support from Mark Foundation, Cepheid; consulting fees from Hologic; educational evens payment from Hologic.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424 [Published correction appears in CA Cancer J Clin 2020;70(4):313.].
2. Skaane P, Sebuødegård S, Bandos AI, et al. Performance of breast cancer screening using digital breast tomosynthesis: results from the prospective population-based Oslo Tomosynthesis Screening Trial. Breast Cancer Res Treat 2018;169(3):489–496.
3. Conant EF, Beaber EF, Sprague BL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: a cohort study within the PROSPR consortium. Breast Cancer Res Treat 2016;156(1):109–116.
4. Dang PA, Freer PE, Humphrey KL, Halpern EF, Rafferty EA. Addition of tomosynthesis to conventional digital mammography: effect on image interpretation time of screening examinations. Radiology 2014;270(1):49–56.
5. Richman IB, Hoag JR, Xu X, et al. Adoption of digital breast tomosynthesis in clinical practice. JAMA Intern Med 2019;179(9):1292–1295.
6. Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 2019;292(2):331–342.
7. Kyono T, Gilbert FJ, van der Schaar M. Improving workflow efficiency for mammography using machine learning. J Am Coll Radiol 2020;17(1 Pt A):56–63.
8. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94 [Published correction appears in Nature 2020;586(7829):E19.].
9. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol 2019; 29(9):4825–4832.
10. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. Radiology 2019;293 (1):38–46.
11. Conant EF, Toledano AY, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. Radiol Artif Intell 2019;1(4):e180096.
12. Benchmarks for Abnormal Screening Mammography Interpretations. BCSC. https://www.bcsc-research.org/statistics/screening-performance-benchmarks/abnormal-scrn-benchmarks. Accessed December 4, 2020.
13. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. Radiology 2021;300(1):57–65.
14. Wasserman L. All of Statistics: A Concise Course in Statistical Inference. New York, NY: Springer Science+Business Media, 2004.
15. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014;34(5):502–508.
16. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubbl R Istitut Super di Scien Econom Commerc Firenz 1936;8:1–62. https://www.scienceopen.com/document?vid=35962296-b63d-4dac-8c75-777d5d9cc0dd.
17. Mansournia MA, Altman DG. Inverse probability weighting. BMJ 2016;352:i189.
18. Breast Cancer Surveillance Consortium (BCSC). https://www.bcsc-research.org/. Accessed September 24, 2020.
19. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med 1994;13(5-7):499–508.
20. Macskassy S, Provost F. Confidence bands for ROC curves: methods and an empirical study. http://archive.nyu.edu/handle/2451/27802. Published August 2004. Accessed September 24, 2020.
21. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.
22. Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas 1981;41(3):687–699.