Radiology

ORIGINAL RESEARCH • BREAST IMAGING

A Deep Learning Model to Triage Screening Mammograms: A Simulation Study

Adam Yala, MEng • Tal Schuster, MS • Randy Miles, MD • Regina Barzilay, PhD • Constance Lehman, MD, PhD

From the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Mass (A.Y., T.S., R.B.); and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit St, WAC 240, Boston, Mass 02114-2698 (R.M., C.L.). Received December 21, 2018; revision requested February 25; revision received June 5; accepted June 18. **Address correspondence to** C.L. (e-mail: *clehman@partners.org*).

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

See also the editorial by Kontos and Conant in this issue.

Radiology 2019; 00:1–9 • https://doi.org/10.1148/radiol.2019182908 • Content codes: BR IN

Background: Recent deep learning (DL) approaches have shown promise in improving sensitivity but have not addressed limitations in radiologist specificity or efficiency.

Purpose: To develop a DL model to triage a portion of mammograms as cancer free, improving performance and workflow efficiency.

Materials and Methods: In this retrospective study, 223 109 consecutive screening mammograms performed in 66 661 women from January 2009 to December 2016 were collected with cancer outcomes obtained through linkage to a regional tumor registry. This cohort was split by patient into 212 272, 25 999, and 26 540 mammograms from 56 831, 7021, and 7176 patients for training, validation, and testing, respectively. A DL model was developed to triage mammograms as cancer free and evaluated on the test set. A DL-triage workflow was simulated in which radiologists skipped mammograms triaged as cancer free (interpreting them as negative for cancer) and read mammograms not triaged as cancer free by using the original interpreting radiologists' assessments. Sensitivities, specificities, and percentage of mammograms read were calculated, with and without the DL-triage–simulated workflow. Statistics were computed across 5000 bootstrap samples to assess confidence intervals (CIs). Specificities were compared by using a two-tailed t test ($P < .05$) and sensitivities were compared by using a one-sided t test with a noninferiority margin of 5% ($P < .05$).

Results: The test set included 7176 women (mean age, 57.8 years ± 10.9 [standard deviation]). When reading all mammograms, radiologists obtained a sensitivity and specificity of 90.6% (173 of 191; 95% CI: 86.6%, 94.7%) and 93.5% (24 625 of 26 349; 95% CI: 93.3%, 93.9%). In the DL-simulated workflow, the radiologists obtained a sensitivity and specificity of 90.1% (172 of 191; 95% CI: 86.0%, 94.3%) and 94.2% (24 814 of 26 349; 95% CI: 94.0%, 94.6%) while reading 80.7% (21 420 of 26 540) of the mammograms. The simulated workflow improved specificity ($P = .002$) and obtained a noninferior sensitivity with a margin of 5% ($P < .001$).

Conclusion: This deep learning model has the potential to reduce radiologist workload and significantly improve specificity without harming sensitivity.

Online supplemental material is available for this article.

Mammography is the only imaging modality shown to reduce breast cancer mortality in randomized trials (1–8). Despite its benefits, challenges include variation in interpretive performance and the scarcity of specialized radiologists (9,10). A recent report of mammography screening performance in U.S. community practice demonstrated that radiologists' diagnostic performance ranged from 66.7% to 98.6% for sensitivity and from 71.2% to 96.9% for specificity (11). False-negative examinations can result in delayed diagnosis, and false-positive examinations can lead to unnecessary procedures, impacting both patient experience and overall costs. Moreover, the ability of specialized radiologists to serve the global population of women eligible for breast cancer screening is limited by workflow inefficiencies (12,13).

Both technologic and workflow solutions have been proposed to improve radiologist interpretive performance and efficiency. Computer-aided detection (CAD), which detects and marks suspicious findings on mammograms, aims to improve radiologist sensitivity. Although traditional approaches have not demonstrated improved radiologist performance in sensitivity or specificity in clinical practice (14–16), a more recent deep learning approach to CAD has shown promise in improving sensitivity in a reader study (17). However, this does not address limitations in radiologist specificity or efficiency. Double reading (ie, having two radiologists interpret the same mammogram) has also been implemented to improve radiologist performance. Although some studies demonstrate slight improvements in sensitivity, double reading worsens workflow efficiency and increases false-positive examination results (18,19).

We hypothesized that a deep learning model trained to triage mammograms as cancer free can improve radiologist efficiency and specificity without harming sensitivity. Specifically, we trained a model to predict cancer

Abbreviations

AUC = area under the receiver operating characteristic curve, CAD = computer-aided detection, CI = confidence interval, DL = deep learning

Summary

In a simulation study, a deep learning model to triage mammograms as cancer free improves workflow efficiency and significantly improves specificity while maintaining a noninferior sensitivity.

Key Results

- After training and validation on 238 271 mammograms, a deep learning model triaged 19% of screening mammograms as cancer free, improving specificity (93.5%–94.3%; $P = .002$) and obtaining a noninferior sensitivity (90.6%–90.1%; $P < .001$) in a retrospective simulation.
- The deep learning model had similar predictive accuracies for all age groups; the area under the receiver operating characteristic curve (AUC) for women in their 40s, 50s, 60s, and 70s or older were 0.80, 0.83, 0.82, 0.79, and 0.86, respectively.
- The deep learning model was effective for women with a range of breast density (AUCs of 0.82, 0.81, 0.85, and 0.71 for women with fatty, scattered, heterogeneously, and extremely dense breasts, respectively).

directly from full-resolution mammograms and chose a high sensitivity threshold to identify a subset of cancer-free patients with near-perfect accuracy. We simulated the scenario in which all patient examinations below this threshold are interpreted as negative for cancer and those above the threshold are read by radiologists who specialize in breast imaging.

Materials and Methods

Our retrospective study was approved by our institutional review board (with a waiver for the need to obtain written informed consent) and was compliant with the Health Insurance Portability and Accountability Act. Mammograms from 60 886 of the 80 818 women in our patient population were previously reported (20,21). Our previously published work focused on the development of breast density and 5-year breast cancer risk algorithms, whereas this article focuses on a deep learning model to triage a subset of mammograms as cancer free.

Data Collection

We collected consecutive digital screening mammograms (Selenia Dimensions and Selenia; Hologic, Bedford, Mass) from 80 818 patients screened between January 1, 2009, and December 31, 2016, at a large tertiary academic medical center. Outcomes were obtained through linkage to tumor registries of five hospitals (academic and general) within our health care system, supplemented with pathologic findings from our mammography information system electronic medical record (MagView, version 8.0.143; Magview, Burtonsville, Md). Outcomes were not linked to the state tumor registry.

Among the initial 80 818 patients, we selected women who had either a diagnosis of breast cancer within 1 year or imaging follow-up for at least 1 year from the date of the index mammogram. We excluded 14 056 women lacking sufficient follow-up and 101 women because they had another form of cancer in their breast. We did not perform exclusions based on prior surgery, age, implants, atypical lesions, or prior cancers. The remaining 66 661 women were randomly assigned to 56 831 for training, 7021 for validation, and 7176 for testing. This resulted in training, validation, and test sets of 212 272, 25 999, and 26 540 mammograms, respectively (Fig 1). We emphasize that we split our data set by patients, and so each woman contributed mammograms to only one set.

Development of Deep Learning Model

In-depth information about our deep learning (DL) model and its training is presented in Appendix E1 (online), and code is available for research use at *http://learningtocure.csail.mit.edu*. In brief, we implemented our model as a deep convolutional neural network (ResNet18 [22]) with PyTorch (version 0.31; *https://pytorch.org*). Given a 1664 × 2048 pixel view of a breast, the model was trained to predict whether or not that breast would develop breast cancer within 1 year. The model makes independent predictions for each view, and we took the maximum predicted score across views to get the prediction for the examination.

To leverage the trained probabilistic model to triage mammograms, we chose a high-sensitivity threshold on the validation set. Specifically, we set the model threshold to the minimum probability score of a radiologist true-positive assessment on the validation set. This procedure maximizes the mammograms triaged while not decreasing sensitivity on the validation set.

Evaluating Our Model for Independent Prediction

We evaluated the overall discrimination accuracy of the DL model when used independently through the area under the receiver operating curve (AUC) and evaluated model calibration through observed-to-expected ratios. We computed AUCs and observed-to-expected ratios across the entire held-out test set as well as subgroups based on age, race, and Breast Imaging Reporting and Data System, or BI-RADS, density category. Specifically, we reported the AUC and observed-to-expected ratios in patients in their 40s, 50s, 60s, and 70s or older; patients who are African American, Asian or Pacific Islander, white, or other; and patients with fatty, scattered, heterogeneously dense, or extremely dense breasts.

Evaluating Our Model for Triage

All mammograms in our test set were read by one of 23 fellowship-trained or equivalent breast imaging radiologists with between 1 year to 31 years of experience through routine clinical operations between 2009 to 2016. To evaluate our model for triage, we simulated the scenario in which the radiologist did not read any mammogram below the model's high-sensitivity "cancer-free" threshold, and read the rest of the mammograms as before. Specifically, all mammograms below the model threshold were assessed as negative for cancer in the simulation, and mammograms above the threshold were given the original interpreting radiologist's BI-RADs assessment as obtained from our electronic medical record. This simulation scenario is illustrated in Figure 2. We calculated the overall sensitivity, specificity, and the percentage of mammograms read by the radiologists in the simulated DL-triage workflow versus the standard workflow in which radiologists interpreted all mammograms.

To illustrate the relationship between radiologist assessments and model probabilities, we provide histograms of the radiologists' true-positive, true-negative, false-positive, and false-negative assessments as ranked by the model-assessed probability of cancer, and highlight examinations triaged as cancer free by the DL model. Finally, we computed the demographics for patients bellow and above the model threshold and calculated demographic-specific sensitivities and specificities in the DL-triage workflow versus the standard workflow.

Statistical Analysis

We used scikit-learn (version 0.19.1; *https://scikit-learn.org)* for our statistical analyses and computed all statistics across 5000 bootstrap samples to obtain confidence intervals (CIs). To account for patients appearing in our test set multiple times, we used the cluster bootstrap procedure (23). We compared demographics and specificities with and without triage by using a two-tailed *t* test ($P < .05$). We compared sensitivities with and without triage by using a one-sided *t* test with a 5% inferiority margin ($P < .05$).

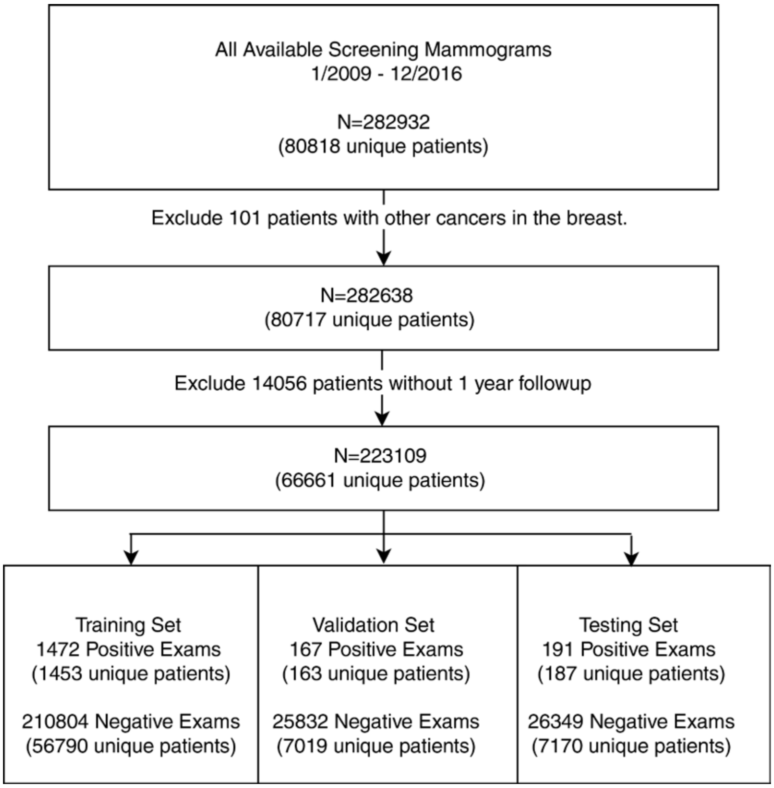

Figure 1: Flowchart shows cohort selection. From 282 638 consecutive screening mammograms performed between January 1, 2009, and December 31, 2016, a set of 223 109 examinations was selected after excluding examinations of patients who developed other cancers in breast or lacked 1-year follow-up. Positive examinations correspond to those followed by cancer diagnosis within 1 year and negative examinations correspond to examinations that were not followed by cancer diagnosis within 1 year. Additional exclusions were not performed.

Results

Patient Demographics and Outcomes

Detailed patient demographics and outcomes for the training, validation, and held-out test sets are shown in Table 1. The training set, validation set, and testing set consisted of 56 831, 7021, and 7176 patients and 212 276, 25 841, and 26 540 mammograms, respectively. The training, validation, and test sets had mean follow-ups of 3.76 years ± 2.12, 3.73 years ± 2.12, and 3.74 years ± 2.11, respectively. A total of 0.7% (1472 of 212 276), 0.6% (167 of 25 841), and 0.7% (191 of 26 540) of mammograms were followed by a cancer diagnosis within 1 year, respectively.

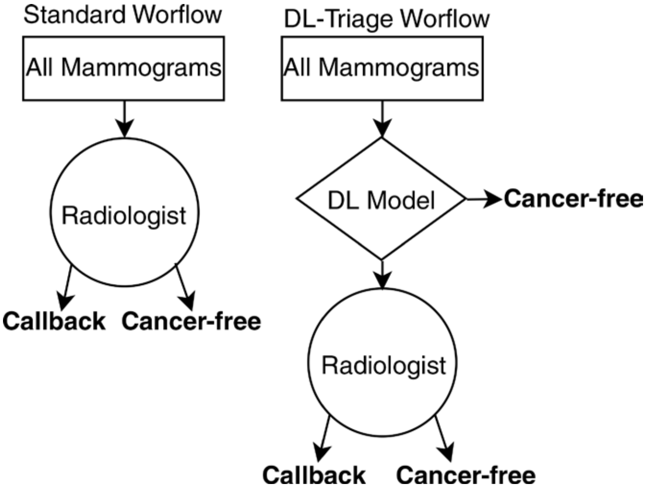

Figure 2: Diagram illustrates experimental setup for triage analysis. In standard scenario, radiologists read all mammograms. In deep learning (DL)–triage scenario, radiologists only read mammograms above model cancer-free threshold. To simulate both scenarios, original interpreting radiologist's assessment on test set was used for radiologist read.

Evaluating Our Model for Independent Prediction

The model obtained an AUC of 0.82 (95% CI: 0.80, 0.85) and an observed-to-expected ratio of 1.15 (95% CI: 0.97, 1.31) on our test set (Table 2). The receiver operating characteristic curve is shown in Figure 3. The model had similar predictive accuracies for all age groups and races (Table 2). The AUC of women in their 40s, 50s, 60s, and 70s or older was 0.80 (95% CI: 0.73, 0.89), 0.83 (95% CI: 0.77, 0.89), 0.82 (95% CI: 0.77, 0.88), 0.79 (95% CI: 0.71, 0.87), and 0.86 (95% CI: 0.75, 1.00), respectively. Similarly, the AUCs for women who were African American, Asian or Pacific Islander, white, and other were 0.86 (95% CI: 0.73, 1.00), 0.80 (95% CI: 0.60, 1.00), 0.82 (95% CI: 0.79, 0.86), and 0.81 (95% CI: 0.66, 0.99), respectively. Moreover, the model was discriminative for both women with and women without dense breasts, with AUCs of 0.82 (95% CI: 0.71, 0.94),

Table 1: Patient Demographics and Outcomes in Training, Validation, and Testing Sets

Variable	Training Examinations	Validation Examinations	Test Examinations	*P* Values*
Total	212 276, 1472 (100, 0.7)	25 841, 167 (100, 0.6)	26 540, 191 (100, 0.7)	N/A
Age (y)				
Less than 40	5856, 26 (2.8, 0.4)	719, 3 (2.8, 0.4)	732, 3 (2.8, 0.4)	>.99, .87
40–50	56 336, 299 (26.5, 0.5)	6868, 35 (26.6, 0.5)	7162, 36 (27.0, 0.5)	.12, .29
50–60	63 747, 386 (30.0, 0.6)	7825, 48 (30.3, 0.6)	8022, 52 (30.2, 0.6)	.51, .89
60–70	55 285, 426 (26.0, 0.8)	6729, 49 (26.0, 0.7)	6923, 62 (26.1, 0.9)	.89, .91
70–80	25 555, 249 (12.0, 1.0)	3058, 27 (11.8, 0.9)	3100, 30 (11.7, 1.0)	.09, .59
Greater than 80	5496, 86 (2.6, 1.6)	642, 5 (2.5, 0.8)	601, 8 (2.3, 1.3)	.002, .10
Race				
African American	9976, 55 (4.7, 0.6)	1215, 5 (4.7, 0.4)	1225, 7 (4.6, 0.6)	.54, .64
Asian or Pacific Islander	9538, 54 (4.5, 0.6)	1245, 6 (4.8, 0.5)	1260, 7 (4.7, 0.6)	.06, .71
White	172 625, 1265 (81.3, 0.7)	20 864, 142 (80.7, 0.7)	21 609, 168 (81.4, 0.8)	.70, .06
Other	20 137, 98 (9.5, 0.5)	2517, 14 (9.7, 0.6)	2446, 9 (9.2, 0.4)	.16, .04
Density				
1, fatty	20 581, 82 (9.7, 0.4)	2453, 13 (9.5, 0.5)	2519, 10 (9.5, 0.4)	.29, >.99
2, scattered	102 734, 698 (48.4, 0.7)	12 596, 75 (48.7, 0.6)	12 851, 83 (48.4, 0.6)	.94, .46
3, heterogeneously dense	79 477, 623 (37.4, 0.8)	9546, 70 (36.9, 0.7)	10 007, 91 (37.7, 0.9)	.40, .07
4, extremely dense	9371, 66 (4.4, 0.7)	1235, 9 (4.8, 0.7)	1151, 7 (4.3, 0.6)	.56, .02
Original radiologist’s BI-RADS assessment				
0	13 818, 1254 (6.5, 9.1)	1688, 132 (6.5, 7.8)	1848, 172 (7.0, 9.3)	.005, .05
1, 2	197 993, 208 (93.2, 0.1)	24 091, 35 (93.2, 0.1)	24 625, 17 (92.8, 0.1)	.003, .05

Note.—Unless otherwise specified, data are the count of all mammograms and count of mammograms positive for breast cancer, with the percentage of data set and percentage of cancers in parentheses. BI-RADS = Breast Imaging Reporting and Data System, N/A = not available.

* Indicates training versus test set, and validation versus test set.

0.81 (95% CI: 0.76, 0.86), 0.85 (95% CI: 0.81, 0.89), and 0.71 (95% CI: 0.50, 0.95) for women with fatty, scattered, heterogeneously, and extremely dense breasts, respectively. Although the AUC for women with extremely dense breast appeared lower at 0.71 than the rest of the subgroups, this population was relatively small with seven cancers in 1151 examinations as reflected by the wide CI from 0.50 to 0.95. We note that the receiver operating characteristic curve in Figure 3 and the AUCs above depict the model being used for independent prediction, that is, considering all examinations above a threshold as positive and those below the threshold as negative for all possible thresholds. This is distinct from our triage assessment, which leverages radiologist assessments above the model threshold.

Evaluating Our Model for Triage

When reading 100% of the mammograms, the original interpreting radiologists obtained a sensitivity and specificity of 90.6% (173 of 191; 95% CI: 86.6%, 94.7%) and 93.5% (24 625 of 26 349; 95% CI: 93.3%, 93.9%), respectively, within our test set. In our simulation of using the DL model for triage, the radiologists would have obtained a sensitivity and specificity of 90.1% (172 of 191; 95% CI: 86.0%, 94.3%) and 94.2% (24 814 of 26 349; 95% CI: 94.0%, 94.6%), respectively, while reading 80.7% (21 420 of 26 540; 95% CI: 80.0%, 81.5%) of all mammograms (Table 3). The increase in specificity was statistically significant (P = .002) and the sensitivity was noninferior by a margin of 5% ($P < .001$). Further analysis for alternative choices of triage threshold is available in Table E2 and Appendix E1 (online).

Radiologist Assessments by Model Probability of Cancer

The relationship between the DL model probabilities, the choice of threshold, and radiologists’ true-positive, true-negative, false-positive, and false-negative assessments are plotted in Figure 4. Of the 5120 mammograms triaged as cancer free, one was a radiologist false-negative assessment and one was a radiologist true-positive assessment. In total, 96.6% (4947 of 5120) of triaged mammograms were radiologist true-negative assessments and 3.4% (171 of 5120) were radiologist false-positive assessments. Four random examples of mammograms triaged below and above the DL model threshold are shown in Figure 5. Of radiologists’ false-negative examinations, 66.7% (12 of 18) were assigned probability of cancer in the upper quartile of all examinations (75% and higher) by the model.

Demographics by Triage: Age

Demographics of patients below and above the DL models’ high-sensitivity threshold on our test set are shown in Table 4. Of the 5120 mammograms triaged as cancer free, there was one case of cancer in a patient in their 40s and one other case of a patient in their 50s. The overall age distributions for patients below and above the threshold were similar. Among patients triaged as cancer free, 29.2% (1497 of 5120) and 32.1% (1641 of 5120) were in their 40s and 50s, respectively, compared with 26.4% (5665 of 21 420) and 29.8% (6381 of 21 420) of women above the DL model threshold. Although more young patients were triaged as cancer free, this trend was expected given the relatively lower incidence of cancer in these age groups on the test set, with 0.5% (36 of 7162) and 0.6% (52 of 8022) diagnosed with cancer, respectively, compared with the overall rate of 0.7% (191 of 26 540) (Table 1).

Demographics by Triage: Race

Of the 5120 mammograms triaged as cancer free, the two cases of cancer were in white patients. The overall distribution by race was similar below and above the threshold. The patients below the model threshold were 78.6% (4016 of 5120) white, 5% (255 of 5120) African American, 5% (257 of 5120) Asian or Pacific Islander, and 11.5% (592 of 5120) other, compared with 82.1% (17 593 of 26 540), 4.5% (970 of 26 540), 4.7% (1009 of 26 540), and 8.6% (1848 of 26 540) above the threshold (Table 4). Although fewer white women were triaged as cancer free by the DL model, this was expected given that white women had the highest cancer incidence on the test set, with incidences of 0.8%, 0.6%, 0.6%, 0.7%, and 0.4% for women who were white, African American, Asian or Pacific Islander, and other, respectively (Table 1).

Table 2: AUCs and Observed-to-Expected Ratios for Different Cohorts in the Test Set

Cohort	AUC	Observed-to-Expected Ratio
Full test set	0.82 (0.80, 0.85)	1.15 (0.97, 1.31)
Age (y)		
40–50	0.80 (0.73, 0.89)	1.00 (0.65, 1.32)
50–60	0.83 (0.77, 0.89)	0.98 (0.68, 1.24)
60–70	0.82 (0.77, 0.88)	1.38 (1.04, 1.70)
70–80	0.79 (0.71, 0.87)	1.22 (0.79, 1.60)
Greater than 80	0.86 (0.75, 1.00)	1.27 (0.37, 2.05)
Race		
African American	0.86 (0.73, 1.00)	0.93 (0.19, 1.57)
Asian or Pacific Islander	0.80 (0.60, 1.00)	1.16 (0.22, 1.95)
White	0.82 (0.79, 0.86)	1.21 (1.03, 1.39)
Other	0.81 (0.66, 0.99)	0.67 (0.22, 1.05)
Density		
1, fatty	0.82 (0.71, 0.94)	0.97 (0.31, 1.53)
2, scattered	0.81 (0.76, 0.86)	1.05 (0.82, 1.27)
3, heterogeneously dense	0.85 (0.81, 0.89)	1.32 (1.04, 1.58)
4, extremely dense	0.71 (0.50, 0.97)	0.91 (0.12, 1.52)

Note.— Test set consists of 26 540 examinations from 7176 patients. Data in parentheses are 95% confidence intervals. AUC = area under receiver operator characteristic curve.

Demographics by Triage: Breast Density

Of the mammograms falling below the threshold, 20.6% (1057 of 5120) were assessed as fatty breast density and 54.3% (2779 of 5120) were assessed as scattered fibroglandular density, compared with 6.8% (1462 of 26 540) and 47.0% (10 072 of 26 540), respectively, for mammograms above the threshold (Table 4). This also matched to a relatively lower incidence of breast cancer for women with nondense breasts in our test set, with incidence rates of 0.4% and 0.6% for women with fatty and scattered fibroglandular breast densities, respectively, compared with the overall incidence of 0.7% (191 of 26 540) (Table 1). Of the 5120 mammograms triaged as cancer free, the two cases of cancer had mammographic density assessments of heterogeneously dense and scattered fibroglandular density.

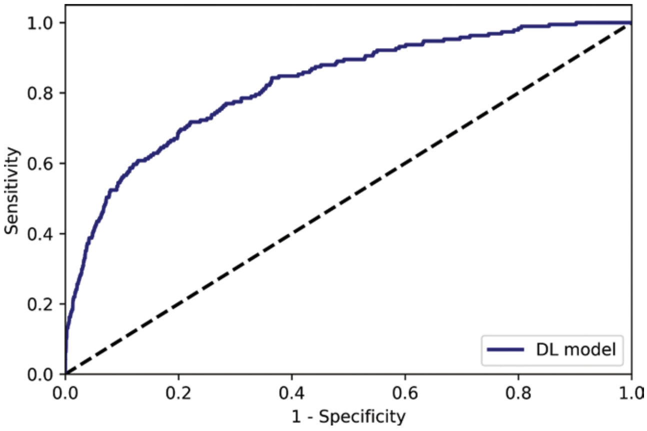

Figure 3: Graph shows receiver operating characteristic curve of deep learning (DL) model making independent predictions on test set. Area under receiver operator characteristic curve of DL model is 0.82 (95% confidence interval: 0.80, 0.85).

Table 3: Sensitivity, Specificity, and Portion of Mammograms Read on the Test Set

Setting	Sensitivity (%)	Specificity (%)	Mammograms Read (%)
Original interpreting radiologist reading all images	90.6 (173/191) [86.6, 94.7]	93.6 (24 625/26 349) [93.3, 93.9]	100 (26 540/26 540) [100, 100]
Original interpreting radiologist + deep learning triage (reading mammograms not triaged as cancer free)	90.1 (172/191) [86.0, 94.3]	94.3 (24 814/26 349) [94.0, 94.6]	80.7 (21 420/26 540)[80.0, 81.5]

Note.— Data in parentheses are numerators and denominators, with 95% confidence intervals in brackets.

Evaluating Triage by Demographic

Demographic-wise sensitivities and specificities as well as *P* values regarding their differences are reported in Table E1 (Appendix E1 [online]). Because only one triaged mammogram was a radiologist true-positive assessment, the sensitivity was identical in almost all subgroups. Although the specificity was improved in all demographic subgroups, this improvement was significant for patients who were in their 40s (*P* = .02), were white (*P* = .01), or with breast density assessment of scattered fibroglandular density (*P* = .01).

Discussion

We developed a deep learning (DL) model to triage mammograms as cancer free and to improve radiologist efficiency and specificity without impacting sensitivity. In our simulated triage workflow, in which radiologists would only read mammograms above the cancer-free threshold, our model showed a workload reduction of 19.3% (5120 of 26 540), a significant improvement in specificity (93.5%–94.2%; *P* = .002), and a noninferior sensitivity (90.6%–90.1%; *P* < .001). Our model was discriminative across all age groups, races, and breast density categories, suggesting the model may be widely applicable to diverse patient populations.

DL models for whole image classification are uniquely suited to triage mammograms due to their ability to both detect local areas of cancer (as shown in a DL-based computer-aided detection [CAD] model) (17) and their ability to assess breast cancer risk (22). By simultaneously leveraging cues of both present cancers and future cancers that may not be visible, DL models are able to access a large portion of the population as cancer free without access to additional information commonly available to radiologists, such as prior mammograms.

This work takes a substantial departure from prior work on CAD (14–16). Instead of annotating images to draw added attention to potentially malignant findings (to improve sensitivity), we propose to triage cancer-free mammograms from the workflow to improve both specificity and efficiency. For example, a reader study showed that a DL-based CAD model could improve sensitivity from 83% to 86% without impacting specificity or slowing image reading time (17). In contrast, our model improved specificity, improved efficiency, and did not impact sensitivity. These two approaches (CAD and triage) may be complimentary, giving more attention to mammograms that warrant it and removing attention from those that do not. The idea of triaging or "preselecting" mammograms with a DL model has been

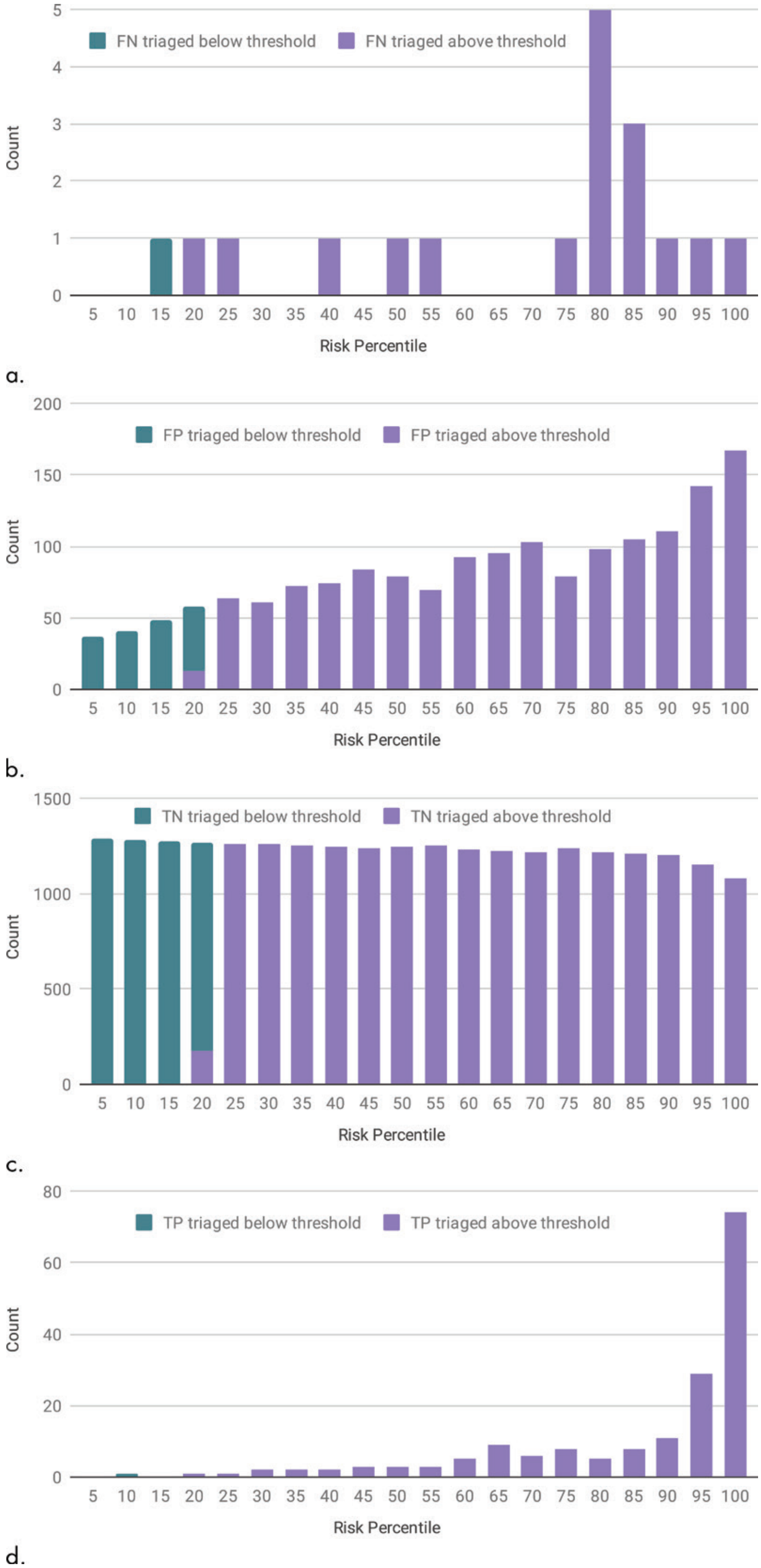

Figure 4: Graphs show relationship between radiologist assessment, risk percentile of deep learning model, and cancer-free threshold in test set. **(a)** Radiologist false-negative (FN) assessments triaged below and above cancer-free threshold by risk decile. **(b)** Radiologist false-positive (FP) assessments triaged below and above cancer-free threshold by risk decile. **(c)** Radiologist true-negative (TN) assessments triaged below and above cancer-free threshold by risk decile. **(d)** Radiologist true-positive (TP) assessments triaged below and above cancer-free threshold by risk decile.

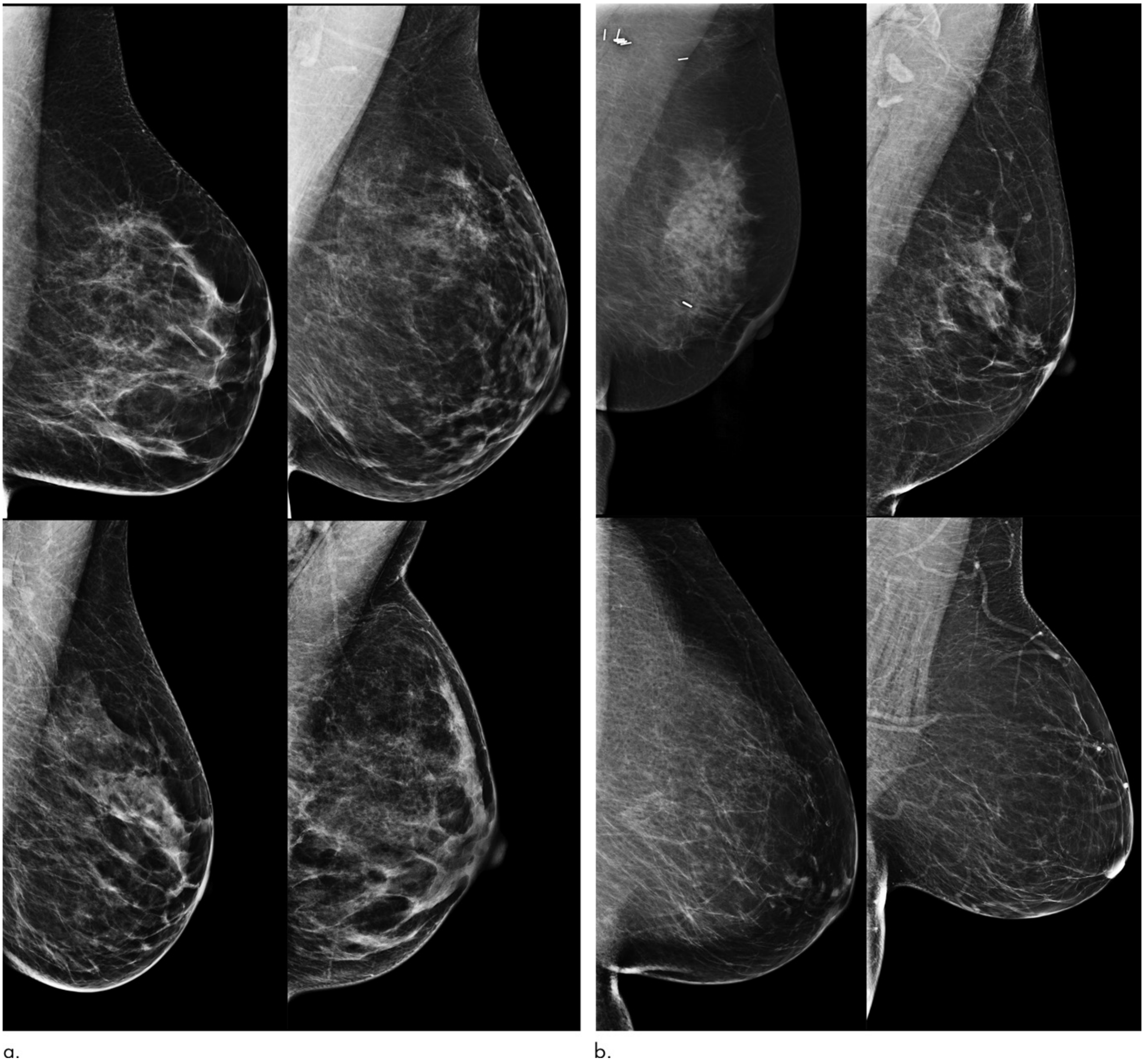

Figure 5: Images show randomly selected left mediolateral-oblique views. **(a)** Examples of mammograms triaged as cancer free on test set. **(b)** Examples of mammograms not triaged as cancer free on test set.

concurrently explored in a recent reader study by Rodriguez-Ruiz et al (24), which used a commercial system and showed a simulated workload reduction of 17% with a drop of 1% in sensitivity and a noninferior radiologist AUC of 0.05 on an enriched data set. In contrast, we showed a workload reduction of 19%, a noninferior sensitivity, and a significant improvement in specificity in a natural screening data set (ie, not enriched). Our results are qualitatively similar and support the same hypothesis through different methods.

Our DL model produces a probability of cancer, and although we identified a threshold to triage mammograms as cancer free, our model could be leveraged in a diversity of other paradigms with different thresholds. In Europe, where double reading is more common (18,19), our model could identify a subset of cases suitable for single-human assessment. In more resource-constrained scenarios with insufficient specialized radiologists, our model could identify the patients with highest risk and support more efficient resource allocation. In all paradigms, reducing workload through examination triage can free radiologists to provide care in other critical areas not currently supported by artificial intelligence, such as performing image-guided procedures, diagnostic testing, and patient interaction. In this realm, the model serves to enhance the overall impact of the radiologist on improved patient care.

In addition to its strong potential to improve workflow efficiency, our model supports improved diagnostic performance by triaging false-positive mammograms as cancer free. Recall rates from screening mammography have increased steadily in the United States, with current average recall rates estimated above 10%, and ranging widely among U.S. radiologists from 3.4% to 30.7% (11). False-positive results are associated with unnecessary additional testing and biopsies, and are estimated to add more than $2.8 billion annually to health care costs in the United States alone (25). The clinical significance of the improved specificity of the model has considerable impact on both patients and health care systems.

Table 4: Patient Demographics for Examinations Triaged Below and Above the Cancer-free Threshold of the Deep Learning Model

Cohort	Triaged Below Threshold	Triaged Above Threshold	*P* Value
Full set	5120, 2 (100.0, 0.0)	21 420, 189 (100.0, 0.9)	N/A
Age (y)			
40–50	1497, 1 (29.2, 0.1)	5665, 35 (26.4, 0.6)	<.001
50–60	1641, 1 (32.1, 0.1)	6381, 51 (29.8, 0.8)	.002
60–70	1308, 0 (25.5, 0.0)	5615, 62 (26.2, 1.1)	.33
70–80	437, 0 (8.5, 0.0)	2663, 30 (12.4, 1.1)	<.001
Greater than 80	49, 0 (1.0, 0.0)	552, 8 (2.6, 1.4)	<.001
Race			
African American	255, 0 (5.0, 0.0)	970, 7 (4.5, 0.7)	.16
Asian or Pacific Islander	257, 0 (5.0, 0.0)	1009, 7 (4.7, 0.7)	.35
White	4016, 2 (78.4, 0.0)	17 593, 166 (82.1, 0.9)	<.001
Other	592, 0 (11.6, 0.0)	1848, 9 (8.6. 0.5)	<.001
Density			
1, fatty	1057, 0 (20.6, 0.0)	1462, 10 (6.8, 0.7)	<.001
2, scattered	2779, 1 (54.3, 0.0)	10 072, 82 (47.0, 0.8)	<.001
3, heterogeneously dense	1166, 1 (22.8, 0.1)	8841, 90 (41.3, 1.0)	<.001
4, extremely dense	116, 0 (2.3, 0.0)	1035, 7 (4.8, 0.7)	<.001
Original radiologist's BI-RADS assessment			
0	172, 1 (3.4, 0.6)	1676, 171 (7.8, 10.2)	<.001
1/2	4931, 1 (96.3, 0.0)	19 694, 16 (91.9, 0.1)	<.001

Note.— Test set consists of 26 540 examinations from 7176 patients. Unless otherwise specified, data are the count of all mammograms and count of mammograms positive for breast cancer, with the percentage of data set and percentage of cancers in parentheses. BI-RADS = Breast Imaging Reporting and Data System, N/A = not available.

Our study had several limitations. First, our analysis of the potential impact of using the model to triage mammograms was retrospective and assumed that radiologists would have read the remaining images the same way, whether those marked as cancer free were included in their routine worklist or not. This simulation method is only able to show a decrease in sensitivity or an increase in specificity. It is possible the knowledge of risk assessment of the mammogram and/or a reduced workload could improve radiologist sensitivity and/or harm the specificity of interpretations of the remaining mammograms. It is also possible that reading in a population with higher incidence of cancer (because a large fraction of noncancers were triaged) in itself will impact reading performance. Although our preliminary results are promising, a prospective trial is needed to confirm the impact of our model in clinical practice across a diversity of radiologists. Moreover, our model was developed at a single tertiary academic institution. Further external validation with diverse populations will be required prior to regulatory approval and widespread clinical implementation. Lastly, our model was developed and tested by using mammograms from a single vendor (Hologic) and more study is needed to determine the performance of our model in examinations obtained with mammography units from diverse vendors. To this end, we make our code and trained model available for research use at *http://learningtocure.csail.mit.edu*.

In summary, we developed a deep learning model to triage mammograms as cancer free and showed that our model could improve radiologist efficiency and specificity without harming sensitivity. This work is a first step to using deep learning to triage mammograms in routine clinical care.

Acknowledgment: The authors are grateful to Susan G. Komen and the Breast Cancer Alliance for their support of this project.

Author contributions: Guarantors of integrity of entire study, A.Y., R.B., C.L.; 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, R.M., R.B., C.L.; clinical studies, R.M., C.L.; statistical analysis, A.Y., T.S.; and manuscript editing, A.Y., R.M., R.B., C.L.

Disclosures of Conflicts of Interest: **A.Y.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: institution has pending patent for algorithm described in this work. **T.S.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: completed a paid internship at Massachusetts General Hospital. Other relationships: disclosed no relevant relationships. **R.M.** disclosed no relevant relationships. **R.B.** Activities related to the present article: institution received grant from Breast Cancer Alliance and Susan G. Komen. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **C.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a member of GE Healthcare advisory board; has grants/grants pending with GE Healthcare. Other relationships: disclosed no relevant relationships.

References

1. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet 2012;380(9855):1778–1786.
2. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. J Natl Cancer Inst 1982;69(2):349–355.
3. Andersson I, Janzon L, Sigfússon BF. Mammographic breast cancer screening: a randomized trial in Malmö, Sweden. Maturitas 1985;7(1):21–29.

4. Tabár L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985;1(8433):829–832.
5. Roberts MM, Alexander FE, Anderson TJ, et al. The Edinburgh randomised trial of screening for breast cancer: description of method. Br J Cancer 1984;50(1):1–6.
6. Frisell J, Glas U, Hellström L, Somell A. Randomized mammographic screening for breast cancer in Stockholm: design, first round results and comparisons. Breast Cancer Res Treat 1986;8(1):45–54.
7. Miller AB, Howe GR, Wall C. The National Study of Breast Cancer Screening Protocol for a Canadian randomized controlled trial of screening for breast cancer in women. Clin Invest Med 1981;4(3-4):227–258.
8. Bjurstam N, Björneld L, Duffy SW, et al. The Gothenburg breast screening trial: first results on mortality, incidence, and mode of detection for women ages 39-49 years at randomization. Cancer 1997;80(11):2091–2099.
9. Wing P, Langelier MH. Workforce shortages in breast imaging: impact on mammography utilization. AJR Am J Roentgenol 2009;192(2):370–378.
10. United States General Accountability Office Website. Mammography: current nationwide capacity is adequate, but access problems may exist in certain locations. Report no. 06-724. Washington, DC: GAO, 2006. https://www.gao.gov/new.items/d06724.pdf. Accessed December 20, 2018.
11. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology 2017;283(1):49–58.
12. Salazar G, Quencer K, Aran S, Abujudeh H. Patient satisfaction in radiology: qualitative analysis of written complaints generated over a 10-year period in an academic medical center. J Am Coll Radiol 2013;10(7):513–517.
13. Amir T, Lee B, Woods RW, Mullen LA, Harvey SC. A pilot of data-driven modeling to assess potential for improved efficiency in an academic breast-imaging center. J Digit Imaging 2019;32(2):221–227.
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. 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.
16. Bahl M. Detecting breast cancers with mammography: will AI succeed where traditional CAD failed? Radiology 2019;290(2):315–316.
17. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290(2):305–314.
18. 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.
19. Posso M, Puig T, Carles M, Rué M, Canelo-Aybar C, Bonfill X. Effectiveness and cost-effectiveness of double reading in digital mammography screening: a systematic review and meta-analysis. Eur J Radiol 2017;96:40–49.
20. Lehman CD, Yala A, Schuster T, et al. Mammographic breast density assessment using deep learning: clinical implementation. Radiology 2019;290(1):52–58.
21. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 2019 May 7:182716 [Epub ahead of print].
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; 770–778.
23. Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc Series B Stat Methodol 2007;69(3):369–390.
24. 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 Apr 16 [Epub ahead of print].
25. Ong MS, Mandl KD. National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. Health Aff (Millwood) 2015;34(4):576–583.