ORIGINAL RESEARCH • BREAST IMAGING

Radiology

# A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction

*Adam Yala, MEng* • *Constance Lehman, MD, PhD* • *Tal Schuster, MS* • *Tally Portnoi, BS* • *Regina Barzilay, PhD*

From the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 32 Vassar St, 32-G484, Cambridge, MA 02139 (A.Y., T.S., T.P., R.B.); and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Mass (C.L.). Received November 28, 2018; revision requested January 18, 2019; revision received March 14; accepted March 18. **Address correspondence to** A.Y. (e-mail: *adamyala@csail.mit.edu*).

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

See also the editorial by Sitek and Wolfe in this issue.

**Radiology 2019; 292:60–66 • https://doi.org/10.1148/radiol.2019182716 • Content code: BR**

***Background:*** Mammographic density improves the accuracy of breast cancer risk models. However, the use of breast density is limited by subjective assessment, variation across radiologists, and restricted data. A mammography-based deep learning (DL) model may provide more accurate risk prediction.

***Purpose:*** To develop a mammography-based DL breast cancer risk model that is more accurate than established clinical breast cancer risk models.

***Materials and Methods:*** This retrospective study included 88 994 consecutive screening mammograms in 39 571 women between January 1, 2009, and December 31, 2012. For each patient, all examinations were assigned to either training, validation, or test sets, resulting in 71 689, 8554, and 8751 examinations, respectively. Cancer outcomes were obtained through linkage to a regional tumor registry. By using risk factor information from patient questionnaires and electronic medical records review, three models were developed to assess breast cancer risk within 5 years: a risk-factor-based logistic regression model (RF-LR) that used traditional risk factors, a DL model (image-only DL) that used mammograms alone, and a hybrid DL model that used both traditional risk factors and mammograms. Comparisons were made to an established breast cancer risk model that included breast density (Tyrer-Cuzick model, version 8 [TC]). Model performance was compared by using areas under the receiver operating characteristic curve (AUCs) with DeLong test ($P < .05$).

***Results:*** The test set included 3937 women, aged 56.20 years ± 10.04. Hybrid DL and image-only DL showed AUCs of 0.70 (95% confidence interval [CI]: 0.66, 0.75) and 0.68 (95% CI: 0.64, 0.73), respectively. RF-LR and TC showed AUCs of 0.67 (95% CI: 0.62, 0.72) and 0.62 (95% CI: 0.57, 0.66), respectively. Hybrid DL showed a significantly higher AUC (0.70) than TC (0.62; $P < .001$) and RF-LR (0.67; $P = .01$).

***Conclusion:*** Deep learning models that use full-field mammograms yield substantially improved risk discrimination compared with the Tyrer-Cuzick (version 8) model.



*Online supplemental material is available for this article.*

Since the creation of the Gail model in 1989 (1), risk models have supported risk-adjusted screening and prevention and their continued evolution has been a central pillar of breast cancer research (1–8). Previous research (2,3) explored multiple risk factors related to hormonal and genetic information. Mammographic breast density, which relates to the amount of fibroglandular tissue in a woman's breast, is a risk factor that received substantial attention. Brentnall et al (8) incorporated mammographic breast density into the Gail risk model and Tyrer-Cuzick model (TC), improving their areas under the receiver operating characteristic curve (AUCs) from 0.55 and 0.57 to 0.59 and 0.61, respectively.

The use of breast density as a proxy for the detailed information embedded on the mammogram is limited because breast density assessment is a subjective assessment and varies widely across radiologists (9), and breast density summarizes the information contained in the digital images into a single value. Same-age patients who are assigned the same density score can have drastically different mammography with vastly different outcomes. Whereas previous studies (10–12) explored automated methods to assess breast density, these efforts reduced the mammographic input into a few statistics largely related to volume of glandular tissue that are not sufficient to distinguish patients who will and will not develop breast cancer.

We hypothesize that there are subtle but informative cues on mammograms that may not be discernible by humans or simple volume-of-density measurements, and deep learning (DL) can leverage these cues to yield improved risk models. Therefore, we developed a DL model that operates over a full-field mammographic image to assess a patient's future breast cancer risk. Rather than manually identifying discriminative image patterns, we rely on our machine learning model to discover these patterns directly from the data. Specifically, our model is provided with full-field mammograms and the outcome of interest, namely whether or not the patient developed breast cancer within 5 years from the date of the examination. In addition to our image-only model, we developed

## Abbreviations

AUC = area under the receiver operating characteristic curve, CI = confidence interval, DL = deep learning, RF-LR = risk-factor-based logistic regression, TC = Tyrer-Cuzick model version 8


## Summary

We developed a deep learning model that uses full-field mammograms and traditional risk factors, and found that our model was more accurate than the Tyrer-Cusick model (version 8), a current clinical standard.


## Key Points

- A deep learning (DL) mammography-based model identified women at high risk for breast cancer and placed 31% of all patients with future breast cancer in the top risk decile compared with only 18% by the Tyrer-Cuzick model (version 8).
- Our hybrid DL model is equally accurate for white and African American women (area under the receiver operating characteristic curve [AUC], 0.71 for both ethnicities) whereas the Tyrer-Cuzick model AUC was 0.62 and 0.45 for women who were white and African American, respectively; the AUC improvement was significant for women who were white ($P < .001$) and African American ($P < .01$).
- When our hybrid DL model was compared with breast density, we found that patients with nondense breasts and model-assessed high risk had 3.9 times the cancer incidence of patients with dense breasts and model-assessed low risk.

two additional models in the same cohort: a logistic regression model that operates on the basis of traditional risk factors and that provides a strong baseline for our population, and a hybrid model that operates on both the full-field mammogram and traditional risk factors. We compare all three to TC (4), a popular risk model that includes breast density and is routinely used in clinical practice.

## Materials and Methods

Our retrospective study was approved by our institutional review board with a waiver for written informed consent. It was compliant with the Health Insurance Portability and Accountability Act. Mammograms in 39 272 of the 60 886 women in our patient population were previously studied in our development of a breast density assessment algorithm (10).

### Data Collection

We collected consecutive digital screening mammograms (Hologic, Bedford, Mass) in 60 886 patients between January 1, 2009, and December 31, 2012, at a large tertiary academic medical center. For each patient, we obtained outcomes through linkage to tumor registries for 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). We collected detailed risk factors, including those used by the TC model, from provider-entered information and patient-entered questionnaires in the electronic medical record. We associated each mammogram with patient risk factors manifest at the time of mammography.

Of the initial 60 886 patients, we included women who had either a diagnosis of breast cancer (ductal carcinoma in situ or invasive breast carcinoma) within 5 years, or imaging follow-up for at least 5 years from the date of index mammography. We note that each woman may have undergone several mammographic examinations, and we considered each mammographic examination as the index mammography independently for inclusion. We excluded 21 328 women because they lacked sufficient follow-up or had another form of cancer in their breast. We did not exclude on the basis of previous operations, age, implants, atypical lesions, or previous cancers. The remaining 39 558 women were randomly assigned as follows: 31 806 women, training; 3804 women, validation; and 3978 women, testing. To restrict our evaluation to a negative-for-cancer screening population, we excluded 41 women who were diagnosed with cancer within 1 year of index mammography. This resulted in training, validation, and test sets of 71 689, 8554, and 8751 mammographic examinations, respectively (Fig 1). We split our data set by patients, therefore each woman only contributed mammograms to one set, and no mammographic examinations in the test set were followed by a cancer diagnosis within 1 year.

### Model Development and Evaluation

In-depth information about all developed models, model selection, and calibration is in Appendix E1 (online). We obtained TC risk assessments by using the Command-Line version of the IBIS Breast Cancer Risk Evaluation Tool (version 8; IBIS, London, England, *http://www.ems-trials.org/riskevaluator/*).

We implemented our risk-factor-only model as a logistic regression model (risk factor logistic regression model [RF-LR]) with scikit-learn (version 0.19.1, *scikit-learn.org*). We trained the RF-LR model to map a patient's risk factors at the time of mammography to whether or not the patient developed cancer within 5 years.

For the image-only DL model, we implemented a deep convolutional neural network (ResNet18 [13]) with PyTorch (version 0.31; *pytorch.org*). Given a 1664 × 2048 pixel view of a breast, the DL model was trained to predict whether or not that breast would develop breast cancer within 5 years. We did not exclude any views, and the model used the entire image at full field.

We also developed a hybrid DL model to combine both image information and risk factors used in the RF-LR model.

To evaluate the models, we computed the AUC, and the portion of all cancers placed in the top risk decile and in the bottom risk decile for all models on the full test set. Next, we calculated each model's AUC for the following subgroups: white and African American women, premenopausal and postmenopausal women, and women with and without a family history of breast or ovarian cancer. To measure the ability of the models to capture long-term future risk, we calculated each model's AUC in distinguishing patients who developed cancer within 3–5 years from patients who did not develop cancer within 5 years.

### Confusion Matrix Analysis

We computed a confusion matrix for examinations with different combinations of breast density and hybrid DL risk. Each examination in the test set was placed in a cell by breast density (row) and hybrid DL risk (column). Rows correspond to non-

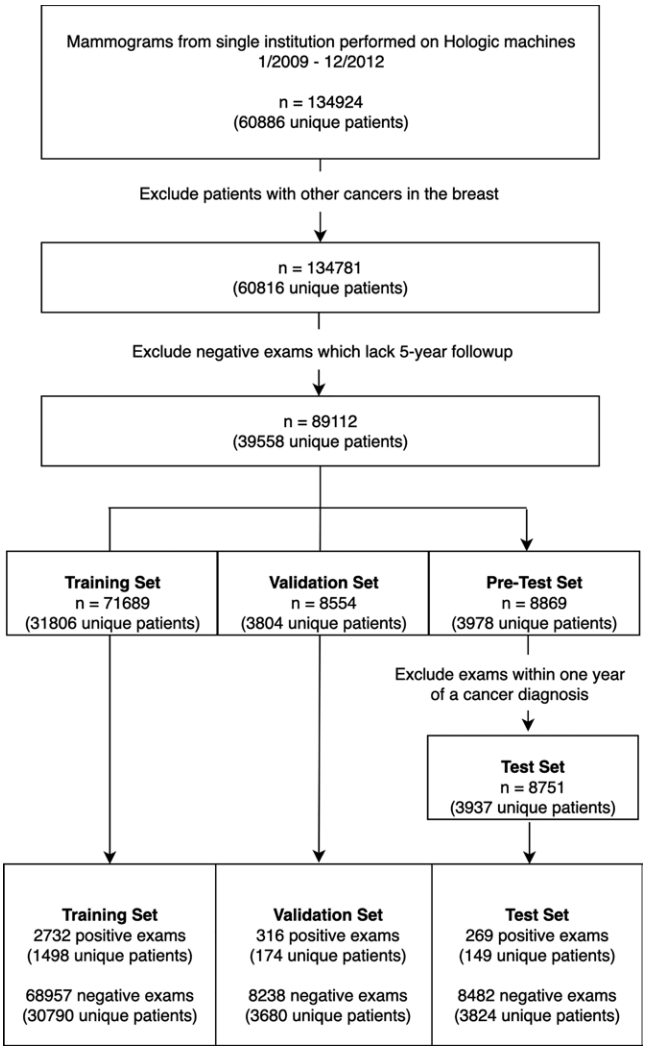



**Figure 1:** Cohort selection flowchart. There were 134 924 consecutive screening mammograms performed between January 1, 2009, and December 31, 2012. Examinations were defined as positive for cancer if they were followed by a cancer diagnosis within 5 years and negative for cancer if they were not. To restrict the test set to a negative screening population, we excluded examinations that were followed by cancer within 1 year.

dense (Breast Imaging Reporting and Data System categories a and b) and dense (Breast Imaging Reporting and Data System categories c and d) breasts, and columns correspond to the lowest and highest 50% risk examinations as ranked by hybrid DL. In each cell, we reported the fraction of examinations that developed cancer within 5 years. We repeated a similar analysis to compare against TC, in which rows represented the low-, middle-, and high-risk thirds by TC and columns represented the low-, middle-, and high-risk thirds by hybrid DL. Last, we provide example mammography of each cell in the confusion matrices.

## Statistical Analysis

We used the pROC (14) package in R (version 3.5.2; R Project for Statistical Computing, *https://www.r-project.org*) to compare AUCs with DeLong test (15) ($P < .05$ indicated statistical significance) and used scikit-learn (version 0.19.1; *scikit-learn.org*) for all other statistical analyses. We computed statistics across 5000 clustered bootstrap samples (16) to obtain confidence intervals (CIs).

# Results

We generated a detailed breakdown of available risk factor information and outcomes for the training, validation, and test sets (Tables 1 and E1 [online]). Risk factors used in TC, RF-LR, and hybrid DL included age, weight, height, menarche age, menopausal status, detailed family history of breast and ovarian cancer, *BRCA* mutation status, history of atypical hyperplasia, history of lobular carcinoma in situ, and breast density. Of the 80 243 mammographic examinations used for training and validation, 3045 (3.8%) were followed by a cancer diagnosis within 5 years. Of the 8751 mammographic examinations used for testing, 269 (3.1%) were followed by a cancer diagnosis within 5 years.

## Model Evaluation

*Full test set.*—The TC, RF-LR, image-only DL, and hybrid DL models showed AUCs of 0.62 (95% CI: 0.57, 0.66), 0.67 (95% CI: 0.62, 0.72), 0.68 (95% CI: 0.64, 0.73), and 0.70 (95% CI: 0.66, 0.75), respectively (Table 2). Hybrid DL had a significantly higher AUC than TC ($P < .001$) and RF-LR ($P = .01$). Image-only DL had a significantly higher AUC than TC ($P < .01$) but not RF-LR ($P = .40$). The receiver operating characteristic curves of the four models are shown in Figure 2.

Hybrid DL showed the best decile performance, placing 31.2% (84 of 269; 95% CI: 24.2%, 38.2%) of cancers in the top decile and 3.0% (eight of 269; 95% CI: 0.3%, 5.0%) of cancers in the bottom decile, compared with 18.2% (49 of 269; 95% CI: 11.3%, 24.3%) and 4.8% (13 of 269; 95% CI: 1.3%, 7.7%) in the top and bottom deciles, respectively, by TC.

*Subgroups by race, menopausal status, and family history.*—Hybrid DL showed AUCs of 0.71 (95% CI: 0.67, 0.74) and 0.71 (95% CI: 0.57, 0.87) for patients who were white and African American, respectively, compared with AUCs of 0.62 (95% CI: 0.58, 0.65) and 0.45 (95% CI: 0.26, 0.64), respectively, at TC (Table 3, Fig 3). Both improvements were significant ($P < .001$ and $< .01$, respectively, for white and African American patients).

The hybrid DL model showed the highest AUC for both pre- and postmenopausal women (AUCs, 0.79 [95% CI: 0.67, 0.97] and 0.70 [95% CI: 0.65, 0.75], respectively; Fig 3, Table 4). However, TC showed AUCs of 0.73 (95% CI: 0.57, 0.90) and 0.58 (95% CI: 0.53, 0.64) for pre- and postmenopausal women, respectively. The improvement was not significant for premenopausal women ($P = .40$) but was significant for postmenopausal women ($P < .001$).

For patients with any family history of breast or ovarian cancer, hybrid DL showed the highest AUC (AUC, 0.70; 95% CI: 0.64, 0.76) compared with image-only DL (AUC, 0.65; 95% CI: 0.59, 0.71) and TC (AUC, 0.59; 95% CI: 0.52, 0.67) (Fig 3,

**Table 1: Patient Characteristics and Outcomes in Training, Development Validation, and Test Sets**

| | Training Examinations | | Validation Examinations | | Test Examinations | |
|---|---|---|---|---|---|---|
| Characteristic | Data Set | Cancer | Data Set | Cancer | Data Set | Cancer |
| All patients | 71 689 (100) | 2729 (3.8) | 8554 (100) | 316 (3.7) | 8751 (100) | 269 (3.1) |
| Age (y) | | | | | | |
| >40 | 2360 (3.3) | 49 (2.1) | 268 (3.1) | 6 (2.2) | 290 (3.3) | 1 (0.3) |
| 40–50 | 20 640 (28.8) | 596 (2.9) | 2464 (28.8) | 73 (3.0) | 2620 (29.9) | 51 (1.9) |
| 50–60 | 22 630 (31.6) | 686 (3.0) | 2750 (32.1) | 87 (3.2) | 2778 (31.7) | 88 (3.2) |
| 60–70 | 18 937 (26.4) | 896 (4.7) | 2247 (26.3) | 96 (4.3) | 2277 (26.0) | 72 (3.2) |
| 70–80 | 6347 (8.9) | 382 (6.0) | 741 (8.7) | 44 (5.9) | 731 (8.4) | 46 (6.3) |
| >80 | 775 (1.1) | 120 (15.5) | 84 (1.0) | 10 (11.9) | 55 (0.6) | 11 (20.0) |
| Density | | | | | | |
| Almost entirely fatty | 6073 (8.5) | 158 (2.6) | 698 (8.2) | 25 (3.6) | 737 (8.4) | 6 (0.8) |
| Scattered areas of fibroglandular tissue | 34 143 (47.6) | 1306 (3.8) | 4156 (48.6) | 141 (3.4) | 4126 (47.1) | 117 (2.8) |
| Heterogeneously dense | 27 897 (38.9) | 1132 (4.1) | 3240 (37.9) | 131 (4.0) | 3473 (39.7) | 135 (3.9) |
| Extremely dense | 3530 (4.9) | 132 (3.7) | 454 (5.3) | 19 (4.2) | 411 (4.7) | 11 (2.7) |
| Unknown | 46 (0.1) | 1 (2.2) | 6 (0.1) | 0 (0) | 4 (0) | 0 (0) |

Note.—Data are the number of examinations in each group; data in parentheses are percentages.

**Table 2: Risk Test Set for All 5-year Risk Assessment Models**

| Model | AUC | Top Decile Hazard Ratio | Bottom Decile Hazard Ratio | Portion of Cancers in Top Decile | Portion of Cancers in Bottom Decile |
|---|---|---|---|---|---|
| TC | 0.62 (0.57, 0.66) | 1.89 (0.91, 2.63) | 0.50 (0.08, 0.81) | 0.18 (0.11, 0.24) | 0.05 (0.01, 0.08) |
| RF-LR | 0.67 (0.62, 0.72) | 3.69 (2.25, 4.94) | 0.41 (0, 0.72) | 0.31 (0.23, 0.38) | 0.03 (0, 0.06) |
| Image-only DL | 0.68 (0.64, 0.73) | 2.31 (1.46, 3.02) | 0.40 (0.09, 0.61) | 0.22 (0.16, 0.27) | 0.04 (0.01, 0.06) |
| Hybrid DL | 0.70 (0.66, 0.75) | 3.80 (2.45, 4.91) | 0.36 (0.01, 0.60) | 0.31 (0.24, 0.38) | 0.03 (0, 0.05) |

Note.—Data in parentheses are 95% confidence intervals. There were a total of 3937 patients, 8751 examinations, and 269 cancers. AUC = area under receiver operator characteristic curve, DL = deep learning, RF-LR = risk-factor-based logistic regression, TC = Tyrer-Cuzick.

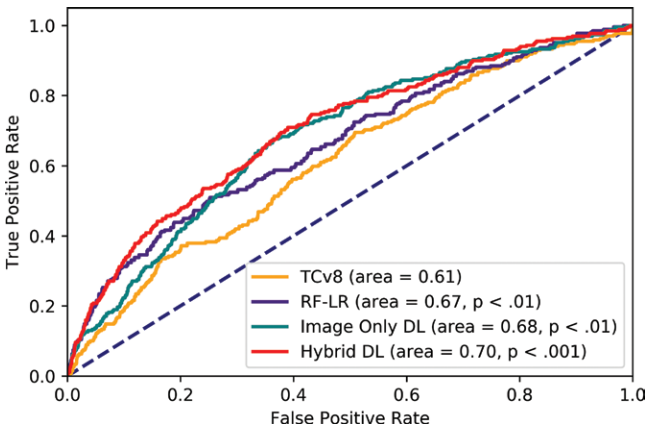


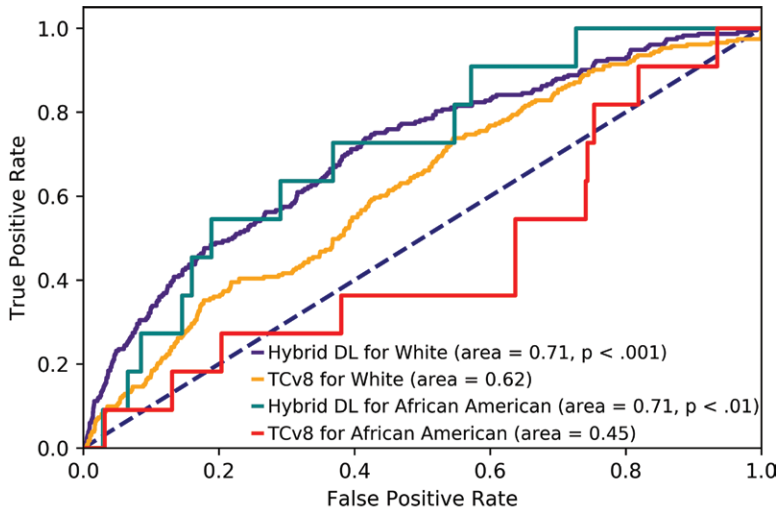

**Figure 2:** Receiver operating characteristic curve of all models on the test set. All *P* values are comparisons with Tyrer-Cuzick version 8 (TCv8). DL = deep learning, hybrid DL = DL model that uses both imaging and the traditional risk factors in risk factor logistic regression, RF-LR = risk factor logistic regression.

**Table 3: Risk Test Set for 5-year Risk Assessment Models by Ethnicity**

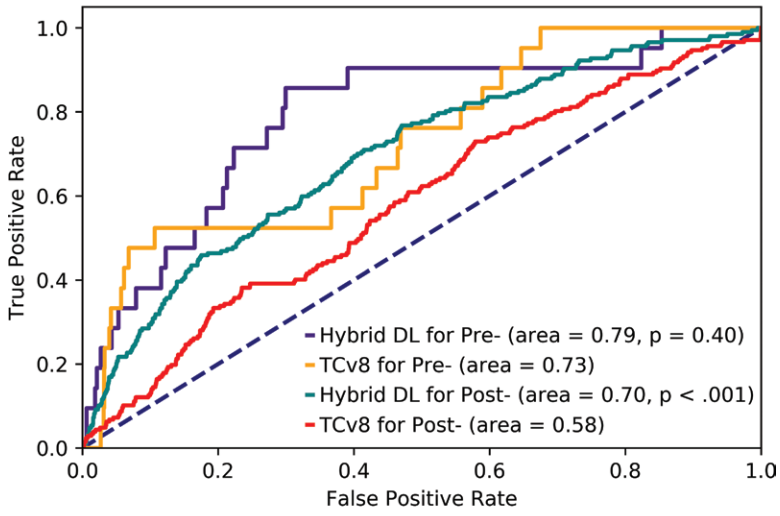
| Parameter | AUC |
|---|---|
| Ethnicity | |
| White | |
| TC | 0.62 (0.57, 0.67) |
| RF-LR | 0.66 (0.61, 0.72) |
| Image-only DL | 0.69 (0.65, 0.74) |
| Hybrid DL | 0.71 (0.66, 0.75) |
| African American | |
| TC | 0.45 (0.21, 0.66) |
| RF-LR | 0.58 (0.33, 0.81) |
| Image-only DL | 0.69 (0.55, 0.92) |
| Hybrid DL | 0.71 (0.55, 0.89) |

Note.—Data in parentheses are 95% confidence intervals. In the 3157 patients who were white, there were 7107 examinations and 233 cancers; in the 202 patients who were African American, there were 424 examinations and 11 cancers. AUC = area under receiver operator characteristic curve, DL = deep learning, RF-LR = risk-factor-based logistic regression, TC = Tyrer-Cuzick.

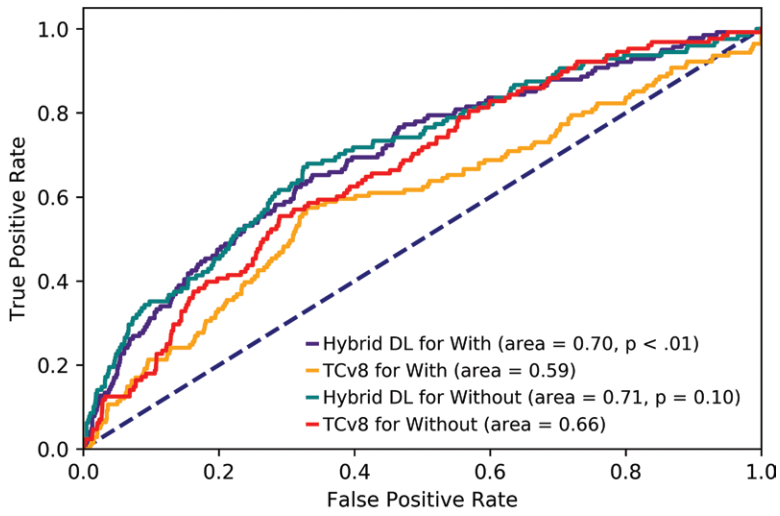
Table E3 [online]). The improvement of hybrid DL over TC was significant ($P < .01$). For patients without a family history of breast or ovarian cancer, hybrid DL and image-only DL showed similar discrimination accuracies (AUCs, 0.71 [95% CI: 0.65, 0.77] and 0.71 [95% CI: 0.66, 0.77] respectively), and compared with an AUC of 0.66 (95% CI: 0.60, 0.73) at TC. The improvement of hybrid DL and image-only DL over TC was




a.




b.




c.

**Figure 3:** Receiver operating characteristic curve for Tyrer-Cuzick version 8 (TCv8) and hybrid deep learning (DL) for different subgroups of patients: **(a)** patients who are white and African American, **(b)** pre- and postmenopausal women, and **(c)** women with and without any family history of breast or ovarian cancer. All *P* values are relative to TCv8 for the same subgroup.

not significant for women without a family history (hybrid DL vs TC, $P = .08$; image-only DL vs TC, $P = .10$).

*Assessing the risk of breast cancer 3–5 years after mammography.*—To distinguish a model's ability to predict future cancer development from its ability to detect cancers on the basis of the current mammography, we compared models on a subgroup of the test set by excluding mammography from women in whom cancer was diagnosed in less than 3 years. We observed that our models showed similar performance when predicting future risk (image-only DL and hybrid DL AUCs, 0.68 [95% CI: 0.63, 0.73] and 0.72 [95% CI: 0.67, 0.78], respectively; Table E4 [online]). This suggested that our image-based models were able to learn features associated with long-term risk and did not only perform early detection. Moreover, RF-LR, image-only DL, and hybrid DL ($P < .01$, $< .01$, and $P < .001$, respectively) significantly outperformed TC (AUC, 0.60; 95% CI: 0.54, 0.67).

## Confusion Matrix Analysis

*Hybrid DL versus breast density.*—When examining different combinations of density category and hybrid DL risk category, we observed that a patient's risk assessed at hybrid DL was more informative than their breast density category (Fig 4). For example, patients who were assessed as low risk but had dense breasts had a low incidence (1.4%; 23 of 1634) but patients who were assessed as high risk and had nondense breasts had a high incidence (5.5%; 123 of 2250). The cancer incidence substantially changed by column (ie, hybrid DL assessment) and not by row (ie, breast density).

*Hybrid DL versus TC.*—By examining different combinations of hybrid DL risk thirds and TC risk thirds, we observed the same findings: hybrid DL was more informative than was TC (Fig 5). By observing disagreements, hybrid DL was more accurate. For example, patients who were assessed as high risk by TC but assessed as low risk by hybrid DL had a low incidence of cancer (1.6%; eight of 516), whereas patients who were assessed as low risk by TC and high risk by hybrid DL had a high incidence of cancer (3.7%; 18 of 492).

# Discussion

We developed a deep learning (DL) model (hybrid DL) that used full-field mammograms in addition to traditional risk factor information to assess breast cancer risk. Hybrid DL was significantly more

**Table 4: Risk Test Set for All 5-year Risk Assessment Models by Menopausal Status**

| Parameter | AUC |
|---|---|
| Premenopausal patients | |
| TC | 0.73 (0.57, 0.90) |
| RF-LR | 0.71 (0.59, 0.85) |
| Image-only DL | 0.72 (0.57, 0.92) |
| Hybrid DL | 0.79 (0.67, 0.97) |
| Postmenopausal patients | |
| TC | 0.58 (0.53, 0.64) |
| RF-LR | 0.64 (0.58, 0.70) |
| Image-only DL | 0.69 (0.65, 0.73) |
| Hybrid DL | 0.70 (0.65, 0.75) |

Note.—Data in parentheses are 95% confidence intervals. There were 3095 premenopausal patients, in whom there were 1649 examinations and 62 cancers. There were 2513 postmenopausal patients, in whom there were 5656 examinations and 207 cancers. AUC = area under receiver operator characteristic curve, DL = deep learning, RF-LR = risk-factor-based logistic regression, TC = Tyrer-Cuzick.

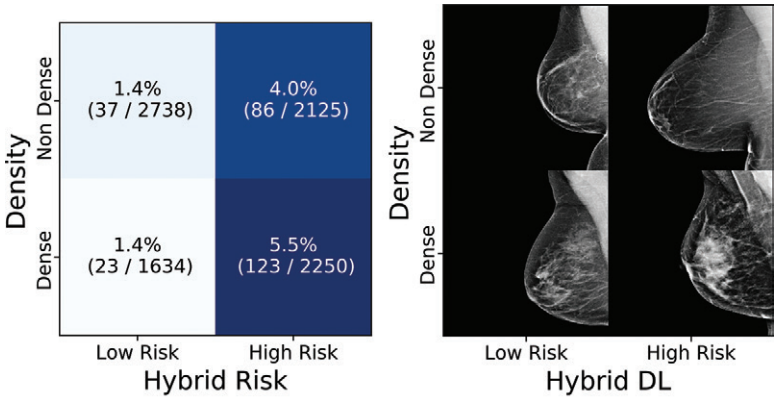



**Figure 4:** Cancer incidences partitioned by density value and hybrid deep learning (DL) risk assessment. **(a)** Each tile shows the percent and numerators/denominators of women with examinations within a specific density and risk group who developed cancer within 5 years. **(b)** Examples of screenings, sampled randomly from all examinations in that group.

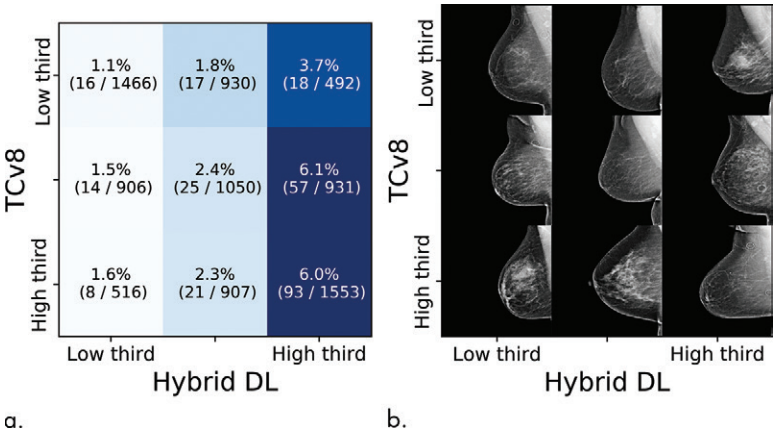



**Figure 5:** Cancer incidences partitioned by Tyrer-Cuzick risk assessment model (TCv8) and hybrid deep learning (DL) risk assessment. **(a)** Each tile shows the percent and numerators/denominators of women with examinations within a specific risk range that developed cancer within 5 years. **(b)** Examples of screenings, sampled randomly from all examinations in that group.

accurate than the Tyrer-Cuzick model (TC), a model used in clinical practice (area under the receiver operating characteristic curve [AUC], 0.70 vs 0.62, respectively). This improved AUC indicated that hybrid DL was better at identifying high-risk cohorts: hybrid DL placed 31.2% (84 of 269) of patients with cancer within the top risk decile versus TC, which placed 18.2% (49 of 269) of patients with cancer within the top risk decile.

The majority of existing risk models were developed on predominantly white populations (1,3,4) and have known limitations in predicting risk for other racial groups (17–20). Our hybrid DL model outperformed TC in both white and African American populations; this performance gap was especially pronounced for African American women, in whom TC obtained an AUC that was lower than that of hybrid DL (AUC, 0.45 vs 0.71, respectively). Moreover, hybrid DL was more accurate than TC in other subgroups (eg, women with a family history of breast or ovarian cancer and postmenopausal women). We found that in cases in which hybrid DL disagreed with TC on the risk of a patient, hybrid DL was more accurate.

Whereas hybrid DL was the best model overall, our DL model on the basis of mammograms alone (ie, image-only DL) also out-performed TC and it provided accurate risk assessment when traditional risk factor information was unavailable. This can be especially beneficial to patients who do not know their family history of breast or ovarian cancer. In addition, image-only DL risk assessment could be rapidly implemented into breast imaging screening programs, with patient risk automatically assessed from the mammogram alone. With current breast density legislation in 37 U.S. states, almost half of all women screened are told that they are at increased risk of breast cancer on the basis of their dense breast tissue. Although well intentioned, sharing dense breast tissue as an indicator of higher risk can lead many women to understandably believe that they are at high risk. At the same time, this practice can mislead women who do not have dense breast tissue to believe they are not at increased risk for breast cancer. Image-only DL would provide more precise information to help inform decisions regarding supplemental imaging and prevention strategies at the individual level. For centers equipped to collect additional patient information, the hybrid DL risk model could be used.

Our results demonstrated that full-field images and traditional risk factors contain complementary information, as illustrated by the AUC improvement of hybrid DL over image-only DL and a logistic regression model that used only traditional risk factor information. In future work, we will explore which risk factors are subsumed by the image and which are complementary. Because hybrid DL incorporated information from heterogeneous sources, we also hope that this approach will scale to incorporate other rich sources of information, such as large gene panels.

It will be important to investigate what kind of imaging patterns hybrid DL relies on to predict cancer risk. When we observed mammography from the cases in which hybrid DL and TC disagreed on the risk of a patient, we found that the model was not relying on a simple density measurement to determine risk. We speculate that the model may rely on different fine-grain tissue patterns and relative orientations of those patterns depending on global patterns in a patient's breast, and that there are distinguishing patterns for both women with dense and nondense breasts. Whereas methods exist (21–24) for obtaining saliency maps at the instance level (ie, an explanation specific to an individual mammogram), further work will be required to obtain the patterns that are most informative across the entire test set.

Our study had limitations. We used patient data from a single tertiary academic institution and mammograms captured by using a single vendor (Hologic). Also, some patients were missing risk factor information, though this limitation is common in both clinical practice and previous studies (1,3–5).

In conclusion, a deep learning (DL) model that directly leverages full-field mammograms in addition to traditional risk factors outperforms the Tyrer-Cuzick model (version 8) by a large margin; this improvement is consistent across demographic subgroups. These results support the hypothesis that mammography contains informative indicators of risk not captured by traditional risk factors, and DL models can deduce these patterns from the data. These models have the potential to replace conventional risk prediction models. Further research is required to validate our model across institutions and vendors before it can be broadly implemented, and to this end, we made our trained model and code available for research (*learningtocure.csail.mit.edu*).

**Acknowledgments:** The authors are grateful to Thomas Schultz, BS (Partners Enterprise Medical Imaging, Boston, Mass), Matthew Melesky (MagView Health care Information Solutions), and the members of the Partners Enterprise Medical Imaging team for their support of this project in image and data management.

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

**Disclosures of Conflicts of Interest:** **A.Y.** Activities related to the present article: MIT and MGH have patents filed on the deep learning models. 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: MIT and MGH have patents filed on the deep learning models. Activities not related to the present article: disclosed money paid to author for consultancy from GE Healthcare; disclosed money to author's institution for grants/grants pending from GE Healthcare; disclosed that MIT and MGH have patents filed related to the work. Other relationships: disclosed no relevant relationships. **T.S.** Activities related to the present article: MIT and MGH have patents filed on the deep learning models. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **T.P.** Activities related to the present article: MIT and MGH have patents filed on the deep learning models. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **R.B.** Activities related to the present article: MIT and MGH have patents filed on the deep learning models; disclosed money to author's institution for grant from Breast Cancer Alliance, Susan G. Komen. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

## References

1. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81(24):1879–1886.
2. Ross RK, Paganini-Hill A, Wan PC, Pike MC. Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. J Natl Cancer Inst 2000;92(4):328–332.
3. Claus EB, Risch N, Thompson WD. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. Breast Cancer Res Treat 1993;28(2):115–120.
4. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004;23(7):1111–1130.
5. Tice JA, Cummings SR, Ziv E, Kerlikowske K. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. Breast Cancer Res Treat 2005;94(2):115–122.
6. Reiner AS, Sisti J, John EM, et al. Breast cancer family history and contralateral breast cancer risk in young women: an update from the women's environmental cancer and radiation epidemiology study. J Clin Oncol 2018;36(15):1513–1520.
7. Collins A, Politopoulos I. The genetics of breast cancer: risk factors for disease. Appl Clin Genet 2011;4:11–19.
8. Brentnall AR, Harkness EF, Astley SM, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. Breast Cancer Res 2015;17(1):147.
9. Sprague BL, Conant EF, Onega T, et al. Variation in mammographic breast density assessments among radiologists in clinical practice: a multicenter observational study. Ann Intern Med 2016;165(7):457–464.
10. Lehman CD, Yala A, Schuster T, et al. Mammographic breast density assessment using deep learning: clinical implementation. Radiology 2019;290(1):52–58.
11. Boyd NF, Byng JW, Jong RA, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. J Natl Cancer Inst 1995;87(9):670–675.
12. Brandt KR, Scott CG, Ma L, et al. Comparison of clinical and automated breast density measurements: implications for risk prediction and supplemental screening. Radiology 2016;279(3):710–719.
13. 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.
14. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.
15. 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.
16. Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc Series B Stat Methodol 2007;69(3):369–390.
17. Boggs DA, Rosenberg L, Adams-Campbell LL, Palmer JR. Prospective approach to breast cancer risk prediction in African American women: the black women's health study model. J Clin Oncol 2015;33(9):1038–1044.
18. Gail MH. Twenty-five years of breast cancer risk models and their applications. J Natl Cancer Inst 2015;107(5):djv042.
19. Gail MH, Costantino JP, Pee D, et al. Projecting individualized absolute invasive breast cancer risk in African American women. J Natl Cancer Inst 2007;99(23):1782–1792.
20. Matsuno RK, Costantino JP, Ziegler RG, et al. Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. J Natl Cancer Inst 2011;103(12):951–961.
21. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365. https://arxiv.org/abs/1703.01365. Published March 4, 2017. Accessed April 4, 2019.
22. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; 1135–1144.
23. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), 2017; 618–626.
24. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. arXiv preprint arXiv:1704.02685. https://arxiv.org/abs/1704.02685. Published April 10, 2017. Accessed April 4, 2019.