CANCER

Toward robust mammography-based models for breast cancer risk

Adam Yala[1,2]*, Peter G. Mikhael[1,2], Fredrik Strand[3,4], Gigin Lin[5], Kevin Smith[6,7], Yung-Liang Wan[5], Leslie Lamb[8], Kevin Hughes[9], Constance Lehman[8†], Regina Barzilay[1,2†]

Improved breast cancer risk models enable targeted screening strategies that achieve earlier detection and less screening harm than existing guidelines. To bring deep learning risk models to clinical practice, we need to further refine their accuracy, validate them across diverse populations, and demonstrate their potential to improve clinical workflows. We developed Mirai, a mammography-based deep learning model designed to predict risk at multiple timepoints, leverage potentially missing risk factor information, and produce predictions that are consistent across mammography machines. Mirai was trained on a large dataset from Massachusetts General Hospital (MGH) in the United States and tested on held-out test sets from MGH, Karolinska University Hospital in Sweden, and Chang Gung Memorial Hospital (CGMH) in Taiwan, obtaining C-indices of 0.76 (95% confidence interval, 0.74 to 0.80), 0.81 (0.79 to 0.82), and 0.79 (0.79 to 0.83), respectively. Mirai obtained significantly higher 5-year ROC AUCs than the Tyrer-Cuzick model ($P < 0.001$) and prior deep learning models Hybrid DL ($P < 0.001$) and Image-Only DL ($P < 0.001$), trained on the same dataset. Mirai more accurately identified high-risk patients than prior methods across all datasets. On the MGH test set, 41.5% (34.4 to 48.5) of patients who would develop cancer within 5 years were identified as high risk, compared with 36.1% (29.1 to 42.9) by Hybrid DL ($P = 0.02$) and 22.9% (15.9 to 29.6) by the Tyrer-Cuzick model ($P < 0.001$).

INTRODUCTION

It is estimated that 39 million mammograms are performed in the United States every year (*1*, *2*), with $1.1 billion dollars being spent by Medicare alone (*3*). Despite the wide adoption of breast cancer screening, the practice is riddled with controversy. Proponents of more aggressive screening strategies aim to maximize the benefits of early detection (*4*–*9*), whereas advocates of less frequent screening aim to reduce the false-positive assessments, anxiety, and costs for the patients who will never develop breast cancer (*10*–*14*). As a result, in the United States, there are multiple guidelines with different recommendations about when to start screening, how often to get screened, and when supplemental screening is needed (*15*–*20*). We argue that both goals of earlier detection and reducing overtreatment can be achieved by leveraging more accurate risk models. With improved risk-based guidelines, we can offer more sensitive screening to patients who will develop cancer, achieving earlier detection while reducing unnecessary screening and overtreatment for the rest. Moreover, because of the scale of breast cancer screening, even modest improvements in screening guidelines have the potential to benefit a wide patient population.

All guidelines currently in clinical use leverage risk models. Some guidelines (*19*) use risk models as simple as a patient's age to determine whether, and how often, a woman should get screened, whereas others (*16*) combine multiple factors relating to age, hormonal factors, genetics, and mammographic breast density to determine whether supplemental imaging should be considered. However, despite decades of effort, the accuracy of risk models used in clinical practice remains modest. For instance, the Tyrer-Cuzick (*21*) and Gail (*22*) models achieved areas under the curve (AUCs) of 0.62 and 0.59, respectively, in a prospective UK screening cohort (*23*). Recently, image-based deep learning models have shown considerable promise (*24*, *25*), obtaining AUCs up to 0.70 for assessing 5-year risk and advancing the state of the art. However, to bring an image-based risk model to the clinic, we not only need to further improve its accuracy but must also validate its performance at scale across diverse populations and clinical settings. Furthermore, we need to demonstrate that it can identify more accurate high-risk cohorts. Here, we aimed to achieve all three of these goals by developing Mirai and studying its performance across multiple populations.

RESULTS

Overview of algorithm

In computational terms, risk assessment can be viewed as a prediction task, where the model is trained to associate features of mammograms with future cancer diagnoses. Although this setup, referred to as supervised learning, is commonly used for medical tasks (*26*–*30*), risk modeling also poses several unique requirements. It requires risk prediction at various time points, the ability to leverage potentially missing nonimage data (such as age and family history), and consistent performance across heterogeneous mammography devices.

Inherent to risk modeling is learning from patients with variable amounts of follow-up and needing to assess risk at different time

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Jameel Clinic, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Breast Radiology Unit, Department of Imaging and Physiology, Karolinska University Hospital, 17164 Solna, Sweden. [4]Department of Oncology-Pathology, Karolinska Institute, 17164 Solna, Sweden. [5]Department of Medical Imaging and Intervention, Chang Gung Memorial Hospital at Linkou, Taoyuan 333, Taiwan. [6]School of Electrical Engineering and Computer, KTH Royal Institute of Technology, 10044 Stockholm, Sweden. [7]Science for Life Laboratory, 17165 Solna, Sweden. [8]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA. [9]Division of Surgical Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA.
*Corresponding author. Email: adamyala@csail.mit.edu
†Joint senior authors.

points. Although it is possible to train separate models to assess risk for each time point based on patients with the corresponding amount of follow-up (1 to 5 years), this approach can result in mutually inconsistent risk assessments. For instance, a model could predict that a patient has a higher risk of developing cancer within 2 years than within 5 years. Moreover, this approach does not leverage the inherent relationship between assessing risk at different time points. We address this by training a single model to predict risk at all time points and by explicitly designing the architecture to produce self-consistent predictions. This formulation also enables the model to learn from data with variable amounts of follow-up.

Although our method primarily focuses on mammograms, we also wanted to leverage nonimage risk factors (for example, age and hormonal factors) if they were available. An obvious mechanism for incorporating nonimage risk factors is to add them as an input to the model jointly with the image. However, this design would prevent hospitals that do not collect this kind of information from using the model. Although we could impute this missing information by using a reference population, that would not take into account the relationship between the mammogram and the risk factors. To address this challenge, we trained our model to predict risk factor values from the mammogram, enriching our original objective with this new prediction task. This formulation enabled the model to benefit from available risk factor data while allowing it to impute the information if it is missing.

To incorporate deep learning risk models into clinical guidelines, the models must be consistent across a range of mammography devices, in other words, they must predict the same risk for a patient regardless of the mammography device. We addressed this challenge by adopting a conditional-adversarial training scheme (*31*). This training regime forces the model to induce image representation in a device-invariant fashion and to produce consistent risk assessments.

Our full model, named Mirai, is depicted in Fig. 1. It takes as input all standard views of a mammogram: left craniocaudal (L CC), left mediolateral-oblique (L MLO), right craniocaudal (R CC), and right mediolateral-oblique (R MLO). Mirai consists of four modules: an image encoder, an image aggregator, a risk factor predictor, and an additive-hazard layer. A run through the model works as follows: first, we pass each mammogram view independently through the image encoder. Next, we take each image representation as well as which view it came from (for example, L CC and R MLO), and pass it into the image aggregation module to combine information across views and obtain a representation of the entire mammogram. Given this rich representation of the mammogram, we then predict a patient's traditional risk factors as used in Tyrer-Cuzick (such as age, weight, and hormonal factors) and refer to this as our risk factor prediction module. If risk factor information is not available at inference time, we then use the predicted values. Next, we take the mammogram representation from our image aggregator, combined with our risk factor information (predicted or given), and predict a patient's risk with an additive-hazard layer. The additive-hazard layer predicts a patient's risk for each year over the next 5 years. Architectural details for each module are presented in the Supplementary Materials and Methods, and all code is released.

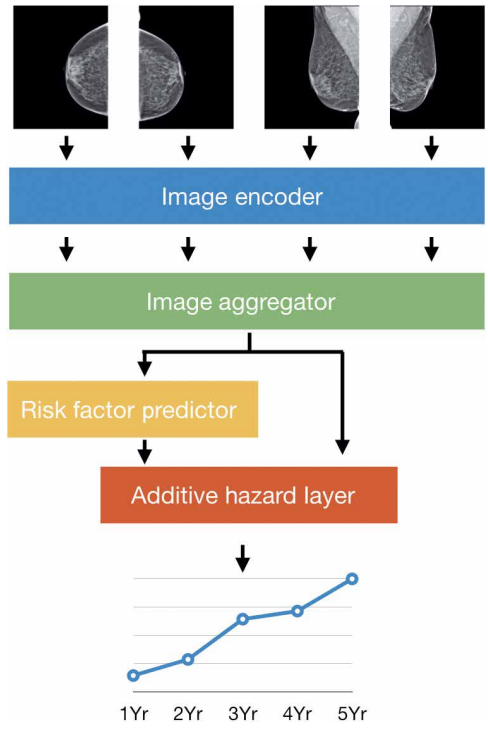

Fig. 1. Schematic description of Mirai. The four standard views of an individual mammogram were fed into Mirai. The image encoder mapped each view to a vector, and the image aggregator combined the four view vectors into a single vector for the mammogram. In this work, we used a single shared ResNet-18 as an image encoder, and a transformer as our image aggregator. The risk factor predictor module predicted all the risk factors used in the Tyrer-Cuzick model, including age, detailed family history, and hormonal factors, from the mammogram vector. The additive hazard layer combined information from both the image aggregator and risk factors (predicted or given) to predict coherent risk assessments across 5 years (Yr).

Training and testing at MGH

We developed Mirai using the Massachusetts General Hospital (MGH) dataset, which consists of 210,819, 25,644, and 25,855 examinations from 56,786, 7020, and 7005 patients, for the training, validation, and test sets, respectively. This dataset contained detailed risk factor information, as used in Tyrer-Cuzick version 8 (TCv8), that was available at the time of mammography. The distribution of clinical risk factors in the MGH dataset, as used by TCv8, is shown in table S1. A flowchart illustrating the construction on the MGH dataset is shown in Fig. 2.

To determine the impact of using predicted risk factors on Mirai's performance, we evaluated the model both when using the electronic health record-based and predicted risk factors, referring to the two scenarios as "Mirai with risk factors" and "Mirai without risk factors," respectively. We compared Mirai against three alternative risk models: Hybrid DL (*25*), Image-Only DL (*25*), and TCv8. Hybrid DL is a deep learning model based on both mammograms and traditional risk factors, and Image-Only DL is a deep learning model based only on mammograms. Hybrid DL requires traditional risk factors to predict risk, whereas Image-Only DL does not use such information. We note that Hybrid DL and Image-Only DL were both developed using the same MGH dataset as Mirai, and so, differences in performance can only be attributed to the algorithm design. Image-Only DL is equivalent to the image encoder component of Mirai trained by itself as a 5-year risk classifier. TCv8 is a traditional risk model that combines a variety of risk factors including age, family history, and hormonal factors and is a current clinical standard. We obtained TCv8 risk assessments using the Command-Line version of the IBIS Breast Cancer Risk Evaluation tool (version 8).

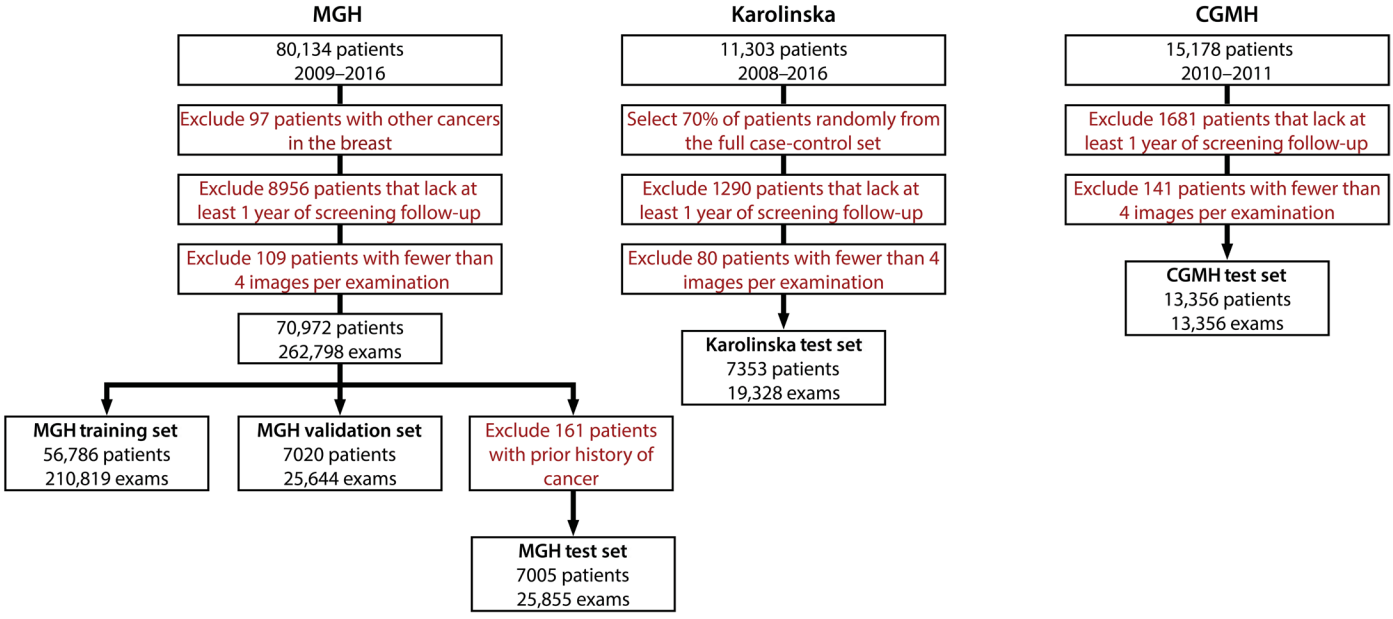

Fig. 2. Dataset construction flowchart. Shown are the MGH (left), Karolinska (middle), and CGMH (right) datasets.

To better investigate the connection between risk estimation and cancer detection, we also compared Mirai with retrospective radiologist BI-RADS (Breast Imaging-Reporting and Data System) assessments and a recently proposed cancer detection model, Image-and-Heatmaps (*32*), on the MGH test set. Image-and-Heatmaps is a convolutional neural network trained on a large dataset from New York University (NYU) using both pixel-level and whole-image annotations to predict cancer within 120 days. We obtained Image-and-Heatmaps cancer predictions using their publicly available GitHub (*33*) and did not use test-time data augmentations or model ensembling.

On the 25,855 examinations (588 positive) in the MGH test set, Mirai with and without risk factors obtained C-indices of 0.76 (0.74 to 0.80) and 0.75 (0.72 to 0.78) compared with C-indices of 0.72 (0.69 to 0.75), 0.72 (0.69 to 0.75), and 0.64 (0.60 to 0.67) by Hybrid DL, Image-Only DL, and TCv8, respectively. The full results on the MGH dataset are summarized in Table 1, and receiver operating characteristic (ROC) curves for each time point are shown in Fig. 3. Mirai with risk factors had a significantly higher 5-year AUC than Hybrid DL, Image-Only DL, and TCv8 with P values of <0.001, <0.001, and <0.001, respectively. Mirai with risk factors did not have a significantly higher 5-year AUC than Mirai without risk factors ($P = 0.27$). We also present an analysis of model performance excluding cancers identified within 6 months of the screening mammogram, resulting in 25,708 examinations (441 positive) (table S2). In this setting, Mirai with risk factors had a significantly higher 5-year AUC than Hybrid DL, Image-Only DL, and TCv8, with P values of <0.001, 0.02, and <0.001, respectively, and did not have a significantly higher 5-year AUC than Mirai without risk factors ($P = 0.27$). We also evaluated the performance of radiologist BI-RADS assessments and Image-and-Heatmaps (*32*) in Table 1. Radiologists obtained ROC AUCs of 0.92 (0.90 to 0.95) and 0.75 (0.72 to 0.78) at 1 and 2 years, respectively, compared with 0.84 (0.81 to 0.88) and 0.80 (0.76 to 0.83) by Mirai. We found that Image-and-Heatmaps obtained a 1-year AUC of 0.78 (0.73 to 0.82) and a C-index of 0.68 (0.65 to 0.72).

We performed an ablation study of Mirai to investigate the effects of different design choices on overall performance and mammography device bias (table S3 and fig. S1). To evaluate the mammography device bias of a risk model, we trained a classifier to predict which machine was used to acquire a mammogram from the model's corresponding risk assessment and measured the AUC of this device-identity classifier on the MGH test set. We found that an ablation of Mirai without risk factors that removed conditional adversarial training obtained a device-identity AUC of 0.76 (0.75 to 0.76), reflecting large device bias. With the addition of conditional adversarial training, Mirai without risk factors obtained a device-identity AUC of 0.50 (0.50 to 0.50), effectively removing the bias. We evaluated the saliency of each risk factor in Mirai's predictions across the MGH test set in fig. S2. The most important risk factors were a patient's *BRCA* status, if they had any family history (binary family history), and if they had had any children (parous), with average saliency scores of 0.07 (0.07, 0.07), 0.04 (0.04, 0.04), and 0.03 (0.03, 0.03), respectively. In contrast, mammograms had an average saliency score of 2.19 (2.17, 2.22). We note that the mammogram obtained a 30-fold higher saliency score than the most important clinical factor, *BRCA* status. This finding is consistent both with the reported performance of Mirai with and without risk factors shown in Table 1 and the result that Mirai with risk factors did not obtain a significantly higher 5-year AUC than Mirai without risk factors ($P = 0.27$).

Generalization to additional populations

For Mirai to be useful to the larger community, it must be validated in a diverse set of clinical environments and patient populations. To this end, we tested the model on a dataset from the Karolinska University Hospital in Sweden consisting of 19,328 examinations (1413 positive) from 7353 patients and a dataset from the Chang Gung Memorial Hospital (CGMH) in Taiwan consisting of 13,356 examinations (244 positive) from 13,356 patients. A dataset construction flowchart for both datasets is shown in Fig. 2. Traditional risk factors were not available in either dataset. As a result, we tested

Table 1. ROC AUCs and C-indices for Mirai and prior risk models on all test sets. On the MGH test set, we also evaluated Image-and-Heatmaps (*32*) and radiologist BI-RADS assessments. All metrics are followed by their 95% confidence interval.

Model	Use risk factors	C-index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
MGH test set: 25,855 exams, 558 followed by cancer diagnosis							
Tyrer-Cuzick Version 8 (TCv8) (*21*)	No	0.64 (0.60–0.67)	0.66 (0.61–0.71)	0.65 (0.61–0.69)	0.64 (0.60–0.68)	0.63 (0.59–0.67)	0.62 (0.59–0.66)
Radiologist BI-RADS	NA	0.67 (0.65–0.70)	0.92 (0.90–0.95)	0.75 (0.72–0.78)	0.68 (0.65–0.70)	0.64 (0.62–0.67)	0.62 (0.60–0.65)
Image-and-Heatmaps (*32*)	No	0.68 (0.65–0.72)	0.78 (0.73–0.82)	0.73 (0.70–0.77)	0.69 (0.66–0.73)	0.67 (0.63–0.70)	0.64 (0.60–0.68)
Image-Only DL (*25*)	No	0.72 (0.69–0.75)	0.79 (0.75–0.83)	0.75 (0.71–0.78)	0.73 (0.70–0.77)	0.73 (0.70–0.76)	0.73 (0.70–0.77)
Hybrid DL (*25*)	Yes	0.72 (0.69–0.75)	0.78 (0.75–0.82)	0.74 (0.71–0.78)	0.72 (0.68–0.75)	0.72 (0.68–0.75)	0.72 (0.69–0.76)
Mirai (ours)	No	0.75 (0.72–0.78)	0.84 (0.80–0.87)	0.78 (0.75–0.82)	0.77 (0.74–0.80)	0.76 (0.73–0.79)	0.76 (0.73–0.79)
	Yes	0.76 (0.74–0.80)	0.84 (0.81–0.88)	0.80 (0.76–0.83)	0.78 (0.75–0.81)	0.76 (0.73–0.80)	0.76 (0.73–0.80)
Karolinska test set: 19,328 examinations, 1413 followed by cancer diagnosis							
Image-Only DL (*25*)	No	0.75 (0.73–0.77)	0.83 (0.81–0.86)	0.79 (0.77–0.81)	0.75 (0.73–0.77)	0.73 (0.71–0.75)	0.71 (0.69–0.73)
Mirai (ours)	No	0.81 (0.79–0.82)	0.90 (0.89–0.92)	0.86 (0.84–0.88)	0.82 (0.80–0.84)	0.80 (0.79–0.82)	0.78 (0.76–0.80)
CGMH test set: 13,356 examinations, 244 followed by cancer diagnosis							
Image-Only DL (*25*)	No	0.70 (0.66–0.74)	0.80 (0.75–0.64)	0.76 (0.71–0.80)	0.72 (0.67–0.76)	0.71 (0.67–0.75)	0.70 (0.66–0.73)
Mirai (ours)	No	0.79 (0.76–0.83)	0.90 (0.87–0.93)	0.86 (0.83–0.90)	0.82 (0.78–0.85)	0.80 (0.77–0.84)	0.79 (0.75–0.82)

Mirai (without risk factors) and Image-Only DL but not TCv8 or Hybrid DL, which require risk factors.

On the Karolinska dataset, Mirai obtained a C-index of 0.81 (0.79 to 0.82) compared with a C-index of 0.75 (0.73 to 0.77) by Image-Only DL. Mirai performed similarly on the CGMH test set, obtaining a C-index of 0.79 (0.76 and 0.83) compared with a C-index of 0.70 (0.66 and 0.74) by Image-Only DL. The full results on the Karolinska and CGMH test sets are summarized in Table 1, and ROC curves for each time point are displayed in Fig. 3. In both Karolinska and CGMH, Mirai had a significantly higher 5-year AUC than Image-Only DL with P values of <0.001 and <0.001, respectively. We note that Mirai obtained similar 5-year AUCs across all test sets, achieving AUCs of 0.76 (0.73 to 0.80), 0.78 (0.76 to 0.80), and 0.79 (0.75 to 0.82) for the MGH, Karolinska, and CGMH test sets, respectively. We also present an analysis excluding cancers identified within 6 months of the screening mammogram in table S2. In this setting, Mirai had a significantly higher 5-year AUC than Image-Only, with P values of <0.001 and <0.001 on the Karolinska and CGMH test sets, respectively.

Subgroup analysis

We also validated all risk models for different clinical subgroups of interest. In the MGH test set, we computed model C-indices for patients of different races (White, African American, and Asian American), different age groups, different density categories, and different mammography devices. We found that Mirai performed similarly across all groups. This information is available in table S4. We note that the C-indices for Mirai with risk factors for White, Asian American, and African American patients were 0.75 (0.72 to 0.78), 0.80 (0.68 to 0.95), and 0.71 (0.55 to 0.90), respectively, compared with 0.64 (0.60 to 0.68), 0.54 (0.36 to 0.75), and 0.62 (0.44 to 0.84) for TCv8. The consistent performance for Asian Americans is further supported by the C-index of 0.79 (0.76 to 0.83) in the CGMH dataset. In the Karolinska dataset, we computed Mirai C-indices by future cancer subtype (invasive, HER2 status, and so on) in table S5. The distribution of cancer subtypes is reported in table S6. We found that Mirai obtained similar C-indices across different subtypes, which is further supported by a t-SNE (t-distributed stochastic neighbor embedding) (*34*) analysis (fig. S3) showing that the model learns similar representations for mammograms regardless of the subtype of the future cancer.

Identifying high-risk cohorts

Our next objective was to investigate whether improved risk models can advance early detection. A wide range of guidelines already exist to offer either supplemental screening (*15*–*20*, *35*) or chemoprevention (*36*, *37*) for patients at high risk of future cancer. To improve these guidelines, it is necessary to improve our definitions of who is at "high risk." To this end, we evaluated the ability of different risk models to identify high-risk patients. We restricted our analysis to patients in the MGH, Karolinska, and CGMH test sets who were screening negative and had either cancer within 5 years or had 5 years of negative follow-up. We did not have access to radiologist BI-RADS assessments for all datasets, so we defined a screening negative examination as not receiving a cancer diagnosis within 6 months.

The MGH, Karolinska, and CGMH 5-year cohorts had 3957, 5707, and 11,167 patients and consisted of 9284 examinations with 441 future cancers, 5707 examinations with 869 future cancers, and 11,167 examinations with 139 future cancers, respectively. Intuitively, we wanted a risk model to identify the most future cancers (high sensitivity) without directing unnecessary interventions to patients without future cancer (high specificity). We considered four possible methods of determining high-risk patients: Tyrer-Cuzick lifetime risk, Image-Only DL, Hybrid DL, and Mirai. For Tyrer-Cuzick lifetime risk, we used a high-risk threshold of 20%, which is used in current guidelines for supplemental screening by the American Cancer

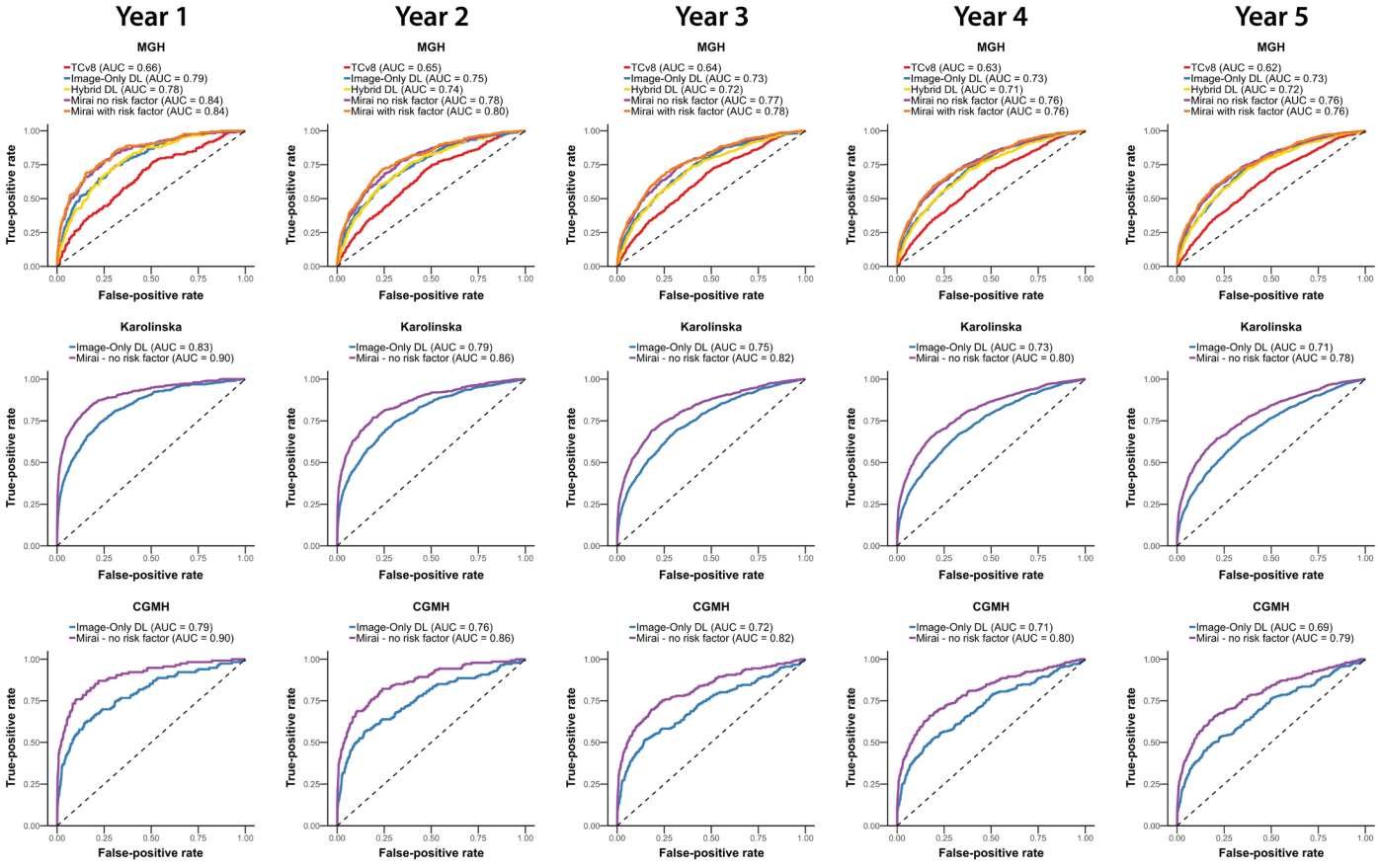

Fig. 3. ROCs for model predictions on MGH, Karolinska, and CGMH test sets. Results are shown in the top, middle, and bottom rows, respectively. The curves are arranged left to right from 1- to 5-year outcomes.

Society, the American College of Radiology, and the National Comprehensive Cancer Network (*19*, *20*, *35*). For Image-Only DL, Hybrid DL, and Mirai, we chose high-risk thresholds to match the specificity of Tyrer-Cuzick lifetime risk model on the MGH development set. On the Karolinska and CGMH 5-year cohorts, we evaluated the performance of Mirai and Image-Only DL using the thresholds computed on the MGH test. To enable more direct comparison, we also computed the performance of Image-Only DL when we chose a threshold to match Mirai's specificity on each dataset.

The full results of our analysis across the MGH, Karolinska, and CGMH test sets are in Table 2, and the ROC curves of all models in this setting are shown in Fig. 4. On the MGH 5-year cancer cohort, the Tyrer-Cuzick lifetime risk >20% guideline obtained a sensitivity and specificity of 22.9% (15.9 to 29.6) and 85.4% (84.1 to 86.6), respectively. Although obtaining similar specificity, Mirai with risk factors, Mirai without risk factors, Hybrid DL, and Image-Only DL obtained sensitivities of 41.5% (34.2 to 48.5), 39.7% (32.9 to 46.5), 36.1% (29.1 to 42.9), and 32.9% (26.1 to 39.4), respectively. Moreover, the sensitivity of Mirai with risk factors was significantly higher than that of Hybrid DL, Image-Only DL, and Tyrer-Cuzick lifetime risk, with P values of $P = 0.02$, $P < 0.001$, and $P < 0.001$, respectively. The sensitivity of Mirai with risk factors was not significantly higher than Mirai without risk factors ($P = 0.37$). We present a supplementary analysis of the MGH dataset in table S7, where we used radiologist BI-RADS assessments in determining who was screening negative. On the Karolinska and CGMH test sets, Mirai without risk factors obtained sensitivities of 26.0% (22.4 to 29.6) and 37.4% (29.3 to 45.5) and specificities of 93.1% (92.4 to 93.9) and 88.5% (88.0 to 89.2), respectively. We found that Image-Only DL performed poorly when using the risk threshold identified on the MGH test set. When calibrated to obtain the same specificities as Mirai, Image-Only DL obtained sensitivities of 18.9% (15.6 to 22.1) and 24.5% (16.9 to 31.3), respectively. Mirai obtained significantly higher sensitivities than Image-Only DL in both datasets ($P < 0.001$ and $P < 0.001$).

DISCUSSION

We developed a risk model, Mirai, to assess breast cancer risk from screening mammograms. Mirai demonstrated improved discriminatory capacity over the state-of-the-art clinically adopted Tyrer-Cuzick and prior deep learning approaches Hybrid DL and Image-Only DL. Moreover, we found that Mirai, which was trained at MGH, maintained its performance on datasets from both Karolinska in Sweden and CGMH in Taiwan without additional training. Externally validating our model across diverse clinical settings is especially important given recent negative findings for the generalization of other proposed mammography-based models for cancer risk (*38*). We evaluated Mirai across races, ages, and breast density categories in the MGH test set and across cancer subtypes on the Karolinska dataset and found that it performed similarly across all subgroups. We also demonstrated

Table 2. Sensitivity and specificity of different risk models in identifying high-risk cohorts. We excluded screening positive mammograms. In this analysis, we defined a screening positive mammogram as one followed by a cancer diagnosis within 6 months. Thresholds were chosen to match the specificity of the Tyrer-Cuzick lifetime risk on the MGH development set. Thresholds marked with * were chosen to best match the specificity of Mirai on the respective test set. All metrics are followed by their 95% confidence interval.

Method	Use risk factors	High risk threshold	Sensitivity	Specificity
Dataset	MGH: 9284 examinations from 3957 patients; 441 examinations followed by future cancer			
Tyrer-Cuzick lifetime risk	Yes	20%	22.9% (15.9–29.6)	85.4% (84.1–86.6)
Image-Only DL (*25*)	No	3.4%	32.9% (26.1–39.4)	85.9% (84.8–86.9)
Hybrid DL (*25*)	Yes	3.4%	36.1% (29.1–42.9)	86.0% (84.9–87.1)
Mirai 5-year risk	No	2.6%	39.7% (32.9–46.5)	85.2% (84.1–86.4)
	Yes	3.0%	41.5% (34.4–48.5)	85.6% (84.5–86.8)
Dataset	Karolinska: 7194 examinations from 5707 patients; 869 examinations followed by future cancer			
Image-Only DL (*25*)	No	3.4%	0.6% (0.0–0.7)	99.9% (99.9–100.0)
		1.3%*	18.9% (15.6–22.1)	93.1% (92.4–93.8)
Mirai 5-year risk	No	2.6%	26.0% (22.4–29.6)	93.1% (92.4–93.9)
Dataset	CGMH: 11,167 examinations from 11,167 patients; 139 examinations followed by future cancer			
Image-Only DL (*25*)	No	3.4%	2.2% (0.7–4.3)	99.9% (99.9–100.0)
		1.2%*	24.5% (16.9–31.3)	88.5% (87.9–89.1)
Mirai 5-year risk	No	2.6%	37.4% (29.3–45.5)	88.5% (88.0–89.2)

how Mirai could be implemented in current clinical pipelines focused on identifying high-risk patients and showed that it improved over existing risk models such as Tyrer-Cuzick lifetime risk and Image-Only DL.

Risk models in clinical practice today, namely, breast density and traditional statistical models, are the foundation of current guidelines for personalized screening and prevention. For instance, breast density, which was described as early as 1967 (*39*), was the first to recognize that a woman's imaging could inform her future cancer risk; the DENSE trial (*40*) later showed that giving women with extremely dense breasts supplemental magnetic resonance imaging could substantially reduce interval cancers. Traditional statistical risk models, like the Gail and Tyrer-Cuzick models (*21*, *22*), have long recognized that combining multiple sources of information can yield better predictions, and they are the base of both current supplemental imaging and chemoprevention guidelines (*16*, *19*, *35*–*37*). Our research builds upon their seminal works. We hypothesize that developing more accurate risk models will enable further guideline personalization and thus lead to better outcomes.

The performance of Mirai can be attributed to how its design captures unique characteristics of breast cancer risk estimation. Specifically, the model architecture jointly reasons over both different views of the mammogram and multiple time points of risk assessment. Moreover, we demonstrated how to incorporate nonimage risk factors such as age or hormonal factors to further refine accuracy, while enabling the model to impute this information if it is not provided. Last, we used a conditional adversarial training regime to learn image representations that are device invariant.

Our work is also related to the large volume of work (*32*, *41*–*53*) focused on developing deep learning models for breast cancer detection. Although the tasks of cancer detection and future cancer risk are distinct, we hypothesize that some of the technical lessons from the two tasks can be complementary. For instance, we hypothesize that aggressive model ensembling strategies used by (*32*, *53*, *54*) and the use of detailed cancer region annotations could be used to improve image-based risk models. Moreover, we hypothesize that our mechanisms for predicting risk at multiple time points, optionally using risk factors, and learning representations that are invariant to

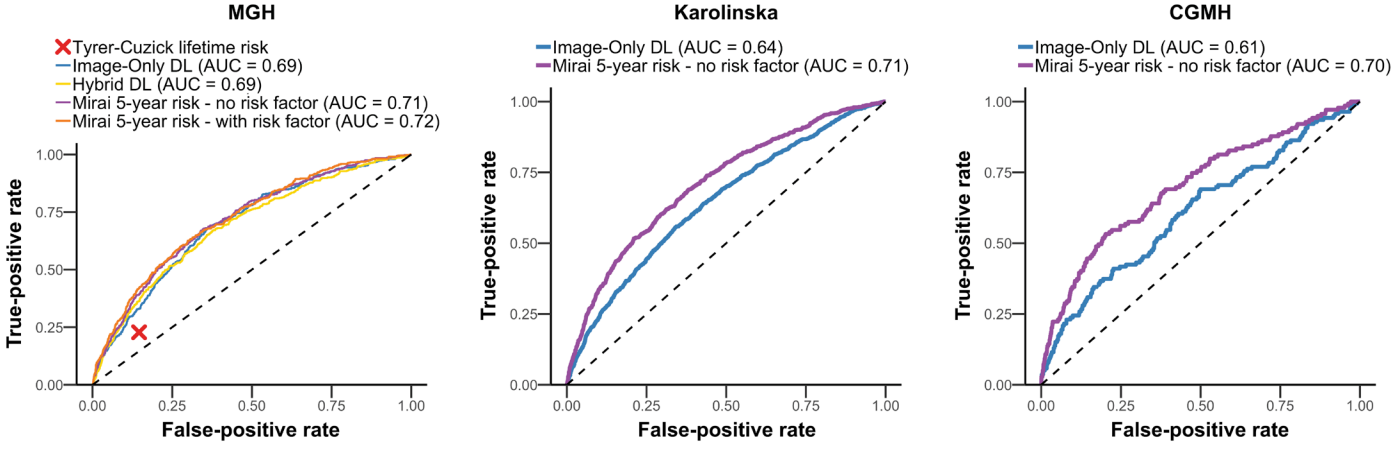

Fig. 4. ROCs of different risk models in identifying high-risk cohorts. MGH (left), Karolinska (middle), and CGMH (right) cohorts are shown. These datasets are restricted to include patients who were screening negative and had either cancer within 5 years or 5 years of negative follow-up. The orange and purple curves refer to Mirai with and without risk factors, respectively.

mammography machines could be used to improve the current state of the art in cancer detection systems.

Although Mirai can be tested as a cancer detection system, direct comparison to prior work in cancer detection is difficult due to a lack of publicly available code (*53*, *55*) and the lack of common benchmarks. Not directly comparable, we note that Mirai obtained a 1-year AUC of 0.90 on the Karolinska test set, similar to the top single-model AUC 0.90 on a separate Karolinska test set reported by (*54*). We also evaluated Image-and-Heatmaps (*32*), a recently proposed cancer detection model trained to predict cancer within 120 days, on a large dataset from NYU. Image-and-Heatmaps obtained a 120-day AUC of 0.89 on the NYU test set (*32*), and it obtained a 1-year AUC of 0.78 on the MGH test set. We note that it is difficult to compare this model with our own because of the difference in study objectives and training datasets. These results further highlight the importance of creating common benchmarks with standardized evaluation to enable direct comparison between models. We believe that sharing trained models is important for the continued development of cancer detection and risk assessment systems, and to this end, we are releasing our code and models for public research use.

There are multiple directions for future work that can further improve the accuracy and utilization of the imaging-based models for cancer risk. Although our model only considers a patient's current mammogram agnostic of previous imaging, it is known that changes in imaging over time contain a wealth of information. A natural next step is to develop methods that can effectively use a patient's full

Table 3. Detailed demographics for the MGH dataset. For each demographic, we report the number of corresponding mammography examinations and the percentage they constitute of the total. All cancer counts reflect cancer within 5 years.

	MGH training set		MGH validation set		MGH test set	
Characteristics	**All**	**Cancer**	**All**	**Cancer**	**All**	**Cancer**
All examinations	210,819 (100%)	5379 (100%)	25,644 (100%)	612 (100%)	25,855 (100%)	588 (100%)
Age						
<40	5,812 (2.8%)	84 (1.6%)	711 (2.8%)	7 (1.1%)	724 (2.8%)	7 (1.1%)
40–50	55,905 (26.5%)	1113 (20.7%)	6,821 (26.6%)	142 (23.2%)	7,025 (27.2%)	95 (16.2%)
50–60	63,314 (30.0%)	1348 (25.1%)	7,762 (30.3%)	166 (27.1%)	7,829 (30.3%)	188 (32.0%)
60–70	54,925 (26.1%)	1770 (32.9%)	6,674 (26.0%)	179 (29.3%)	6,708 (25.9%)	182 (31.0%)
70–80	25,401 (12.0%)	816 (15.2%)	3,037 (11.8%)	102 (16.7%)	3,001 (11.6%)	94 (16.0%)
>80	5,461 (2.6%)	248 (4.5%)	639 (2.5%)	16 (2.6%)	568 (2.2%)	22 (3.7%)
Density						
Almost entirely fatty	20,411 (9.7%)	315 (5.9%)	2,429 (9.5%)	53 (8.7%)	2,474 (9.6%)	31 (5.3%)
Scattered areas of fibroglandular tissue	102,112 (48.4%)	2623 (48.8%)	12,519 (48.8%)	261 (42.7%)	12,490 (48.3%)	264 (44.9%)
Heterogeneously dense	78,892 (37.4%)	2196 (40.8%)	9,461 (36.9%)	263 (43.0%)	9,751 (37.7%)	271 (46.1%)
Extremely dense	9,293 (4.4%)	242 (4.5%)	1,225 (4.8%)	35 (5.7%)	1,129 (4.4%)	22 (3.7%)
BI-RADS						
0–additional imaging needed	13,810 (6.6%)	1579 (29.4%)	1,686 (6.6%)	164 (26.8%)	1,785 (6.9%)	186 (31.6%)
1–negative or 2–benign	196,797 (93.3%)	3786 (70.4%)	23,932 (93.3%)	447 (73.0%)	24,043 (93.0%)	400 (68.0%)
Other	47 (0.02%)	9 (0.2%)	3 (0.01%)	1 (0.2%)	4 (0.01%)	1 (0.2%)
Race						
White	171,509 (81.4%)	4646 (86.4%)	20,710 (80.8%)	518 (84.6%)	21,006 (81.2%)	512 (87.1%)
African American	9,883 (4.7%)	209 (3.9%)	1,209 (4.7%)	26 (4.3%)	1,204 (4.7%)	21 (3.6%)
Asian or Pacific Islander	9,477 (4.5%)	160 (3.0%)	1,231 (4.8%)	17 (2.8%)	1,238 (4.8%)	26 (4.4%)
Hispanic	2,266 (1.1%)	63 (1.2%)	260 (1.0%)	5 (0.8%)	225 (0.9%)	6 (1.0%)
Other race	11,423 (5.4%)	138 (2.6%)	1,439 (5.6%)	20 (3.3%)	1,486 (5.7%)	15 (2.6%)
Device						
Lorad Selenia	81,106 (38.5%)	2009 (37.4%)	9,850 (38.4%)	216 (35.29%)	9,937 (38.4%)	241 (41.0%)
Selenia Dimensions	129,493 (61.4%)	3150 (58.6%)	15,767 (61.5%)	369 (60.29%)	15,882 (61.4%)	311 (52.9%)
Unknown	220 (0.1)%	220 (4.1%)	27 (0.1%)	27 (4.4%)	36 (0.1%)	36 (6.1%)

Table 4. Demographics of Karolinska and CGMH test sets. For each demographic, we report the number of corresponding mammography examinations and the percentage they constitute of the total. All cancer counts reflect cancer within 5 years.

	Karolinska dataset		Chang Gung Memorial dataset	
Characteristics	**All**	**Cancer**	**All**	**Cancer**
All examinations	19,328 (100%)	1413 (100%)	13,356 (100%)	244 (100%)
Age				
40–50	7,814 (40.4%)	364 (25.8%)	4,008 (30.0%)	74 (33.3%)
50–60	5,477 (28.3%)	387 (27.4%)	6,301 (47.2%)	115 (47.1%)
60–70	5,174 (26.8%)	563 (39.8%)	3,042 (22.8%)	55 (22.5%)
70–80	863 (4.4%)	99 (7.0%)	0 (0.0%)	0 (0.0%)
Density as assessed by deep learning model (*57*)				
Almost entirely fatty	933 (4.8%)	39 (2.7%)	51 (0.4%)	0 (0.0%)
Scattered areas of fibroglandular tissue	9,767 (50.5%)	682 (48.3%)	3,272 (24.5%)	40 (16.4%)
Heterogeneously dense	8,057 (41.7%)	655 (46.4%)	9,278 (69.5%)	194 (79.5%)
Extremely dense	571 (3.0%)	37 (2.6%)	755 (5.7%)	10 (4.1%)

history of imaging. In a similar fashion, expanding the model to use tomosynthesis is likely to yield further performance improvements. Beyond work in improving accuracy, additional research is required to determine how to adapt image-based risk models to different mammography devices across multiple vendors. Although our conditional adversarial training scheme enabled us to obtain consistent risk assessments across mammography devices where we have training data, we did not evaluate whether our models can generalize to unseen mammography devices. In addition, although our own evaluation focused on defining high-risk cohorts, other methods are required to design more fine-grained risk-based guidelines.

This study has limitations. Although our analysis showed Mirai obtained strong performance across different races, our datasets contained few African American and Hispanic women, making up 5 and 1% of the MGH test set, respectively. More work is needed to further validate the model in large Hispanic and African American screening populations. Moreover, prospective trials are necessary to measure the impact of these models on clinical care before widespread adoption.

MATERIALS AND METHODS

Study design

The primary objectives of this study were to develop a model to assess breast cancer risk and to validate its performance across diverse populations and clinical settings. We designed and benchmarked our algorithm, Mirai, against the Tyrer-Cuzick model and other deep learning models trained on the same MGH dataset, namely, Image-Only DL and Hybrid DL, in predicting future risk. Although Mirai was trained to predict both first-time cancer cases and recurrences, we limited our analysis to patients without a prior history of breast cancer to enable a fair comparison against the Tyrer-Cuzick model. Our secondary objective was to demonstrate the ability of Mirai to identify high-risk cohorts and to compare it with alternative risk models.

To develop Mirai, we collected consecutive screening mammograms from 80,134 patients screened between 1 January 2009 and 31 December 2016 at the MGH under approval of the MGH's Institutional Review Board and in compliance with the Health Portability and Accountability Act. Mammograms were taken either on a Selenia Dimensions device (Hologic) or a Lorad Selenia device (Hologic). We obtained outcomes through linkage to a local five-hospital registry in the Massachusetts General Brigham healthcare system, alongside pathology findings from MGH's mammography electronic medical record. We excluded patients who did not have at least 1 year of screening follow-up who were diagnosed with other cancers such as sarcomas of the breast, or who did not have all four views (L CC, L MLO, R CC, and R MLO), to identify 70,972 patients. Patients were randomly split into n = 56,786 for training, n = 7020 for development, and n = 7166 for testing. To enable fair comparison against the Tyrer-Cuzick model, we excluded 161 patients with prior history of breast cancer from the test set, leaving 7005 patients. Because each patient had multiple examinations, this resulted in 210,819, 25,644, and 25,855 examinations for training, development, and testing, respectively. We refer to an examination as "positive" if it was followed by a pathology-confirmed cancer diagnosis within 5 years. We collected detailed risk factors, including those used by the TCv8, from provider- and patient-entered information in the mammography reporting system and associated each mammogram with patient risk factors as they were present at the time of mammography. Detailed demographics are shown in Table 3, and our data collection procedure is illustrated in Fig. 4.

To evaluate the ability of Mirai to generalize to additional populations, we collected the Karolinska and CGMH datasets under approval of the relevant institutional review boards. The Karolinska dataset was extracted from the Cohort of Screen-Aged Women (*56*). All women aged 40 to 74 within the Karolinska University uptake area who had attended screening and were diagnosed with breast cancer, without implants and without prior breast cancer, from 2008 to 2016 were included, as well as a random sample of controls with at least 2 years of follow-up from the same time period. The full Karolinska case-control dataset included 11,301 women, and 70% of both cases and controls were randomly selected for inclusion in this study. We included all mammograms, acquired on Hologic machines, from 2008 to 2016 for the included women that contained

all four views (L CC, L MLO, R CC, and R MLO), resulting in 19,328 examinations from 7353 patients. To create the CGMH dataset, we selected random women undergoing screening mammography there between 2010 and 2011 who were aged 45 to 70 or were aged 40 to 44 and had a family history of breast cancer, resulting in 13,356 examinations from 13,356 patients. Cancer outcomes were obtained from the national cancer registry. In both datasets, we excluded patients who did not have at least 1 year of screening follow-up or did not have all four views (L CC, L MLO, R CC, and R MLO). We obtained mammographic breast density assessments for both the Karolinska and CGMH datasets using a clinically validated deep learning model trained on the MGH dataset (*57*, *58*). More details about these datasets are available in Table 4 and Fig. 4. We emphasize that the Karolinska and CGMH datasets were only used for testing.

Statistical analysis

We evaluated all models by the AUC for 1- to 5-year outcomes. For instance, to compute the 3-year AUC, we considered a mammogram as positive if it was followed by a cancer diagnosis within 3 years and negative if it had at least 3 years of screening follow-up. Table S8 describes the distribution of follow-up and cancer times for each dataset. We also calculated Uno's C-index (*59*), which offers a generalized AUC across all time points. To address that patients may have multiple examinations, we used a clustered bootstrap approach with 5000 samples to calculate confidence intervals. To assess the significance of the difference between two AUCs, we used the paired DeLong's test (*60*) as implemented in the pROC package in R (*61*). To assess the significance of the difference between two ratios, we used a two-tailed *t* test as implemented in R (*62*). For both tests, we used a predefined $P < 0.05$ for significance.

SUPPLEMENTARY MATERIALS

stm.sciencemag.org/cgi/content/full/13/578/eaba4373/DC1
Materials and Methods
Fig. S1. t-SNE plot for Mirai's hidden representation (left) without and (right) with adversarial training on 5000 random samples from the MGH test set.
Fig. S2. Saliency scores of images and all clinical risk factors across the MGH test set.
Fig. S3. t-SNE plots for Mirai's hidden representation colored by cancer subtype factors on 1000 random positive examinations from the Karolinska test set.
Table S1. The distribution of clinical risk factors in the MGH dataset.
Table S2. ROC AUCs and C-indices for Mirai and prior risk models on all test sets excluding cancers confirmed within 6 months of the screening mammogram.
Table S3. Ablation study of Mirai on the MGH datasets.
Table S4. C-index for different models on different subpopulations in the MGH test set.
Table S5. C-indices and ROC AUCs for Mirai in predicting cancers of different subtypes in the Karolinska test set.
Table S6. Number of examinations per cancer type in the Karolinska dataset.
Table S7. Sensitivity and specificity of different risk models in identifying high-risk cohorts at MGH, excluding mammograms with a BI-RADS 0 assessment that were followed by a cancer diagnosis within 1 year.
Table S8. Distribution of follow-up times and times until cancer diagnosis for examinations in the MGH, Karolinska, and CGMH test sets.
Data file S1. Primary data from figures.
References (*63*–*72*)

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. U. S. Food and Drug Administration, Mammography Quality Standards Act and Program. Retrieved from https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics.
2. C. D. Lehman, R. D. Wellman, D. S. M. Buist, K. Kerlikowske, A. N. A. Tosteson, D. L. Miglioretti; Breast Cancer Surveillance Consortium, Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern. Med.* **175**, 1828–1837 (2015).
3. C. P. Gross, J. B. Long, J. S. Ross, M. M. Abu-Khalaf, R. Wang, B. K. Killelea, H. T. Gold, A. B. Chagpar, X. Ma, The cost of breast cancer screening in the Medicare population. *JAMA Intern. Med.* **173**, 220–226 (2013).
4. L. Tabar, M. F. Yen, B. Vitak, H. H. T. Chen, R. A. Smith, S. W. Duffy, Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *Lancet* **361**, 1405–1410 (2003).
5. Swedish Organised Service Screening Evaluation Group, Reduction in breast cancer mortality from organized service screening with mammography: 1. Further Confirmation with Extended Data. *Cancer Epidemiol. Biomarkers Prev.* **15**, 45–51 (2006).
6. A. Hackshaw, The benefits and harms of mammographic screening for breast cancer: Building the evidence base using service screening programmes. *J. Med. Screen.* **19**, 1–2 (2012).
7. M. G. Marmot, D. G. Altman, D. A. Cameron, J. A. Dewar, S. G. Thompson, M. Wilcox, The benefits and harms of breast cancer screening: An independent review. *Br. J. Cancer* **108**, 2205–2240 (2013).
8. A. Coldman, N. Phillips, C. Wilson, K. Decker, A. M. Chiarelli, J. Brisson, B. Zhang, J. Payne, G. Doyle, R. Ahmad, Pan-Canadian study of mammography screening and mortality from breast cancer. *J. Natl. Cancer Inst.* **106**, dju261 (2014).
9. L. Tabár, A. M.-F. Yen, W. Y.-Y. Wu, S. L.-S. Chen, S. Y.-H. Chiu, J. C.-Y. Fann, M. M.-S. Ku, R. A. Smith, S. W. Duffy, T. H.-H. Chen, Insights from the breast cancer screening trials: How screening affects the natural history of breast cancer and implications for evaluating service screening programs. *Breast J.* **21**, 13–20 (2015).
10. A. N. A. Tosteson, N. K. Stout, D. G. Fryback, S. Acharyya, B. A. Herman, L. G. Hannah, E. D. Pisano; DMIST Investigators, Cost-effectiveness of digital mammography breast cancer screening. *Ann. Intern. Med.* **148**, 1–10 (2008).
11. T. Salz, A. R. Richman, N. T. Brewer, Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psychooncology* **19**, 1026–1034 (2010).
12. M. Bond, T. Pavey, K. Welch, C. Cooper, R. Garside, S. Dean, C. Hyde, Systematic review of the psychological consequences of false-positive screening mammograms. *Health Technol. Assess.* **17**, 1–170 (2013).
13. J. Brodersen, V. D. Siersma, Long-term psychosocial consequences of false-positive screening mammography. *Ann. Fam. Med.* **11**, 106–115 (2013).
14. A. N. A. Tosteson, D. G. Fryback, C. S. Hammond, L. G. Hanna, M. R. Grove, M. Brown, Q. Wang, K. Lindfors, E. D. Pisano, Consequences of false-positive screening mammograms. *JAMA Intern. Med.* **174**, 954–961 (2014).
15. T. J. Wilt, R. P. Harris, A. Qaseem; High Value Care Task Force of the American College of Physicians, Screening for cancer: Advice for high-value care from the American College of Physicians. *Ann. Intern. Med.* **162**, 718–725 (2015).
16. K. C. Oeffinger, E. T. H. Fontham, R. Etzioni, A. Herzig, J. S. Michaelson, Y.-C. T. Shih, L. C. Walter, T. R. Church, C. R. Flowers, S. J. LaMonte, A. M. D. Wolf, C. De Santis, J. Lortet-Tieulent, K. Andrews, D. Manassaram-Baptiste, D. Saslow, R. A. Smith, O. W. Brawley, R. Wender; American Cancer Society, Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. *JAMA* **314**, 1599–1614 (2015).
17. A. L. Siu; U.S. Preventive Services Task Force, Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* **164**, 279–296 (2016).
18. D. L. Monticciolo, M. S. Newell, R. E. Hendrick, M. A. Helvie, L. Moy, B. Monsees, D. B. Kopans, P. R. Eby, E. A. Sickles, Breast cancer screening for average-risk women: Recommendations from the ACR Commission on Breast Imaging. *J. Am. Coll. Radiol.* **14**, 1137–1143 (2017).
19. D. L. Monticciolo, M. S. Newell, L. Moy, B. Niell, B. Monsees, E. A. Sickles, Breast cancer screening in women at higher-than-average risk: Recommendations from the ACR. *J. Am. Coll. Radiol.* **15**, 408–414 (2018).
20. R. A. Smith, K. S. Andrews, D. Brooks, S. A. Fedewa, D. Manassaram-Baptiste, D. Saslow, R. C. Wender, Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. *CA Cancer J. Clin.* **69**, 184–210 (2019).
21. J. Tyrer, S. W. Duffy, J. Cuzick, A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **23**, 1111–1130 (2004).
22. M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, J. J. Mulvihill, Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81**, 1879–1886 (1989).
23. A. R. Brentnall, E. F. Harkness, S. M. Astley, L. S. Donnelly, P. Stavrinos, S. Sampson, L. Fox, J. C. Sergeant, M. N. Harvie, M. Wilson, U. Beetles, S. Gadde, Y. Lim, A. Jain, S. Bundred, N. Barr, V. Reece, A. Howell, J. Cuzick, D. G. R. Evans, Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Res.* **17**, 147 (2015).
24. K. Dembrower, Y. Liu, H. Azizpour, M. Eklund, K. Smith, P. Lindholm, F. Strand, Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* **294**, 265–272 (2020).

25. A. Yala, C. Lehman, T. Schuster, T. Portnoi, R. Barzilay, A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60–66 (2019).
26. J. Lao, Y. Chen, Z.-C. Li, Q. Li, J. Zhang, J. Liu, G. Zhai, A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci. Rep.* **7**, 10353 (2017).
27. R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, D. R. Webster, Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
28. D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, S. Shetty, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
29. P. Courtiol, C. Maussion, M. Moarii, E. Pronier, S. Pilcer, M. Sefta, P. Manceron, S. Toldo, M. Zaslavskiy, N. Le Stang, N. Girard, O. Elemento, A. G. Nicholson, J. Y. Blay, F. Galateau-Sallé, G. Wainrib, T. Clozel, Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
30. A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, A. Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
31. M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, M. T. Bianchi, Learning sleep stages from radio signals: A conditional adversarial architecture. *Int. J. Mach. Learn.* , 4100–4109 (2017).
32. N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzębski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Young Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, K. J. Geras, Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* **39**, 1184–1194 (2020).
33. K. Geras, Deep neural networks improve radiologists' performance in breast cancers screening (2020); https://www.github.com/nyukat/breast_cancer_classifier.
34. L. V. D. van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
35. T. B. Bevers, J. H. Ward, B. K. Arun, G. A. Colditz, K. H. Cowan, M. B. Daly, J. E. Garber, M. L. Gemignani, W. J. Gradishar, J. A. Jordan, L. A. Korde, N. Kounalakis, H. Krontiras, S. Kumar, A. Kurian, C. Laronga, R. M. Layman, L. S. Loftus, M. C. Mahoney, S. D. Merajver, I. M. Meszoely, J. Mortimer, L. Newman, E. Pritchard, S. Pruthi, V. Seewaldt, M. C. Specht, K. Visvanathan, A. Wallace, M. Bergman, R. Kumar, Breast Cancer risk reduction, version 2.2015. *J. Natl. Compr. Canc. Netw.* **13**, 880–915 (2015).
36. US Preventive Services Task Force, D. K. Owens, K. W. Davidson, A. H. Krist, M. J. Barry, M. Cabana, A. B. Caughey, C. A. Doubeni, J. W. Epling Jr., M. Kubik, C. S. Landefeld, C. M. Mangione, L. Pbert, M. Silverstein, C.-W. Tseng, J. B. Wong, Medication use to reduce risk of breast cancer: US Preventive Services Task Force recommendation statement. *JAMA* **322**, 857–867 (2019).
37. K. Visvanathan, P. Hurley, E. Bantug, P. Brown, N. F. Col, J. Cuzick, N. E. Davidson, A. DeCensi, C. Fabian, L. Ford, J. Garber, M. Katapodi, B. Kramer, M. Morrow, B. Parker, C. Runowicz, V. G. Vogel III, J. L. Wade, S. M. Lippman, Use of pharmacologic interventions for breast cancer risk reduction: American Society of Clinical Oncology clinical practice guideline. *J. Clin. Oncol.* **31**, 2942–2962 (2013).
38. C. Wang, A. R. Brentnall, J. Mainprize, M. Yaffe, J. Cuzick, J. A. Harvey, External validation of a mammographic texture marker for breast cancer risk in a case–control study. *J. Med. Imaging (Bellingham)* **7**, 014003 (2020).
39. J. N. Wolfe, A study of breast parenchyma by mammography in the normal woman and those with benign and malignant disease. *Radiology* **89**, 201–205 (1967).
40. M. F. Bakker, S. V. de Lange, R. M. Pijnappel, R. M. Mann, P. H. M. Peeters, E. M. Monninkhof, M. J. Emaus, C. E. Loo, R. H. C. Bisschops, M. B. I. Lobbes, M. D. F. de Jong, K. M. Duvivier, J. Veltman, N. Karssemeijer, H. J. de Koning, P. J. van Diest, W. P. T. M. Mali, M. A. A. J. van den Bosch, W. B. Veldhuis, C. H. van Gils; DENSE Trial Study Group, Supplemental MRI screening for women with extremely dense breast tissue. *N. Engl. J. Med.* **381**, 2091–2102 (2019).
41. K. J. Geras, S. Wolfson, Y. Shen, N. Wu, S. G. Kim, E. Kim, L. Heacock, U. Parikh, L. Moy, K. Cho, High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv*:1703.07047 (2017).
42. W. Lotter, G. Sorensen, D. Cox, A Multi-scale CNN and Curriculum Learning Strategy for Mammogram Classification, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, M. Jorge Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, João Manuel R.S. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, Z. Lu, Eds. (Springer, Cham., 2017), vol. 10553, pp. 169–177.
43. E. Wu, K. Wu, D. Cox, W. Lotter, Conditional Infilling GANs for Data Augmentation in Mammogram Classification, in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, D. Stoyanov, Z. Taylor, B. Kainz, G. Maicas, R. R. Beichel, A. Martel, L. Maier-Hein, K. Bhatia, T. Vercauteren, O. Oktay, G. Carneiro, A. P. Bradley, J. Nascimento, H. Min, M. S. Brown, C. Jacobs, B. Lassen-Schmidt, K. Mori, J. Petersen, R. San José Estépar, A. Schmidt-Richberg, C. Veiga, Eds. (Springer, Cham., 2018); pp. 98–106.
44. L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep.* **9**, 12495 (2019).
45. A. Yala, T. Schuster, R. Miles, R. Barzilay, C. Lehman, A deep learning model to triage screening mammograms: A simulation study. *Radiology* **293**, 38–46 (2019).
46. A. Akselrod-Ballin, M. Chorev, Y. Shoshan, A. Spiro, A. Hazan, R. Melamed, E. Barkan, E. Herzel, S. Naor, E. Karavani, G. Koren, Y. Goldschmidt, V. Shalev, M. Rosen-Zvi, M. Guindy, Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* **292**, 331–342 (2019).
47. K. Wu, E. Wu, Y. Wu, H. Tan, G. Sorensen, M. Wang, B. Lotter, Validation of a deep learning mammography model in a population with low screening rates. *arXiv preprint arXiv*:1911.00364 (2019).
48. A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M. G. Wallis, I. Andersson, S. Zackrisson, R. M. Mann, I. Sechopoulos, Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. *J. Natl. Cancer Inst.* **111**, 916–922 (2019).
49. D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with deep learning. *Sci. Rep.* **8**, 4165 (2018).
50. T. Févry, J. Phang, N. Wu, S. G. Kim, L. Moy, K. Cho, K. J. Geras, Improving localization-based approaches for breast cancer screening exam classification. *arXiv preprint arXiv*:1908.00615 (2019).
51. Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho, K. J. Geras, An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv preprint arXiv*:2002.07613 (2020).
52. T. Kyono, F. J. Gilbert, M. van der Schaar, MAMMO: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *arXiv preprint arXiv*:1811.02661 (2018).
53. S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, S. Shetty, International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
54. T. Schaffter, D. S. M. Buist, C. I. Lee, Y. Nikulin, D. Ribli, Y. Guan, W. Lotter, Z. Jie, H. Du, S. Wang, J. Feng, M. Feng, H.-E. Kim, F. Albiol, A. Albiol, S. Morrell, Z. Wojna, M. E. Ahsen, U. Asif, A. J. Yepes, S. Yohanandan, S. Rabinovici-Cohen, D. Yi, B. Hoff, T. Yu, E. C. Neto, D. L. Rubin, P. Lindholm, L. R. Margolies, R. B. McBride, J. H. Rothstein, W. Sieh, R. Ben-Ari, S. Harrer, A. Trister, S. Friend, T. Norman, B. Sahiner, F. Strand, J. Guinney, G. Stolovitzky; DM DREAM Consortium, L. Mackey, J. Cahoon, L. Shen, J. H. Sohn, H. Trivedi, Y. Shen, L. Buturovic, J. C. Pereira, J. S. Cardoso, E. Castro, K. T. Kalleberg, O. Pelka, I. Nedjar, K. J. Geras, F. Nensa, E. Goan, S. Koitka, L. Caballero, D. D. Cox, P. Krishnaswamy, G. Pandey, C. M. Friedrich, D. Perrin, C. Fookes, B. Shi, G. C. Negrie, M. Kawczynski, K. Cho, C. S. Khoo, J. Y. Lo, A. G. Sorensen, H. Jung, Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw. Open* **3**, e200265 (2020).
55. H. E. Kim, H. H. Kim, B. K. Han, K. H. Kim, K. Han, H. Nam, E. H. Lee, E. K. Kim, Changes in cancer detection and false-positive recall in mammography using artificial intelligence: A retrospective, multireader study. *Lancet Digit Health* **2**, e138–e148 (2020).
56. K. Dembrower, P. Lindholm, F. Strand, A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the Cohort of Screen-Aged Women (CSAW). *J. Digit. Imaging*, 408–413 (2020).
57. C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, R. Barzilay, Mammographic breast density assessment using deep learning: Clinical implementation. *Radiology* **290**, 52–58 (2019).
58. B. N. Dontchos, A. Yala, R. Barzilay, J. Xiang, C. D. Lehman, External validation of a deep learning model for predicting mammographic breast density in routine clinical practice. *Acad. Radiol.* **S1076-6332**, 30626–30629 (2020).
59. H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, L. J. Wei, On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).
60. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 837–845 (1988).
61. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
62. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018).
63. I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, vol. 1 (MIT Press, 2016).

64. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Las Vegas, NV, 2016), pp. 770–778.
65. N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer. *arXiv preprint arXiv*:1802.05751 (2018).
66. O. O. Aalen, T. H. Scheike, Aalen's additive regression model, in *Encyclopedia of Biostatistics*, P. Armitage, T. Colton, Eds. (Wiley StatsRef: Statistics Reference Online, 2005).
67. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv*:1502.03167 (2015).
68. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
69. J. C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in *Advances in Large Margin Classifiers* (MIT Press, 1999), vol. 10, pp. 61–74.
70. M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks. *Int. J. Mach. Learn.* **70**, 3319–3328 (2017).
71. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
72. J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).

Acknowledgments: We are grateful to Y. Liu (KTH Royal Institute of Technology) and I. Beronius-Magras (Science for Life Laboratory) for help in preparing the Karolinska dataset. We are also grateful to the Cancer Center of Linkou Chang Gung Memorial Hospital for assisting in data collection under IRB no. 201901491B0C601 and to R. Yang, J. Huang, and their team (Quanta Computer) for providing technical and computing support in analyzing the CGMH dataset. **Funding:** We are grateful to be supported by grants from Susan G. Komen (A.Y., P.G.M., and R.B.), an anonymous foundation (A.Y., P.G.M., C.L., and R.B.), the Breast Cancer Research Foundation (A.Y., P.G.M., C.L., and R.B.), Quanta Computing (A.Y., P.G.M., and R.B.), and the MIT J-Clinic (A.Y., P.G.M., and R.B.). This work was also supported by the Chang Gung Medical Foundation grant SMRPG3K0051 (G.L. and Y.W.) and by the Stockholm Läns Landsting HMT Grant 201708002 (K.S.). **Author contributions:** A.Y., C.L., and R.B. designed the research goals and aims. A.Y. and R.B. designed the model. A.Y., C.L., F.S., G.L., and R.B. designed the evaluation methodology. A.Y. and P.G.M. wrote the software. C.L., L.L., K.S., F.S., Y.-L.W., and G.L. curated the datasets. A.Y. and P.G.M. performed the analysis. P.G.M. created the visualizations. Computational resources were provided by R.B., C.L., K.S., F.S., Y.-L.W., and G.L. All authors contributed to the writing of the manuscript. R.B. and C.L. supervised the project. **Competing interests:** MIT and MGH have filed patents regarding Mirai, namely, "Machine-Learning Model for Predicting Breast Cancer Risk" (M0437.70158US00) and "Robust Imaging-based risk models" (M0437.70156US00). A.Y. has consulted for Janssen and Outcomes4Me. F.S. has consulted for Collective Minds Radiology AB and for Lunit. K.H. has received honoraria from Hologic (Surgical implant for radiation planning with breast conservation and wire free breast biopsy) and Myriad Genetics. K.H. has financial interests in CRA Health and Ask2Me.Org. K.S. has consulted for Institut Produits Synthése (IPSEN) AB. R.B. has consulted for Janssen, Bayer, Novartis, ImmunAI, Intersystems, Moderna, Vertex, Miliporesigma, Merck, OSDI, and McKenzie. C.L. has consulted for GE Healthcare and is a founder of Clarity Inc. The other authors declare that they have no competing interests. **Data and materials availability:** All code used for training and developing the models is available at learningtocure.csail.mit.edu (DOI: 10.5281/zenodo.4291202). The trained Mirai model is available upon request for research use. All datasets were used under license to the respective hospital system for the current study and are not publicly available. All data associated with this study are present in the paper or the Supplementary Materials.

Submitted 4 May 2020
Resubmitted 24 July 2020
Accepted 21 December 2020
Published 27 January 2021
10.1126/scitranslmed.aba4373

Citation: A. Yala, P. G. Mikhael, F. Strand, G. Lin, K. Smith, Y.-L. Wan, L. Lamb, K. Hughes, C. Lehman, R. Barzilay, Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* **13**, eaba4373 (2021).