Science Translational Medicine
AAAS

stm.sciencemag.org/cgi/content/full/13/578/eaba4373/DC1

Supplementary Materials for

Toward robust mammography-based models for breast cancer risk

Adam Yala*, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, Regina Barzilay

*Corresponding author. Email: adamyala@csail.mit.edu

Published 27 January 2021, *Sci. Transl. Med.* **13**, eaba4373 (2021)
DOI: 10.1126/scitranslmed.aba4373

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at stm.sciencemag.org/cgi/content/full/13/578/eaba4373/DC1)

Data file S1 (Microsoft Excel format). Primary data from figures.

Supplementary Material

Materials and Methods

Image Preprocessing

All of the mammograms used in this study were captured using either the Hologic Selenia or Selenia Dimensions mammography devices. We converted presentation view dicoms to PNG16 files using the DCMTK library. We used the dcmj2pnm program (v3.6.1, 2015) with +on2 and–min-max-window flags. We used torchvision (version 0.2.1) and Pillow (version 5.2.0) python libraries for image preprocessing and data augmentations. First, we resized each mammogram view to 1664 by 2048 pixels. Following standard practice [63], we normalized our images to have zero mean and unit variance. To this end, we calculated the pixel mean and standard deviation across the training set and normalized each image by this mean and standard deviation before feeding it into the model. We used the training set image mean and standard deviation for all images, including those on the testing and development sets at MGH, and on the test sets from Karolinska and CGMH.

Architecture Details

We encoded each view of the mammogram independently using ResNet-18 [64], with a global max pooling layer at the end, to compress the image representation to a 512-dimensional vector, *x*. We refer to this as our Image Encoder. We note that this is akin to the Image-Only model from [25]. To aggregate the information from different views, we took the image representation from each view, and conditioned it on a learned view and laterality embedding, to obtain view-specific representations. To condition a vector *x* by an embedding *e,* we used the following expression:

$$h = (W_{scale}e) \circ x + (W_{shift}e)$$

We then took these view-specific representations and passed them into a Transformer network [65] with attention-pooling to obtain a 512-dimensional mammogram level representation. We refer to this component as our Image Aggregator.

Given the mammogram-level representation, we trained the model to independently predict each risk factor as used in TCv8. We minimized the combined cross-entropy loss of predicting each risk factor, weighted by a hyperparameter lambda, and the log-likelihood loss of predicting future cancer. We note that the risk factor prediction module can be thought of as a generative model that uses the mammogram to impute missing risk factors, and thus allows the model to be run using the mammogram alone. We refer to this component as our Risk Factor Predictor.

The additive-hazard layer first took in a patient's features, *m*, from the mammogram representation and the traditional risk factors (predicted or given), and predicted a patient's baseline risk, *B(m)* using a small network (in our case a linear layer). To predict risk at *k* years

away from the mammogram, it separately predicted the positive 0-1 year marginal hazard (i.e., the additional risk of getting cancer in the next year) using network H_0, and the 1-2 year hazard using network H_1, etc. Each marginal hazard network, e.g H_1, is implemented as a linear layer followed by a ReLU. To obtain the overall risk at year *k,* the additive-hazard layer summed the baseline risk and the marginal hazards up to year *k.* This is summarized in equation 1, where P(*Y=1, T=k | m)* refers to a patient being diagnosed with cancer within *k* years. We note that this modeling objective follows seminal work [66] in linear additive-hazard survival models.

$$(1) \quad P(t_{cancer} = k \mid m) = B(m) + \Sigma H_i(m)$$

The architecture of our additive-hazard layer ensured that risk predictions were always monotonic (that is, a patients two-year risk is always higher than their one-year risk) and enabled us to easily optimize our model by maximizing the log-likelihood of the observed data in our training set. For patients with less than five years of screening followup, we leveraged their data to supervise the prediction over the years for which we know their outcomes.

The device discriminator took as input the mammogram level representation from the Image Aggregator, as well as the predicted risk over time, and aimed to predict the identity of the device that took the mammogram, Hologic Selenia or Selenia Dimensions. This function is implemented as a two-layer multilayer perceptron with a batch-normalization [67] and ReLU nonlinearities.

Model Training

We trained Mirai in two phases; first, we trained the image encoder in conjunction with the risk factor predictor and additive hazard layer to predict breast cancer independently from each view without using conditional adversarial training. In this stage, we intialialized our image encoder with weights from ImageNet [68], and augmented our training set with random flips and rotations of the original images. We found that adding an adversarial loss at this stage or training the whole architecture end-to-end prevented the model from converging. In the second stage of training, we froze our image encoder, and trained the image aggregation module, the risk factor prediction module, the additive hazard layer, and the device discriminator in a conditional adversarial training regime [31]. We trained our adversary for three steps for every one step of training Mirai. In each stage, we performed small hyperparameter searches and chose the model that obtained the highest C-index on the development set.

Model Calibration

To obtain absolute probabilities of cancer, we utilized the Platt method [69] to calibrate the predicted probabilities of cancer on the development set. We calibrated each year's risk prediction separately. For instance, to calibrate our predictions for 5-year cancer risk, we restricted our calibrator to match the incidence seen for exams with at least five years of followup on the development set.

Saliency Analysis

Saliency scores for the model inputs were calculated with the integrated gradients method [70]. Specifically, the Image Aggregator of Mirai 'with risk factors' was passed the image representation from each view along with the patient risk factors from the MGH test set. The gradient of the 5-year logit score was then computed with respect to each individual input. The integral over the gradients was approximated using 150 steps and a baseline vector of all zeroes. Last, the saliency score was obtained by summing the attributions of each input, averaging over the entire test set, and taking the absolute value of the resulting mean.

Measuring Device Bias

To investigate the impact of different mammography devices on model calibration, we trained a device-identity classifier to recover which device an exam was taken from (Selenia Dimensions vs Lorad Selenia) from the risk assessment alone for each model, and report the ROC AUC of this classifier. Specifically, we trained a logistic regression model on the risk assessments of each model on the MGH validation set, and tested its ability to predict the correct mammography device on the MGH test set. If there exists a systematic bias in risk assessments by mammography device, then the device-identity classifier can leverage this signal to obtain a high AUC on the test set. For models that do not contain any device-related bias in their risk assessments, the device-identity classifier obtains an AUC of 0.50.

Both Hybrid DL [25] and ImageOnly DL [25] formulated five-year cancer risk prediction as a classification task and so they were trained on the 2009-2012 subset of the MGH dataset with five-years of followup. MGH only utilized one mammography machine during this time, Lorad Selenia, and as a result, ImageOnly DL and Hybrid DL did not learn device specific bias. In contrast, all Mirai ablation variants shown in table S3 were able the full MGH training set because they used a survival formulation of cancer risk (additive hazard or Cox), thus were able to learn device-related bias.

t-SNE Analysis

For all t-SNE analysis, we used the final image hidden representation from Mirai and visualized it in two dimensions with the t-SNE function in sklearn.manifold module of scikit-learn 0.21.3 [71] with default parameters.

Ablation Analysis

To study the effect of our design decisions, we report a detailed ablation study of Mirai's components in table S3. Moreover, to study the importance of our Additive Hazard formulation, we compare an Image Encoder with our Additive Hazard layer to an Image Encoder trained with a Cox Proportional Hazard's layer. The Cox proportional hazard layer predicted a single relative hazard per patient, analogous to B(x), and this model was optimized to maximize the Cox partial likelihood objective, similar to prior work in deep Cox survival models [72].

Supplementary Figures

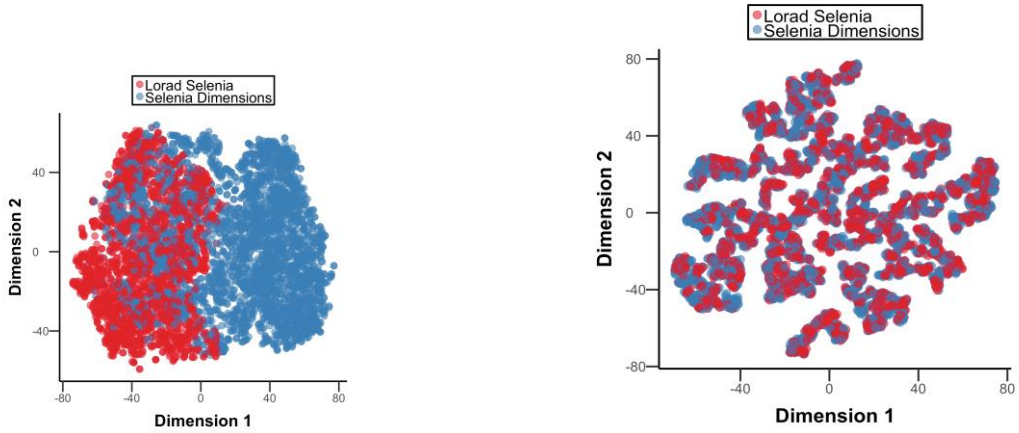

Fig. S1. t-SNE plot for Mirai's hidden representation (left) without and (right) with adversarial training on 5000 random samples from the MGH test set. Samples are colored by mammography device.

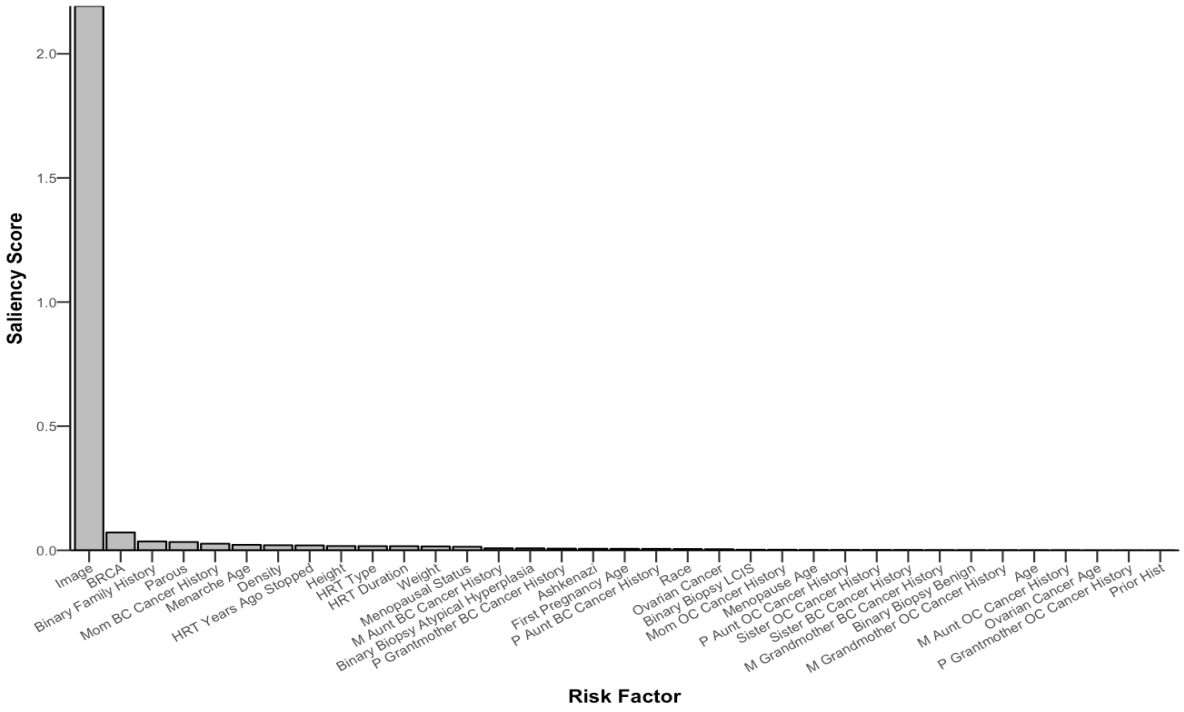

Fig. S2. Saliency scores of images and all clinical risk factors across the MGH test set.

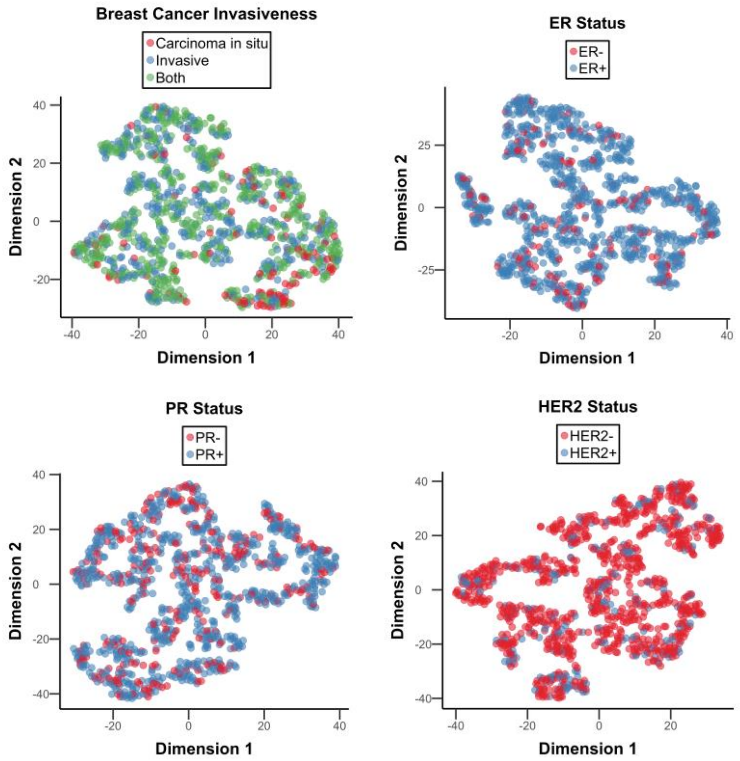

Fig. S3. t-SNE plots for Mirai's hidden representation colored by cancer subtype factors on 1000 random positive examinations from the Karolinska test set.

Supplementary Tables

Table S1. The distribution of clinical risk factors in the MGH dataset. For each demographic, we report the number of corresponding mammography exams the percentage of they constitute of the total.

	MGH Training Set		MGH Validation Set		MGH Test Set	
	All (%)	**Cancer (%)**	**All (%)**	**Cancer (%)**	**All (%)**	**Cancer (%)**
All exams	210819 (100.0%)	5379 (100.0%)	25644 (100.0%)	612 (100.0%)	25855 (100.0%)	588 (100.0%)
Age						
<40	5812 (2.8%)	84 (1.6%)	711 (2.8%)	7 (1.1%)	724 (2.8%)	7 (1.2%)
40-50	55905 (26.5%)	1113 (20.7%)	6821 (26.6%)	142 (23.2%)	7025 (27.2%)	95 (16.2%)
50-60	63314 (30.0%)	1348 (25.1%)	7762 (30.3%)	166 (27.1%)	7829 (30.3%)	188 (32.0%)
60-70	54925 (26.1%)	1770 (32.9%)	6674 (26.0%)	179 (29.2%)	6708 (25.9%)	182 (31.0%)
70-80	25401 (12.0%)	816 (15.2%)	3037 (11.8%)	102 (16.7%)	3001 (11.6%)	94 (16.0%)
80<	5461 (2.6%)	248 (4.6%)	639 (2.5%)	16 (2.6%)	568 (2.2%)	22 (3.7%)
Prior History						
Negative	207299 (98.3%)	4961 (92.2%)	25170 (98.2%)	555 (90.7%)	25855 (100.0%)	588 (100.0%)
Positive	3520 (1.7%)	418 (7.8%)	474 (1.8%)	57 (9.3%)	0 (0.0%)	0 (0.0%)
BRCA						
Negative	38 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
BRCA1	218 (0.1%)	26 (0.5%)	43 (0.2%)	7 (1.1%)	49 (0.2%)	4 (0.7%)
BRCA2	1262 (0.6%)	115 (2.1%)	126 (0.5%)	16 (2.6%)	68 (0.3%)	10 (1.7%)
Never or Unknown	209301 (99.3%)	5238 (97.4%)	25475 (99.3%)	589 (96.2%)	25738 (99.5%)	574 (97.6%)
Density						

Almost entirely fatty	20411 (9.7%)	315 (5.9%)	2429 (9.5%)	53 (8.7%)	2474 (9.6%)	31 (5.3%)
Scattered areas of fibroglandular tissue	102112 (48.4%)	2623 (48.8%)	12519 (48.8%)	261 (42.6%)	12490 (48.3%)	264 (44.9%)
Heterogeneously dense	78892 (37.4%)	2196 (40.8%)	9461 (36.9%)	263 (43.0%)	9751 (37.7%)	271 (46.1%)
Extremely dense	9293 (4.4%)	242 (4.5%)	1225 (4.8%)	35 (5.7%)	1129 (4.4%)	22 (3.7%)
Benign Biopsy						
Negative	207903 (98.6%)	5262 (97.8%)	25253 (98.5%)	587 (95.9%)	25475 (98.5%)	574 (97.6%)
Positive	2916 (1.4%)	117 (2.2%)	391 (1.5%)	25 (4.1%)	380 (1.5%)	14 (2.4%)
LCIS Biopsy						
Negative	208325 (98.8%)	5182 (96.3%)	25307 (98.7%)	581 (94.9%)	25583 (98.9%)	575 (97.8%)
Positive	2494 (1.2%)	197 (3.7%)	337 (1.3%)	31 (5.1%)	272 (1.1%)	13 (2.2%)
Atypical Hyperplasia Biopsy						
Negative	205764 (97.6%)	5047 (93.8%)	25005 (97.5%)	564 (92.2%)	25259 (97.7%)	547 (93.0%)
Positive	5055 (2.4%)	332 (6.2%)	639 (2.5%)	48 (7.8%)	596 (2.3%)	41 (7.0%)
Ovarian Cancer						
Negative	208594 (98.9%)	5353 (99.5%)	25413 (99.1%)	605 (98.9%)	25630 (99.1%)	588 (100.0%)
Positive	2225 (1.1%)	26 (0.5%)	231 (0.9%)	7 (1.1%)	225 (0.9%)	0 (0.0%)
Ovarian Cancer Age						
<30	116 (0.1%)	4 (0.1%)	18 (0.1%)	0 (0.0%)	11 (0.0%)	0 (0.0%)
30-40	210 (0.1%)	0 (0.0%)	23 (0.1%)	3 (0.5%)	9 (0.0%)	0 (0.0%)
40-50	365 (0.2%)	4 (0.1%)	42 (0.2%)	4 (0.7%)	45 (0.2%)	0 (0.0%)
50-60	476 (0.2%)	5 (0.1%)	25 (0.1%)	0 (0.0%)	38 (0.1%)	0 (0.0%)
60-70	181 (0.1%)	2 (0.0%)	16 (0.1%)	0 (0.0%)	25 (0.1%)	0 (0.0%)
<70	42 (0.0%)	0 (0.0%)	8 (0.0%)	0 (0.0%)	4 (0.0%)	0 (0.0%)

Family History						
Negative	114161 (54.2%)	2369 (44.0%)	13864 (54.1%)	258 (42.2%)	14324 (55.4%)	274 (46.6%)
Positive	96658 (45.8%)	3010 (56.0%)	11780 (45.9%)	354 (57.8%)	11531 (44.6%)	314 (53.4%)
Breast Cancer Family History						
Negative (Mother)	180018 (85.4%)	4260 (79.2%)	21751 (84.8%)	489 (79.9%)	22215 (85.9%)	475 (80.8%)
Positive (Mother)	30801 (14.6%)	1119 (20.8%)	3893 (15.2%)	123 (20.1%)	3640 (14.1%)	113 (19.2%)
Negative (Sister)	189623 (89.9%)	4573 (85.0%)	22834 (89.0%)	502 (82.0%)	23354 (90.3%)	503 (85.5%)
Positive (Sister)	21196 (10.1%)	806 (15.0%)	2810 (11.0%)	110 (18.0%)	2501 (9.7%)	85 (14.5%)
Negative (M. Aunt)	178621 (84.7%)	4412 (82.0%)	21666 (84.5%)	488 (79.7%)	21931 (84.8%)	473 (80.4%)
Positive (M. Aunt)	32198 (15.3%)	967 (18.0%)	3978 (15.5%)	124 (20.3%)	3924 (15.2%)	115 (19.6%)
Negative (M. Grandmother)	193744 (91.9%)	4900 (91.1%)	23546 (91.8%)	565 (92.3%)	23794 (92.0%)	540 (91.8%)
Positive (M. Grandmother)	17075 (8.1%)	479 (8.9%)	2098 (8.2%)	47 (7.7%)	2061 (8.0%)	48 (8.2%)
Negative (P. Aunt)	189298 (89.8%)	4646 (86.4%)	23086 (90.0%)	511 (83.5%)	23235 (89.9%)	514 (87.4%)
Positive (P. Aunt)	21521 (10.2%)	733 (13.6%)	2558 (10.0%)	101 (16.5%)	2620 (10.1%)	74 (12.6%)
Negative (P. Grandmother)	199500 (94.6%)	5031 (93.5%)	24244 (94.5%)	567 (92.6%)	24448 (94.6%)	553 (94.0%)
Positive (P. Grandmother)	11319 (5.4%)	348 (6.5%)	1400 (5.5%)	45 (7.4%)	1407 (5.4%)	35 (6.0%)
Ovarian Cancer Family History						
Negative (Mother)	209714 (99.5%)	5365 (99.7%)	25469 (99.3%)	612 (100.0%)	25695 (99.4%)	588 (100.0%)
Positive (Mother)	1105 (0.5%)	14 (0.3%)	175 (0.7%)	0 (0.0%)	160 (0.6%)	0 (0.0%)
Negative (Sister)	210155 (99.7%)	5372 (99.9%)	25579 (99.7%)	612 (100.0%)	25796 (99.8%)	588 (100.0%)
Positive (Sister)	664 (0.3%)	7 (0.1%)	65 (0.3%)	0 (0.0%)	59 (0.2%)	0 (0.0%)
Negative (M. Aunt)	209882 (99.6%)	5343 (99.3%)	25497 (99.4%)	612 (100.0%)	25781 (99.7%)	588 (100.0%)
Positive (M. Aunt)	937 (0.4%)	36 (0.7%)	147 (0.6%)	0 (0.0%)	74 (0.3%)	0 (0.0%)

Negative (M. Grandmother)	210317 (99.8%)	5370 (99.8%)	25594 (99.8%)	612 (100.0%)	25798 (99.8%)	588 (100.0%)
Positive (M. Grandmother)	502 (0.2%)	9 (0.2%)	50 (0.2%)	0 (0.0%)	57 (0.2%)	0 (0.0%)
Negative (P. Aunt)	210264 (99.7%)	5372 (99.9%)	25571 (99.7%)	607 (99.2%)	25784 (99.7%)	588 (100.0%)
Positive (P. Aunt)	555 (0.3%)	7 (0.1%)	73 (0.3%)	5 (0.8%)	71 (0.3%)	0 (0.0%)
Negative (P. Grandmother)	210604 (99.9%)	5372 (99.9%)	25607 (99.9%)	612 (100.0%)	25815 (99.8%)	588 (100.0%)
Positive (P. Grandmother)	215 (0.1%)	7 (0.1%)	37 (0.1%)	0 (0.0%)	40 (0.2%)	0 (0.0%)
HRT						
Combined	309 (0.1%)	14 (0.3%)	24 (0.1%)	0 (0.0%)	49 (0.2%)	2 (0.3%)
Estrogen	71315 (33.8%)	1896 (35.2%)	8606 (33.6%)	205 (33.5%)	8494 (32.9%)	231 (39.3%)
Unknown	2064 (1.0%)	77 (1.4%)	272 (1.1%)	5 (0.8%)	256 (1.0%)	0 (0.0%)
HRT Duration						
<1	12335 (5.9%)	331 (6.2%)	1513 (5.9%)	28 (4.6%)	1501 (5.8%)	40 (6.8%)
1-3	9725 (4.6%)	216 (4.0%)	1164 (4.5%)	29 (4.7%)	1102 (4.3%)	34 (5.8%)
3-5	9031 (4.3%)	291 (5.4%)	1132 (4.4%)	25 (4.1%)	1008 (3.9%)	23 (3.9%)
5-7	5768 (2.7%)	143 (2.7%)	690 (2.7%)	23 (3.8%)	613 (2.4%)	15 (2.6%)
7<	24264 (11.5%)	685 (12.7%)	2916 (11.4%)	80 (13.1%)	2994 (11.6%)	92 (15.6%)
HRT Years Ago Stopped						
<1	2163 (1.0%)	57 (1.1%)	271 (1.1%)	9 (1.5%)	266 (1.0%)	3 (0.5%)
1-3	4332 (2.1%)	104 (1.9%)	540 (2.1%)	15 (2.5%)	517 (2.0%)	13 (2.2%)
3-7	4634 (2.2%)	112 (2.1%)	603 (2.4%)	16 (2.6%)	543 (2.1%)	22 (3.7%)
5-7	5365 (2.5%)	156 (2.9%)	703 (2.7%)	18 (2.9%)	628 (2.4%)	21 (3.6%)
7<	33957 (16.1%)	1010 (18.8%)	3891 (15.2%)	78 (12.7%)	3842 (14.9%)	109 (18.5%)
Menopausal Status						

Pre	14581 (6.9%)	302 (5.6%)	1685 (6.6%)	43 (7.0%)	1818 (7.0%)	32 (5.4%)
Peri	5724 (2.7%)	134 (2.5%)	689 (2.7%)	21 (3.4%)	680 (2.6%)	9 (1.5%)
Post	142869 (67.8%)	3947 (73.4%)	17375 (67.8%)	454 (74.2%)	17308 (66.9%)	453 (77.0%)
Unknown	47645 (22.6%)	996 (18.5%)	5895 (23.0%)	94 (15.4%)	6049 (23.4%)	94 (16.0%)
Menopause Age						
<45	32737 (15.5%)	830 (15.4%)	3914 (15.3%)	93 (15.2%)	3797 (14.7%)	67 (11.4%)
45-50	49212 (23.3%)	1326 (24.7%)	5883 (22.9%)	152 (24.8%)	6193 (24.0%)	195 (33.2%)
50-55	53352 (25.3%)	1482 (27.6%)	6569 (25.6%)	164 (26.8%)	6290 (24.3%)	144 (24.5%)
55-60	12322 (5.8%)	428 (8.0%)	1593 (6.2%)	66 (10.8%)	1588 (6.1%)	51 (8.7%)
60<	971 (0.5%)	15 (0.3%)	105 (0.4%)	0 (0.0%)	120 (0.5%)	5 (0.9%)
Parous						
Negative	56440 (26.8%)	1491 (27.7%)	6898 (26.9%)	181 (29.6%)	6644 (25.7%)	150 (25.5%)
Positive	154379 (73.2%)	3888 (72.3%)	18746 (73.1%)	431 (70.4%)	19211 (74.3%)	438 (74.5%)
First Pregnancy Age						
<20	23145 (11.0%)	524 (9.7%)	2843 (11.1%)	47 (7.7%)	3085 (11.9%)	57 (9.7%)
20-25	44408 (21.1%)	1283 (23.9%)	5377 (21.0%)	139 (22.7%)	5083 (19.7%)	124 (21.1%)
25-30	40668 (19.3%)	1022 (19.0%)	5004 (19.5%)	120 (19.6%)	5044 (19.5%)	102 (17.3%)
30-35	29365 (13.9%)	674 (12.5%)	3520 (13.7%)	71 (11.6%)	3768 (14.6%)	99 (16.8%)
35-45	13247 (6.3%)	301 (5.6%)	1513 (5.9%)	45 (7.4%)	1740 (6.7%)	49 (8.3%)
40<	3241 (1.5%)	77 (1.4%)	452 (1.8%)	10 (1.6%)	455 (1.8%)	5 (0.9%)
Menarche Age						
<10	8435 (4.0%)	236 (4.4%)	998 (3.9%)	14 (2.3%)	900 (3.5%)	11 (1.9%)
10-12	67365 (32.0%)	1782 (33.1%)	8012 (31.2%)	222 (36.3%)	8302 (32.1%)	214 (36.4%)

12-14	93721 (44.5%)	2382 (44.3%)	11638 (45.4%)	274 (44.8%)	11266 (43.6%)	252 (42.9%)
14-16	29748 (14.1%)	646 (12.0%)	3534 (13.8%)	83 (13.6%)	3924 (15.2%)	84 (14.3%)
16<	5850 (2.8%)	146 (2.7%)	725 (2.8%)	12 (2.0%)	748 (2.9%)	6 (1.0%)
Weight						
<100	2958 (1.4%)	60 (1.1%)	348 (1.4%)	6 (1.0%)	300 (1.2%)	4 (0.7%)
100-130	48546 (23.0%)	995 (18.5%)	5860 (22.9%)	115 (18.8%)	6034 (23.3%)	113 (19.2%)
130-160	77645 (36.8%)	1958 (36.4%)	9231 (36.0%)	186 (30.4%)	9449 (36.5%)	211 (35.9%)
160-190	44907 (21.3%)	1274 (23.7%)	5522 (21.5%)	169 (27.6%)	5747 (22.2%)	153 (26.0%)
190-220	18392 (8.7%)	574 (10.7%)	2301 (9.0%)	83 (13.6%)	2249 (8.7%)	47 (8.0%)
220-250	7153 (3.4%)	199 (3.7%)	982 (3.8%)	34 (5.6%)	789 (3.1%)	18 (3.1%)
250<	3692 (1.8%)	77 (1.4%)	425 (1.7%)	7 (1.1%)	394 (1.5%)	15 (2.6%)
Height						
<50	2516 (1.2%)	32 (0.6%)	278 (1.1%)	3 (0.5%)	269 (1.0%)	1 (0.2%)
50-55	828 (0.4%)	16 (0.3%)	92 (0.4%)	2 (0.3%)	114 (0.4%)	1 (0.2%)
55-60	19298 (9.2%)	428 (8.0%)	2232 (8.7%)	30 (4.9%)	2427 (9.4%)	44 (7.5%)
60-65	122499 (58.1%)	2992 (55.6%)	14965 (58.4%)	353 (57.7%)	14859 (57.5%)	340 (57.8%)
65-75	55956 (26.5%)	1600 (29.7%)	6790 (26.5%)	182 (29.7%)	6956 (26.9%)	161 (27.4%)
70-75	1766 (0.8%)	49 (0.9%)	232 (0.9%)	20 (3.3%)	243 (0.9%)	14 (2.4%)
75<	282 (0.1%)	10 (0.2%)	52 (0.2%)	5 (0.8%)	53 (0.2%)	3 (0.5%)
Ashkenazi						
Negative	195636 (92.8%)	4993 (92.8%)	23737 (92.6%)	552 (90.2%)	24101 (93.2%)	526 (89.5%)
Positive	15183 (7.2%)	386 (7.2%)	1907 (7.4%)	60 (9.8%)	1754 (6.8%)	62 (10.5%)

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. Note, 1-year AUC is not defined for the Karolinska dataset because there were no cancers diagnosed in the interval between six months and one year from the mammogram. 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: 25,708 Exams, 441 followed by cancer diagnosis							
Tyrer-Cuzick Version 8 (TCv8) [21]	Yes	0.62 (0.58, 0.67)	0.65 (0.51, 0.79)	0.64 (0.59, 0.70)	0.63 (0.58, 0.67)	0.62 (0.58, 0.66)	0.62 (0.57, 0.66)
Radiologist BI-RADs	NA	0.53 (0.50, 0.55)	0.74 (0.63, 0.86)	0.55 (0.51, 0.58)	0.53 (0.50, 0.55)	0.52 (0.50, 0.54)	0.52 (0.50, 0.53)
Image-And-Heathmaps [32]	No	0.63 (0.59, 0.67)	0.71 (0.61, 0.84)	0.68 (0.63, 0.73)	0.64 (0.60, 0.69)	0.62 (0.58, 0.66)	0.59 (0.55, 0.63)
Image-Only DL [25]	No	0.67 (0.64, 0.71)	0.64 (0.53, 0.76)	0.67 (0.62, 0.73)	0.68 (0.64, 0.72)	0.68 (0.65, 0.73)	0.70 (0.66, 0.73)
Hybrid DL [25]	Yes	0.67 (0.63, 0.71)	0.63 (0.51, 0.76)	0.68 (0.63, 0.73)	0.67 (0.62, 0.71)	0.67 (0.63, 0.72)	0.69 (0.65, 0.73)
Mirai (Ours)	No	0.69 (0.66, 0.73)	0.71 (0.60, 0.84)	0.71 (0.66, 0.76)	0.71 (0.67, 0.75)	0.71 (0.67, 0.75)	0.71 (0.68, 0.75)
	Yes	0.70 (0.66, 0.74)	0.72 (0.61, 0.84)	0.72 (0.67, 0.78)	0.72 (0.68, 0.76)	0.71 (0.68, 0.75)	0.72 (0.68, 0.76)
Karolinska Test Set: 18,811 Exams, 896 followed by cancer diagnosis							
Image-Only DL [25]	No	0.67 (0.64, 0.69)	NA	0.66 (0.61, 0.71)	0.68 (0.65, 0.70)	0.66 (0.64, 0.69)	0.64 (0.62, 0.67)
Mirai (Ours)	No	0.71 (0.69, 0.74)	NA	0.72 (0.67, 0.77)	0.73 (0.71, 0.76)	0.73 (0.70, 0.75)	0.71 (0.69, 0.73)

Chang Gung Memorial Hospital Test Set: 13,251 Exams, 139 followed by cancer diagnosis							
Image-Only DL [25]	No	0.61 (0.56, 0.66)	0.72 (0.56, 0.92)	0.63 (0.52, 0.73)	0.60 (0.53, 0.67)	0.62 (0.56, 0.68)	0.61 (0.56, 0.66)
Mirai (Ours)	No	0.70 (0.66, 0.75)	0.84 (0.72, 0.99)	0.76 (0.68, 0.84)	0.71 (0.64, 0.77)	0.71 (0.66, 0.76)	0.70 (0.66, 0.75)

Table S3. Ablation study of Mirai on the MGH datasets. We report the C-Index for each model on the MGH validation and test sets, as well as the AUC of the Device-Identity Classifier on the test set. All metrics are followed by 95% confidence intervals.

Model	Use Risk Factors	MGH Validation Set C-Index	MGH Test Set C-Index	Device-Identity Classifier AUC on MGH Test Set
Tyrer-Cuzick Version 8 (TCv8) [21]	Yes	0.63 (0.59, 0.67)	0.64 (0.60, 0.67)	0.50 (0.50, 0.50)
ImageOnly DL [25]	No	0.69 (0.66, 0.73)	0.72 (0.69, 0.75)	0.51 (0.50, 0.51)
Hybrid DL [25]	Yes	0.71 (0.68, 0.75)	0.72 (0.69, 0.75)	0.50 (0.50, 0.50)
Image Encoder + Cox Proportional Hazard Layer	No	0.64 (0.60, 0.67)	0.63 (0.60, 0.67)	0.74 (0.73, 0.74)
Image Encoder + Additive Hazard Layer	No	0.71 (0.68, 0.75)	0.73 (0.70, 0.76)	0.77 (0.76, 0.77)
Image Encoder + Additive Hazard + Predict Risk Factors	No	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	0.68 (0.67, 0.69)
	Yes	0.75 (0.72, 0.79)	0.74 (0.72, 0.77)	0.68 (0.67, 0.69)
Image Encoder + Additive Hazard + Image Aggregator + Predict Risk Factors	No	0.75 (0.72, 0.78)	0.75 (0.73, 0.78)	0.76 (0.75, 0.76)
	Yes	0.77 (0.74, 0.80)	0.75 (0.72, 0.78)	0.74 (0.73, 0.74)
Mirai = Image Encoder + Additive	No	0.73 (0.70, 0.77)	0.75 (0.72, 0.78)	0.50 (0.50, 0.50)
	Yes	0.76 (0.73, 0.80)	0.76 (0.74, 0.80)	0.50 (0.50, 0.50)

Hazard + Image Aggregator + Predict Risk Factors + Adversarial Training				

Table S4. C-index for different models on different subpopulations in the MGH test set. All metrics are followed by their 95% confidence interval.

Model	TCv8	ImageOnly	Hybrid DL	Mirai without Risk Factors	Mirai with Risk Factors
Race					
African American	0.62 (0.44, 0.84)	0.72 (0.61, 0.89)	0.73 (0.59, 0.88)	0.72 (0.56, 0.89)	0.71 (0.55, 0.90)
Asian	0.54 (0.36, 0.75)	0.68 (0.53, 0.85)	0.67 (0.50, 0.85)	0.77 (0.64, 0.92)	0.80 (0.68, 0.95)
White	0.64 (0.60, 0.68)	0.73 (0.69, 0.76)	0.72 (0.68, 0.75)	0.75 (0.71, 0.78)	0.75 (0.72, 0.78)
Age					
<50	0.63 (0.56, 0.71)	0.66 (0.59, 0.74)	0.68 (0.60, 0.77)	0.71 (0.63, 0.78)	0.71 (0.55, 0.90)
50-70	0.64 (0.60, 0.69)	0.71 (0.67, 0.74)	0.71 (0.68, 0.75)	0.74 (0.71, 0.78)	0.80 (0.68, 0.95)
>70	0.54 (0.46, 0.62)	0.76 (0.69, 0.83)	0.71 (0.63, 0.89)	0.74 (0.67, 0.82)	0.75 (0.72, 0.78)
Density					
Non-Dense	0.63 (0.58, 0.68)	0.71 (0.67, 0.76)	0.70 (0.66, 0.75)	0.74 (0.70, 0.78)	0.75 (0.71, 0.79)
Dense	0.64 (0.59, 0.69)	0.73 (0.69, 0.77)	0.73 (0.69, 0.78)	0.76 (0.72, 0.80)	0.76 (0.72, 0.80)
Mammography Device					
Lorad Selenia	0.65 (0.61, 0.70)	0.71 (0.67, 0.75)	0.71 (0.67, 0.76)	0.73 (0.69, 0.77)	0.74 (0.68, 0.78)
Selenia Dimensions	0.62 (0.57, 0.67)	0.74 (0.71, 0.78)	0.73 (0.69, 0.77)	0.77 (0.74, 0.81)	0.78 (0.74, 0.82)

Table S5. C-indices and ROC AUCs for Mirai in predicting cancers of different subtypes in the Karolinska test set. For each row in the table, we evaluate the ability of the model to discriminate between patients who developed the specific subtype of cancer (e.g., HER2-) from those who did not develop cancer. All metrics are followed by their 95% confidence interval.

Subtype	C-Index	1-year AUC	2-year AUC	3-year AUC	4-year AUC	5-year AUC
Invasive	0.80 (0.78, 0.82)	0.90 (0.88, 0.92)	0.85 (0.83, 0.87)	0.81 (0.79, 0.83)	0.8 (0.78, 0.82)	0.77 (0.75, 0.79)
DCIS	0.81 (0.79, 0.84)	0.92 (0.9, 0.94)	0.88 (0.85, 0.91)	0.83 (0.81, 0.86)	0.81 (0.79, 0.84)	0.78 (0.76, 0.81)
ER+	0.81 (0.79, 0.83)	0.91 (0.89, 0.93)	0.87 (0.85, 0.89)	0.82 (0.81, 0.84)	0.81 (0.79, 0.83)	0.78 (0.76, 0.81)
ER-	0.75 (0.70, 0.80)	0.87 (0.82, 0.94)	0.79 (0.73, 0.85)	0.76 (0.71, 0.82)	0.75 (0.69, 0.80)	0.73 (0.68, 0.78)
PR+	0.80 (0.78, 0.82)	0.9 (0.88, 0.93)	0.86 (0.83, 0.88)	0.81 (0.79, 0.84)	0.8 (0.78, 0.82)	0.78 (0.75, 0.80)
PR-	0.81 (0.78, 0.84)	0.9 (0.87, 0.94)	0.86 (0.82, 0.90)	0.83 (0.79, 0.86)	0.81 (0.78, 0.85)	0.78 (0.74, 0.81)
HER2+	0.79 (0.75, 0.84)	0.92 (0.87, 0.97)	0.87 (0.82, 0.93)	0.83 (0.78, 0.88)	0.79 (0.74, 0.85)	0.75 (0.70, 0.81)
HER2-	0.81 (0.79, 0.83)	0.9 (0.88, 0.92)	0.86 (0.83, 0.88)	0.82 (0.80, 0.84)	0.81 (0.79, 0.83)	0.78 (0.76, 0.81)

Table S6. Number of examinations per cancer type in the Karolinska dataset. These define the positive samples for the results obtained in table S4.

Cancer Type	Number of Exams
Invasive	1243
DCIS	760
ER+	1093
ER-	183
PR+	934
PR-	341
HER2+	156
HER2-	884

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. Thresholds were chosen to match the specificity of 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.

Dataset	MGH Cohort: 9,274 exams, 431 cancers within five years			
Method	Use Risk Factors	High Risk Threshold	Sensitivity	Specificity
Tyrer-Cuzick Lifetime Risk	Yes	20%	23.4% (16.4, 30.0)	85.4% (84.1, 86.6)
ImageOnly DL [25]	No	3.4%	32.5% (25.5, 38.9)	85.9% (84.8, 87.0)
Hybrid DL [25]	Yes	3.4%	36.0% (28.8, 43.0)	85.9% (84.9, 87.1)
Mirai 5-Year Risk	No	2.6 %	38.5% (31.8, 45.1)	85.2% (84.1, 86.3)
	Yes	3.0 %	41.1% (34.0, 48.2)	85.6% (84.5, 86.8)

Table S8. Distribution of follow-up times and times until cancer diagnosis for examinations in the MGH, Karolinska, and CGMH test sets.

	MGH Test Set	Karolinska Test Set	CGMH Test Set
Exams with at least X years of screening followup			
X = 1	25855	19328	13356
X = 2	21534	16148	12779
X = 3	16702	12873	12249
X = 4	12525	9578	11658
X = 5	8911	6530	11060
Exams followed by a cancer diagnosis within X years			
X = 1	173	517	116
X = 2	301	681	141
X = 3	424	1040	182
X = 4	520	1181	212
X = 5	588	1413	244