Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model

Adam Yala, MEng[1,2]; Peter G. Mikhael, BS[1,2]; Fredrik Strand, MD, PhD[3,4]; Gigin Lin, MD, PhD[5]; Siddharth Satuluru, BS[6];
Thomas Kim, MS[7]; Imon Banerjee, PhD[8]; Judy Gichoya, MD, MS[9]; Hari Trivedi, MD[9]; Constance D. Lehman, MD, PhD[10];
Kevin Hughes, MD[11]; David J. Sheedy, MPH[12]; Lisa M. Matthis, RN, NHA[12]; Bipin Karunakaran, MS, MBA[12]; Karen E. Hegarty, PhD[13];
Silvia Sabino, MD, MS[14]; Thiago B. Silva, PhD[14]; Maria C. Evangelista, MS[14]; Renato F. Caron, MS[14]; Bruno Souza, BS[14];
Edmundo C. Mauad, MD, PhD[14]; Tal Patalon, MD, MBA[15]; Sharon Handelman-Gotlib, MA[15]; Michal Guindy, MD[16]; and
Regina Barzilay, PhD[1,2]

PURPOSE Accurate risk assessment is essential for the success of population screening programs in breast cancer. Models with high sensitivity and specificity would enable programs to target more elaborate screening efforts to high-risk populations, while minimizing overtreatment for the rest. Artificial intelligence (AI)-based risk models have demonstrated a significant advance over risk models used today in clinical practice. However, the responsible deployment of novel AI requires careful validation across diverse populations. To this end, we validate our AI-based model, Mirai, across globally diverse screening populations.

METHODS We collected screening mammograms and pathology-confirmed breast cancer outcomes from Massachusetts General Hospital, USA; Novant, USA; Emory, USA; Maccabi-Assuta, Israel; Karolinska, Sweden; Chang Gung Memorial Hospital, Taiwan; and Barretos, Brazil. We evaluated Uno's concordance index for Mirai in predicting risk of breast cancer at one to five years from the mammogram.

RESULTS A total of 128,793 mammograms from 62,185 patients were collected across the seven sites, of which 3,815 were followed by a cancer diagnosis within 5 years. Mirai obtained concordance indices of 0.75 (95% CI, 0.72 to 0.78), 0.75 (95% CI, 0.70 to 0.80), 0.77 (95% CI, 0.75 to 0.79), 0.77 (95% CI, 0.73 to 0.81), 0.81 (95% CI, 0.79 to 0.82), 0.79 (95% CI, 0.76 to 0.83), and 0.84 (95% CI, 0.81 to 0.88) at Massachusetts General Hospital, Novant, Emory, Maccabi-Assuta, Karolinska, Chang Gung Memorial Hospital, and Barretos, respectively.

CONCLUSION Mirai, a mammography-based risk model, maintained its accuracy across globally diverse test sets from seven hospitals across five countries. This is the broadest validation to date of an AI-based breast cancer model and suggests that the technology can offer broad and equitable improvements in care.

J Clin Oncol 40:1732-1740. © 2021 by American Society of Clinical Oncology

INTRODUCTION

Accurate risk assessment is essential for the success of population screening programs in breast cancer. American Cancer Society and National Comprehensive Cancer Network (NCCN) guidelines currently leverage statistical risk models to determine eligibility for magnetic resonance imaging (MRI) screening.[1,2] Risk models are also leveraged by NCCN and US Food and Drug Administration guidelines to recommend chemoprevention. The Tyrer-Cuzick (TC) model is a widely adopted risk model used by American Cancer Society and NCCN guidelines that leverage patient demographics, detailed family history, and breast density to predict breast cancer risk.[3] However, the TC model only provides a global risk prediction and has limited accuracy for individuals and for specific timeframes. Improved short-term risk prediction (ie, within 5 years) would enable programs to target more effective screening and prevention efforts to high-risk populations, while minimizing overtreatment for the rest.

Recent work has demonstrated that Mirai, an AI model to predict 5-year cancer risk from screening mammograms, has shown considerable promise, obtaining area under the curves (AUCs) of 0.76, 0.81, and 0.79 on independent test sets from Mass General Hospital, Karolinska, and Chang Gung Memorial hospital, respectively.[4] This performance constitutes significant advancement over the traditional risk models used in clinical practice today, such as the TC model, which obtained a 5-year AUC of 0.62 on the Mass General Hospital test set. In addition, traditional risk models have been shown to exhibit bias when applied to minority populations.[5-7] These concerns motivate the introduction of AI-based models for population screening.

To ensure equitable improvements in care, the responsible deployment of novel AI requires careful validation across diverse screening populations. Multiple studies have demonstrated that transferability of AI tools should not be taken for granted,[8-10] especially when the training population exhibits differences

CONTEXT

Key Objective

Improved breast cancer risk models would enable screening programs to improve early detection and reduce overtreatment. This study explored the robustness of an AI breast cancer risk model, Mirai, across globally diverse test sets from Massachusetts General Hospital, USA; Novant, USA; Emory, USA; Maccabi-Assuta, Israel; Karolinska, Sweden; Chang Gung Memorial Hospital, Taiwan; and Barretos, Brazil. This constitutes the broadest validation to date of an AI-based breast cancer model.

Knowledge Generated

Mirai maintained its accuracy across the globally diverse test sets. In a retrospective analysis, guidelines based on Mirai significantly outperformed existing guidelines based on Tyrer-Cuzick lifetime risk for selecting patients for supplemental screening MRI.

Relevance

Mirai has the potential to replace current risk models used in guidelines for MRI screening, offering broad and equitable improvements in care. Prospective trials are needed to confirm the benefit of identifying improved high-risk cohorts and to establish Mirai-based guidelines.

from the population to which the models are applied. Moreover, when AI tools are not carefully designed, they can capture and propagate bias. The need for equitable AI models is especially pronounced in breast cancer, where discrepancies in outcomes have been a long-standing concern for the field. To this end, we evaluate and compare the performance of Mirai across seven hospital systems across the United States, Israel, Sweden, Taiwan, and Brazil and study the impact of Mirai-based risk guidelines across these globally diverse cohorts.

METHODS

Our retrospective study was approved by the institutional review board of each clinical institution with a waiver for written informed consent and was compliant with the Health Insurance Portability and Accountability Act. We collected data sets from Massachusetts General Hospital (MGH), USA; Novant, USA; Emory, USA; Maccabi-Assuta, Israel; Karolinska, Sweden; Chang Gung Memorial Hospital (CGMH), Taiwan; and Barretos, Brazil. The MGH training set was previously used to develop Mirai, and AUC analyses on the MGH, Karolinska, and CGMH test sets were previously evaluated.[4] These three data sets are included for new analyses and for completeness. Across all data sets, we collected mammograms from a large subset of patients and leveraged the mammograms to obtain Mirai risk assessments. We did not use additional risk factors for Mirai risk assessments. Mirai was trained using Hologic images, and all mammograms included in this study were taken using a Hologic machine.

Description of Cohorts

To collect the MGH data set, we collected consecutive screening mammograms from 80,134 patients screened between January 1, 2009, and December 31, 2016, at MGH. 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 without at least 1 year of screening follow-up, who were diagnosed with other cancers (eg, sarcoma) in the breast or did not have all four views available, to identify 70,972 patients, following the previous work on the MGH data set.[4] Seven thousand one hundred sixty-six patients were randomly selected for the test set. We excluded 161 patients with history of breast cancer from the test set, leaving 7,005 patients with 25,855 examinations.

To collect the Novant data set, we selected 7,238 patients randomly from the cohort of all patients age 40-69 years screened at a Novant Health clinic between January 1, 2012, and December 31, 2016. We included all mammograms across this time period and obtained outcomes by querying both a local cancer registry and the Novant electronic medical record. We excluded patient examinations that did not have at least 1 year of screening follow-up with prior cancer or whose mammogram did not include all four standard views to identify 14,157 examinations from 5,887 patients.

To collect the Emory data set, we extracted 8 years of mammograms from an institutional database of all comers for screening mammography from 2013 to 2020 and randomly selected 30% of women from this database, totaling 75,010 examinations from 28,994 patients. We collected outcomes from pathology findings from Emory's institutional database using Magview software (Fulton, MD). As with other data sets, we excluded patients' examinations that did not have at least 1 year of screening follow-up, with prior cancer or whose mammogram did not contain all four standard views to identify 44,008 examinations from 16,495 patients.

To collect the Maccabi-Assuta data set, we selected all comers for screening mammography at Maccabi-Assuta during 2015 age 30 years or older, resulting in 9,775

examinations from 9,775 women. For each patient, we obtained dates of first breast cancer diagnosis from the Maccabi-Assuta electronic medical records and a regional registry. We excluded examinations from non-Hologic machines and patients with a history of breast cancer to identify 6,189 examinations from 6,189 patients.

The Karolinska data set was extracted from the Cohort of Screen-Aged Women.[11] All women age 40-74 years 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 data set included 11,303 women, and 70% of both cases and controls were randomly selected for inclusion in this study, resulting in 19,328 examinations from 7,353 patients.

To collect the CGMH data set, which consisted of 13,356 examinations from 13,356 patients, we selected random women undergoing screening mammography there between 2010 and 2011 who were age 45-70 years. Following local guidelines, we also included women age 40-44 years who had a family history of breast cancer. Cancer outcomes were obtained from the national cancer registry.

To collect the Barretos test set, we selected all women age 40 to 69 years who received screening mammograms at the Fernanópolis and Campo Grande units from January 2, 2014, to June 30, 2015, to obtain a cohort of 6,206 mammograms from 6,206 patients. Cancer outcomes were obtained from patient medical records at Barretos Cancer Hospital. We excluded mammograms without all four standard views, with prior cancer, and with insufficient follow-up to identify 5,900 examinations from 5,900 patients.

Across all data sets, we defined a cancer-positive outcome as a pathology-confirmed diagnosis of either invasive breast carcinoma or ductal carcinoma in situ. We used screening follow-up to define when patients were cancer-negative. For instance, we considered a patient negative for 3 years if they had screening follow-up for at least 3 years without a cancer diagnosis. For all data sets, except the CGMH data set, we excluded patients with prior cancer to enable fair comparison against the TC model, which does not assess risk for this population. We did not perform this exclusion for the CGMH data set because of difficulties in manual data curation.

Model Evaluation

We evaluated the overall accuracy of Mirai across all tests using area under the Curve (AUC) for 1- to 5-year outcomes. For instance, to compute the 3-year AUC, we considered the outcome 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 without a diagnosis. We also computed Uno's concordance index (C-index),[12] which offers a generalized AUC across time-points. We computed this analysis on the entirety of all test sets and on

subgroups of the Emory data set, which contained detailed demographic information and reflects a large representation of African-American women. Specifically, we studied the performance of Mirai on White and African-American or Black Women and on women younger than 50 years, between 50 and 70 years, and older than 70 years.

To evaluate the clinical significance of Mirai's performance, we evaluated its ability to identify high-risk cohorts that may benefit from supplemental screening. To perform this analysis, we restricted our attention to patients who were initially screening negative and had at least 5 years of screening follow-up. We defined an examination as screening negative if it was not followed by a cancer diagnosis within 6 months. We defined the sensitivity of a guideline as the percentage of all patients who would develop cancer within 5 years included within the high-risk cohort and thus may benefit from supplemental screening. We defined the specificity of the guidelines as the percentage of all patients who do not develop cancer within 5 years not included in the high-risk cohort and thus may avoid overtreatment. We compared three guidelines for identifying high-risk patients: 20% lifetime risk by TC (TC guideline), Mirai at the specificity of the TC guideline, and Mirai at the sensitivity of the TC guideline. We studied Mirai at TC specificity and TC sensitivity to evaluate the potential of Mirai to improve early detection for a fixed cost (ie, specificity) and the potential to reduce costs for a fixed level of early detection (ie, sensitivity), respectively. The Mirai at TC specificity and Mirai at TC sensitivity guidelines were chosen to match the specificity and sensitivity of the TC guideline on the MGH development set. We only evaluated the TC model on the MGH test set as the necessary risk factors were not available at the other six institutions. We performed this analysis on all test sets and subgroups of the Emory data set by race. To illustrate the full spectrum of possible operating points for this use case, we also plot receiver operating curves for Mirai for each institution.

Statistical Analysis

We analyzed the performance of Mirai on the basis of all the mammograms in the held-out test sets. To address that patients may have multiple examinations in a test set, we used the clustered bootstrap with 5,000 samples to calculate CIs. To assess the significance of difference between two sensitivities or specificities, we used a two-tailed t-test as implemented in R with a predefined $P < .05$ for significance.

RESULTS

Cohort Demographics

The demographics for all test sets are reported in Table 1, and the data set creation process is illustrated in flowcharts in Figure 1. The MGH, Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos test sets consisted of 25,855, 14,157, 44,008, 6,189, 19,328, 13,356, and 5,900 examinations from 7,005, 5,887, 16,495, 6,189, 7,353, 13,356, and 5,900 patients of which 588, 235, 1,003, 186,

TABLE 1. Demographics of MGH, Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos Test Sets

Patient Groups	MGH	Novant	Emory	Maccabi-Assuta	Karolinska	CGMH	Barretos
Unique patients	7,005 (233)	5,887 (123)	16,495 (495)	6,189 (186)	7,353 (799)	13,356 (244)	5,900 (146)
All examinations	25,855 (588)	14,157 (235)	44,008 (1,003)	6,189 (186)	19,328 (1,413)	13,356 (244)	5,900 (146)
Age at examination, years							
< 40	724 (7)	0 (0)	410 (11)	23 (1)	0 (0)	0 (0)	0 (0)
40-50	7,025 (95)	3,917 (53)	9,047 (147)	1,589 (47)	7,814 (364)	4,008 (74)	2,810 (41)
50-60	7,829 (188)	5,368 (65)	12,113 (235)	1,232 (23)	5,477 (387)	6,301 (115)	2,114 (59)
60-70	6,708 (182)	4,872 (117)	13,182 (302)	2,232 (57)	5,174 (563)	3,024 (55)	976 (46)
70-80	3,001 (94)	0 (0)	7,638 (285)	1,038 (44)	863 (99)	0 (0)	0 (0)
> 80	568 (22)	0 (0)	1,495 (23)	75 (14)	0 (0)	0 (0)	0 (0)

NOTE. Patient statistics are followed by the number of patients who were diagnosed with breast cancer within 5 years. Examination level statistics, including age demographics, are followed by the number of examinations that were followed by a cancer diagnosis within 5 years.

Abbreviations: CGMH, Chang Gung Memorial Hospital; MGH, Massachusetts General Hospital.

1,413, 244, and 146 examinations were followed by cancer within 5 years, respectively. Detailed demographics of the MGH, Emory, and Novant test sets, including race, are given in Appendix Tables A1-A3 (online only), respectively. The number of patients and examinations used for each AUC computation is shown Appendix Table A4 (online only).

Model Evaluation

The performance of Mirai across all time-points and across all test sets is reported in Table 2. Mirai performed similarly across all test sets, obtaining Uno's C-indices of 0.75 (95% CI, 0.70 to 0.80), 0.77 (95% CI, 0.75 to 0.79), 0.77 (95% CI, 0.73 to 0.81), and 0.84 (95% CI, 0.81 to 0.88) on the Novant, Emory, Maccabi-Assuta, and Barretos test sets, respectively. These results are similar to the previously reported[4] C-indices of 0.75 (95% CI, 0.72 to 0.78), 0.81 (95% CI, 0.79 to 0.82), and 0.79 (95% CI, 0.76 to 0.83) on the MGH, Karolinska, and CGMH test sets. By contrast, TC obtained a C-index of 0.64 (95% CI, 0.60 to 0.67) on the MGH data set.[4] Mirai obtained 1-year AUCs of 0.84 (95% CI, 0.80 to 0.87), 0.78 (95% CI, 0.73 to 0.84), 0.83 (95% CI, 0.81 to 0.86), 0.86 (95% CI, 0.81 to 0.91), 0.90 (95% CI, 0.89 to 0.92), 0.90 (95% CI, 0.87 to 0.93), and 0.89 (95% CI, 0.86 to 0.93) at MGH, Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. Mirai obtained one higher 1-year AUC at Karolinska (0.90), CGMH (0.90), and Barretos (0.89), where screening is biennial, than at MGH (0.84), Novant (0.78), Emory (0.83), and Maccabi-Assuta (0.86), where screening is annual. The performance of Mirai when excluding cancers diagnosed within 6 months is shown in Appendix Table A5 (online only). Here, Mirai obtained C-indices of 0.69 (95% CI, 0.66 to 0.73), 0.72 (95% CI, 0.66 to 0.79), 0.69 (95% CI, 0.66 to 0.72), 0.70 (95% CI, 0.64 to 0.76), 0.71 (95% CI, 0.69 to 0.74), 0.70 (95% CI, 0.66 to 0.75), and 0.78 (95% CI, 0.74 to 0.83) on the MGH, Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos test sets,

respectively, compared with a C-index of 0.62 (95% CI, 0.58 to 0.67) obtained by TC on the MGH test set.

The performance of Mirai on different subgroups of the Emory test set is shown in Appendix Table A6 (online only). Mirai obtained C-indices of 0.75 (95% CI, 0.71 to 0.78) and 0.79 (95% CI, 0.76 to 0.82) for African-American and White patients at Emory, respectively. The model obtained 1-year AUCs of 0.82 (95% CI, 0.78 to 0.85) and 0.85 (95% CI, 0.82 to 0.89) and 5-year AUCs of 0.75 (95% CI, 0.71 to 0.78) and 0.78 (95% CI, 0.75 to 0.82) for African-American and White patients, respectively. It obtained C-indices of 0.78 (95% CI, 0.72 to 0.83), 0.77 (95% CI, 0.74 to 0.80), and 0.74 (95% CI, 0.70 to 0.79) for patients younger than 50 years, between 50 and 70 years, and older than 70 years, respectively.

In evaluating the ability of Mirai to identify high-risk cohorts, we excluded positive examinations followed by a cancer diagnosis within 6 months and negative examinations without at least 5 years of screening follow-up. This resulted in cohorts of 9,284, 7,524, 8,640, 1,385, 7,194, 11,167, and 2,057 examinations from 3,957, 3,617, 5,774, 1,385, 5,707, 11,167, and 2,057 patients of which 441, 140, 632, 107, 869, 139, and 70 were followed by cancer within 5 years from MGH, Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. The performance of Mirai on these different cohorts is shown in Table 3. On the MGH test set, the Mirai at TC specificity guideline obtained a sensitivity of 39.7% (95% CI, 32.9 to 46.5) compared with a sensitivity of 22.9% (95% CI, 15.9 to 29.6) obtained by TC, yielding a significant improvement (P < .001). The Mirai at TC sensitivity obtained a specificity of 94.2% (95% CI, 93.4 to 94.9) compared with 85.4% (95% CI, 84.1 to 86.6) obtained by TC, yielding a significant improvement (P < .001). This performance was maintained across our other institutions. The Mirai at TC specificity guideline obtained sensitivities of 50.0% (95% CI, 38.5 to 61.4), 36.7% (95% CI, 31.6 to 41.8), 40.2% (95% CI, 30.9 to 49.3), 42.9% (95% CI, 38.5 to 47.0), 45.3% (95% CI,
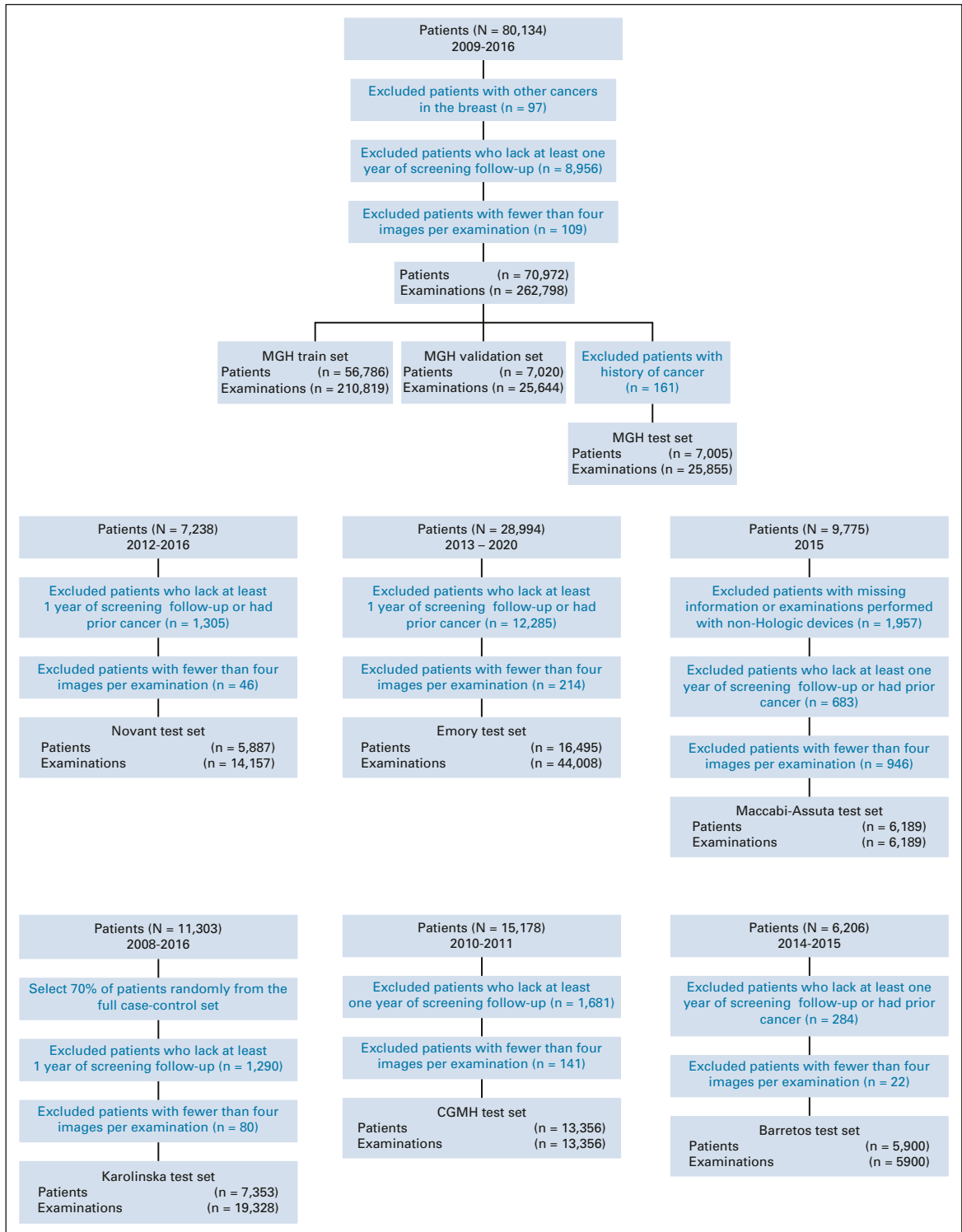
FIG 1. Data set construction flowcharts. CGMH, Chang Gung Memorial Hospital; MGH, Massachusetts General Hospital.

36.7 to 53.5), and 37.1% (95% CI, 25.6 to 48.1) at Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. The Mirai at TC sensitivity guideline obtained specificities of 95.4% (95% CI, 94.7 to 96.0), 91.5% (95% CI, 90.7 to 92.2), 92.5% (95% CI, 91.1 to 94.0), 94.3% (95% CI, 93.7 to 95.0), 94.8% (95% CI, 94.4 to 95.2), and

91.7% (95% CI, 90.51 to 92.92) at Novant, Emory, Maccabi-Assuta, Karolinska, CGMH, and Barretos, respectively. The Mirai receiver operating curves for selecting high-risk cohorts across all test sets are shown in Figure 2.

As shown in Table 4, we found that Mirai performed similarly across different race subgroups of the Emory test

TABLE 2. AUCs for Predicting Cancer in 1-5 Years and Uno's C-Index for Mirai on All Test Sets

Site	C-Index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
MGH, USA[4]	0.75 (0.72 to 0.78)	0.84 (0.80 to 0.87)	0.78 (0.75 to 0.82)	0.77 (0.74 to 0.80)	0.76 (0.73 to 0.79)	0.76 (0.73 to 0.79)
Novant, USA	0.75 (0.70 to 0.80)	0.78 (0.73 to 0.84)	0.76 (0.71 to 0.81)	0.76 (0.71 to 0.81)	0.75 (0.70 to 0.80)	0.75 (0.70 to 0.80)
Emory, USA	0.77 (0.75 to 0.79)	0.83 (0.81 to 0.86)	0.79 (0.77 to 0.82)	0.77 (0.75 to 0.80)	0.77 (0.75 to 0.79)	0.76 (0.74 to 0.79)
Maccabi-Assuta, Israel	0.77 (0.73 to 0.81)	0.86 (0.81 to 0.91)	0.81 (0.76 to 0.87)	0.79 (0.75 to 0.84)	0.77 (0.73 to 0.81)	0.75 (0.71 to 0.79)
Karolinska, Sweden[4]	0.81 (0.79 to 0.82)	0.90 (0.89 to 0.92)	0.86 (0.84 to 0.88)	0.82 (0.80 to 0.84)	0.80 (0.79 to 0.82)	0.78 (0.76 to 0.80)
CGMH, Taiwan[4]	0.79 (0.76 to 0.83)	0.90 (0.87 to 0.93)	0.86 (0.83 to 0.90)	0.82 (0.78 to 0.85)	0.80 (0.77 to 0.84)	0.79 (0.75 to 0.82)
Barretos, Brazil	0.84 (0.81 to 0.88)	0.89 (0.86 to 0.93)	0.87 (0.84 to 0.91)	0.86 (0.83 to 0.90)	0.85 (0.81 to 0.89)	0.82 (0.78 to 0.86)

NOTE. All metrics are followed by their 95% CI.

Abbreviations: AUC, area under the curve; CGMH, Chang Gung Memorial Hospital; MGH, Massachusetts General Hospital.

set. The Mirai at TC specificity guideline obtained sensitivities of 33.9% (95% CI, 26.3 to 41.0) and 40.0% (95% CI, 32.0 to 47.2) for African-American and White patients, respectively.

The Mirai at TC sensitivity guideline obtained specificities of 90.7% (95% CI, 89.6 to 91.9) and 91.9% (95% CI, 90.8 to 93.0) for African-American and White patients, respectively.

TABLE 3. High-Risk Cohort Analysis for All Test Sets

Method	Sensitivity, % (95% CI)	Specificity, % (95% CI)
MGH, USA: 9,284 examinations from 3,957 patients, 441 examinations followed by future cancer		
TC lifetime risk >20%[4]	22.9 (15.9 to 29.6)	85.4 (84.1 to 86.6)
Mirai at TC specificity[4]	39.7 (32.9 to 46.5)	85.2 (84.1 to 86.4)
Mirai at TC sensitivity	20.0 (14.2 to 25.1)	94.2 (93.4 to 94.9)
Novant, USA: 7,524 examinations from 3,617 patients, 140 examinations followed by future cancer		
Mirai at TC specificity	50.0 (38.5 to 61.4)	84.6 (83.3 to 85.7)
Mirai at TC sensitivity	23.6 (14.1 to 32.1)	95.4 (94.7 to 96.0)
Emory, USA: 8,640 examinations from 5,774 patients, 632 examinations followed by future cancer		
Mirai at TC specificity	36.7 (31.6 to 41.8)	84.9 (84.0 to 85.9)
Mirai at TC sensitivity	22.0 (17.4 to 26.3)	91.5 (90.7 to 92.2)
Maccabi-Assuta, Israel: 1,385 examinations from 1,385 patients, 107 examinations followed by future cancer		
Mirai at TC specificity	40.2 (30.9 to 49.3)	84.8 (82.9 to 86.8)
Mirai at TC sensitivity	22.4 (14.6 to 30.2)	92.5 (91.1 to 94.0)
Karolinska, Sweden: 7,194 examinations from 5,707 patients, 869 examinations followed by future cancer		
Mirai at TC specificity	42.9 (38.5 to 47.0)	85.0 (84.0 to 86.0)
Mirai at TC sensitivity	21.9 (18.4 to 25.2)	94.3 (93.7 to 95.0)
CGMH, Taiwan: 11,167 examinations from 11,167 patients, 139 examinations followed by future cancer		
Mirai at TC specificity	45.3 (36.7 to 53.5)	84.6 (83.9 to 85.2)
Mirai at TC sensitivity	23.0 (15.8 to 29.8)	94.8 (94.4 to 95.2)
Barretos, Brazil: 2,057 examinations from 2,057 patients, 70 examinations followed by future cancer		
Mirai at TC specificity	37.1 (25.6 to 48.1)	85.2 (83.6 to 86.8)
Mirai at TC sensitivity	21.4 (11.1 to 30.4)	91.7 (90.51 to 92.92)

NOTE. For each test set, we restricted our analysis to patients who were initially screening negative and had at least 5 years of screening follow-up. We defined an examination as screening negative if it was not followed by a cancer diagnosis within 6 months. We defined a future cancer as a pathology-confirmed breast cancer diagnosis within 5 years of the mammogram. Mirai thresholds (ie, at TC sensitivity and at TC specificity) were chosen to match the performance of the TC model on the development MGH set.

Abbreviations: CGMH, Chang Gung Memorial Hospital; MGH, Massachusetts General Hospital; TC, Tyrer-Cuzick.
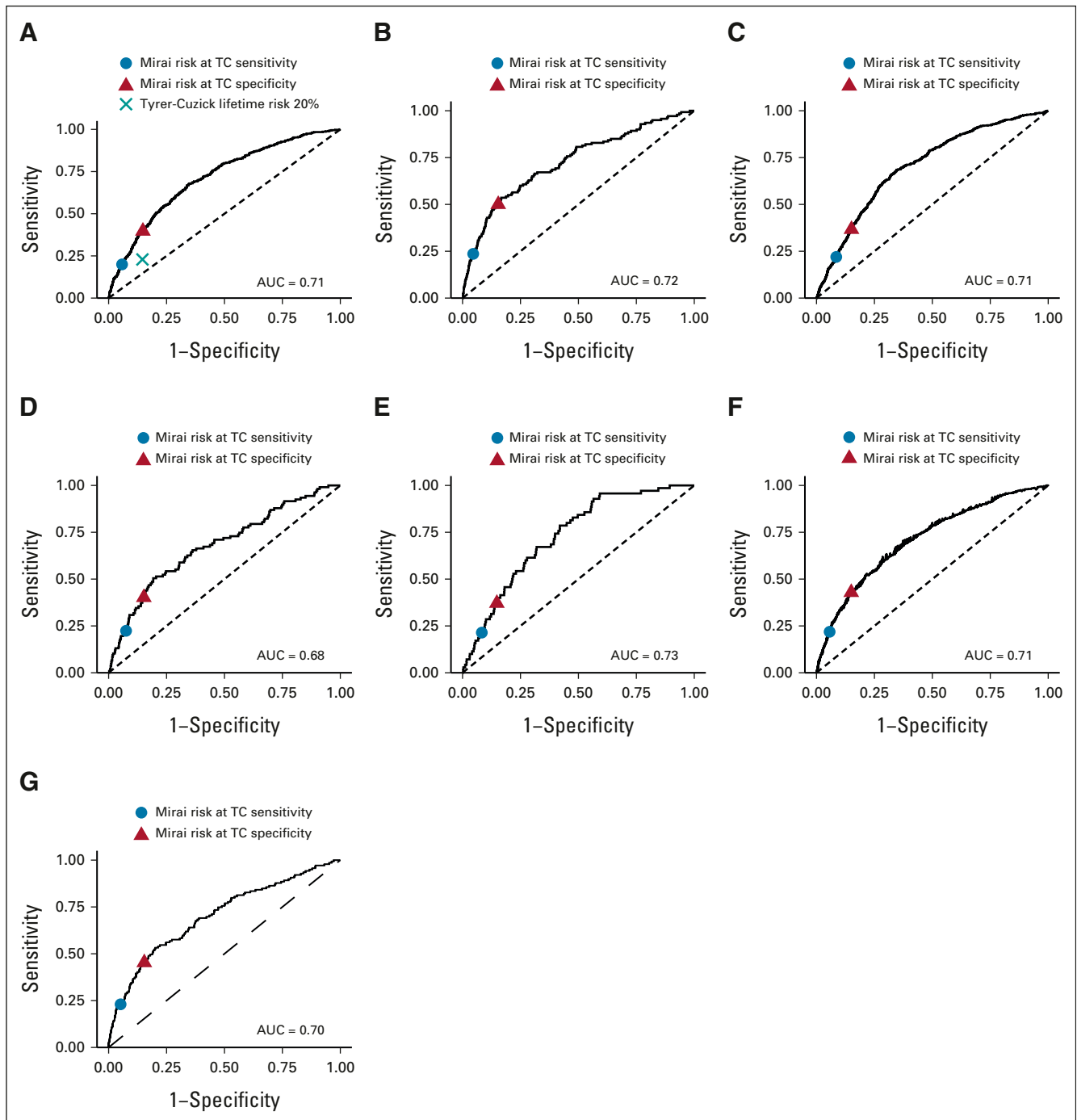
FIG 2. Receiver operating curves for Mirai in selecting high-risk cohorts across all test sets: (A) MGH, (B) Novant, (C) Emory, (D) Maccabi-Assuta, (E) Barretos, (F) Karolinska, and (G) CGMH. These data sets are restricted to include patients who were screening negative and either had cancer within 5 years or 5 years of negative follow-up. AUC, area under the curve; CGMH, Chang Gung Memorial Hospital; MGH, Massachusetts General Hospital.

DISCUSSION

Our study explored the robustness of an AI breast cancer risk model, Mirai, across globally diverse populations. We validated Mirai on test sets from seven hospitals across five countries. Across all test sets, we found that Mirai obtained the same C-index as on the MGH test set or higher, ranging from 0.75 at Novant to 0.84 at Barretos. The model obtained higher performance in hospitals with biennial screening such as Barretos than hospitals with annual screening such as MGH, because of differences in screening patterns. We demonstrated that Mirai can be used to accurately select high-risk cohorts across all test sets. Moreover, Mirai-based guidelines performed similarly across both African-American and White patients in the Emory test set.

TABLE 4. High-Risk Cohort Analysis for Subgroups of the Emory Data Set by Race

Method	Sensitivity, % (95% CI)	Specificity, % (95% CI)
African American: 4,311 examinations from 2,831 patients, 301 examinations followed by future cancer		
Mirai at TC specificity	33.9 (26.3 to 41.0)	83.7 (82.3 to 85.1)
Mirai at TC sensitivity	22.9 (16.2 to 29.3)	90.7 (89.6 to 91.9)
White: 3,728 examinations from 2,501 patients, 306 examinations followed by future cancer		
Mirai at TC specificity	40.0 (32.0 to 47.2)	85.5 (84.0 to 87.0)
Mirai at TC sensitivity	20.6 (14.6 to 26.2)	91.9 (90.8 to 93.0)

NOTE. We restricted our analysis to patients who were initially screening negative and had at least 5 years of screening follow-up. We defined an examination as screening negative if it was not followed by a cancer diagnosis within 6 months. We defined a future cancer as a pathology-confirmed breast cancer diagnosis within 5 years of the mammogram. Mirai thresholds (ie, at TC sensitivity and at TC specificity) were chosen to match the performance of the TC model on the development MGH set.

Abbreviations: MGH, Massachusetts General Hospital; TC, Tyrer-Cuzick.

Accurate short-term risk prediction (ie, within 5 years) is essential for early detection efforts in breast cancer. Traditional risk models, such as the TC model, are already widely implemented and support existing supplemental screening guidelines by the American Cancer Society, the American College of Radiology, and the National Comprehensive Cancer Network.[1,2,13-15] However, these models only provide a global risk prediction for large groups of patients, limiting their predictive accuracy for individuals and for specific time frames. Moreover, current guidelines for MRI eligibility[1,2] leverage lifetime TC risk, which ignores a patient's short-term risk of breast cancer and further limits the model's predictive utility. Our retrospective analysis across multiple test sets suggests that Mirai has the potential to replace current risk models (eg, TC) in guidelines for MRI screening, improving early detection and reducing overtreatment. For instance, we found that Mirai could obtain 70% relative improvement in sensitivity over the TC-based guideline at MGH while maintaining the same specificity. Moreover, we anticipate that AI models for breast cancer risk prediction will continue to improve, as risk models begin to leverage richer patient information like tomosynthesis. We expect that these algorithmic improvements will, in turn, yield further improved risk-based screening guidelines.

Our study had limitations. Our analysis of the benefit of different screening guidelines was retrospective. Prospective clinical trials are needed to confirm the clinical benefit of identifying improved high-risk cohorts using Mirai and to establish Mirai guidelines. Moreover, Mirai was only developed and tested using Hologic mammograms. Future work will be needed to test and adapt this technology to more mammography vendors and to tomosynthesis images. Moreover, although Mirai provides a risk assessment for cancer in either breast, it does not provide a risk estimate for each breast.

In conclusion, Mirai, a mammography-based risk model, maintained its accuracy across globally diverse test sets from MGH, USA; Novant, USA; Emory, USA; Maccabi-Assuta, Israel; Karolinska, Sweden; CGMH, Taiwan; and Barretos, Brazil. Moreover, guidelines based on Mirai significantly outperformed the existing clinical guidelines based on the TC model at MGH and maintained their performance across all test sets. This is the broadest validation to date of an AI-based breast cancer model and demonstrates that the technology can offer broad and equitable improvements in care. Prospective clinical trials of this technology are warranted.

AFFILIATIONS

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA

[2]Jameel Clinic, Massachusetts Institute of Technology, Cambridge, MA

[3]Breast Radiology Unit, Department of Imaging and Physiology, Karolinska University Hospital, Stockholm, Sweden

[4]Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden

[5]Department of Medical Imaging and Intervention, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan

[6]Department of Computer Science, University of California Los Angeles, Los Angeles, CA

[7]Department of Computer Science, Georgia Institute of Technology, Atlanta, GA

[8]Department of Biomedical Informatics, Emory University, Atlanta, GA

[9]Department of Radiology, Emory University, Atlanta, GA

[10]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA

[11]Division of Surgical Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA

[12]Analytics and Informatics Department, Novant Health, Winston-Salem, NC

[13]Digital Product and Services, Novant Health, Winston-Salem, NC

[14]Department of Cancer Prevention, Barretos Cancer Hospital, Barretos, Brazil

[15]Maccabitech, Maccabi Health Services, Tel Aviv, Israel

[16]Department of Imaging, Assuta Medical Centers, Tel Aviv, Israel

CORRESPONDING AUTHOR

Adam Yala, MEng, Jameel Clinic, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139; e-mail: adamyala@csail.mit.edu.

AUTHOR CONTRIBUTIONS

Conception and design: Adam Yala, Constance D. Lehman, Tal Patalon, Regina Barzilay
Financial support: Constance D. Lehman, Regina Barzilay

Administrative support: Constance D. Lehman, Bipin Karunakaran, Sharon Handelman-Gotlib, Michal Guindy
Provision of study materials or patients: Fredrik Strand, Hari Trivedi, Constance D. Lehman, Bipin Karunakaran, Thiago B. Silva, Renato F. Caron, Sharon Handelman-Gotlib, Michal Guindy
Collection and assembly of data: Adam Yala, Fredrik Strand, Gigin Lin, Siddharth Satuluru, Thomas Kim, Imon Banerjee, Judy Gichoya, Hari Trivedi, Constance D. Lehman, David J. Sheedy, Lisa M. Matthis, Karen E. Hegarty, Silvia Sabino, Thiago B. Silva, Maria C. Evangelista, Renato F. Caron, Bruno Souza, Edmundo C. Mauad, Tal Patalon, Sharon Handelman-Gotlib, Michal Guindy
Data analysis and interpretation: Adam Yala, Peter G. Mikhael, Fredrik Strand, Imon Banerjee, Hari Trivedi, Constance D. Lehman, Kevin Hughes, Bipin Karunakaran, Thiago B. Silva, Regina Barzilay
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

REFERENCES

1. Smith RA, Andrews KS, Brooks D, et al: 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

2. Bevers TB, Ward JH, Arun BK, et al: Breast cancer risk reduction, version 2. 2015. J Natl Compr Cancer Netw 13:880-915, 2015

3. Tyrer J, Duffy SW, Cuzick J: A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 23:1111-1130, 2004

4. Yala A, Mikhael PG, Strand F, et al: Toward robust mammography-based models for breast cancer risk. Sci Transl Med 13:eaba4373, 2021

5. Gail MH, Costantino JP, Pee D, et al: Projecting individualized absolute invasive breast cancer risk in African American women. J Natl Cancer Inst 99: 1782-1792, 2007

6. Matsuno RK, Costantino JP, Ziegler RG, et al: Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American Women. J Natl Cancer Inst 103:951-961, 2011

7. Boggs DA, Rosenberg L, Adams-Campbell LL, et al: Prospective approach to breast cancer risk prediction in African American women: The black women's health study model. J Clin Oncol 33:1038-1044, 2015

8. Mårtensson G, Ferreira D, Granberg T, et al: The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. Med Image Anal 66:101714, 2020

9. AlBadawy EA, Saha A, Mazurowski MA: Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. Med Phys 45: 1150-1158, 2018

10. Zech JR, Badgeley MA, Liu M, et al: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross sectional study. PLoS Med 15:e1002683, 2018

11. Dembrower K, Lindholm P, Strand F: A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks-the cohort of screenaged women (CSAW). J Digit Imaging 33:408-413, 2020

12. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30: 1105-1117, 2011

13. Monticciolo DL, Newell MS, Hendrick RE, et al: Breast cancer screening for average-risk women: Recommendations from the ACR commission on breast imaging. J Am Coll Radiol 14:1137-1143, 2017

14. Monticciolo DL, Newell MS, Moy L, et al: Breast cancer screening in women at higher-than-average risk: Recommendations from the ACR. J Am Coll Radiol 15: 408-414, 2018

15. Bakker MF, de Lange SV, Pijnappel RM, et al: Supplemental MRI screening for women with extremely dense breast tissue. N Engl J Med 381:2091-2102, 2019

- - -

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/authors/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Adam Yala
Honoraria: Sanofi
Consulting or Advisory Role: Janssen Research & Development

Peter G. Mikhael
Consulting or Advisory Role: Outcomes4Me

Fredrik Strand
Honoraria: Lunit
Uncompensated Relationships: Lunit

Gigin Lin
Research Funding: Quanta Computer (Inst)

Siddharth Satuluru
Employment: Amerimed EMS
Stock and Other Ownership Interests: Sorrento Therapeutics (I), Moderna Therapeutics (I)

Constance D. Lehman
Honoraria: GE Healthcare
Consulting or Advisory Role: GE Healthcare, Clairity, Inc
Research Funding: GE Healthcare, Hologic
Travel, Accommodations, Expenses: Clairity, Inc
Other Relationship: Clairity, Inc

Kevin Hughes
Stock and Other Ownership Interests: CRA Health
Honoraria: Hologic, Myriad Genetics
Consulting or Advisory Role: Targeted Medical Education, Inc, MedNeon,
Other Relationship: Ask2Me.Org

Regina Barzilay
Consulting or Advisory Role: J&J, Bayer, Moderna Therapeutics, Amgen, Vertex
Research Funding: Bayer
Travel, Accommodations, Expenses: J&J, Bayer, Novo Nordisk

No other potential conflicts of interest were reported.

APPENDIX

TABLE A1. Detailed Demographics of the MGH Test Set

Characteristic	MGH Test Set	
	All, No. (%)	Cancer, No. (%)
All examinations	25,855 (100.0)	588 (100.0)
Age, years		
< 40	724 (2.8)	7 (1.19)
40-50	7,025 (27.17)	95 (16.16)
50-60	7,829 (30.28)	188 (31.97)
60-70	6,708 (25.94)	182 (30.95)
70-80	3,001 (11.61)	94 (15.99)
> 80	568 (2.2)	22 (3.74)
Density		
Almost entirely fatty	2,474 (9.57)	31 (5.27)
Scattered areas of fibroglandular tissue	12,490 (48.31)	264 (44.9)
Heterogeneously dense	9,751 (37.71)	271 (46.09)
Extremely dense	1,129 (4.37)	22 (3.74)
BI-RADS		
0—additional imaging needed	1,785 (6.9)	186 (31.63)
1—negative or 2—benign	24,043 (93.0)	400 (68.03)
Others	4 (0.01)	1 (0.17)
Race		
White	21,006 (81.25)	512 (87.07)
African American	1,204 (4.66)	21 (3.57)
Asian or Pacific Islander	1,238 (4.79)	26 (4.42)
Hispanic	225 (0.87)	6 (1.02)
Other race	1,486 (5.75)	15 (2.55)
Device		
Lorad Selenia	9,937 (38.43)	241 (40.99)
Selenia Dimensions	15,882 (61.43)	311 (52.89)
Unknown	36 (0.14)	36 (6.12)

Abbreviations: BI-RADS, Breast Imaging-Reporting and Data System; MGH, Massachusetts General Hospital.

TABLE A2. Detailed Demographics of the Novant Test Set

Characteristic	Novant Data Set	
	All, No. (%)	Cancer, No. (%)
All examinations	14,157 (100.0)	235 (100.0)
Age, years		
40-50	3,917 (27.67)	53 (22.55)
50-60	5,368 (37.92)	65 (27.66)
60-70	4,872 (34.41)	117 (49.79)
Race		
White	10,555 (74.56)	185 (78.72)
African American	2,687 (18.98)	44 (18.72)
Asian	220 (1.55)	0 (0.0)
Hispanic	391 (2.76)	5 (2.13)
American Indian or Alaskan Native	28 (0.2)	0 (0.0)
Time to cancer, years		
0-1	95 (0.67)	95 (40.43)
1-2	52 (0.37)	52 (22.13)
2-3	48 (0.34)	48 (20.43)
3-4	31 (0.22)	31 (13.19)
4-5	9 (0.06)	9 (3.83)

TABLE A3. Detailed Demographics of the Emory Test Set

	Emory Data Set	
Characteristic	All, No. (%)	Cancer, No. (%)
All examinations	44,008 (100.0)	1,003 (100.0)
Age, years		
< 40	410 (0.93)	11 (1.1)
40-50	9,047 (20.56)	147 (14.66)
50-60	12,113 (27.52)	235 (23.43)
60-70	13,182 (29.95)	302 (30.11)
70-80	7,638 (17.36)	285 (28.41)
> 80	1,495 (3.4)	23 (2.29)
Race		
White	19,587 (44.51)	492 (49.05)
African American	19,918 (45.26)	463 (46.16)
Asian	2,251 (5.11)	27 (2.69)
Native Hawaiian or Other Pacific Islander	296 (0.67)	7 (0.7)
American Indian or Alaskan Native	47 (0.11)	0 (0.0)
Multiple	103 (0.23)	0 (0.0)
Time to cancer, years		
0-1	410 (0.93)	410 (40.88)
1-2	220 (0.5)	220 (21.93)
2-3	177 (0.4)	177 (17.65)
3-4	113 (0.26)	113 (11.27)
4-5	83 (0.19)	83 (8.28)

TABLE A4. No. of Patients Followed by the No. of Examinations Used to Compute Yearly AUC Values

Site	No. of Patients and Examinations for Computing 1-Year AUC	No. of Patients and Examinations for Computing 2-Year AUC	No. of Patients and Examinations for Computing 3-Year AUC	No. of Patients and Examinations for Computing 4-Year AUC	No. of Patients and Examinations for Computing 5-Year AUC
MGH, USA[4]	7,005 (173) 25,855 (173)	6,258 (211) 21,707 (301)	5,442 (225) 17,003 (424)	4,685 (228) 12,949 (520)	4,002 (233) 9,431 (588)
Novant, USA	5,887 (95) 14,157 (95)	5,475 (108) 13,264 (147)	5,018 (118) 12,212 (195)	4,459 (122) 10,529 (226)	3,659 (123) 7,619 (235)
Emory, USA	16,495 (406) 44,008 (410)	13,141 (457) 31,862 (630)	10,532 (482) 22,460 (807)	8,098 (490) 14,864 (920)	5,937 (495) 9,011 (1,003)
Maccabi-Assuta, Israel	6,189 (79) 6,189 (79)	6,079 (101) 6,079 (101)	5,609 (147) 5,609 (147)	4,551 (165) 4,551 (165)	1,464 (186) 1,464 (186)
Karolinska, Sweden[4]	7,353 (517) 19,328 (517)	7,148 (634) 16,665 (681)	6,773 (747) 13,554 (1,040)	6,379 (771) 10,618 (1,181)	5,931 (799) 7,711 (1,413)
CGMH, Taiwan[4]	13,356 (116) 13,356 (116)	12,895 (141) 12,895 (141)	12,390 (182) 12,390 (182)	11,840 (212) 11,840 (212)	11,272 (244) 11,272 (244)
Barretos, Brazil	5,900 (84) 5,900 (84)	5,394 (100) 5,394 (100)	4,569 (114) 4,569 (114)	3,689 (128) 3,689 (128)	2,133 (146) 2,133 (146)

NOTE. Values in parentheses are the number of patients and examinations developing cancer within that number of years following an examination.
Abbreviations: AUC, area under the curve; CGMH, Chang Gung Memorial Hospital; MGH, Massachusetts General Hospital.

TABLE A5. AUCs for Predicting Cancer in 1-5 Years and Uno's C-Index for Mirai on All Test Sets Excluding Cancers Diagnosed With 6 Months of the Mammogram

Site	C-Index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
MGH, USA[4]	0.69 (0.66 to 0.73)	0.71 (0.60 to 0.84)	0.71 (0.66 to 0.76)	0.71 (0.67 to 0.75)	0.71 (0.67 to 0.75)	0.71 (0.68 to 0.75)
Novant, USA	0.72 (0.66 to 0.79)	NA	0.71 (0.63 to 0.80)	0.73 (0.66 to 0.80)	0.72 (0.65 to 0.79)	0.72 (0.66 to 0.79)
Emory, USA	0.69 (0.66 to 0.72)	0.74 (0.66 to 0.84)	0.71 (0.68 to 0.75)	0.70 (0.67 to 0.73)	0.71 (0.68 to 0.74)	0.71 (0.68 to 0.74)
Maccabi-Assuta, Israel	0.70 (0.64 to 0.76)	NA	0.67 (0.53 to 0.83)	0.72 (0.66 to 0.79)	0.70 (0.63 to 0.76)	0.68 (0.62 to 0.74)
Karolinska, Sweden[4]	0.71 (0.69 to 0.74)	NA	0.72 (0.67 to 0.77)	0.73 (0.71 to 0.76)	0.73 (0.70 to 0.75)	0.71 (0.69 to 0.73)
CGMH, Taiwan[4]	0.70 (0.66 to 0.75)	0.84 (0.72 to 0.99)	0.76 (0.68 to 0.84)	0.71 (0.64 to 0.77)	0.71 (0.66 to 0.76)	0.70 (0.66 to 0.75)
Barretos, Brazil	0.78 (0.74 to 0.83)	0.87 (0.80 to 0.94)	0.82 (0.76 to 0.89)	0.81 (0.76 to 0.87)	0.79 (0.74 to 0.84)	0.75 (0.70 to 0.80)

NOTE. All metrics are followed by their 95% CI.

Abbreviations: AUC, area under the curve; C-index, concordance index; CGMH, Chang Gung Memorial Hospital; MGH, Massachusetts General Hospital; NA, not available.

TABLE A6. ROC AUCs and Uno's C-Index for Mirai for Subgroups of the Emory Test Set by Race and Age

Subgroup	C-Index	1-Year AUC	2-Year AUC	3-Year AUC	4-Year AUC	5-Year AUC
Race						
African American	0.75 (0.71 to 0.78)	0.82 (0.78 to 0.85)	0.77 (0.74 to 0.81)	0.74 (0.71 to 0.78)	0.75 (0.71 to 0.78)	0.75 (0.71 to 0.78)
White	0.79 (0.76 to 0.82)	0.85 (0.82 to 0.89)	0.81 (0.78 to 0.84)	0.80 (0.77 to 0.83)	0.79 (0.76 to 0.82)	0.78 (0.75 to 0.82)
Age, years						
< 50	0.78 (0.72 to 0.83)	0.85 (0.79 to 0.91)	0.80 (0.74 to 0.87)	0.78 (0.72 to 0.85)	0.76 (0.70 to 0.82)	0.75 (0.69 to 0.81)
50-70	0.77 (0.74 to 0.80)	0.82 (0.79 to 0.86)	0.79 (0.76 to 0.82)	0.77 (0.74 to 0.80)	0.77 (0.74 to 0.80)	0.77 (0.74 to 0.80)
> 70	0.74 (0.70 to 0.79)	0.83 (0.79 to 0.87)	0.76 (0.72 to 0.81)	0.74 (0.70 to 0.79)	0.73 (0.68 to 0.78)	0.73 (0.68 to 0.78)

Abbreviations: AUC, area under the curve; C-index, concordance index.