

# UCSF

## UC San Francisco Previously Published Works

### Title

Comparison of Mammography AI Algorithms with a Clinical Risk Model for 5-year Breast Cancer Risk Prediction: An Observational Study.

### Permalink

<https://escholarship.org/uc/item/8mp5r66q>

### Journal

Radiology, 307(5)

### Authors

Arasu, Vignesh

Habel, Laurel

Achacoso, Ninah

et al.

### Publication Date

2023-06-01

### DOI

10.1148/radiol.222733

Peer reviewed


# Comparison of Mammography AI Algorithms with a Clinical Risk Model for 5-year Breast Cancer Risk Prediction: An Observational Study

Vignesh A. Arasu, MD, PhD • Laurel A. Habel, PhD • Ninah S. Achacoso, MS • Diana S. M. Buist, PhD • Jason B. Cord, MD • Laura J. Esserman, MD • Nola M. Hylton, PhD • M. Maria Glymour, ScD • John Kornak, PhD • Lawrence H. Kushi, ScD • Donald A. Lewis, MS • Vincent X. Liu, MD • Caitlin M. Lydon, MPH • Diana L. Miglioretti, PhD • Daniel A. Navarro, MD • Albert Pu, MS • Li Shen, PhD • Weiva Sieh, MD, PhD • Hyo-Chun Yoon, MD, PhD • Catherine Lee, PhD

From the Division of Research, Kaiser Permanente Northern California, 2000 Broadway, Oakland, CA 94612 (V.A.A., L.A.H., N.S.A., L.H.K., V.X.L., C.M.L., C.L.); Department of Radiology, Kaiser Permanente Northern California, Vallejo Medical Center, Vallejo, Calif (V.A.A.); Kaiser Permanente Washington Health Research Institute, Seattle, Wash (D.S.M.B.); Department of Radiology, Southern California Permanente Medical Group, Orange County, Irvine, Calif (J.B.C.); Department of Surgery (L.J.E.), Department of Radiology and Biomedical Imaging (N.M.H.), and Department of Epidemiology and Biostatistics (M.M.G., J.K.), University of California–San Francisco, San Francisco, Calif; Department of Medical Imaging Technology and Informatics, Southern California Permanente Medical Group, Pasadena, Calif (D.A.L.); Department of Biostatistics, University of California–Davis, Davis, Calif (D.L.M.); The Technology Group, The Permanente Medical Group, Oakland, Calif (D.A.N.); KP Information Technology, Kaiser Foundation Health Plan Inc and Kaiser Foundation Hospitals, Oakland, Calif (A.P.); Department of Artificial Intelligence and Human Health and Nash Family Department of Neuroscience (L.S.) and Department of Population Health Science and Policy, Department of Genetics and Genomic Sciences (W.S.), Icahn School of Medicine at Mount Sinai, New York, NY; and Department of Radiology, Hawaii Permanente Medical Group, Moanalua Medical Center, Honolulu, Hawaii (H.C.Y.). Received October 26, 2022; revision requested December 7; revision received April 5, 2023; accepted April 18. **Address correspondence to** V.A.A. (email: [Vignesh.a.arasu@kp.org](mailto:Vignesh.a.arasu@kp.org)).

Supported by the Permanente Medical Group (TPMG) Delivery Science and Applied Research Physician Researcher Program and the National Cancer Institute (R01CA264987). N.M.H. supported by grant from National Institutes of Health (U01 CA225427).

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

Radiology 2023; 307(5):e222733 • <https://doi.org/10.1148/radiol.222733> • Content codes: 

**Background:** Although several clinical breast cancer risk models are used to guide screening and prevention, they have only moderate discrimination.

**Purpose:** To compare selected existing mammography artificial intelligence (AI) algorithms and the Breast Cancer Surveillance Consortium (BCSC) risk model for prediction of 5-year risk.

**Materials and Methods:** This retrospective case-cohort study included data in women with a negative screening mammographic examination (no visible evidence of cancer) in 2016, who were followed until 2021 at Kaiser Permanente Northern California. Women with prior breast cancer or a highly penetrant gene mutation were excluded. Of the 324 009 eligible women, a random subcohort was selected, regardless of cancer status, to which all additional patients with breast cancer were added. The index screening mammographic examination was used as input for five AI algorithms to generate continuous scores that were compared with the BCSC clinical risk score. Risk estimates for incident breast cancer 0 to 5 years after the initial mammographic examination were calculated using a time-dependent area under the receiver operating characteristic curve (AUC).

**Results:** The subcohort included 13 628 patients, of whom 193 had incident cancer. Incident cancers in eligible patients (additional 4391 of 324 009) were also included. For incident cancers at 0 to 5 years, the time-dependent AUC for BCSC was 0.61 (95% CI: 0.60, 0.62). AI algorithms had higher time-dependent AUCs than did BCSC, ranging from 0.63 to 0.67 (Bonferroni-adjusted  $P < .0016$ ). Time-dependent AUCs for combined BCSC and AI models were slightly higher than AI alone (AI with BCSC time-dependent AUC range, 0.66–0.68; Bonferroni-adjusted  $P < .0016$ ).

**Conclusion:** When using a negative screening examination, AI algorithms performed better than the BCSC risk model for predicting breast cancer risk at 0 to 5 years. Combined AI and BCSC models further improved prediction.

© RSNA, 2023

*Supplemental material is available for this article.*

Breast cancer risk models are used to evaluate and guide clinical considerations such as hereditary risk, supplemental screening, and risk-reducing medications (1). Risk models are also undergoing active investigation for broader management in the population, such as risk-based personalized screening (2,3) or capacity management (4). Several models have been developed to assess the risk for breast cancer in the general population, including Breast Cancer Risk Assessment Tool (Gail Model; 5), Breast Cancer Surveillance Consortium (BCSC; 6,7), and International Breast Cancer Intervention Study (Tyrer-Cuzick Risk

Model; 8). These models include age, clinical factors (eg, family history of breast cancer, race and/or ethnicity, and previous breast biopsy with benign results), genetic factors, and mammographic breast density but have only moderate discrimination for predicting 5- or 10-year risk of breast cancer (area under the receiver operating characteristic curve [AUC] range, 0.62–0.66) (5–8).

Computer vision–based artificial intelligence (AI) models can potentially improve risk prediction beyond clinical risk factors. These models quantitatively extract imaging biomarkers that represent underlying pathophysiologic

## Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, BCSC = Breast Cancer Surveillance Consortium, BI-RADS = Breast Imaging Reporting and Data System

## Summary

Negative screening mammographic examinations were analyzed with five artificial intelligence (AI) algorithms; all predicted breast cancer risk to 5 years better than the Breast Cancer Surveillance Consortium (BCSC) clinical risk model, and combining AI and BCSC models further improved prediction.

## Key Results

- Five artificial intelligence (AI) algorithms were used to generate continuous risk scores from retrospectively acquired screening mammographic examinations negative for cancer in 18 019 women.
- AI predicted incident cancers at 0 to 5 years better than the Breast Cancer Surveillance Consortium (BCSC) clinical risk model (AI time-dependent area under the receiver operating characteristic curve [AUC] range, 0.63–0.67; BCSC time-dependent AUC, 0.61; Bonferroni-adjusted  $P < .0016$ ).
- Combining AI algorithms with BCSC slightly improved the time-dependent AUC versus AI alone (AI with BCSC time-dependent AUC range, 0.66–0.68; Bonferroni-adjusted  $P < .0016$ ).

mechanisms and phenotypes (9). Breast density is the imaging biomarker most commonly incorporated into clinical risk models, but recent advances in AI deep learning (10) provide the ability to extract hundreds to thousands of additional mammographic features. However, most mammography-based AI algorithms have only been trained to assist radiologists by flagging cancer visible at screening mammography (computer-aided diagnosis or computer-aided detection) and not to predict future risk several years after mammography with negative results (11). A few studies (12,13) have evaluated the ability of mammography-trained AI algorithms to predict future risk of breast cancer, which demonstrated substantial improvements in risk prediction versus clinical risk models alone. To our knowledge, it is unknown whether currently available computer-aided detection or diagnosis AI algorithms trained for shorter time horizons (ie, the time over which risk is assessed) and representing the majority of mammography AI algorithms can also predict longer-term risk. The ability for computer-aided detection or diagnosis to provide personalized future risk prediction would expand the applications into the realm of breast cancer risk models.

This study used screening mammography negative for breast cancer at final assessment from a large community-based cohort in the United States to compare five commercial and academic mammography AI algorithms with each other and with the BCSC clinical model. This study also assessed whether combining the AI and BCSC risk models improved risk prediction compared with either model type alone.

## Materials and Methods

### Study Cohort and Design

This is a retrospective case-cohort study of women who had a bilateral screening mammographic examination (two-dimensional

digital mammography) in 2016 at Kaiser Permanente Northern California (ie, the index mammogram) that was negative at final imaging assessment. Specifically, screening mammographic examinations were selected if they were assessed with screening Breast Imaging Reporting and Data System (BI-RADS) (14) as category 1 or 2, or a screening BI-RADS 0 and diagnostic BI-RADS 1 or 2 in 90 days or less, or a screening BI-RADS 0 and diagnostic BI-RADS 4 or 5 and radiologic-pathologic concordant benign biopsy in 90 days or less. Patients were excluded if they had a history of breast cancer or a high-penetrance breast cancer susceptibility gene as defined by the National Comprehensive Cancer Network guidelines (15). A case-cohort study design was used, which is a hybrid of the cohort and case-control study designs and has the advantage of allowing direct unbiased estimate of cumulative incidence (or absolute risk) and analysis of multiple outcomes, similar to a cohort study (16). This study was approved by the Kaiser Permanente Northern California institutional review board for Health Insurance Portability and Accountability Act compliance and followed the Strengthening the Reporting of Observational Studies in Epidemiology guidelines (17,18).

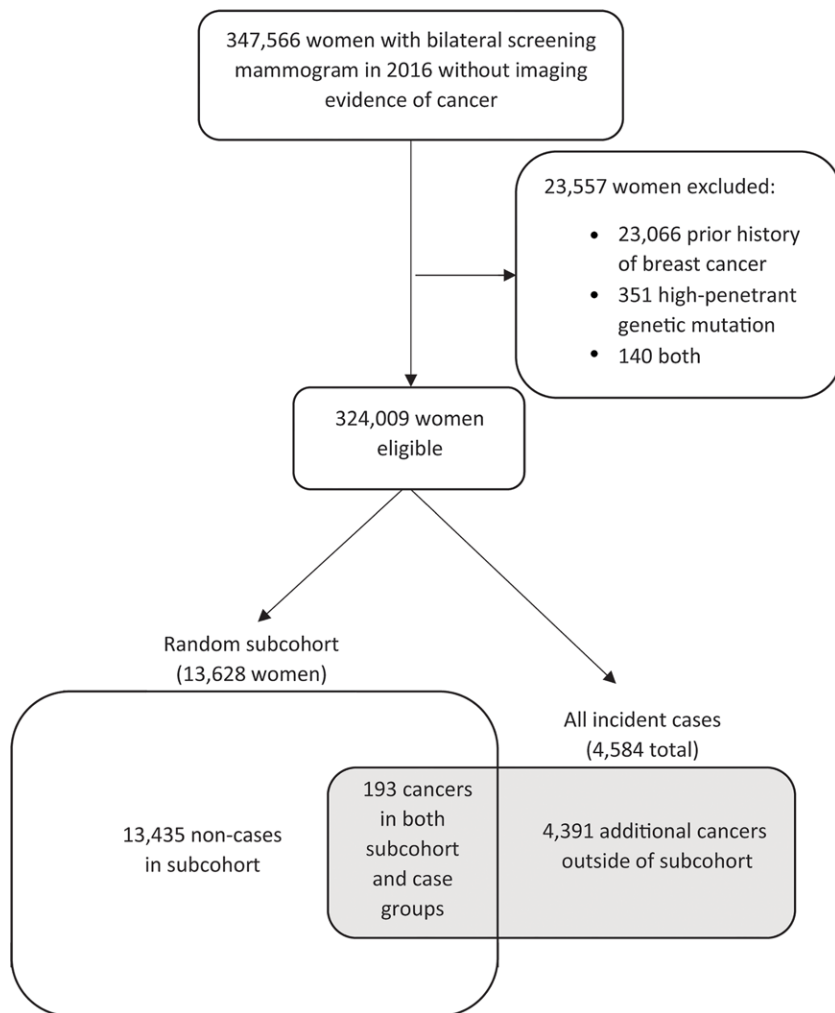
### Data Collection and Imaging Procedures

Screening mammographic examinations in 2016 were identified by Current Procedural Terminology examination code 77057. Incident breast cancer, detected either symptomatically or on a subsequent mammogram, was defined as pathology-confirmed invasive carcinoma or ductal carcinoma in situ. Cancers were confirmed by using the Kaiser Permanente Northern California Breast Cancer Tracking System (19) quality assurance program. This tracking system has a 99.8% concordance with the Kaiser Permanente Northern California tumor registry that reports to the National Cancer Institute Surveillance, Epidemiology, and End Results program, but identifies incident cancers more rapidly (within 1 month of diagnosis) while using manual verification. Women were followed from their index mammogram to date of breast cancer diagnosis, death, health plan disenrollment (allowing up to a 3-month gap in enrollment), or the end of the study (August 31, 2021), whichever occurred first.

The full-field digital mammograms were evaluated in their archived processed form.

### Deriving AI Risk Score from Screening Mammograms

AI scores were generated from five deep-learning computer vision algorithms that use screening mammograms as their input and then produce patient-level predicted scores. Candidate algorithms were chosen from an ongoing institutional AI operational evaluation. Further details on the AI algorithms and the underlying architecture are available in Appendix S1. Briefly, this study evaluated two academic algorithms freely available for research, Mirai (13) and Globally-Aware Multiple Instance Classifier (20), and three commercially available algorithms, MammoScreen (21), ProFound AI (22), and Mia (23). Because computer-aided detection or diagnosis algorithms themselves are trained at various time horizons between 3 months and 2 years, each algorithm's trained time horizon and their ability to predict future risk up to 5 years were displayed. When any algorithm



**Figure 1:** Patient selection flowchart. No imaging evidence of cancer: Screening examination Breast Imaging Reporting and Data System (BI-RADS) 1 or 2, or screening BI-RADS 0 and diagnostic BI-RADS 1 or 2 in 90 days or fewer, or screening BI-RADS 0 and diagnostic BI-RADS 4 or 5 and benign biopsy in 90 days or fewer.

failed to process an individual mammogram, the missing score was imputed using the algorithm's specific overall median score (missing data by algorithm are in Table S1; evaluation by complete scored mammograms is in Table S2).

The ability of the models to predict 5-year breast cancer risk was divided into three periods after the index screening mammographic examination: interval cancer risk was defined as incident cancers diagnosed between 0 and 1 years, future cancer risk was defined as incident cancers diagnosed from at least 1 to 5 years, and all cancer risk was defined as incident cancers diagnosed between 0 and 5 years.

### BCSC Clinical Risk Score Generation

The BCSC clinical 5-year risk model version 2 (6,24) was used as the comparator for the AI models. The BCSC model predicts risk for women without a history of breast cancer or *BRCA1/2* mutation based on age, ethnicity, first-degree family history of breast cancer, prior benign breast biopsy, and mammographic breast density. For risk score generation, clinical data from at or before the first screening mammographic examination in 2016

were obtained from the Kaiser Permanente Northern California electronic health record, regardless of prior membership in the Kaiser Permanente Northern California health system. Breast density was based on the clinical interpretation of the index mammogram using the BI-RADS classification system. Whereas the Breast Cancer Tracking System database prospectively classifies atypia and lobular carcinoma in situ, it does not distinguish proliferative benign pathology from otherwise benign pathology, so these outcomes were conservatively classified as nonproliferative lesions.

### Statistical Analysis

Statistical software (R version 4.0.2, R Program for Statistical Computing; 25) was used for all statistical analyses (C.L.). All statistical tests were two sided, with a threshold for statistical significance using a Bonferroni correction of the significance level for the 30 tests performed for a threshold  $\alpha$  level of  $.05/30 = .0016$ . Therefore, estimated differences in AUCs with  $P < .0016$  indicated statistical significance after accounting for multiple comparisons.

Kaplan-Meier was used to estimate the overall 5-year cumulative incidence of breast cancer within strata of each risk score (>90th percentile, middle 80 percentiles, and <10th percentile) as hypothetical thresholds for risk groups. Design weights were included for case-cohort sampling. Model performance was evaluated using the time-dependent AUC, for the dynamic definition of patients with or without breast cancer at any given

time when handling time-to-event outcomes (26), and for censoring and sampling distribution using inverse probability of censoring weights and case-cohort sampling (27). Corresponding 95% CIs were obtained using bootstrapping with 1000 bootstrap samples (28). To compare time-dependent AUC estimates from two separate risk scores, the difference in estimates and corresponding bootstrapped 95% CI was calculated. CIs that did not include 0 indicated that the difference in time-dependent AUC estimates was statistically significant ( $\alpha < .05$ ) (29).

A Cox model was fitted to predict 5-year risk by using the combined AI-predicted score and the BCSC score. The Cox models accounted for the case-cohort sampling with design weights and included both the AI score and BCSC score flexibly by using restricted cubic splines with four knots (30,31). Five-fold cross-validation was used to estimate the time-dependent AUC estimator (27) and presented the average value across the five folds. Corresponding 95% CIs for the average cross-validation–time-dependent AUC were obtained through bootstrapping with 1000 bootstrap samples. Time-dependent AUC was

the outcome for comparison. Based on the number of patients with and without cancer, the statistical power for the smallest detectable improvement with the AI model would be .02 compared with reference BCSC (AUC = .60), assuming 80% power,  $\alpha$  level of .05, and two-sided tests.

Similar analyses were performed in post hoc subgroups: patients with invasive cancer or ductal carcinoma in situ, patients with complete scores available across all models, patients with BI-RADS 1 or 2 results on screening mammograms, patients with BI-RADS 0 results on screening mammograms, BI-RADS 1 or 2 results at diagnostic imaging or biopsy, and mammograms acquired on equipment manufactured by Hologic or GE Healthcare.

This study assessed the 5-year calibration (ie, the ratio of expected values to observed values) of Mirai and BCSC risk within pre-specified strata of 5-year risk based on thresholds established by BCSC. The observed number of patients with cancer during the 5-year study period was compared with the expected number of patients with cancer by calculating the total of the cumulative hazard estimates over all individuals in the study (32). The ratio of observed to expected cases was reported with exact 95% CIs (33). The incidence rates (cases per 1000 person-years) and incident rate ratios with 95% CIs were calculated based on a Poisson distribution. All expected incidence estimates incorporated design weights that accounted for the case-cohort sampling.

## Results

### Characteristics of the Study Sample

Figure 1 shows the patient selection process for the case-cohort design. Of 347 566 women with a negative screening mammographic examination in 2016, 23 557 were excluded. Of 324 009 women who met eligibility criteria, a simple random cohort of 13 628 women (4.2%) including 193 women with incident breast cancer was selected for analyses. An additional 4391 patients from the complete cohort who were diagnosed with cancer within 5 years of the index mammography in 2016 were also included (4584 total patients; 100%). This sample size was based on the maximum cohort size feasible for AI algorithm evaluation with the resources available. Women younger than 50 years made up 23.3% (3170 of 13 628) of this group, and 51.1% (6970 of 13 628) were non-Hispanic White women (Table 1). Median follow-up was 5.0 years (IQR, 4.7–5.3). Of the 13 435

**Table 1: Cohort Characteristics**

Parameter	Patients in Subcohort ( <i>n</i> = 13 628)	Patients with Breast Cancer* ( <i>n</i> = 4584)	All Eligible Patients ( <i>n</i> = 324 009)
<b>Age (y)</b>			
<40	84 (1)	19 (< 1)	1991 (1)
40–49	3086 (23)	713 (16)	73 331 (23)
50–59	4694 (34)	1305 (28)	112 716 (35)
60–69	4147 (30)	1777 (39)	97 746 (30)
≥70	1617 (12)	770 (17)	38 225 (12)
<b>Race/ethnicity</b>			
Asian or Pacific Islander	2557 (19)	861 (19)	60 732 (19)
Black or non-Hispanic	975 (7)	327 (7)	23 513 (7)
Hispanic	2350 (17)	561 (12)	56 440 (17)
Multiracial	493 (4)	158 (3)	10 826 (3)
Native American	51 (< 1)	18 (< 1)	1364 (< 1)
White, non-Hispanic	6970 (51)	2643 (58)	166 014 (51)
Missing	232 (2)	16 (< 1)	5120 (2)
<b>No. of first-degree relatives with history of breast cancer</b>			
0	11 920 (87)	3676 (80)	283 147 (87)
1	1620 (12)	853 (19)	38 854 (12)
≥ 2	88 (1)	55 (1)	2008 (1)
<b>No. previous benign breast biopsies</b>			
0	12 871 (94)	4076 (89)	305 185 (94)
≥1	757 (6)	508 (11)	18 824 (6)
<b>BI-RADS breast density</b>			
Almost entirely fat	1324 (10)	244 (5)	30 499 (9)
Scattered fibroglandular densities	6314 (46)	1986 (43)	151 810 (47)
Heterogeneously dense	5227 (38)	2083 (45)	123 572 (38)
Extremely dense	702 (5)	231 (5)	16 420 (5)
Missing	61 (< 1)	40 (1)	1708 (1)
<b>Cancer type</b>			
Invasive	166 (86)	3783 (83)	3783 (83)
DCIS	27 (14)	801 (17)	801 (17)
Median follow-up interval (y) <sup>†</sup>	5.0 (4.7–5.3)	2.8 (2.0–4.1)	5.0 (4.7–5.3)
Median length of health care enrollment before index date (y) <sup>†</sup>	17.9 (9.6–19.4)	18.9 (10.7–19.5)	17.6 (9.2–19.4)

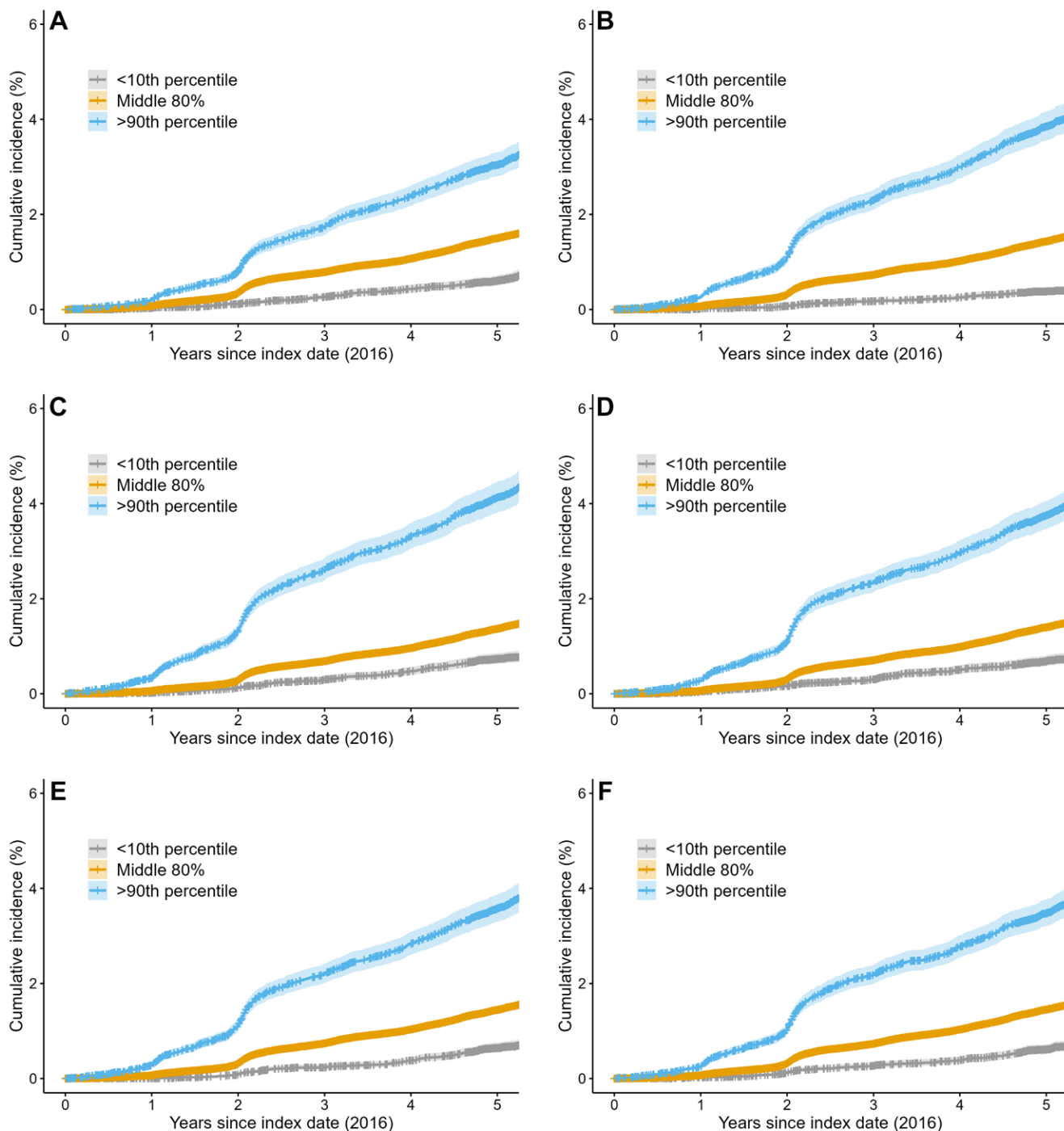
Note.—Unless otherwise indicated, data are numbers; data in parentheses are percentages. BI-RADS = Breast Imaging Reporting and Data System, DCIS = ductal carcinoma in situ.

\* Includes cases within subcohort (*n* = 193) as well as cases outside subcohort (*n* = 4391).

<sup>†</sup> Data are medians, with IQRs in parentheses.

women in the subcohort who did not develop cancer, 12 226 (91.9%) women were censored because of end of follow-up period, 940 (7.0%) because of disenrollment, and 269 (2.0%) because of death. Of the mammograms, 87.0% were acquired with Hologic units (11 856 of 13 628; Hologic) and 13.0% were acquired with GE units (1772 of 13 628; GE Healthcare).





**Figure 2:** Cumulative risk of breast cancer by risk model type at 5 years. Kaplan-Meier curves for (A) the clinical Breast Cancer Surveillance Consortium (BCSC) risk model and for the mammography-trained artificial intelligence (AI) risk models (B) Mirai, (C) MammoScreen, (D) ProFound, (E) Mia, and (F) Globally-Aware Multiple Instance Classifier. Women with a BCSC risk greater than 90th percentile accounted for 21% of all cancers by 5 years, whereas women with less than 10th percentile risk accounted for 3% of all cancers. Women with AI risk greater than 90th percentile accounted for 24%–28% of all cancers by 5 years, whereas women with less than 10th percentile risk accounted for approximately 2%–5% of cancers across all AI algorithms. The blue line represents women with a risk score greater than 90th percentile, the orange line represents women with a risk score in the middle 80 percentile, and the gray line represents women with a risk score in the less than 10th percentile. Shading surrounding the line is the 95% CI.

**Cumulative Incidence Rates of BCSC Clinical Risk Model and AI Algorithm Scores**

The average cumulative incidence rate at 5 years was 30.4 per 1000 person-years (95% CI: 28.1, 33.1) for women with a BCSC risk score greater than 90th percentile, 15.0 per 1000 person-years (95% CI: 14.4, 15.6) for women with a BCSC risk

score in the middle 80 percentiles, and 6.1 per 1000 person-years (95% CI: 5.1, 7.2) for women with a BCSC score in the less than 10th percentile (Fig 2). The incidence rate ratio of greater than 90th percentile risk to less than 10th percentile risk was 5.5. Women with a BCSC risk greater than 90th percentile accounted for 20.0% (919 of 4584) of all cancers by 5 years,

whereas women with less than 10th percentile risk accounted for 3.2% (149 of 4584) of all cancers.

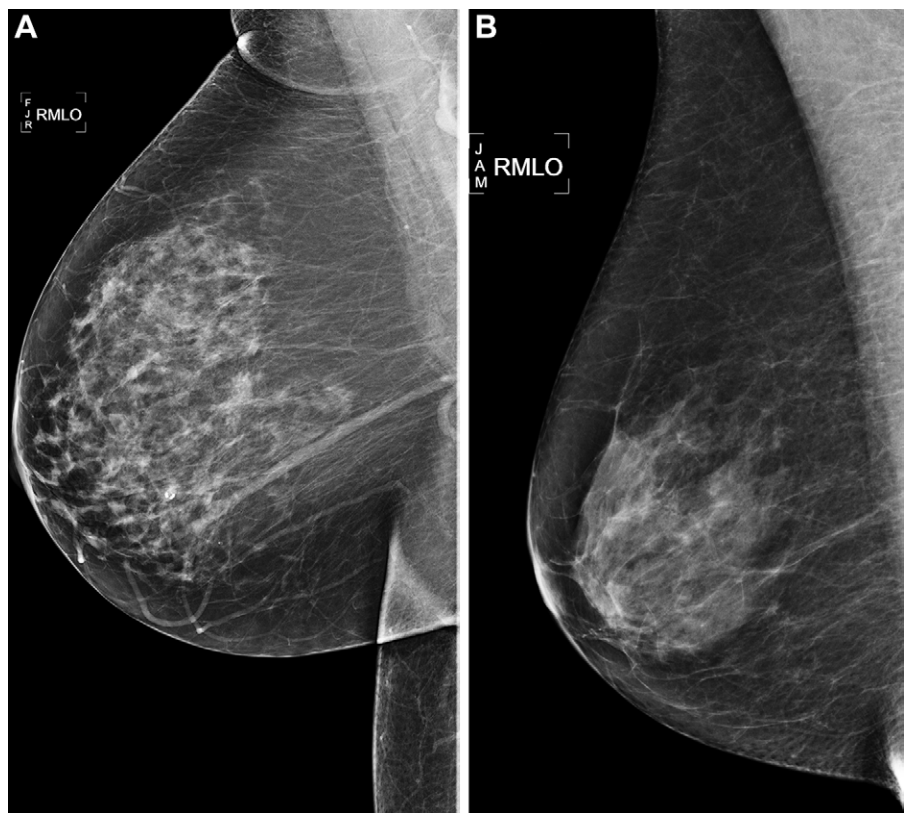
For AI algorithms, the average cumulative incidence rate at 5 years ranged from 34.9 to 41.3 per 1000 person-years for women with a risk score greater than 90th percentile, 13.7 to 14.5 per 1000 person-years for women with a risk score in the middle 80 percentiles, and 3.8 to 7.4 per 1000 person-years for women with a risk score less than 10th percentile. The incidence rate ratio of the risk greater than 90th percentile to risk less than 10th percentile ranged between 5.8 and 11.7. Women with risk greater than 90th percentile accounted for 24%–28% of all cancers at 5 years, whereas women with less than 10th percentile risk accounted for approximately 2%–5% of cancers across all AI algorithms. Examples of women who did and did not develop cancer within 5 years of follow-up are shown in Figure 3.

### Discrimination and Calibration of BCSC Clinical Risk Model and AI Algorithm Scores

When evaluating discrimination for interval cancer risk (Table 2), BCSC demonstrated a time-dependent AUC of 0.62 (95% CI: 0.59, 0.66), whereas the AI algorithms' time-dependent AUCs ranged from 0.67 to 0.71, with only Mammoscreen (time-dependent AUC, 0.71; 95% CI: 0.68, 0.75) and Mia (time-dependent AUC, 0.71; 95% CI: 0.67, 0.74) significantly higher than BCSC (Bonferroni-adjusted  $P < .0016$ ). For the 5-year future cancer risk, BCSC demonstrated a time-dependent AUC of 0.61 (95% CI: 0.60, 0.62), whereas the AI algorithm time-dependent AUCs ranged from 0.63 to 0.67, with all algorithms but Mia significantly higher than BCSC (Bonferroni-adjusted  $P < .0016$ ). For all cancer risk, BCSC demonstrated a time-dependent AUC of 0.61 (95% CI: 0.60, 0.62), whereas the AI algorithms' time-dependent AUCs ranged from 0.63 to 0.67, all significantly higher than BCSC (Bonferroni-adjusted  $P < .0016$ ).

The combined AI and BCSC models' time-dependent AUCs for interval cancer risk ranged from 0.67 to 0.73 (Table 3), although none were significantly higher than the corresponding AI algorithm alone when using Bonferroni-adjusted  $P$  values. The combined models' time-dependent AUCs for 5-year future cancer risk ranged from 0.66 to 0.68 and were significantly higher than all individual AI algorithms. Similarly, the combined models' time-dependent AUCs for 5-year all-cancer risk ranged from 0.66 to 0.68 and were higher than all individual AI algorithms.

Additional subgroup analyses (Tables S2–S8) demonstrated comparable performance to the primary results shown in Table 2 for complete scores available across all models (Table S2), in



**Figure 3:** Right medial lateral oblique (RML) screening mammograms show negative results from 2016 in (A) a 73-year-old woman with Mirai artificial intelligence (AI) risk score with more than 90th percentile risk who developed right breast cancer in 2021 at 5 years of follow-up and (B) a 73-year-old woman with Mirai AI risk score with less than 10th percentile risk who did not develop cancer at 5 years after 5 years of follow-up.

women with invasive breast cancer (Table S3), BI-RADS 1 or 2 only on screening mammograms (Table S5), and on mammograms acquired by using Hologic equipment (Table S7). Performance was mixed for some algorithms in women with ductal carcinoma in situ (Table S4), mammograms acquired on GE equipment only (Table S6), and in women with BI-RADS 0 on screening mammograms (Table S8), although interpretation was limited because of small sample size.

The 5-year calibration of the BCSC ranged from 1.02 to 1.08 depending on the prespecified BCSC risk threshold ranges, whereas that of the Mirai algorithm ranged from 0.49 to 0.76 (Table S9). Absolute differences in time-dependent AUC were also derived (Table S10 representing Table 2, and Table S11 representing Table 3).

### Discussion

We tested several mammography artificial intelligence (AI) models, many of which have been trained for shorter time horizons (ie, time over which risk is assessed), to determine whether they can predict future risk better than the commonly used Breast Cancer Surveillance Consortium (BCSC) clinical risk model (6,7) when used either alone or in combination with the BCSC model. AI algorithms showed a significantly higher discrimination of breast cancer risk than did the BCSC clinical risk model for predicting 5-year risk (AI time-dependent area under the receiver operating characteristic curve [AUC] range, 0.63–0.67,

**Table 2: Comparative Time-Dependent AUC Performance of Combined AI and BCSC Clinical Risk Model Using Negative Index Screening Mammography for Prediction of Invasive Cancer and DCIS**

Model*	Model Training Time Horizon Range (y)	Interval Cancer Risk 0–1 Years ( <i>n</i> = 259)	Future Cancer Risk (Excluding Interval Cancers at 0 to 1 Year)				All Cancer Risk 0–5 Years ( <i>n</i> = 4348)
			>1–2 Years ( <i>n</i> = 869)	>1–3 Years ( <i>n</i> = 2190)	>1–4 Years ( <i>n</i> = 3033)	>1–5 Years ( <i>n</i> = 4089)	
BCSC (Clinical)	0.5–5	0.62 (0.59, 0.66)	0.62 (0.60, 0.63)	0.63 (0.61, 0.64)	0.62 (0.60, 0.63)	0.61 (0.60, 0.62)	0.61 (0.60, 0.62)
Mirai (MIT, AI)	0–5	0.68 (0.65, 0.72) [.002] <sup>†</sup>	0.69 (0.67, 0.70) [<.001]	0.69 (0.68, 0.70) [<.001]	0.69 (0.68, 0.70) [<.001]	0.67 (0.66, 0.68) [<.001]	0.67 (0.66, 0.68) [<.001]
MammoScreen Therapixel, AI)	0–2	0.71 (0.68, 0.75) [<.001]	0.69 (0.67, 0.71) [<.001]	0.68 (0.67, 0.70) [<.001]	0.67 (0.66, 0.68) [<.001]	0.65 (0.64, 0.66) [<.001]	0.65 (0.64, 0.66) [<.001]
ProFound AI (iCAD, AI)	0–2	0.67 (0.63, 0.70) [.02] <sup>†</sup>	0.67 (0.65, 0.69) [<.001]	0.68 (0.67, 0.69) [.001]	0.66 (0.65, 0.67) [<.001]	0.65 (0.64, 0.66) [<.001]	0.65 (0.64, 0.66) [<.001]
Mia (Kheiron, AI)	0–1	0.71 (0.67, 0.74) [<.001]	0.66 (0.64, 0.68) [<.001]	0.66 (0.64, 0.67) [<.001]	0.64 (0.63, 0.65) [<.001]	0.63 (0.62, 0.64) [.002] <sup>†</sup>	0.63 (0.62, 0.64) [<.001]
GMIC (NYU, AI)	0–0.25	0.68 (0.64, 0.71) [.01] <sup>†</sup>	0.66 (0.64, 0.68) [<.001]	0.67 (0.66, 0.68) [<.001]	0.66 (0.65, 0.67) [<.001]	0.64 (0.63, 0.65) [<.001]	0.64 (0.63, 0.65) [<.001]

Note.—Data are time-varying areas under the receiver operating characteristic curve (AUCs), with 95% CIs in parentheses. Data in brackets are *P* values. Unless otherwise indicated, *P* values are Bonferroni corrected ( $P < .0016$ ), accounting for 30 tests, compared with the Breast Cancer Surveillance Consortium (BCSC) clinical risk model for the same time horizon. Missing data were imputed with cohort median value for each model. Further details are in Appendix S1. AI = artificial intelligence, DCIS = ductal carcinoma in situ, GMIC = Globally-aware Multiple Instance Classifier, MIT = Massachusetts Institute of Technology, NYU = New York University.

\* Information in parentheses is model manufacturer and type of model (ie, clinical or AI).

<sup>†</sup> Not meeting Bonferroni-corrected statistical significance.

vs BCSC time-dependent AUC, 0.61; Bonferroni-adjusted  $P < .0016$ ). This difference was most pronounced for interval cancer risk for certain algorithms, which highlights the strength of AI to identify missed or aggressive interval cancers. Furthermore, we demonstrated that AI algorithms trained for short time horizons can predict future risk of cancer up to 5 years when no cancer is clinically detected at mammography. As expected, performance improved for algorithms trained for longer time horizons. Combining BCSC and AI further improves risk prediction versus AI alone and decreases the differences in future risk performance across AI algorithms (for all cancer risk: time-dependent AUC range, 0.66–0.68; Bonferroni-adjusted  $P < .0016$ ).

Mammography AI algorithms provide an approach for improving breast cancer risk prediction beyond clinical variables such as age, family history, or the traditional imaging risk biomarker of breast density. The absolute increase in the AUC for the best mammography AI relative to BCSC was 0.09 for interval cancer risk and 0.06 for overall 5-year risk, a substantial and clinically meaningful improvement. The overall performance improvement remained when restricting the analysis to invasive cancer only. In order for an AI model to achieve an AUC of approximately 0.7, the model must have predictors that are two to three times more informative than clinical models such as the BCSC with an AUC of approximately 0.6 (1). Although we focus on AUC as an

accepted metric to compare general performance of risk models, a further approach to understand clinical significance is provided by our estimates for cancer yield or incidence rate ratios using hypothetical percentile cutoffs. For example, for a high-risk group defined at greater than 90th percentile risk, AI predicted up to 28% of cancers versus 21% with BCSC. However, because of the numerous use cases in which risk models are applied, clinical impact ultimately depends on the context and specific approach in which risk stratification is implemented. Continued strong predictive performance at 1–5 years is surprising and suggests that AI is not only identifying missed cancers but may identify breast tissue features that help predict future cancer development. This is analogous to high breast density independently predicting both tissue masking and future cancer risk (34).

We evaluated risk at different time horizons because each has distinct clinical implications. Certain AI algorithms excelled at predicting patients at high risk of interval cancer, which are often aggressive cancers (34,35) and may require a second reading of mammograms, supplementary screening (eg, with breast MRI), or short-interval follow-up. We also found AI algorithms predicted future risk, which may lead to more frequent and intensive screening or risk counseling for primary prevention. Overall, algorithms maintained robust performance in subgroup analyses for invasive cancer only.



**Table 3: Comparative Time-Dependent AUC Performance of Combined AI and BCSC Clinical Risk Models Using Negative Index Screening Mammography**

Model	Model Training Time Horizon Range (y)	Interval Cancer Risk 0–1 Year (n = 259)	Future Cancer Risk (Excluding Interval Cancers at 0 to 1 Year)				All Cancer Risk 0–5 Years (n = 4348)
			>1–2 Years (n = 869)	>1–3 Years (n = 2190)	>1–4 Years (n = 3033)	>1–5 Years (n = 4089)	
BCSC and Mirai	0–5	0.69 (0.66, 0.72) [.04]*	0.70 (0.68, 0.72) [<.001]	0.70 (0.69, 0.71) [<.001]	0.70 (0.69, 0.71) [<.001]	0.68 (0.67, 0.69) [<.001]	0.68 (0.67, 0.69) [<.001]
BCSC and MammoScreen	0–2	0.73 (0.70, 0.76) [.09]*	0.71 (0.69, 0.73) [<.001]	0.71 (0.69, 0.72) [<.001]	0.69 (0.68, 0.70) [<.001]	0.67 (0.66, 0.68) [<.001]	0.68 (0.67, 0.69) [<.001]
BCSC and ProFound AI	0–2	0.67 (0.64, 0.71) [.14]*	0.68 (0.66, 0.70) [.05]*	0.69 (0.67, 0.70) [.009]*	0.67 (0.66, 0.68) [<.001]	0.66 (0.65, 0.67) [.001]	0.66 (0.65, 0.67) [.001]
BCSC and Mia	0–1	0.72 (0.69, 0.75) [.13]*	0.69 (0.68, 0.71) [<.001]	0.69 (0.68, 0.70) [<.001]	0.68 (0.66, 0.69) [<.001]	0.66 (0.65, 0.67) [<.001]	0.66 (0.66, 0.67) [<.001]
BCSC and GMIC	0–0.25	0.69 (0.66, 0.72) [.03]*	0.69 (0.67, 0.71) [<.001]	0.70 (0.68, 0.71) [<.001]	0.68 (0.67, 0.69) [<.001]	0.67 (0.66, 0.67) [<.001]	0.67 (0.66, 0.68) [<.001]

Note.—Data are presented as time-varying areas under the receiver operating characteristic curve (AUCs), with 95% CIs in parentheses. Data in brackets are *P* values. Unless otherwise indicated, *P* values are Bonferroni corrected ( $P < .0016$ ), accounting for 30 tests, compared with the corresponding artificial intelligence (AI)-only model (Table 2) for the same time horizon. Combined models were fit using restricted cubic splines. AI = artificial intelligence, BCSC = Breast Cancer Surveillance Consortium, GMIC = Globally-Aware Multiple Instance Classifier.

\* Not meeting Bonferroni-corrected statistical significance.

The BCSC model prediction was originally built using U.S. national incidence rates from the Surveillance, Epidemiology, and End Results Program, and the predictions remained well calibrated with outcomes using our cohort, suggesting that our study population and results are similarly generalizable. However, the Mirai model (the only model that generated absolute risk estimates) overestimated cancer risk by a factor of two across all risk strata (observed to expected ratios of 0.49–0.76; Table S9). This is likely because Mirai was originally calibrated for both diagnostic risk (ie, cancer detected on the index mammogram) and future risk. Although model calibration does not affect the observed discriminative performance, calibration is critical when clinical decisions are based on prespecified risk thresholds. At the same time, AI models trained to predict specific thresholds can be recalibrated to support these decisions.

Beyond improved performance, mammography-based AI risk models provide practical advantages versus traditional clinical risk models. AI uses a single data source (the screening mammographic examination) that is available for most women in whom breast cancer risk prediction is relevant, enabling risk scores to be generated consistently and efficiently across a large population. Mammography AI risk models overcome certain barriers for risk models such as time and cost for combining multiple data elements from potentially different sources, as well as dependence on patient-reported history and susceptibility to missing data or recall bias. However, mammography AI risk models are limited to women who have undergone mammography. Therefore, these models cannot inform decisions regarding when women should

start screening. Moreover, mammography AI risk models also have potential costs (eg, software or hardware) and other technical and workflow considerations for implementation. Some breast imaging practices may already incorporate computer-aided detection AI, and the generated score may simultaneously be used for future risk stratification. Before AI is applied, it should be evaluated in the local patient populations for validity and potential hidden biases or disparities (36).

Our study had limitations. It was unable to evaluate all existing mammography AI algorithms, which are numerous (11,37) and may have produced different results than the five algorithms we evaluated. However, we provided a robust sample of one-third of the U.S. Food and Drug Administration- and Conformité Européenne-cleared commercial algorithms and well-known open-source algorithms. We were also unable to assess the extent to which family history was missing. However, the prevalence of family history was comparable to national estimates (7), suggesting reasonably complete ascertainment. Thus, our estimated BCSC AUC was likely valid and was indeed similar to previously published studies (38,39). Previously reported (13,40) Mirai algorithm performances were higher than those in our results, but this was because those studies evaluated combined diagnostic and future risk performance.

In conclusion, mammography artificial intelligence (AI) algorithms provided prediction of breast cancer risk to 5 years that was better than the Breast Cancer Surveillance Consortium (BCSC) clinical risk model, and the combination of AI and BCSC models further improved prediction. Our results imply

that mammography AI algorithms alone may provide a clinically meaningful improvement compared with current clinical risk models at early time horizons (ie, time during which risk is assessed), with further improvements in prediction when AI and clinical risk models are combined. Although AI algorithm performance declines with longer time horizons, most of the algorithms evaluated have not yet been trained to predict longer-term outcomes, suggesting a rich opportunity for further improvement. Evaluating a larger sample of the numerous AI mammography algorithms that are available remains for future efforts (11,37), although we examined multiple U.S. Food and Drug Administration– and Conformité Européenne–cleared commercial algorithms and well-known open-source algorithms. Moreover, AI provides a powerful way to stratify women for clinical considerations that necessitate shorter time horizons such as risk-based screening and supplemental imaging. The impact of AI models on clinical decisions requiring risk prediction beyond 5 years requires further study in cohorts with longer follow-up.

**Acknowledgments:** The authors thank Jane Bethard-Tracy, MA, and the Kaiser Permanente Northern California Breast Cancer Tracking System staff, Bing Lee, MS, Wei Yu, MS, Naomi Ruff, PhD, ELS, and Seth Selkow, AAS, for their additional contributions in supporting this study. The authors also thank the patients, mammography facilities, and radiologists for the data they provided for this study.

**Author contributions:** Guarantor of integrity of entire study, V.A.A.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, V.A.A., D.S.M.B., J.B.C., L.J.E., L.S.; clinical studies, V.A.A., J.B.C., D.A.L., D.L.M., D.A.N., A.P.; experimental studies, V.A.A., N.M.H., A.P.; statistical analysis, V.A.A., J.B.C., M.M.G., J.K., D.L.M., A.P., L.S., C.L.; and manuscript editing, V.A.A., L.A.H., N.S.A., D.S.M.B., J.B.C., L.J.E., N.M.H., M.M.G., J.K., L.H.K., V.X.L., C.M.L., L.S., W.S., H.C.Y., C.L.

**Disclosures of conflicts of interest:** V.A.A. No relevant relationships. L.A.H. No relevant relationships. N.S.A. No relevant relationships. D.S.N.B. DataSafety Monitoring Board membership for WISDOM study; stock/stock options in Grail (a subsidiary of Illumina); employed by Grail. J.B.C. No relevant relationships. L.J.E. Board of directors for Quantum Leap Healthcare Collaborative; grant funding from Quantum Leap Healthcare Collaborative for the I-Spy trial; grant from Merck; fees from UpToDate for writing; member of Blue Cross/Blue Shield Medical Advisory Panel and is reimbursed for travel and time. N.M.H. Research funding paid to institution from Kheiron Medical. M.M.G. Royalties from Oxford University Press; DataSafety Monitoring Board/Advisory Board membership for Study of Women Across the Nation. J.K. No relevant relationships. L.H.K. No relevant relationships. D.A.L. No relevant relationships. V.X.L. No relevant relationships. C.M.L. No relevant relationships. D.L.M. Grants paid to author's institution from NCI, PCORI. A.P. No relevant relationships. L.S. No relevant relationships. W.S. No relevant relationships. H.C.Y. No relevant relationships. C.L. No relevant relationships.

## References

- Gail MH, Pfeiffer RM. Breast Cancer Risk Model Requirements for Counseling, Prevention, and Screening. *J Natl Cancer Inst* 2018;110(9):994–1002.
- Pashayan N, Antoniou AC, Ivanus U, et al. Personalized early detection and prevention of breast cancer: ENVISION consensus statement. *Nat Rev Clin Oncol* 2020;17(11):687–705. [Published correction appears in *Nat Rev Clin Oncol* 2020;17(11):716.]
- Shieh Y, Eklund M, Madlensky L, et al. Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. *J Natl Cancer Inst* 2017;109(5):djw290.
- Miglioretti DL, Bissell MCS, Kerklikowske K, et al. Assessment of a Risk-Based Approach for Triaging Mammography Examinations During Periods of Reduced Capacity. *JAMA Netw Open* 2021;4(3):e211974.
- Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;81(24):1879–1886.
- Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerklikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med* 2008;148(5):337–347.
- Tice JA, Miglioretti DL, Li CS, Vachon CM, Gard CC, Kerklikowske K. Breast Density and Benign Breast Disease: Risk Assessment to Identify Women at High Risk of Breast Cancer. *J Clin Oncol* 2015;33(28):3137–3143.
- Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* 2004;23(7):1111–1130.
- Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* 2018;2(1):36.
- Geras KJ, Mann RM, Moy L. Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. *Radiology* 2019;293(2):246–259.
- Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open* 2020;3(3):e200265.
- Eriksson M, Czene K, Strand F, et al. Identification of Women at High Risk of Breast Cancer Who Need Supplemental Screening. *Radiology* 2020;297(2):327–333.
- Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. *Sci Transl Med* 2021;13(578):eaba4373.
- American College of Radiology. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. 5th ed. Reston, Va: American College of Radiology, 2013.
- Daly MB, Pal T, Berry MP, et al. Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2021;19(1):77–102.
- Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986;73(1):1–11.
- Sharp SJ, Poulaliou M, Thompson SG, White IR, Wood AM. A review of published analyses of case-cohort studies and recommendations for future reporting. *PLoS One* 2014;9(6):e101176.
- von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 2008;61(4):344–349.
- Callahan M, Sanderson J. A breast cancer tracking system. *Perm J* 2000;4:36–39. <https://www.thepermanentejournal.org/doi/10.7812/TPP/00.930>.
- Shen Y, Wu N, Phang J, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med Image Anal* 2021;68:101908.
- Mammocare by Therapixel. <https://www.mammocare.com/about>. Accessed January 7, 2023.
- iCAD. <https://www.icadmed.com/>. Accessed January 7, 2023.
- Kheiron Medical Technologies. <https://www.kheironmed.com/>. Accessed January 22, 2021.
- Breast Cancer Surveillance Consortium. Breast Cancer Surveillance Consortium Risk Calculator. <https://tools.bccs-cc.org/BC5yearRisk/sourcecode.htm>. Accessed August 18, 2021.
- R Core Team. The R Project for Statistical Computing. <https://www.r-project.org/>. Published 2019. Accessed March 6, 2020.
- Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000;56(2):337–344.
- Liu D, Cai T, Zheng Y. Evaluating the predictive value of biomarkers with stratified case-cohort design. *Biometrics* 2012;68(4):1219–1227.
- Efron B, Tibshirani R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat Sci* 1986;1(1):54–75.
- Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 2013;32(30):5381–5397.
- Harrell FE Jr. Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer, 2015.
- Harrell FE Jr. rms: Regression Modeling Strategies. <https://CRAN.R-project.org/package=rms>. Published 2020. Accessed March 6, 2020.
- Brentnall AR, Cuzick J, Buist DSM, Bowles EJA. Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncol* 2018;4(9):e180174.
- Liddell FD. Simple exact analysis of the standardised mortality ratio. *J Epidemiol Community Health* 1984;38(1):85–88.

34. Kerlikowske K, Zhu W, Tosteson AN, et al. Identifying women with dense breasts at high risk for interval cancer: a cohort study. *Ann Intern Med* 2015;162(10):673–681.
35. Porter PL, El-Bastawissi AY, Mandelson MT, et al. Breast tumor characteristics as predictors of mammographic detection: comparison of interval- and screen-detected cancers. *J Natl Cancer Inst* 1999;91(23):2020–2028.
36. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25(9):1337–1340 [Published correction appears in *Nat Med* 2019;25(10):1627.].
37. Dreyer K, Wald C, Allen B, Agarwal S, Gichoya J, Patti J. American College of Radiology Data Science Institute AI Central. <https://aicentral.acrdsi.org/>. Accessed April 11, 2022.
38. McCarthy AM, Liu Y, Ehsan S, et al. Validation of Breast Cancer Risk Models by Race/Ethnicity, Family History and Molecular Subtypes. *Cancers (Basel)* 2021;14(1):45.
39. Tice JA, Bissell MCS, Miglioretti DL, et al. Validation of the breast cancer surveillance consortium model of breast cancer risk. *Breast Cancer Res Treat* 2019;175(2):519–523.
40. Yala A, Mikhael PG, Strand F, et al. Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *J Clin Oncol* 2022; 40(16):1732–1740.