Radiology

ORIGINAL RESEARCH • BREAST IMAGING

Evaluation of an AI Model to Assess Future Breast Cancer Risk

Celeste Damiani, PhD • Grigorios Kalliatakis, PhD • Muthyala Sreenivas, MBBS, MRCS, FRCR • Miaad Al-Attar, MBBS, MRCS, FRCR • Janice Rose • Clare Pudney • Emily F. Lane, BSc GradStat • Jack Cuzick, PhD • Giovanni Montana, PhD • Adam R. Brentnall, PhD

From the Center for Human Technologies, Istituto Italiano di Tecnologia, Via Melen 83, Genoa 16152, Italy (C.D.); Wolfson Institute of Population Health, Queen Mary University of London, London, UK (C.D., E.F.L., J.C., A.R.B.); Institute of Computer Science (ICS), Foundation of Research and Technology Hellas, Heraklion, Crete, Greece (G.K.); Joint for Director Breast Screening, University Hospitals Coventry and Warwickshire NHS Trust Coventry, Coventry, UK (M.S.); Department of Oncoplastic Breast Surgery, University Hospitals of Leicester NHS Trust, Leicester, UK (M.A.A.); Consumer member at National Cancer Research Institute, Breast Group, London, UK (J.R., C.P.); and University of Warwick, WMG, Coventry, UK (G.M.). Received October 26, 2022; revision requested January 12, 2023; revision received March 22; accepted April 7. **Address correspondence to** C.D. (email: *celeste.damiani@iit.it*).

This work was supported by Cancer Research UK (C49757/A28689).

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

See also the editorial by Mann and Sechopoulos in this issue.

Radiology 2023; 307(5):e222679 • https://doi.org/10.1148/radiol.222679 • Content codes: BR AI

Background: Accurate breast cancer risk assessment after a negative screening result could enable better strategies for early detection.

Purpose: To evaluate a deep learning algorithm for risk assessment based on digital mammograms.

Materials and Methods: A retrospective observational matched case-control study was designed using the OPTIMAM Mammography Image Database from the National Health Service Breast Screening Programme in the United Kingdom from February 2010 to September 2019. Patients with breast cancer (cases) were diagnosed following a mammographic screening or between two triannual screening rounds. Controls were matched based on mammography device, screening site, and age. The artificial intelligence (AI) model only used mammograms at screening before diagnosis. The primary objective was to assess model performance, with a secondary objective to assess heterogeneity and calibration slope. The area under the receiver operating characteristic curve (AUC) was estimated for 3-year risk. Heterogeneity according to cancer subtype was assessed using a likelihood ratio interaction test. Statistical significance was set at $P < .05$.

Results: Analysis included patients with screen-detected (median age, 60 years [IQR, 55–65 years]; 2044 female, including 1528 with invasive cancer and 503 with ductal carcinoma in situ [DCIS]) or interval (median age, 59 years [IQR, 53–65 years]; 696 female, including 636 with invasive cancer and 54 with DCIS) breast cancer and 1:1 matched controls, each with a complete set of mammograms at the screening preceding diagnosis. The AI model had an overall AUC of 0.68 (95% CI: 0.66, 0.70), with no evidence of a significant difference between interval and screen-detected (AUC, 0.69 vs 0.67; $P = .085$) cancer. The calibration slope was 1.13 (95% CI: 1.01, 1.26). There was similar performance for the detection of invasive cancer versus DCIS (AUC, 0.68 vs 0.66; $P = .057$). The model had higher performance for advanced cancer risk (AUC, 0.72 ≥stage II vs 0.66 <stage II; $P = .037$). The AUC for detecting breast cancer in mammograms at diagnosis was 0.89 (95% CI: 0.88, 0.91).

Conclusion: The AI model was found to be a strong predictor of breast cancer risk for 3–6 years following a negative mammographic screening.

Supplemental material is available for this article.

There is increasing discussion about replacing one-size-fits-all mammographic screening for breast cancer with risk-adapted screening, where both frequency and modality of screening are chosen based on the risk of breast cancer (1). Several models to assess the risk of breast cancer have been developed (2) and some are being evaluated in trials of risk-based screening (3–5). These largely combine classic hormonal and reproductive risk factors with family history, genetic testing that includes polygenic risk scores, and mammographic density.

Mammography is the primary breast screening test. The interpretation of screening mammograms by trained readers and subsequent recall, alongside early diagnosis and treatment, reduces breast cancer mortality (6,7). Mammographic density, which may be derived from a screening mammogram, measures the amount of fibroglandular tissue in the breast and is one of the strongest risk factors for breast cancer (8).

There might be additional information in the mammogram beyond density for breast cancer risk assessment. Data to support this hypothesis include analysis of computer-aided detection (CAD) software flags (9) and CAD suspicion scores as predictors of interval cancer (10) alone or in combination with breast density (11). Such information may be most useful over a short follow-up period. One reason might be that readers occasionally miss cancers, or there might be subtle early signs of cancer on the mammogram.

Recently, an artificial intelligence (AI) model based on digital screening mammograms, called Mirai, has been designed for risk assessment over a 5-year period after a negative screening result (12,13). This algorithm appears to be

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, DCIS = ductal carcinoma in situ, ER = estrogen receptor, NHS = National Health Service

Summary

A breast cancer artificial intelligence system was a stronger predictor for risk stratification over 3–6 years than mammographic density and could help determine screening intervals.

Key Results

- In this matched case-control study of female patients with screen-detected (n = 2044) or interval (n = 696) breast cancer, the artificial intelligence model was a strong predictor of breast cancer risk (area under the receiver operating characteristic curve [AUC], 0.68) when using earlier triannual screening-round mammograms.
- The model performance for interval and screen-detected cancers was similar (AUC, 0.69 vs 0.67).
- Performance was higher for advanced breast cancer (≥stage II) compared with earlier stage disease (AUC, 0.72 vs 0.66).

a stronger risk predictor than mammographic density during this time period, but it has not been externally validated in an English screening population or by using multiple mammograms over time. The latter is important to know whether the model tracks signs of cancer on the mammogram (eg, like a computer-aided detection system) or is more static like breast density. To our knowledge, the model has also not been assessed for disease staging and grading (13), which is needed to help determine if the cancers identified early are likely to be fatal without additional interventions.

The purpose of our study was to evaluate a deep learning algorithm for risk assessment of future breast cancer based on digital mammograms obtained in the National Health Service (NHS) Breast Screening Programme (14). Furthermore, we evaluated the potential heterogeneity of performance for interval versus screen-detected cancers, according to age group and cancer type (invasive cancer vs ductal carcinoma in situ [DCIS]).

Materials and Methods

Patients

Female patients included in this retrospective observational matched case-control study attended the United Kingdom NHS Breast Screening Programme from February 2010 to September 2019 at OPTIMAM Mammography Image Database sites (Appendix S1) (15). Previous studies using this data are listed at *https://medphys.royalsurrey.nhs.uk/omidb/publications/*. Data were fully anonymized and received ethical approval for research (REC reference: 19/SC/0284, IRAS reference: 265403). The current study reports new work using this database (16).

Patients aged 46–74 years were eligible for inclusion if they had undergone a standard four-view mammographic screening examination with "for presentation" (ie, processed) images from a Hologic mammography system (Lorad Selenia or Selenia Dimensions) at two screening sites and did not have breast implants. Patients were excluded if they had benign disease.

Prognostic Model

Mirai (version 0.3.1, *https://github.com/yala/Mirai*) was developed by Yala et al (12) using a U.S. cohort and the computer code is freely available (17). The model estimates risk annually for 5 years. It was run following the developer instructions (17) using mammograms up to 6 years (two screening visits) prior to diagnosis. The percentage of mammographic breast density as determined using Volpara (Volpara Health) software was provided from the OPTIMAM registry for all screening examinations with unprocessed Digital Imaging and Communications in Medicine, or DICOM, images.

Study Design

The primary end point of the study was diagnosis of biopsy-confirmed invasive carcinoma or DCIS as recorded in the NHS Breast Screening Programme database from February 2010 to September 2019. The main prognostic factor was 3-year absolute risk. Most female individuals have triennial screening from 50 to 70 years of age in England, with some starting at 47 years or ending at 73 years of age during the study period due to a trial (ISRCTN registry no. ISRCTN33292440). All included patients had at least 3 years of follow-up after the index mammogram. Patients with breast cancer (cases) were included if they attended a screening visit and *(a)* the cancer was screen-detected on mammograms at a screening visit 3 years prior or *(b)* they were diagnosed with interval cancer between two screening visits and a mammogram was available from the screening visit less than 3 years earlier. Patients without breast cancer (controls) were matched 1:1 with cases according to age at the time the mammogram was obtained (within 1 year) and the mammography system used, with at least 3 years of follow-up to a subsequent normal screening assessment. Exploratory outcomes were breast cancer subtypes recorded in the NHS Breast Screening Programme database as follows: advanced cancer stage (18,19) (≥stage II, indicated by positive nodal status or tumor size ≥20 mm), cancer grade, estrogen receptor (ER) status, hormone receptor status (positive if ER positive and/or progesterone receptor positive), human epidermal growth factor receptor 2 (HER2; also called ERBB2) status, and triple-negative breast cancer (positive if ER negative, progesterone receptor negative, and HER2 negative).

The study sample size was a priori judged sufficient because the area under the receiver operating characteristic curve (AUC) of breast density is approximately 0.60 (20,21). We expected that the model would be stronger than mammographic density and strongest for interval cancer (12), and we determined the model had approximately 90% power to test for heterogeneity if the AUC is 0.67 for interval cancer versus 0.60 for screen-detected cancer.

Statistical Analysis

A statistical analysis plan was finalized before analysis by two authors (C.D., A.R.B.), with analysis performed by three authors (C.D., A.R.B., E.L.). The primary analysis objective was to assess the strength of the AI model in individuals attending the NHS Breast Screening Programme. The secondary objective was to assess potential heterogeneity for interval

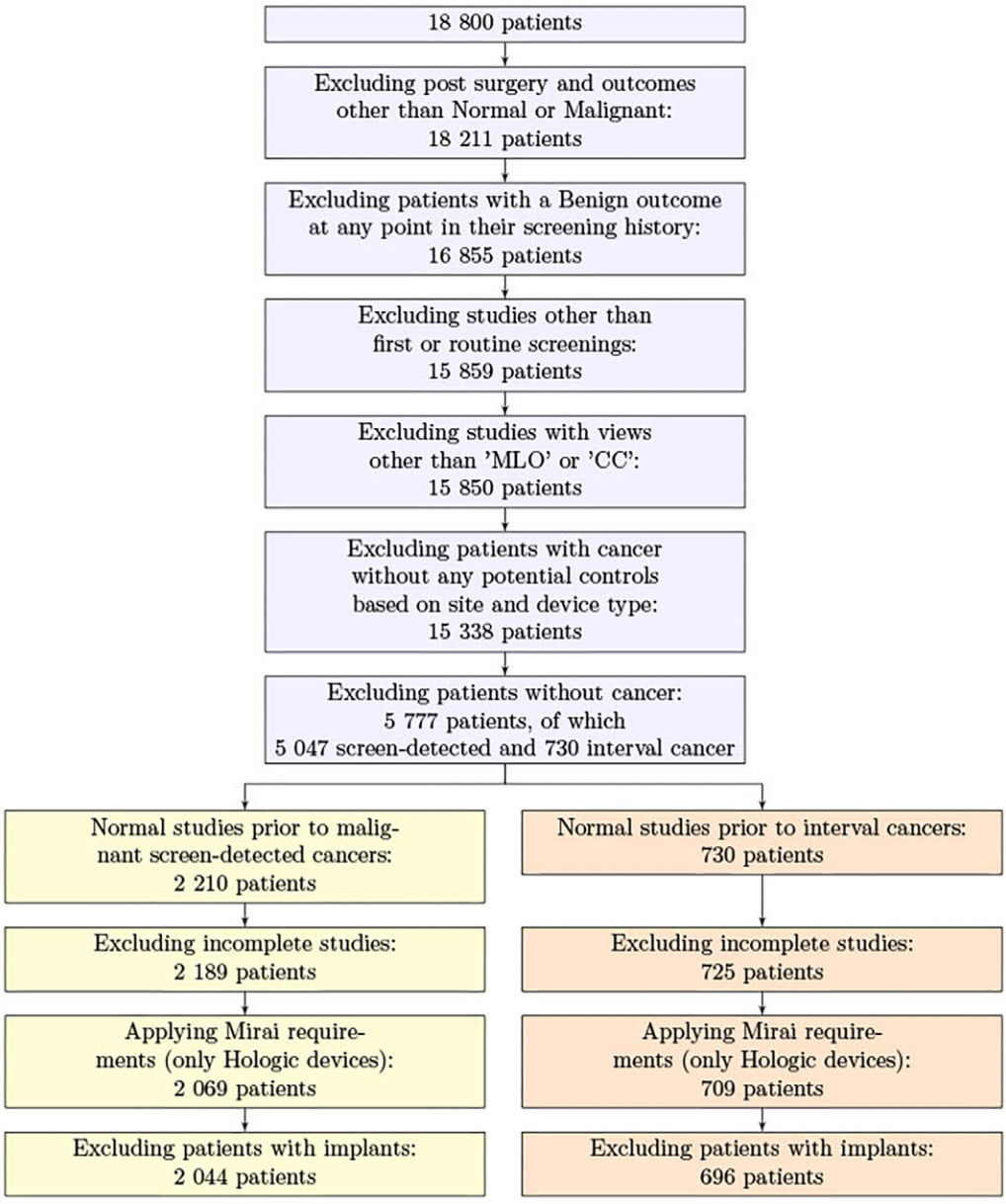

Figure 1: Flowchart shows patient inclusion and exclusion from the OPTIMAM Mammography Image Database. A study is defined as a patient visit where mammograms were obtained. Only mammograms from two screening sites were included because there were no controls available from the third site. CC = craniocaudal, MLO = mediolateral oblique.

versus screen-detected cancers, age, and cancer type (invasive vs DCIS), with calibration of relative risk. Primary analysis was performed on a complete matched-pair basis. The odds ratio per SD in controls of the natural logarithm of absolute 3-year risk was estimated. To reflect the study design, matching was considered in the analysis through use of conditional logistic regression (odds ratios with 95% profile likelihood CIs) and a concordance index adjusted for matching factors (AUCs with 95% CIs from the Wilson method [21]). This AUC indicates the chance that a case has a greater risk score than the matched control. Heterogeneity of performance by subgroup was assessed using likelihood ratio tests for interaction. To assess whether the relative risks of the AI model were accurate, a calibration slope was estimated, which used the logistic regression coefficient associated with log absolute risk such that a perfect calibration would yield a coefficient of 1.0. Calibration of the predicted relative risk was plotted using a generalized additive model fit to continuous values of the AI model relative risk (22). Positive predictive value (or prevalence) was estimated according to risk decile, assuming a prevalence of 8.5 patients per 1000 at screening (14) and interval cancer rates of approximately three per 1000 (23). Other predefined subgroups were cancer type (invasive vs DCIS) and age (<55 years, 55–64 years, ≥65 years). Exploratory analysis assessed performance according to subtypes of invasive cancer using the same methodology. Performance at diagnosis was evaluated in screen-detected cancers and 6 years prior. The AI model was compared with breast

Table 1: Baseline Patient Characteristics according to Case or Control Status

Characteristic	Interval Cancer Group		Screen-detected Cancer Group	
	Control (*n* = 696)	Case (*n* = 696)	Control (*n* = 2044)	Case (*n* = 2044)
Age (y)				
<55	210 (30)	210 (30)	472 (23)	472 (23)
55–64	287 (41)	287 (41)	1027 (50)	1027 (50)
≥65	199 (29)	199 (29)	545 (27)	545 (27)
Median*	59 (53–65) [46–74]	59 (53–65) [46–74]	60 (55–65) [46–73]	60 (55–65) [46–73]
Screening site				
1	605 (87)	605 (87)	1495 (73)	1494 (73)
2	91 (13)	91 (13)	549 (27)	550 (27)
Mammography system				
Lorad Selenia	657 (94)	657 (94)	1964 (96)	1964 (96)
Selenia Dimensions	39 (6)	39 (6)	80 (4)	80 (4)
Cancer type				
DCIS	NA	54 (8)	NA	503 (25)
Invasive	NA	636 (91)	NA	1528 (75)
Unknown	NA	6 (1)	NA	13 (1)

Note.—Except where indicated, data are numbers of patients, with percentages in parentheses. Patients with breast cancer (cases) and those without (controls) were screened at two OPTIMAM Mammography Image Database sites from the United Kingdom National Health Service Breast Screening Programme; only two screening sites were included because there were no controls available from the third site. Lorad Selenia and Selenia Dimensions are manufactured by Hologic. DCIS = ductal carcinoma in situ, NA = not applicable.

* Data are medians, with IQRs in parentheses and ranges in brackets.

Table 2: Predictive Ability of the AI Model for Future Breast Cancer over the 3 Years after a Negative Screening Result with Predefined Subgroup Analysis

Group and Subgroup	Odds Ratio	Matched AUC	Calibration Slope	*P* Value
Overall	1.72 (1.63, 1.83)	0.68 (0.66, 0.70)	1.13 (1.01, 1.26)	.085
Interval cancer	1.87 (1.67, 2.09)	0.69 (0.66, 0.73)	1.30 (1.07, 1.54)	
Screen-detected cancer	1.67 (1.56, 1.79)	0.67 (0.65, 0.69)	1.07 (0.93, 1.21)	
Overall subgroup				
Age (y)				
<55	1.63 (1.44, 1.84)	0.66 (0.62, 0.69)	1.01 (0.76, 1.27)	.6
55–64	1.84 (1.68, 2.01)	0.69 (0.67, 0.72)	1.27 (1.08, 1.45)	
≥65	1.64 (1.48, 1.81)	0.67 (0.63, 0.70)	1.03 (0.82, 1.24)	
Cancer type				
Invasive	1.78 (1.66, 1.90)	0.68 (0.66, 0.70)	1.20 (1.05, 1.34)	.057
DCIS	1.55 (1.38, 1.75)	0.66 (0.61, 0.69)	0.91 (0.67, 1.16)	
Unknown	1.87 (0.94, 3.75)	0.58 (0.36, 0.77)	1.31 (−0.14, 2.75)	
Screen-detected cancer				
Age (y)				
<55	1.43 (1.25, 1.64)	0.63 (0.59, 0.68)	0.75 (0.46, 1.03)	
55–64	1.80 (1.63, 2.00)	0.70 (0.67, 0.72)	1.23 (1.01, 1.44)	
≥65	1.63 (1.45, 1.84)	0.66 (0.62, 0.70)	1.02 (0.77, 1.27)	
Cancer type				
Invasive	1.70 (1.57, 1.84)	0.68 (0.66, 0.70)	1.10 (0.94, 1.27)	
DCIS	1.56 (1.38, 1.77)	0.65 (0.61, 0.69)	0.93 (0.67, 1.18)	
Unknown	2.12 (0.81, 5.53)	0.54 (0.29, 0.77)	1.56 (−0.44, 3.56)	
Interval cancer				
Age (y)				
<55	2.33 (1.75, 3.09)	0.72 (0.65, 0.78)	1.76 (1.17, 2.35)	
55–64	1.95 (1.62, 2.36)	0.69 (0.63, 0.74)	1.39 (1.00, 1.79)	
≥65	1.66 (1.38, 2.00)	0.67 (0.60, 0.73)	1.06 (0.67, 1.44)	
Cancer type				
Invasive	1.96 (1.73, 2.24)	0.69 (0.66, 0.73)	1.41 (1.14, 1.67)	
DCIS	1.48 (1.01, 2.15)	0.67 (0.53, 0.78)	0.81 (0.03, 1.59)	
Unknown	1.55 (0.55, 4.38)	0.67 (0.30, 0.90)	0.91 (−1.25, 3.08)	

Note.—Data in parentheses are 95% CIs. The odds ratio is per SD of log Mirai (AI model) 3-year risk in controls, adjusted for matching factors. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, DCIS = ductal carcinoma in situ.

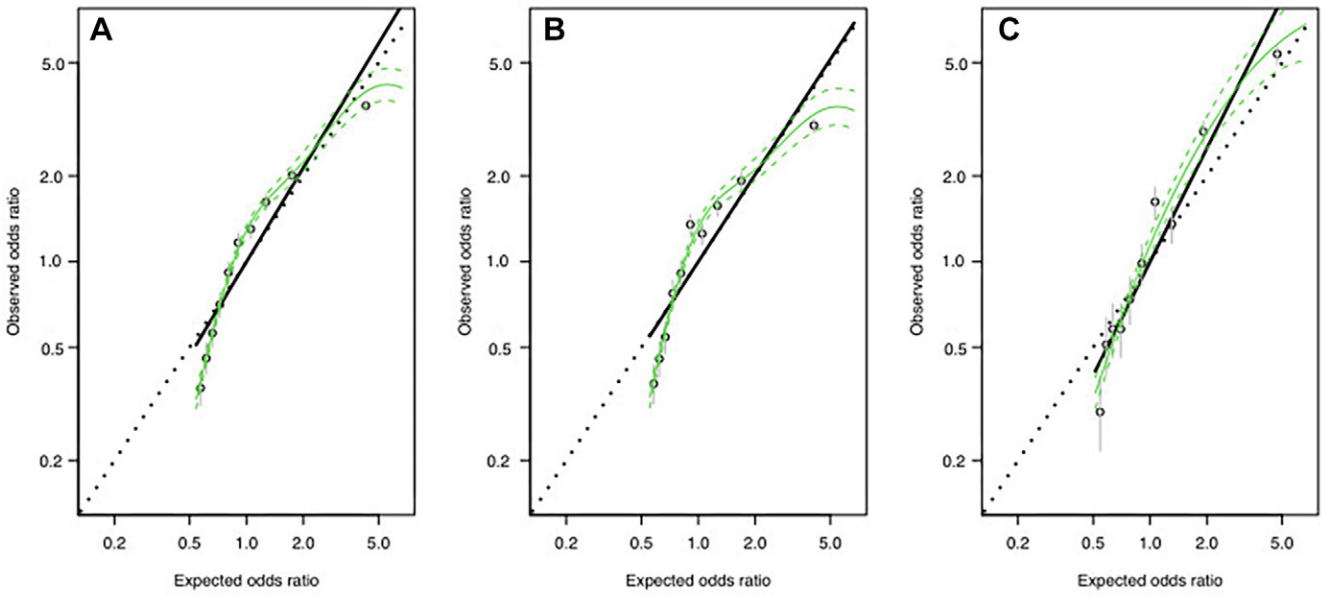

Figure 2: Calibration plots show the odds ratios associated with projected 3-year risk **(A)** overall and for **(B)** screen-detected and **(C)** interval cancers. The solid green line is fitted using a generalized additive model fit, with the standard error as dashed green lines, and the solid black line is a logistic regression fit. The points, with pointwise 95% CIs shown as gray bars, are at deciles of the risk score in both patient cases and controls.

density by using conditional logistic regression χ^2 statistics with matched pairs that had Volpara-determined breast density available for both the case and control. Interval cancer subgroups were examined based on national screening program review from at least two reviewers as *(a)* radiologically normal, *(b)* seen with hindsight but not obvious, or *(c)* obviously malignant (24). This review is conducted in the NHS Breast Screening Programme as part of routine quality assurance, and the identity of the reviewers is anonymized in the database. $P < .05$ was considered indicative of a statistically significant difference. Statistical analysis was performed using R (version 4.1.2; The R Foundation) (25).

Results

Patient Characteristics

Patient inclusion into this study from the OPTIMAM Mammography Image Database is shown in Figure 1. After exclusions, there were 2044 patient cases with screen-detected cancer and 696 patient cases with interval cancer. They were matched 1:1 to controls in our primary analysis data using the factors shown in Table 1. The median ages were 59 years (IQR, 53–65 years) for patients with interval cancer and 60 years (IQR, 55–65 years) for those with screen-detected cancer. Proportionally more interval cancers than screen-detected cancers were invasive (91% vs 75%) (Table 1).

Analysis and Presentation

Correlation with age and breast density.—The AI model was weakly positively correlated with age (Spearman correlation in controls, 0.18) and breast density (Spearman correlation, 0.15; both $P < .001$).

Association with breast cancer risk.—Overall, 3-year risk was strongly associated with development of both future interval cancer and screen-detected cancer (Table 2). There was no evidence of a significant difference in model performance between interval and screen-detected cancers (AUC, 0.69 vs 0.67; $P = .085$) or invasive cancer versus DCIS (AUC, 0.68 vs 0.66; $P = .057$).

Calibration slope.—To evaluate whether the implied relative risks from the AI model were accurate, we estimated the calibration slope of observed versus expected relative risks. This analysis showed that the 3-year risk slightly underestimated the spread of implied relative risks (overall calibration slope [observed vs expected odds ratio], 1.13; 95% CI: 1.01, 1.26) (Fig 2), indicating that the observed performance had slightly more discrimination than was expected by the AI model.

Positive predictive value.—Patients with an AI model risk greater than the top 10% of our sample were estimated to be at more than 10 times the risk of breast cancer over the next 3 years compared with patients with an AI model risk in the bottom 10% (Table S1). At the next screening round, the highest decile had a positive predictive value of approximately 25.2 patients per 1000 screened versus 3.2 per 1000 for the lowest decile (Table S1). The corresponding projected risks for interval cancer were 15.9 patients per 1000 versus 0.9 per 1000, respectively.

Invasive cancer subtypes.—The AI model showed discrimination for all subtypes, although there was some evidence of heterogeneity in strength of effect as follows (Table 3). First, the model was more predictive of stage II or higher cancers than cancers that were less advanced (<stage II) (AUC, 0.72 vs 0.66;

Table 3: Exploratory Subgroup Analysis of the AI Model Outputs Considering Subtypes of Invasive Cancer with Screen-detected and Interval Cancers Combined

Subgroup	Patients (n = 4328)*	Odds Ratio	Matched AUC	Calibration Slope	P Value
Cancer stage					
<II	1852 [43]	1.67 (1.51, 1.84)	0.66 (0.63, 0.69)	1.06 (0.86, 1.27)	.037
≥II	2032 [47]	1.94 (1.75, 2.16)	0.72 (0.69, 0.75)	1.38 (1.17, 1.60)	
Unknown	444 [10]	1.58 (1.32, 1.90)	0.64 (0.58, 0.70)	0.96 (0.57, 1.34)	
Nodal status					
Negative	2760 [64]	1.75 (1.61, 1.91)	0.68 (0.65, 0.70)	1.17 (0.99, 1.35)	.27
Positive	1000 [23]	1.92 (1.66, 2.23)	0.71 (0.67, 0.75)	1.36 (1.06, 1.67)	
Unknown	568 [13]	1.65 (1.39, 1.96)	0.67 (0.61, 0.72)	1.04 (0.68, 1.40)	
Tumor size (mm)					
<10	630 [15]	1.62 (1.37, 1.93)	0.66 (0.61, 0.71)	1.01 (0.65, 1.37)	.6
10–19	1510 [35]	1.85 (1.64, 2.08)	0.67 (0.64, 0.71)	1.28 (1.03, 1.53)	
20–29	890 [21]	1.84 (1.58, 2.15)	0.71 (0.67, 0.75)	1.27 (0.95, 1.59)	
≥30	852 [20]	1.89 (1.62, 2.20)	0.72 (0.68, 0.76)	1.32 (1.01, 1.64)	
Unknown	446 [10]	1.48 (1.24, 1.78)	0.64 (0.58, 0.70)	0.82 (0.44, 1.20)	
Cancer grade					
1	772 [18]	1.77 (1.51, 2.07)	0.68 (0.63, 0.72)	1.19 (0.86, 1.52)	0.24
2	2304 [53]	1.88 (1.70, 2.07)	0.70 (0.68, 0.73)	1.31 (1.11, 1.51)	
3	930 [21]	1.62 (1.41, 1.86)	0.65 (0.61, 0.70)	1.00 (0.72, 1.29)	
Unknown	322 [7]	1.62 (1.29, 2.05)	0.67 (0.59, 0.74)	1.01 (0.53, 1.49)	
ER status					
Negative	392 [9]	1.42 (1.17, 1.72)	0.65 (0.58, 0.71)	0.72 (0.33, 1.12)	.015
Positive	3434 [79]	1.86 (1.72, 2.01)	0.69 (0.67, 0.71)	1.29 (1.12, 1.45)	
Unknown	502 [12]	1.64 (1.37, 1.96)	0.67 (0.61, 0.73)	1.03 (0.65, 1.40)	
Hormone receptor status					
Negative	326 [8]	1.52 (1.22, 1.91)	0.67 (0.59, 0.74)	0.88 (0.41, 1.35)	.15
Positive	3498 [81]	1.83 (1.70, 1.98)	0.69 (0.67, 0.71)	1.26 (1.10, 1.42)	
Unknown	504 [12]	1.61 (1.35, 1.92)	0.67 (0.61, 0.73)	0.99 (0.63, 1.36)	
HER2 status					
Negative	3374 [78]	1.82 (1.68, 1.97)	0.69 (0.67, 0.71)	1.25 (1.08, 1.41)	>.99
Positive	316 [7]	1.83 (1.43, 2.34)	0.68 (0.61, 0.75)	1.26 (0.74, 1.77)	
Unknown	638 [15]	1.59 (1.36, 1.85)	0.66 (0.60, 0.71)	0.96 (0.65, 1.28)	
TNBC					
No	3582 [83]	1.82 (1.69, 1.96)	0.69 (0.66, 0.71)	1.24 (1.09, 1.40)	.31
Yes	222 [5]	1.53 (1.12, 2.09)	0.67 (0.57, 0.75)	0.88 (0.23, 1.54)	
Unknown	524 [12]	1.61 (1.36, 1.92)	0.68 (0.62, 0.73)	1.00 (0.64, 1.36)	

Note.—Data in parentheses are 95% CIs. The odds ratio is per SD of log Mirai (AI model) 3-year risk in controls, adjusted for matching factors. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, ER = estrogen receptor, HER2 = human epidermal growth factor receptor 2, TNBC = triple-negative breast cancer.

* Data are numbers of patients (cases with invasive cancer and matched controls), with percentages in brackets.

P = .037). Second, there was evidence that the model was more predictive for ER-positive cancers (AUC, 0.69 vs 0.65; P = .015), being higher for that of screen-detected cancers (AUC, 0.70 vs 0.61; P = .001) (Tables S2, S3). Third, the model was less predictive for screen-detected grade 3 cancers (AUC, 0.62) than grade 2 cancers (AUC, 0.71; P = .005).

Timing of mammograms.—Mirai was evaluated at the time of cancer diagnosis and showed strong discrimination (AUC, 0.89; 95% CI: 0.88, 0.91) (Table S4), with an approximate 70-fold difference in risk between the top and bottom deciles, suggesting that the AI algorithm is detecting visible signs of breast cancer in the mammogram. Change in the projected 3-year risk over the 6-year period to diagnosis in cases with complete data (n = 1260) is shown graphically in Figure 3. In the subset of cases that had matched controls, the AUC increased from 0.69 (95% CI: 0.63, 0.74) at 6 years and 0.71 (95% CI: 0.65, 0.77) at 3 years prior to diagnosis to 0.88 (95% CI: 0.83, 0.92) at the time of diagnosis (Tables S5, S6).

Comparison with breast density.—A total of 1477 matched case and control pairs (482 with interval cancer and 995 with screen-detected cancer) had Volpara-determined breast density data available (Table S7). Overall, the model had a higher likelihood of predicting a future breast cancer than breast density alone (AUC, 0.67 vs 0.56; P < .001). Breast density improved

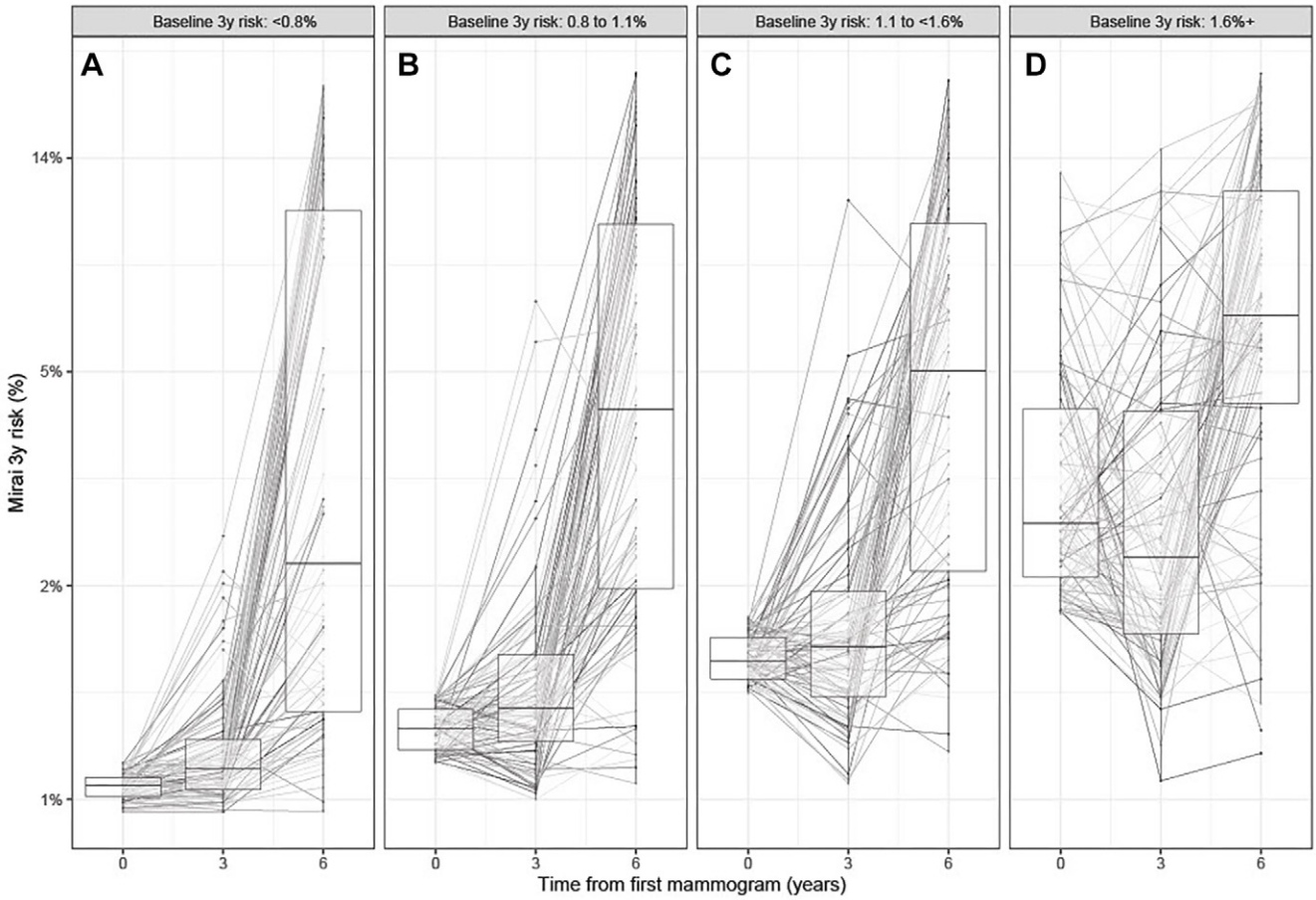

Figure 3: Longitudinal change in the artificial intelligence (AI)–projected 3-year cancer risk in patient cases with full screening mammograms available 6 years prior to diagnosis. **(A–D)** Graphs show the changes according to 3-year risk group (<0.8%, 0.8% to 1.1%, 1.1% to <1.6%, ≥1.6%) from the first mammogram until year 6 when all patients included in the plots were diagnosed with breast cancer. The individual lines track the projected 3-year risk given by the AI model for each patient.

performance of the AI model for interval cancer prediction (AUC, [model alone] 0.69, [combined] 0.71; $P < .001$), but did not significantly increase predictive power for screen-detected cancers ($P = .38$) (Table 4).

Interval cancer reader review subgroups.—Upon review, 81% of interval cancers were "not radiologically apparent" using the previous screening mammograms and only 1% were classified as "radiologically obvious" (Table S8) (24).

Discussion

The objective of our study was to determine performance of an artificial intelligence (AI) model for risk assessment over 3 years using mammograms that were judged healthy in the NHS Breast Screening Programme (14) and to evaluate potential heterogeneity. We found the AI model predicted breast cancer risk (area under the receiver operating characteristic curve [AUC], 0.68) with higher performance than breast density. There was similar performance for invasive cancer and ductal carcinoma in situ (AUC, 0.68 and 0.66), and performance was higher for advanced cancer risk (AUC, 0.72) and screen-detected estrogen receptor–positive cancer (AUC, 0.70). These results suggest the model might have a role in the development of risk-based screening algorithms.

Previous analysis of the model has compared performance with the Tyrer-Cuzick (International Breast Cancer Intervention Study) breast cancer risk model (2,12). The evaluation included mammographic density and questionnaire risk factors, but not polygenic risk scores (12). The AI model has also been evaluated in several other settings where risk factors and breast density were not available (13). These data suggest higher performance than the classic risk model and strong predictive ability across multiple settings. However, limitations of previous evaluations include measures of performance that included cancers diagnosed very close to the mammographic screening, where discrimination of the AI model is likely to be very high due to high performance for detection rather than risk assessment. In addition, interval and screen-detected cancers were not considered separately, and there was no evaluation of longitudinal change in risk or evaluation of heterogeneity by cancer stage and other prognostic factors. Our analysis has contributed to the evidence on the model by reporting analysis in these areas (12,13,26).

Several potential clinical actions might be taken using model risk in the NHS Breast Screening Programme (27). One strategy would be to use the model for detection and recall if the patient is at high risk. This is unreasonable; our estimate of the risk of interval cancer in the highest decile was approximately 1.6%,

Table 4: Direct Comparison of the Performances of the AI Algorithm and Breast Density in Patients with Both Measures

Group and Analysis	Odds Ratio	Matched AUC	LR-χ^2*	*P* Value
Overall				
Univariable				
Volpara	1.26 (1.17, 1.36)	0.56 (0.54, 0.59)	37.81	<.001
Mirai	1.69 (1.56, 1.83)	0.67 (0.65, 0.70)	236.18	<.001
Multivariable				
Combined		0.67 (0.65, 0.70)	253.89	
Volpara	1.19 (1.10, 1.29)		17.71	<.001
Mirai	1.66 (1.53, 1.80)		216.08	<.001
Interval cancer				
Univariable				
Volpara	1.62 (1.41, 1.87)	0.63 (0.58, 0.67)	53.16	<.001
Mirai	1.80 (1.57, 2.07)	0.69 (0.65, 0.73)	106.73	<.001
Multivariable				
Combined		0.71 (0.67, 0.75)	142.87	
Volpara	1.55 (1.34, 1.81)		36.14	<.001
Mirai	1.75 (1.52, 2.01)		89.71	<.001
Screen-detected cancer				
Univariable				
Volpara	1.11 (1.02, 1.22)	0.53 (0.50, 0.56)	5.25	.022
Mirai	1.63 (1.48, 1.80)	0.67 (0.64, 0.70)	130.84	<.001
Multivariable				
Combined		0.67 (0.64, 0.70)	131.62	
Volpara	1.05 (0.95, 1.15)		0.78	.38
Mirai	1.62 (1.47, 1.79)		126.37	<.001

Note.—Data in parentheses are 95% CIs. Volpara (Volpara Health) software was used to determine the percentage of mammographic breast density. Mirai is the AI algorithm (12,17). The odds ratio is per SD of the risk factor (log Mirai 3-year risk, log Volpara percentage density) in controls, adjusted for matching factors. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, LR-χ2 = likelihood ratio χ^2 statistic.

* For univariable analysis, the likelihood ratio χ^2 statistic and *P* value are for Volpara or Mirai alone. For multivariable analysis, the likelihood ratio χ^2 statistic and *P* value are for a model combining Volpara and Mirai (combined, on two degrees of freedom) and the change in the likelihood ratio χ^2 statistic when dropping Volpara or Mirai from the combined model.

but in England, where double reading is practiced, it is approximately 20%–25%. More realistic strategies might include asking radiologists to reassess mammograms if the patient is at high risk or offering a supplemental screening test beyond mammography. Another option is to increase or decrease the frequency of screening. For instance, to offer another screening visit sooner than 3 years to those at highest risk. This might be justified because the positive predictive value of the screening program is approximately 0.85%, which would be exceeded in the top risk decile for interval cancers (1.6%) and the next screening round (2.5%).

A potential concern with early detection interventions is that harms outweigh benefits (6). For example, if the detected cancers were all small and node-negative when detected during the current program, then there would likely be little clinical benefit in earlier detection because prognosis and treatment would likely be the same, while there will be increased costs and risks associated with false-positive results. On the other hand, if advanced cancers are detected earlier then clinical benefits might be large. We found evidence that the model identified high-risk groups of invasive cancers at an advanced stage.

There was some evidence that the model was less predictive of high-grade or ER-negative screen-detected cancers. The potential interaction with ER status has also previously been noted (12). Therefore, it is possible that the model is particularly useful for early identification of individuals at risk for slower-growing ER-positive cancers. This might make it an effective tool to identify those who would gain the most (absolute risk reduction) from preventive therapy because the observed preventive effectiveness of endocrine agents in trials must be predominantly due to suppression of preclinical malignancies.

Our study had some limitations. First, a retrospective and observational case-control design was used, although prospective cohorts would provide stronger evidence. Second, analysis was based on areas in the United Kingdom that were included in the OPTIMAM Mammography Image Database and we do not know the race or ethnicity of those included in the analysis. However, OPTIMAM includes sites with racially and ethnically diverse populations (15). Third, the AI model is largely a black box so that its internal workings are opaque. Fourth, we were unable to compare directly with risk models that use other domains, such as family history or polygenic risk scores.

In conclusion, further work is needed to explain the internal workings of the artificial intelligence model and the relative contribution of distinct mammographic features captured by it, such as in reader studies. In addition, more testing in retrospective and prospective diverse cohorts is needed in comparison with other tools (18). Finally, development of tools for digital breast tomosynthesis is needed.

Acknowledgments: We thank Cancer Research UK Technologies and staff at Royal Surrey NHS Foundation Trust for facilitating this data access.

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

Disclosures of conflicts of interest: **C.D.** Grants from Leverhulme Trust, Cancer Research UK (CRUK), and MISE Ministry of Economical Development; lecture

payments from Queen Mary University. **G.K.** No relevant relationships. **M.S.** No relevant relationships. **M.A.A.** No relevant relationships. **J.R.** No relevant relationships. **C.P.** No relevant relationships. **E.F.L.** No relevant relationships. **J.C.** Institutional grants from the Breast Cancer Research Foundation (CONS-22-001), AstraZeneca (IBIS-II), and CRUK (C569/A5032, C569/A27254); royalties from CRUK for commercial use of the Tyrer-Cuzick (IBIS) breast cancer risk evaluator; consulting fees from Qiagen, Becton Dickinson, and Myriad Genetics. **G.M.** No relevant relationships. **A.R.B.** Grants from National Institute for Health and Care Research (AI_AWARD01816), Breast Cancer Now (2019DecPRl395), and Barts Charity (G-002331); royalties from CRUK for commercial use of the Tyrer-Cuzick (IBIS) breast cancer risk evaluation algorithm; consulting fees from Kings College London; data monitoring committee for B-AHEAD 3; genetic epidemiologist and statistician on the NICE familial ovarian cancer guideline committee; committee member, UK National Screening Committee Research and Methodology Group; member of the CRUK Early Detection and Diagnosis Expert Review Funding Panel; statistical editor, *British Journal Radiology*.

References

1. Harkness EF, Astley SM, Evans DG. Risk-based breast cancer screening strategies in women. Best Pract Res Clin Obstet Gynaecol 2020;65:3–17.
2. Brentnall AR, Cuzick J. Risk Models for Breast Cancer and Their Validation. Stat Sci 2020;35(1):14–30.
3. 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.
4. Paci E, Mantellini P, Giorgi Rossi P, Falini P, Puliti D; TBST Working Group. Tailored Breast Screening Trial (TBST) [in Italian]. Epidemiol Prev 2013;37(4-5):317–327.
5. UNICANCER. International Randomized Study Comparing Personalized, Risk-Stratified to Standard Breast Cancer Screening In Women Aged 40-70. clinicaltrials.gov. https://clinicaltrials.gov/ct2/show/NCT03672331. Published 2022. Accessed October 23, 2022.
6. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. Br J Cancer 2013;108(11):2205–2240.
7. Maroni R, Massat NJ, Parmar D, et al. A case-control study to evaluate the impact of≠ the breast screening programme on mortality in England. Br J Cancer 2021;124(4):736–743.
8. Brentnall AR, van Veen EM, Harkness EF, et al. A case-control evaluation of 143 single nucleotide polymorphisms for breast cancer risk stratification with classical factors and mammographic density. Int J Cancer 2020;146(8):2122–2129.
9. Eriksson M, Czene K, Pawitan Y, Leifland K, Darabi H, Hall P. A clinical model for identifying the short-term risk of breast cancer. Breast Cancer Res 2017;19(1):29.
10. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94. [Published correction appears in Nature 2020;586(7829):E19.
11. Wanders AJT, Mees W, Bun PAM, et al. Interval Cancer Detection Using a Neural Network and Breast Density in Women with Negative Screening Mammograms. Radiology 2022;303(2):269–275.
12. Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. Sci Transl Med 2021;13(578):eaba4373.
13. 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.
14. NHS Digital. Breast Screening Programme, England 2019-20. https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme/england---2019-20. Accessed October 25, 2022.
15. Halling-Brown MD, Warren LM, Ward D, et al. OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. Radiol Artif Intell 2020;3(1):e200103.
16. OPTIMAM. Publications | OMI-DB. https://medphys.royalsurrey.nhs.uk/omidb/publications/. Accessed October 25, 2022.
17. Mirai YA. Mammography-based model for breast cancer risk. https://github.com/yala/Mirai. Published November 13, 2021. Accessed October 25, 2022.
18. Kerlikowske K, Chen S, Golmakani MK, et al. Cumulative Advanced Breast Cancer Risk Prediction Model Developed in a Screening Mammography Population. J Natl Cancer Inst 2022;114(5):676–685.
19. Tabár L, Yen AMF, Wu WYY, et al. Insights from the breast cancer screening trials: how screening affects the natural history of breast cancer and implications for evaluating service screening programs. Breast J 2015;21(1):13–20.
20. Damiani C, Brentnall AR. Statistical Analysis Plan: validation of MIRAI in OMI-DB. https://github.com/celebu/MammoAI_Public/tree/main/Risk_assessment. Published 2022. Accessed October 25, 2022.
21. Brentnall AR, Cuzick J, Field J, Duffy SW. A concordance index for matched case-control studies with applications in cancer risk. Stat Med 2015;34(3):396–405.
22. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Series B Stat Methodol 2011;73(1):3–36.
23. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. Br J Cancer 2011;104(4):571–577.
24. Public Health England. Breast screening: reporting, classification and monitoring of interval cancers and cancers following previous assessment. https://www.gov.uk/government/publications/breast-screening-interval-cancers/breast-screening-reporting-classification-and-monitoring-of-interval-cancers-and-cancers-following-previous-assessment#contents. Published 2021. Accessed February 2, 2023.
25. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Published 2022.
26. Lehman CD, Mercaldo S, Lamb LR, et al. Deep Learning vs Traditional Breast Cancer Risk Models to Support Risk-Based Mammography Screening. J Natl Cancer Inst 2022;114(10):1355–1363.
27. NHS Cancer Screening Programmes. Quality assurance guidelines for breast cancer screening radiology. https://www.gov.uk/government/publications/breast-screening-quality-assurance-standards-in-radiology. Published 2011. Accessed October 25, 2022.