Radiology: Artificial Intelligence

AI IN BRIEF

Impact of Different Mammography Systems on Artificial Intelligence Performance in Breast Cancer Screening

Clarisse F. de Vries, PhD • Samantha J. Colosimo, PhD • Roger T. Staff, PhD • Jaroslaw A. Dymiter, MSc • Joseph Yearsley, MSc • Deirdre Dinneen, MSc • Moragh Boyle, PGDip • David J. Harrison, MD • Lesley A. Anderson, PhD • Gerald Lip, MB, BCH, BAO* • on behalf of the iCAIRD Radiology Collaboration*[1]

From the Aberdeen Centre for Health Data Science, Institute of Applied Health Sciences (C.F.d.V., M.B., L.A.A.), School of Medicine, Medical Science and Nutrition (S.J.C., R.T.S.), and Grampian Data Safe Haven (DaSH), Aberdeen Centre for Health Data Science, Institute of Applied Health Sciences (J.A.D.), University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen AB24 3FX, Scotland; National Health Service Grampian (NHSG), Aberdeen Royal Infirmary, Aberdeen, Scotland (S.J.C., R.T.S., G.L.); Kheiron Medical Technologies, London, England (J.Y., D.D.); and School of Medicine, University of St Andrews, St Andrews, Scotland (D.J.H.). Received July 21, 2022; revision requested August 12; revision received February 14, 2023; accepted March 2. **Address correspondence to** C.F.d.V. (email: *clarisse.devries@abdn.ac.uk*).

[1] Members of the iCAIRD Radiology Collaboration team are listed at the end of this article.

Supported by the Industrial Centre for Artificial Intelligence Research in Digital Diagnostics (iCAIRD), which is funded by Innovate UK on behalf of UK Research and Innovation (UKRI) (project no. 104690).

* L.A.A. and G.L. are co–senior authors.

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

Radiology: Artificial Intelligence 2023; 5(3):e220146 • https://doi.org/10.1148/ryai.220146 • Content codes: AI OI BR

Artificial intelligence (AI) tools may assist breast screening mammography programs, but limited evidence supports their generalizability to new settings. This retrospective study used a 3-year dataset (April 1, 2016–March 31, 2019) from a U.K. regional screening program. The performance of a commercially available breast screening AI algorithm was assessed with a prespecified and site-specific decision threshold to evaluate whether its performance was transferable to a new clinical site. The dataset consisted of women (aged approximately 50–70 years) who attended routine screening, excluding self-referrals, those with complex physical requirements, those who had undergone a previous mastectomy, and those who underwent screening that had technical recalls or did not have the four standard image views. In total, 55 916 screening attendees (mean age, 60 years ± 6 [SD]) met the inclusion criteria. The prespecified threshold resulted in high recall rates (48.3%, 21 929 of 45 444), which reduced to 13.0% (5896 of 45 444) following threshold calibration, closer to the observed service level (5.0%, 2774 of 55 916). Recall rates also increased approximately threefold following a software upgrade on the mammography equipment, requiring per–software version thresholds. Using software-specific thresholds, the AI algorithm would have recalled 277 of 303 (91.4%) screen-detected cancers and 47 of 138 (34.1%) interval cancers. AI performance and thresholds should be validated for new clinical settings before deployment, while quality assurance systems should monitor AI performance for consistency.

Supplemental material is available for this article.

A recent U.K. National Screening Committee review (1,2) concluded that evidence was insufficient to support the implementation of artificial intelligence (AI) in routine breast cancer screening. The review identified limited evidence on sources of variability, impact on interval cancers (ICs) detected between screening cycles, and performance of a preset threshold to classify recall or no recall. In addition, evidence for the transferability of AI models is inconsistent (3–5).

We evaluated commercial AI software (6) by using data from a U.K. screening program to determine whether its performance transferred to an external dataset generated with different mammography equipment. The AI software is Conformité Européenne marked, indicating compliance with applicable European Union regulations. This study evaluates the generalizability of the AI tool by using consecutively acquired clinical data, comparing stand-alone performance to the dual reporting system in the U.K. screening service.

Materials and Methods

Sample

The Proportionate Review Subcommittee of the London-Bloomsbury Research Ethics Committee approved this retrospective study (reference no. 20/LO/0563). Secondary use of de-identified data negated the requirement for individual consent. Public Benefit and Privacy Panel approval was obtained (reference no. 1920–0258).

National Health Service (NHS) Grampian clinical data and mammograms were collected from the Scottish Breast Screening Service (SBSS) (February 12, 2016–March 31, 2020). Full-field digital mammograms were acquired with five mammography units of the same make and model (Selenia Dimensions; Hologic) with no known differences at study commencement. All units conform to NHS breast cancer screening quality standards (7). The standard imaging protocol consisted of two views per breast (craniocaudal and mediolateral oblique). As part of routine screening,

Abbreviations

AI = artificial intelligence, AUC = area under the ROC curve, DaSH = Grampian Data Safe Haven, IC = interval cancer, NHS = National Health Service, ROC = receiver operating characteristic, SBSS = Scottish Breast Screening Service, SHAIP = Safe Haven Artificial Intelligence Platform

Summary

Artificial intelligence (AI) performance in breast cancer screening was affected by mammography equipment and software used, highlighting the importance of local clinical settings and technology for effective AI implementation.

Key Points

- A mammography equipment software upgrade resulted in a threefold increase in the recall rate of a commercially available breast cancer screening artificial intelligence (AI) algorithm.
- Calibration of the AI decision threshold reduced recall rates from 47.7% to 13.0%.
- Implementation of AI into clinical practice requires local retrospective evaluation and ongoing quality assurance.

Keywords

Breast, Screening, Mammography, Computer Applications–Detection/Diagnosis, Neoplasms-Primary, Technology Assessment

two readers interpreted each set of images, with a third reader arbitrating in cases of disagreement. During the study period, mammograms in the screening center were routinely read by a pool of 11 readers with 1 to 20 years of experience each, led by one reader (G.L.).

The evaluation dataset was limited to a 3-year U.K. screening cycle (April 1, 2016–March 31, 2019) of women (aged approximately 50–70 years) attending routine screening. Figure 1 shows exclusions.

Data Processing

SBSS clinical data were transferred to the Grampian Data Safe Haven (DaSH). Mammograms from the breast screening picture archiving and communication system were transferred to the Safe Haven Artificial Intelligence Platform (SHAIP) developed by Canon Medical Research Europe (8). "Hiding in Plain Sight" (9) de-identification was performed.

Mia (version 2.0.1), developed by Kheiron Medical Technologies, the vendor in this study, assessed mammograms in SHAIP for potential malignancies. Mia was previously trained and tested on images acquired with Hologic, GE Healthcare, Siemens, and IMS Giotto mammography equipment. Mia, an ensemble of deep learning algorithms, employs the four standard image views (full-field digital mammography craniocaudal and mediolateral oblique views for each breast) to generate a continuous output ranging from 0 to 1 (malignancy prediction value). The malignancy prediction values were linked to the clinical data in DaSH. Mia's performance was evaluated by using a predefined threshold (≥0.1117 indicates recall) (6) and site-specific threshold.

Mia's performance was evaluated by academic health data scientists (C.F.d.V., J.A.D.) in DaSH (10), which the vendor could not access (meaning authors affiliated with the vendor had no control of the data). The vendor ran Mia within SHAIP with no access to the clinical outcomes to provide the Mia malignancy prediction values. The vendor also provided the Mia decision thresholds.

Threshold Calibration

Mia was not previously evaluated on images from Hologic Selenia Dimensions mammography equipment. The initial evaluation identified variability in algorithm performance. The vendor was provided with a validation dataset (16 204 screens) to generate a site-specific decision threshold. This subset included all screening data from 200 confirmed positive cases (women with histologically confirmed cancer), 4000 confirmed negative cases (women with negative findings for cancer with a negative 3-year follow-up screening and no IC), and 8000 unconfirmed negative cases (Appendix S1).

Statistical Analysis

A receiver operating characteristic (ROC) curve was plotted, and the area under the ROC curve (AUC) and CI (DeLong method) (11) were calculated. Positive screens were defined as histologically confirmed cancers detected through standard screening.

Sensitivity, specificity, and positive and negative predictive values, as well as cancer detection and recall rates of Mia, with CIs (Clopper-Pearson method) (12), were calculated for the prespecified and site-specific thresholds. Cancer detection rate was quantified as the number of screen-detected cancers with a (Mia) recall opinion divided by the total number of screens. The prespecified threshold was evaluated on the entire dataset after exclusions (original dataset) and on the subset not used to calibrate the threshold (test dataset). The site-specific threshold was evaluated using the test dataset. Furthermore, Mia's performance was compared with the performance of the first reader (reader 1). Mia was not compared with the second reader, as in the United Kingdom, the second reader can access the first reader's opinion and therefore does not read independently.

As an exploratory subanalysis, the site-specific threshold performance on the test dataset was stratified by mammography unit. Differences across units were assessed using Pearson χ^2 (specificity, recall, and cancer detection rate) and Fisher exact (sensitivity) tests. Additionally, sensitivity was compared between small (<15 mm) and large (≥15 mm) tumors using a χ^2 test.

ICs (cancers not detected during routine screening but identified between screening rounds) were analyzed separately. Following individual review, all readers in the clinical team regularly met to form a consensus on cancer visibility on prior screening mammograms, using the following categories (13): 1 = no visible lesion, 2 = lesion visible on review in hindsight, 3 = lesion clearly visible, and occult = lesion not visible at screening or subsequent symptomatic imaging. The proportion of IC patients Mia indicated to recall (with the updated threshold) was determined and stratified by consensus opinion.

Statistical analyses were performed in R (version 4.0.3) (Appendix S2). ROC curves, AUCs, and CIs were generated using the pROC package (14). Sample size information is available in

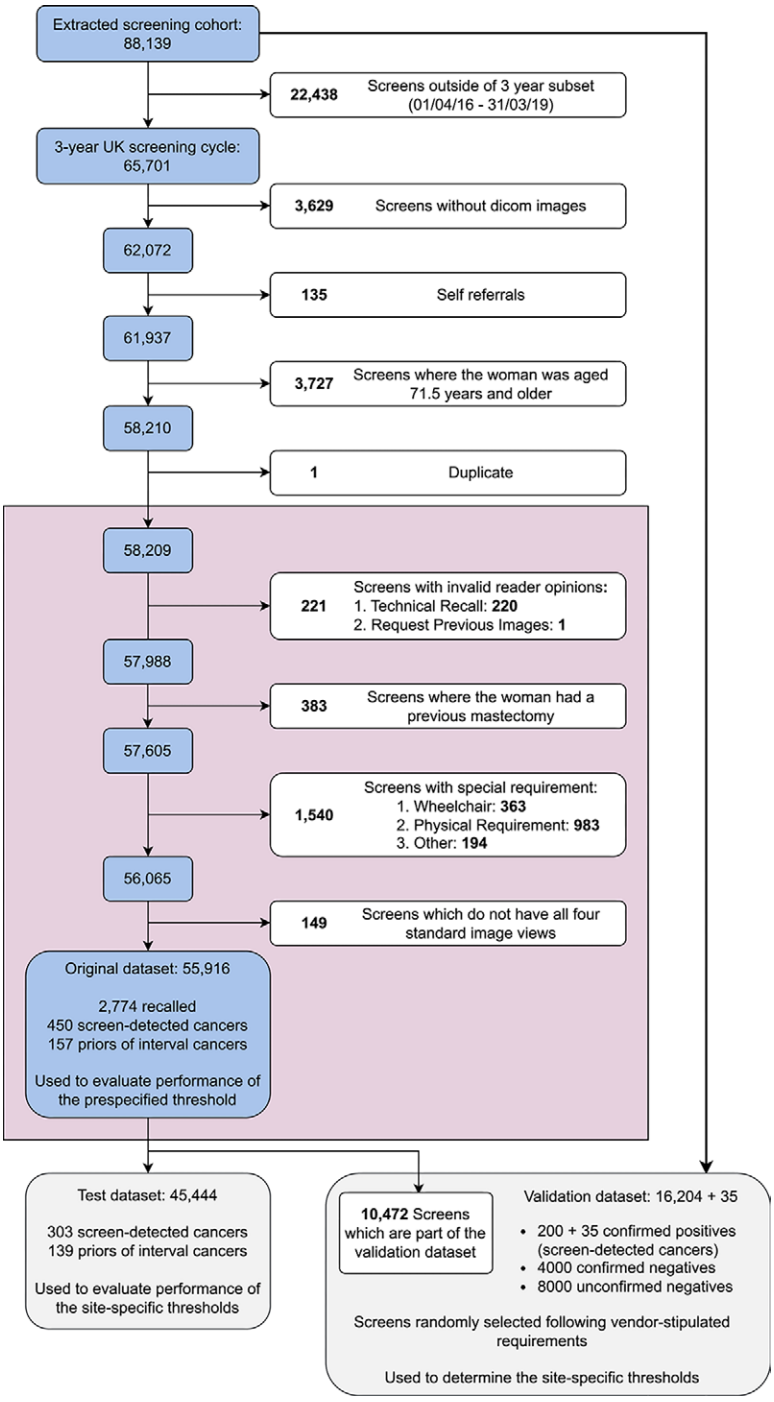

Figure 1: Flow diagram shows the generation and composition of the original, test, and validation datasets. Exclusions are indicated in the white boxes. The vendor-recommended exclusions are indicated in the shaded outer box. Confirmed positive cases are women with histologically confirmed cancer. Confirmed negative cases are women with negative findings for cancer with a negative 3-year follow-up screening and no interval cancer. DICOM = Digital Imaging and Communications in Medicine.

Appendix S3. *P* value less than .05 was considered to indicate a statistically significant difference.

Data Availability

The statistical output alongside the relevant R code is available in Appendix S2. Access to the raw SBSS data and mammograms (with de-identified participant data) is subject to the required approvals (eg, Public Benefit and Privacy Panel, NHS Research & Development, Research Ethics Committee approval) and data agreements being in place. More information can be found on the DaSH website: *https://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven.php*.

Results

Cohort Characteristics

After the application of vendor-recommended exclusions (3.9% [2293 of 58 209]) (15), an evaluation dataset of 55 916 screens was used (Fig 1). Of these, 2774 (5.0%) were recalled.

The mean age was 60 years (SD, 6.0 years); 450 patients had histologically confirmed screen-detected breast cancer, and 156 ICs were detected at follow-up (Table 1).

AI Performance Prethreshold Calibration

Figure 2A shows the Mia ROC curve. The AUC is 0.95 (95% CI: 0.94, 0.96). The Mia precision-recall curve can be found in Appendix S4.

For the prespecified threshold (original dataset: 55 916 screens and 450 cancers), sensitivity and specificity were 97.3% and 52.7%, respectively (Table 2). The recall rate was 47.7% and the cancer detection rate was 7.8 per 1000. For the test dataset (45 444 screens and 303 cancers, excluding screens used for threshold calibration), sensitivity and specificity were 98.3% and 52.1%, respectively; recall rate was 48.3%, and cancer detection rate was 6.6 per 1000.

Threshold Calibration

An initial site-specific threshold of 0.2938 was generated. This threshold revealed a step change in recall rate at set points for each mammography unit (Fig 2B). Review of image headers revealed that the increase in recalls correlated with a mammography unit software update. The AI algorithm was not updated during the study. All units had the same software before the update (version 1.7). The software running on units 1 to 4 was upgraded to version 1.8 at different time points. The monthly recall rate for software version 1.7 ranged from 8.3% (63 of 760) to 13.2% (183 of 1382); for version 1.8, it ranged from 23.8% (79 of 332) to 38.6% (86 of 223). In comparison, the reader 1 monthly recall rate ranged from 3.8% (37 of 966) to 6.9% (84 of 1218) before the software update and from 2.5% (seven of 282) to 7.9% (13 of 164) after the software update. Reader 1 sensitivity and specificity changed from 85.4% (328 of 384) to 87.9% (58 of 66) and from 95.1% (43 075 of 45 276) to 95.6% (9746 of 10 190), respectively.

Table 1: U.K. Breast Screening Program Cohort Characteristics

Original Dataset Characteristic	No.	Percentage (%)
Age (y)		
50–54	14 866	26.6
55–59	14 328	25.6
60–64	12 660	22.6
65–71.5	14 062	25.1
Included special requirement (*n* = 1048, 1.9%)		
Learning difficulty	116	0.2
Language need	304	0.5
Implant	364	0.7
Deafness	182	0.3
Blindness	40	0.07
Special needs	30	0.05
Two special requirements	12	0.02
Screen-detected breast cancer (*n* = 450, 0.8%)		
Type of cancer		
Nonbreast primary tumor	2	0.4
DCIS, preinvasive	101	22.4
Invasive status or grade unknown	5	1.1
Invasive breast cancer	342	76.0
Grade I	68	15.1
Grade II	211	46.9
Grade III	63	14.0
Tumor size		
<15 mm	259	57.6
≥15 mm	169	37.6
Unknown	22	4.9
Interval cancer (*n* = 156, 0.3%)		
Type of cancer		
DCIS	11	7.1
Invasive breast cancer	145	92.9
Grade I	5	3.2
Grade II	72	46.2
Grade III	67	42.9
Grade unknown	1	0.6
Tumor size		
<15 mm	24	15.4
≥15 mm	59	37.8
Unknown	73	46.8
Consensus opinion*		
Category 1	58	37.2
Category 2	15	9.6
Category 3	3	1.9
Occult	10	6.4
Not yet classified	70	44.9

Note.—Dataset comprised 55 916 screening attendees from April 1, 2016, to March 31, 2019. Percentages for screen-detected breast cancers and interval cancers are based on total number of screen-detected breast cancers and interval cancers, respectively. DCIS = ductal carcinoma in situ.

* Consensus opinion has four categories: 1 = no lesion visible on prior screening mammogram, 2 = uncertainty regarding whether a possible lesion was visible, 3 = a visible lesion which was missed, occult = no lesion visible on the prior screening mammogram, nor on the follow-up mammogram. Occult lesions usually manifest as palpable masses not discernible or outside the mammographic image.

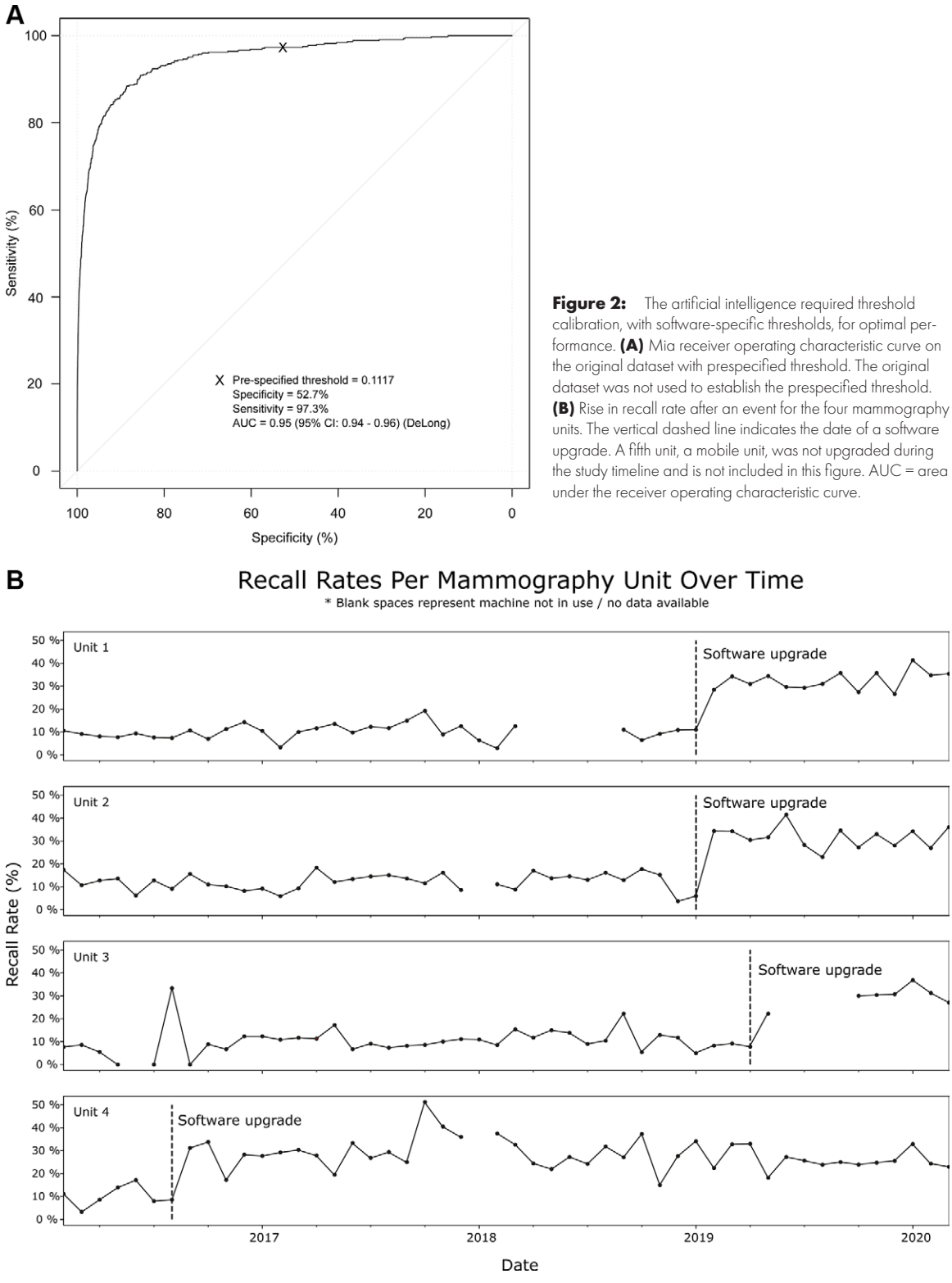

Figure 2: The artificial intelligence required threshold calibration, with software-specific thresholds, for optimal performance. **(A)** Mia receiver operating characteristic curve on the original dataset with prespecified threshold. The original dataset was not used to establish the prespecified threshold. **(B)** Rise in recall rate after an event for the four mammography units. The vertical dashed line indicates the date of a software upgrade. A fifth unit, a mobile unit, was not upgraded during the study timeline and is not included in this figure. AUC = area under the receiver operating characteristic curve.

Per–software version thresholds were generated to ensure stability of recall rates (Appendix S1). Due to a small number of positive studies in the post–software update subset, the vendor was provided with 35 additional positive studies (from mammography unit 4, after software upgrade) to reduce the threshold's susceptibility to noise.

Two site-specific thresholds were generated across all mammography units: 0.2712 before upgrade and 0.4319 after upgrade.

Applying the new thresholds to the test dataset resulted in a sensitivity of 91.4%, specificity of 87.6%, recall rate of 13.0%, and cancer detection rate of 6.1 per 1000 (Table 2). By comparison, reader 1 sensitivity, specificity, recall rate, and

Table 2: Mia Performance on Screen-detected Cancers

Parameter	No. of Data Points	No. of Cancers	Sensitivity (%)	Specificity (%)	Positive Predictive Value (%)	Negative Predictive Value (%)	Recall Rate (%)	Cancer Detection Rate per 1000
	AI and Reader 1 Performance							
Mia, original dataset								
Pre-specified threshold	55 916	450	97.3 (95.4, 98.6) [438/450]	52.7 (52.3, 53.1) [29 233/55 466]	1.6 (1.5, 1.8) [438/26 671]	99.96 (99.93, 99.98) [29 233/29 245]	47.7 (47.3, 48.1) [26 671/55 916]	7.8 (7.1, 8.6) [438/55 916]
Mia, test dataset								
Pre-specified threshold	45 444	303	98.3 (96.2, 99.5) [298/303]	52.1 (51.6, 52.5) [23 510/45 141]	1.4 (1.2, 1.5) [298/21 929]	99.98 (99.95, 99.99) [23 510/23 515]	48.3 (47.8, 48.7) [21 929/45 444]	6.6 (5.8, 7.3) [298/45 444]
Updated thresholds	45 444	303	91.4 (87.7, 94.3) [277/303]	87.6 (87.2, 87.9) [39 522/45 141]	4.7 (4.2, 5.3) [277/5896]	99.93 (99.90, 99.96) [39 522/39 548]	13.0 (12.7, 13.3) [5896/45 444]	6.1 (5.4, 6.9) [277/45 444]
Reader 1, test dataset	45 444	303	86.1 (81.7, 89.8) [261/303]	95.2 (95.0, 95.4) [42 956/45 141]	10.7 (9.5, 12.0) [261/2446]	99.90 (99.87, 99.93) [42 956/42 998]	5.4 (5.2, 6.0) [2446/45 444]	5.7 (5.1, 6.5) [261/45 444]
	AI Performance Split by Mammography Unit							
Unit 1	13 104	94	93.6 (86.6, 97.6) [88/94]	87.8 (87.2, 88.3) [11 421/13 010]	5.2 (4.2, 6.4) [88/1677]	99.95 (99.89, 99.98) [11 421/11 427]	12.8 (12.2, 13.4) [1677/13 104]	6.7 (5.4, 8.3) [88/13 104]
Unit 2	9960	78	92.3 (84.0, 97.1) [72/78]	86.2 (85.5, 86.8) [8514/9882]	5.0 (3.9, 6.3) [72/1440]	99.93 (99.85, 99.97) [8514/8520]	14.5 (13.8, 15.2) [1440/9960]	7.2 (5.7, 9.1) [72/9960]
Unit 3	13 000	95	90.5 (82.8, 95.6) [86/95]	88.7 (88.1, 89.2) [11 445/12 905]	5.6 (4.5, 6.8) [86/1546]	99.92 (99.85, 99.96) [11 445/11 454]	11.9 (11.3, 12.5) [1546/13 000]	6.6 (5.3, 8.2) [86/13 000]
Unit 4	8541	31	83.9 (66.3, 94.5) [26/31]	87.3 (86.6, 88) [7433/8510]	2.4 (1.6, 3.4) [26/1103]	99.93 (99.84, 99.98) [7433/7438]	12.9 (12.2, 13.6) [1103/8541]	3.0 (2.0, 4.5) [26/8541]*
Unit 5	839	5	100.0 (47.8, 100.0) [5/5]	85 (82.4, 87.4) [709/834]	3.8 (1.3, 8.8) [5/130]	100.00 (99.48, 100.00) [709/709]	15.5 (13.1, 18.1) [130/839]	6.0 (1.9, 13.9) [5/839]

Note.—Values in parentheses are 95% CIs; values in brackets are numerators and denominators. χ^2 tests (or Fisher exact tests, when there were small counts in the contingency table) were performed to determine whether the preset threshold performance was significantly different than the site-specific threshold performance, and whether the site-specific threshold performance was significantly different than reader 1 performance on screen-detected cancers. Sensitivity, specificity, recall, and cancer detection rate were significantly different between the preset and site-specific thresholds ($P < .001$). There were significant differences between the site-specific threshold and reader 1 for specificity, recall, and cancer detection rate ($P < .001$), but not for sensitivity ($P = .067$). AI = artificial intelligence.
* Unit 4 was excluded from the per-unit comparison of cancer detection rate. As 35 additional positive studies were provided to the vendor from unit 4 for threshold calibration, the cancer detection rate reported for this unit was artificially low.

cancer detection rate were 86.1%, 95.2%, 5.4%, and 5.7 per 1000, respectively. Reader 1 detected 261 of 303 (86.1%) screening-diagnosed cancers, while Mia would have detected 277 of 303 (91.4%) cancers.

AI Performance Split by Mammography Unit and Lesion Size

Mia performance with the site-specific thresholds was significantly different across mammography units for specificity ($P < .001$) and recall rate ($P < .001$), but not for sensitivity ($P = .51$) or cancer detection rate ($P = .93$) (Table 2). We found no evidence of a difference in the sensitivity of Mia between small and large tumors (91.0% [162 of 178] and 93.7% [104 of 111], respectively; $P = .55$).

IC Recall

The test dataset contained 138 ICs. Using the site-specific thresholds, Mia would have recalled 47 (34.1%) ICs. Mia indicated to recall 15 of 56 category 1 ICs (no visible lesion); four of 14 category 2 ICs (lesion visible on review in hindsight); three of three category 3 ICs (lesion clearly visible on previous screening

mammograms); and two of nine occult ICs. Mia would have recalled a further 24 of 57 ICs not yet categorized by consensus opinion (due to COVID-19–related delays in IC review).

Discussion

AI performance could be affected by different mammography systems, impacting deployment in new settings. In this study, local calibration and per–software version thresholds were required to reduce recall rates from 47.7% to 13.0%. After threshold optimization, Mia had a higher recall rate than reader 1 (13.0% vs 5.4%) but would have detected more cancers (277 vs 261), including those missed by routine dual reporting (47 of 138). The U.K. acceptable recall rate is less than 9% in a double reading setting with arbitration (16). The Mia false-positive rate was higher than that in routine clinical practice, suggesting that Mia would be best used combined with human reader input, as recommended by the vendor. Economic and operational evaluations are required across possible implementation scenarios.

Our results are supported by previous research observing issues relating to the generalizability of radiology AI models (3,5,17). Furthermore, we have established that AI performance can be influenced by different mammography systems. The AI had previously been calibrated on a range of mammography units, including the Hologic Lorad Selenia, an older model of the unit employed in this study (Hologic Selenia Dimensions). The software update applied to the mammography units included several enhancements that may affect image characteristics. Human reader performance was not adversely affected following the update. Independent verification of vendor-reported transferability of thresholds using the same mammography unit and software version elsewhere is needed.

A user-definable threshold could allow centers to perform threshold recalibration themselves. However, many centers would struggle to gather enough data and/or will lack the technological expertise to adjust the thresholds successfully. A national implementation and validation framework for AI in breast cancer screening, alongside representative national datasets, could help set AI decision thresholds and quality assurance standards.

Study strengths included using a retrospective unenriched dataset consecutively acquired in a dual reporting screening setting, with sufficient follow-up to capture screen-detected cancers and ICs. The AI was not trained on the dataset. Exclusions were minimal (3.9%).

Study limitations included the following: the evaluation of one AI product, a single-center setting, a predominantly White patient sample group, and the unavailability of IC information because of COVID-19–related delays. Also, post hoc analyses of performance stratified by mammography unit and lesion size were not adequately powered and require further evaluation in larger studies.

As different mammography systems can substantially affect AI performance, AI performance and decision thresholds should be validated when applied in new clinical settings. Quality assurance systems, including change management, should monitor AI algorithms for consistent performance.

Acknowledgment: We would like to thank the DaSH team, including Joanne Lumsden, PhD, for their technical support.

iCAIRD Radiology Collaborators: Corri Black, Alison D. Murray, and Katie Wilde, University of Aberdeen; James D. Blackwood, NHS Greater Glasgow and Clyde; Claire Butterly and John Zurowski, University of Glasgow; Jon Eilbeck and Colin McSkimming, NHS Grampian; Canon Medical Research Europe–SHAIP platform.

Author contributions: Guarantors of integrity of entire study, **C.F.d.V., R.T.S., J.A.D., L.A.A., G.L.**; 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, **C.F.d.V., R.T.S., D.D., M.B., G.L.**; clinical studies, **R.T.S., G.L.**; experimental studies, **S.J.C., R.T.S., D.J.H.**; statistical analysis, **C.F.d.V., S.J.C., R.T.S., J.A.D.**; and manuscript editing, **C.F.d.V., S.J.C., R.T.S., J.A.D., D.D., M.B., D.J.H., L.A.A., G.L.**

Disclosures of conflicts of interest: C.F.d.V. No relevant relationships. **S.J.C.** No relevant relationships. **R.T.S.** No relevant relationships. **J.A.D.** No relevant relationships. **J.Y.** Employed by Kheiron Medical Technologies; support for attending meetings/travel from Kheiron Medical Technologies; patents planned, issued, or pending with Kheiron Medical Technologies; stock or stock options in Kheiron Medical Technologies. **D.D.** Full-time employee of Kheiron Medical Technologies, supplier of the medical device evaluated in this project; grant from Innovate UK via iCAIRD, the industrial center for AI research in digital diagnostics, all parties received grant monies for the work done; associate member of the Faculty for Clinical Informatics and a health executive in residence for the UCL Global Business School for Health (unpaid, volunteer role); stock or stock options in Kheiron Medical Technologies (employee share options benefit scheme). **M.B.** iCAIRD funded by Innovate UK, under the UK Research and Innovation (UKRI) Industrial Strategy Challenge Fund "From Data to Early Diagnosis in Precision Medicine" challenge. **D.J.H.** Receipt of research award (chief investigator) from Innovate UK/UKRI, this funding underpinned the research infrastructure and some staff time. **L.A.A.** Funding from Innovate UK. **G.L.** No relevant relationships.

References

1. Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021;374:n1872.
2. Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for mammographic image analysis in breast cancer screening. Rapid review and evidence map. 2022. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1057279/UK_NSC_evidence_summary_-_the_use_of_AI_for_mammographic_image_analysis_in_breast_cancer_screening.pdf. Accessed August 26, 2022.
3. Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet Digit Health 2022;4(5):e351–e358.
4. 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.
5. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022;4(3):e210064.
6. Sharma N, Ng AY, James JJ, et al. Large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. medRxiv 2021.02.26.21252537 [preprint] https://doi.org/10.1101/2021.02.26.21252537. Posted March 01, 2021. Accessed December 16, 2021.
7. Workman A, Castellano I, Kulama E, Lawinski CP, Marshall N, Young KC. Commissioning and routine testing of full field digital mammography systems. NHS Cancer Screening Programmes. NHSBSP Equipment Report 0604, Version 3, April 2009. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/441857/nhsbsp-equipment-report-0604.pdf.
8. Canon Medical Research Europe L. Safe Haven Artificial Intelligence Platform (SHAIP). https://research.eu.medical.canon/specialism/technology-research-and-development/shaip. Accessed December 18, 2021.
9. Carrell D, Malin B, Aberdeen J, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. J Am Med Inform Assoc 2013;20(2):342–348.

10. Gao C, McGilchrist M, Mumtaz S, et al. A national network of safe havens: Scottish perspective. J Med Internet Res 2022;24(3):e31684.
11. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
12. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404–413.
13. 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. Updated February 25, 2021. Accessed January 28, 2022.
14. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.
15. Kheiron Medical Technologies. Warnings & Cautions. Mia User Manual, 2021:8.
16. Public Health England. NHS Breast screening programme screening standards valid for data collected from 1 April 2021. https://www.gov.uk/government/publications/breast-screening-consolidated-programme-standards/nhs-breast-screening-programme-screening-standards-valid-for-data-collected-from-1-april-2021#bsp-s07-referral-rate-of-referral-to-assessment. Updated March 31, 2021. Accessed January 28, 2022.
17. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health 2022;4(5):e384–e397. [Published correction appears in Lancet Digit Health 2022;4(6):e405.]