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

Article type: AI in brief

Abbreviations:
AI: Artificial Intelligence
NHS: National Health Service
UK: United Kingdom

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.

Summary statement:

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.

Abstract

Artificial intelligence (AI) tools may assist breast screening mammography programmes, but limited evidence supports their generalisability to new settings. This retrospective study used a three-year dataset (1/04/2016-31/03/2019) from a UK regional screening programme. The performance of a commercially available breast screening AI algorithm was assessed with a pre-specified and a site-specific decision threshold to evaluate whether its performance was transferable to a new clinical site. The dataset consisted of women who attended routine screening (50-70 years), excluding technical recalls, self-referrals, and those with a previous mastectomy, complex physical requirements or without the four standard image views. In total, 55,916 screening attendees (mean age, 60 ± 6 [SD] years) met the inclusion criteria. The pre-specified threshold resulted in high recall rates (48.3%; 21,929/45,444), which reduced to 13.0% (5,896/45,444) following threshold calibration, closer to the observed service level (5.0%; 2,774/55,916). Recall rates also increased approximately three-fold following a software upgrade on the mammography equipment, requiring per-software version thresholds. Using software-specific thresholds, the AI algorithm would have recalled 277/303 (91.4%) screen-detected cancers and 47/138 (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.

Introduction

A recent United Kingdom (UK) National Screening Committee review (3,4) concluded that evidence was insufficient to support the implementation of AI in routine breast cancer screening. The review identified limited evidence on sources of variability, impact on interval cancers detected between screening cycles, and performance of a pre-set threshold to classify recall or no recall. In addition, evidence for the transferability of AI models is inconsistent (5-7).

We evaluated a commercial AI software (8) using data from a UK Screening Programme to determine whether its performance transferred to an external dataset generated with different mammography equipment. The AI software is CE-marked (CE: Conformité Européenne), indicating compliance with applicable European Union (EU) regulations. This study evaluates generalisability of the AI tool using consecutively acquired clinical data, comparing stand-alone performance to the dual reporting system in the UK screening service.

Materials and Methods

Sample

The Proportionate Review Sub-committee of the London - Bloomsbury Research Ethics Committee approved this retrospective study (20/LO/0563). Secondary use of de-identified data negated the requirement for individual consent. Public Benefit and Privacy Panel (PBPP) approval was obtained (1920-0258).

National Health Service (NHS) Grampian clinical data and mammograms were collected from the Scottish Breast Screening Service (SBSS) (12/02/2016-31/03/2020). Full-field digital mammography (FFDM) images were acquired on five mammography X-ray units of the same make and model (make: Hologic; model: Selenia Dimensions) with no known differences at study commencement. All units conform to NHS breast cancer screening quality standards (9). The standard imaging protocol consisted of 2 views per breast [craniocaudal (CC) and mediolateral oblique (MLO)]. As part of routine screening, two readers interpreted each set of images with a third reader arbitrating in cases of disagreement. During the study period, mammograms in the screening centre were routinely read by a pool of 11 readers with 1 to 20 years of experience, led by GL.

The evaluation dataset was limited to a 3-year UK screening cycle (1/04/2016-31/03/2019) of women (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 (PACS) were transferred to the Safe Haven Artificial Intelligence Platform (SHAIP) developed by Canon Medical Research Europe (10). “Hidden in Plain Sight” (11) de-identification was performed.

Mia™ (version 2.0.1), developed by Kheiron Medical Technologies (vendor), assessed mammograms for potential malignancies in SHAIP. Mia™ was previously trained and tested on images acquired on Hologic, GE Healthcare, Siemens and IMS Giotto mammography equipment. Mia™, an ensemble of deep learning algorithms, employs the four standard image views (FFDM CC and MLO 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 using a predefined threshold ($\geq$0.1117 indicates recall) (8) and site-specific threshold.

Mia’s™ performance was evaluated by academic health data scientists (CFDV, JAD) in DaSH (12), which the vendor could not access. 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 positives (women with histologically confirmed cancer), 4000 confirmed negatives (women negative for cancer with a negative 3-year follow-up screening and no interval cancer) and 8000 unconfirmed negatives (Appendix E1).

Statistical Analysis

A receiver operator characteristic (ROC) curve was plotted, and the area under the curve (AUC) and confidence interval (CI; DeLong method (13)) were calculated. Positive screens were defined as histologically confirmed cancers detected through standard screening.

Sensitivity, specificity, positive and negative predictive values (PPV and NPV, respectively), as well as cancer detection and recall rates of Mia™, with CIs (Clopper-Pearson method (14)), were calculated for the pre-specified 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 pre-specified 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 performance of the first reader (Reader 1). Mia™ was not compared with the second reader as, in the UK, they can access the first reader's opinion and therefore do not read independently.

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

Interval cancers (cancers not detected during routine screening but identified between screening rounds) were analysed separately. Following individual review, all readers in the clinical team regularly met to form a consensus on cancer visibility on prior screening mammograms (15): 1 - no visible lesion; 2 - lesion visible on review in hindsight; 3 - lesion clearly visible; and Occult - lesion not visible on screening or subsequent symptomatic imaging. The proportion of interval cancer 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 E3. ROC curves, AUC and CIs were generated using the pROC package (16). Sample size information is available in Appendix E2. P<0.05 was considered to indicate a statistically significant difference.

Data availability

The statistical output alongside the relevant R code is available in Appendix E3. Access to the raw SBSS data and mammograms (de-identified participant data) is subject to the required approvals (e.g. PBPP, NHS R&D, REC 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% (2,293/58,209)] (17), an evaluation dataset of 55,916 screens was used (Figure 1). Of these 2,774 (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 interval cancers were detected in follow-up (Table 1).

AI performance pre-threshold 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 E4.

For the pre-specified 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 thousand. 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 thousand.

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 (Figure 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% [63/760] to 13.2% [183/1,382]; for version 1.8, it ranged from 23.8% [79/332] to 38.6% [86/223]. In comparison, the Reader 1 monthly recall rate ranged from 3.8% [37/966] to 6.9% [84/1,218] pre-software update and from 2.5% [7/282] to 7.9% [13/164] post-software update. Reader 1 sensitivity and specificity changed from 85.4% [328/384] to 87.9% [58/66], and from 95.1% [43,075/45,276] to 95.6% [9,746/10,190], respectively.

Per-software version thresholds were generated to ensure stability of recall rates (Appendix E1). 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, post-software upgrade) to reduce the threshold's susceptibility to noise.

Two site-specific thresholds were generated across all mammography units: 0.2712 pre-upgrade and 0.4319 post-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 thousand (Table 2). By comparison, Reader 1 sensitivity, specificity, recall rate and cancer detection rate were 86.1%, 95.2%, 5.4%, and 5.7 per thousand. Reader 1 detected 261/303 (86.1%) screening diagnosed cancers, while Mia™ would have detected 277/303 (91.4%) cancers.

AI performance split by mammography X-ray unit and lesion size

Mia™ performance with the site-specific thresholds was significantly different across mammography units for specificity (p<0.001) and recall rate (p<0.001), but not for sensitivity (p=0.51) or cancer detection rate (p=0.93), Table 2. We found no evidence of a difference in sensitivity of Mia™ between small and large tumours (91.0% [162/178] and 93.7% [104/111], respectively; p=0.55).

Interval cancers

The test dataset contained 138 interval cancers (ICs). Using the site-specific thresholds, Mia™ would have recalled 47 (34.1%) ICs. Mia™ indicated to recall 15 out of 56 category 1 ICs (no visible lesion); 4 out of 14 category 2 ICs (lesion visible on review in hindsight); 3 out of 3 category 3 ICs (lesion clearly visible on previous screening mammograms); and 2 out of 9 Occult ICs. Mia would have recalled a further 24/57 ICs not yet categorised by consensus opinion (due to Covid-related delays in interval cancer 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%. Mia™ post-threshold optimisation 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 (47/138). The UK acceptable recall rate is <9% in a double reading setting with arbitration (18). The Mia™ false positive rate was higher than 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 generalisability of radiology AI models (5,7,19). 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 (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 centres to perform threshold recalibration themselves. However, many centres 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 include using a retrospective unenriched dataset consecutively acquired in a dual reporting screening setting, with sufficient follow-up to capture screen-detected and interval cancers. The AI was not trained on the dataset. Exclusions were minimal (3.9%).

Study limitations include the following: 1) the evaluation of one AI product, 2) single centre setting, 3) a predominantly white Caucasian sample, and 4) detailed interval cancer information was not available due to Covid-related delays. Post-hoc analyses of performance stratified by mammography unit and lesion size were not adequately powered and require further evaluation in larger studies.

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.

Authors and contributors

DJH, MB and DD contributed to the conception of the study. LAA, GL and RTS designed the study. CFDV, SJC and JAD generated the evaluated dataset (data extraction and quality control) and incorporated the exclusions. CFDV and JAD performed the statistical analyses. JY performed the threshold calibration. GL provided the clinical background. CFDV, LAA, GL, RTS, SJC and JAD interpreted the data and results. All authors contributed to writing the paper, read and approved the final manuscript.

Role of the funding source

The funding source was not involved in study design, data collection, analysis and interpretation, writing of the report or the decision to submit the paper publication.

References

1. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89-94. doi: 10.1038/s41586-019-1799-6

2. Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Stand-Alone Use of Artificial Intelligence for Digital Mammography and Digital Breast Tomosynthesis Screening: A Retrospective Evaluation. Radiology 2021:211590. doi: 10.1148/radiol.211590

3. 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. doi: 10.1136/bmj.n1872

4. 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

5. 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. The Lancet Digital Health 2022;4(5):e351-e358. doi: 10.1016/S2589-7500(22)00004-8

6. Yala A, Mikhael PG, Strand F, et al. Multi-institutional validation of a mammography-based breast cancer risk model. Journal of Clinical Oncology 2021:JCO. 21.01337. doi: 10.1200/JCO.21.01337

7. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiology.Artificial Intelligence 2022;4(3):e210064. doi: 10.1148/ryai.210064

8. 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. doi: 10.1101/2021.02.26.21252537

9. 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 2009;0604

10. 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.

11. 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. Journal of the American Medical Informatics Association 2013;20(2):342-348. doi: 10.1136/amiajnl-2012-001034

12. Gao C, McGilchrist M, Mumtaz S, et al. A National Network of Safe Havens: Scottish Perspective. Journal of Medical Internet Research 2022;24(3):e31684. doi: 10.2196/31684

13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988:837-845. doi: 10.2307/2531595

14. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934:404-413. doi: 10.2307/2331986

15. 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.

16. 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):1-8. doi: 10.1186/1471-2105-12-77

17. Kheiron Medical Technologies. Warnings & Cautions. Mia™ User Manual 2021:8.

18. Public Health England. NHS Breast screening programme screening standards valid for data collected from 1 April 202 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.

19. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. The Lancet Digital Health 2022. doi: 10.1016/S2589-7500(22)00003-6

Figure Legends

Figure 1: Flow diagram showing 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 positives are women with histologically confirmed cancer. Confirmed negatives are women negative for cancer with a negative 3-year follow-up screening and no interval cancer. DICOM = Digital Imaging and Communications in Medicine, UK = United Kingdom

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 pre-specified threshold. The original dataset was not used to establish the pre-specified threshold. **b**: Rise in recall rate after an event for the four mammography X-ray units. The vertical dashed line indicates the date of a software upgrade. A fifth unit, a floating service mobile unit, was not upgraded during the study timeline and is not included in this figure.

Tables

Table 1: Cohort Characteristics – UK Breast Screening Program (01/04/2016-31/03/2019).

Original dataset	**N = 55,916**	**%**
Age (Years)		**(% of cohort)**
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 Requirements	**1,048 (1.9%)**	**(% of cohort)**
Learning Difficulties	116	0.2%
Language Needs	304	0.5%
Implant	364	0.7%
Deaf	182	0.3%
Blind	40	0.07%
Special Needs	30	0.05%
Two special requirements	12	0.02%
Screen-detected Breast Cancers	**450 (0.8%)**	**(% of all screen-detected cancers)**
Type of Cancer		
Non-breast primary tumour	2	0.4%
Ductal Carcinoma in situ (DCIS, pre-invasive)	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%
Tumour size		
<15mm	259	57.6%
≥15mm	169	37.6%
Unknown	22	4.9%
Interval Cancers	**156 (0.3%)**	**(% of all interval cancers)**
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%
Tumour size		
<15mm	24	15.4%
≥15mm	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.—*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 present as palpable masses not discernible or outwith the mammographic image. UK = United Kingdom

Table 2: Mia™ Performance on Screen-detected Cancers

AI and Reader 1 performance	Number of datapoints	Number of cancers	% Sensitivity		% Specificity		% Positive predictive value		% Negative predictive value		% Recall rate		Cancer detection rate per thousand	
			Value	*95% CI*	*Value*	*95% CI*	*Value*	*95% CI*	*Value*	*95% CI*	*Value*	*95% CI*	*Value*	*95% CI*
Mia™ - original dataset														
Pre-specified threshold	55,916	450	97.3 [438/450]	95.4 to 98.6	52.7 [29,233/55,466]	52.3 to 53.1	1.6 [438/26,671]	1.5 to 1.8	99.96 [29,233/29,245]	99.93 to 99.98	47.7 [26,671/55,916]	47.3 to 48.1	7.8 [438/55,916]	7.1 to 8.6
Mia™ - test dataset														
Pre-specified threshold	45,444	303	98.3 [298/303]	96.2 to 99.5	52.1 [23,510/45,141]	51.6 to 52.5	1.4 [298/21,929]	1.2 to 1.5	99.98 [23,510/23,515]	99.95 to 99.99	48.3 [21,929/45,444]	47.8 to 48.7	6.6 [298/45,444]	5.8 to 7.3
Updated thresholds	45,444	303	91.4 [277/303]	87.7 to 94.3	87.6 [39,522/45,141]	87.2 to 87.9	4.7 [277/5,896]	4.2 to 5.3	99.93 [39,522/39,548]	99.90 to 99.96	13.0 [5,896/45,444]	12.7 to 13.3	6.1 [277/45,444]	5.4 to 6.9
Reader 1 - test dataset	45,444	303	86.1 [261/303]	81.7 to 89.8	95.2 [42,956/45,141]	95.0 to 95.4	10.7 [261/2,446]	9.5 to 12.0	99.90 [42,956/42,998]	99.87 to 99.93	5.4 [2,446/45,444]	5.2 to 6.0	5.7 [261/45,444]	5.1 to 6.5
AI performance split by mammography unit														
Unit 1	13,104	94	93.6 [88/94]	86.6 to 97.6	87.8 [11,421/13,010]	87.2 to 88.3	5.25 [88/1,677]	4.2 to 6.4	99.95 [11,421/11,427]	99.89 to 99.98	12.8 [1,677/13,104]	12.2 to 13.4	6.7 [88/13,104]	5.4 to 8.3
Unit 2	9,960	78	92.3 [72/78]	84.0 to 97.1	86.2 [8,514/9,882]	85.5 to 86.8	5.0 [72/1,440]	3.9 to 6.3	99.93 [8,514/8,520]	99.85 to 99.97	14.5 [1,440/9,960]	13.8 to 15.2	7.2 [72/9,960]	5.7 to 9.1
Unit 3	13,000	95	90.5 [86/95]	82.8 to 95.6	88.7 [11,445/12,905]	88.1 to 89.2	5.6 [86/1,546]	4.5 to 6.8	99.92 [11,445/11,454]	99.85 to 99.96	11.9 [1,546/13,000]	11.3 to 12.5	6.6 [86/13,000]	5.3 to 8.2
Unit 4	8,541	31	83.9 [26/31]	66.3 to 94.5	87.3 [7,433/8,510]	86.6 to 88	2.4 [26/1,103]	1.6 to 3.4	99.93 [7,433/7,438]	99.84 to 99.98	12.9 [1,103/8,541]	12.2 to 13.6	3.0 [26/8,541] *****	2.0 to 4.5
Unit 5	839	5	100.0 [5/5]	47.8 to 100.0	85 [709/834]	82.4 to 87.4	3.85 [5/130]	1.3 to 8.8	100.00 [709/709]	99.48-100.00	15.5 [130/839]	13.1 to 18.1	6.0 [5/839]	1.9 to 13.9

Note—Chi-squared tests (or Fisher's exact tests when there were small counts in the contingency table) were performed to determine whether the pre-set threshold performance was significantly different to 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 pre-set and site-specific thresholds ($p < 0.001$). There were significant differences between the site-specific threshold and Reader 1 for specificity, recall and cancer detection rate ($p < 0.001$), but not for sensitivity ($p = 0.067$). AI = artificial intelligence. ***** Unit 4 was excluded from the per-unit comparison of cancer detection rate. Since 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.

Appendix E1 – Threshold calibration

Studies in the validation dataset were randomly selected following exclusions [non-double read, technical recalls, repeated image views, previous breast cancer or malignant operation, four standard image views unavailable, and those which could not be definitively linked to clinical data].

The Mia™ predictions and the clinical outcomes for the confirmed (negative or positive) studies were utilised to generate a receiver operating characteristic (ROC) curve. The updated threshold jointly maximised sensitivity and specificity, i.e. the true positive rate (TPR) was high and the false positive rate (FPR) was low, by choosing the threshold p which satisfies: $arg\ \underset{p}{min}|TPR(p) - 1.0 + FPR(p)|$ (1). arg min (argument of the minimum) returns the p which minimises the function $|TPR(p) - 1.0 + FPR(p)|$.

For the per-software version thresholds, the validation dataset was split to create one dataset before the update and one after the update. Further, the definition for confirmed negative was widened to include any non-positive study (without requirement for non-positive follow-up at least 3 years later) for the post-software update threshold calibration only, to ensure sufficient screens were available. The pre- and post-software update datasets (for threshold calibration) consisted of N = 9,672 screens (118 positive) and N = 6,341 screens (97 positive), respectively.

References

1. Sanchez IE. Optimal threshold estimation for binary classifiers using game theory. F1000Research 2016;5. doi: 10.12688/f1000research.10114.3

Appendix E2 - Sample size calculation

Minimum sample size was calculated by an independent statistician (Quantics, CRO: Veristat). The aimed precision of the 95% confidence interval (CI; exact Clopper-Pearson) was 5% for sensitivity and specificity, and 1% for recall rate (proportion of screens recalled). Assuming sensitivity and specificity at 80%, 264 confirmed positive and 264 confirmed negative screens were required. For an estimated 15% recall rate, 85% power was estimated with 5,150 screens. No sample sizes were calculated for exploratory endpoints.

Appendix E3 – Statistical output with corresponding R code

```
library(tidyverse)
library(pROC)
library(lubridate)

# loads data frame
data_full <- readRDS("dataset_full.rds")

# 3-year subset
data <-  filter(data_full, StudyDate >= as.POSIXct("2016/04/01
") & StudyDate <= as.POSIXct("2019/03/31"))

# pre-specified (or out-of-the-box, OOB) threshold
threshold_OOB <- 0.11169749509569679

# initial site-specific threshold
new_threshold_overall <- 0.29380058497190475

# pre and post software update thresholds
threshold_pre <- 0.271178413182497
threshold_post <- 0.4318937659263611
```

1 Custom Functions

1.1 Calculate sensitivity

```
calc_sens = function(recall, cancer){
  TP <- recall == 1 & cancer == 1
  FN <- recall == 0 & cancer == 1

  # sensitivity --> Number of TP/(Number of TP + Number of FN)
  #Sensitivity <- sum(TP)/(sum(TP) + sum(FN))
  binom_test <- binom.test(c(sum(TP), sum(FN)))
}
```

1.2 Calculate specificity

```
calc_spec = function(recall, cancer){
  TN <- recall == 0 & cancer == 0
  FP <- recall == 1 & cancer == 0

  # specificity --> Number of TN/(Number of TN + Number of FP)
  #Specificity <- sum(TN)/(sum(TN) + sum(FP))
  binom_test <- binom.test(c(sum(TN), sum(FP)))
}
```

1.3 Calculate positive predictive value (PPV)

```
calc_PPV = function(recall, cancer){
  TP <- recall == 1 & cancer == 1
  FP <- recall == 1 & cancer == 0

  # PPV --> Number of TP/(Number of TP + Number of FP)
  #PPV <- sum(TP)/(sum(TP) + sum(FP))
  binom_test <- binom.test(c(sum(TP), sum(FP)))
}
```

1.4 Calculate negative predictive value (PPV)

```
calc_NPV = function(recall, cancer){
  TN <- recall == 0 & cancer == 0
  FN <- recall == 0 & cancer == 1

  # NPV --> Number of TN/(Number of TN + Number of FN)
  #NPV <- sum(TN)/(sum(TN) + sum(FN))
  binom_test <- binom.test(c(sum(TN), sum(FN)))
}
```

1.5 Calculate recall rate

```
calc_RR = function(recall){
  #RR <- sum(recall == 1)/sum(recall == 1 | recall == 0)
  binom_test <- binom.test(c(sum(recall == 1), sum(recall == 0
)))
}
```

1.6 Calculate cancer detection rate

```
calc_CDR = function(recall, cancer){
  #CDR <- sum(cancer[recall == 1])/sum(cancer == 1 | cancer ==
0)
  binom_test_CDR <- binom.test(sum(cancer[recall == 1]), sum(c
ancer == 1 | cancer == 0))
}
```

1.7 Determine threshold

```
determine_threshold <- function(Unit, StudyDate){

  # pre and post software update decision thresholds
  threshold_pre <- 0.271178413182497
  threshold_post <- 0.4318937659263611

  # Estimated dates of software upgrade on the X-ray units (format: yyyy/mm/dd)
  # Unit 5 was not updated during the course of the study
  unit1_upgrade <- as.Date("2019/02/01")
  unit2_upgrade <- as.Date("2019/03/01")
  unit3_upgrade <- as.Date("2019/05/01")
  unit4_upgrade <- as.Date("2016/09/01")

  Mia_threshold = case_when(
    Unit == '1' & StudyDate > unit1_upgrade ~ threshold_post,
    Unit == '2' & StudyDate > unit2_upgrade ~ threshold_post,
    Unit == '3' & StudyDate > unit3_upgrade ~ threshold_post,
    Unit == '4' & StudyDate > unit4_upgrade ~ threshold_post,
    TRUE ~ threshold_pre
  )
}
```

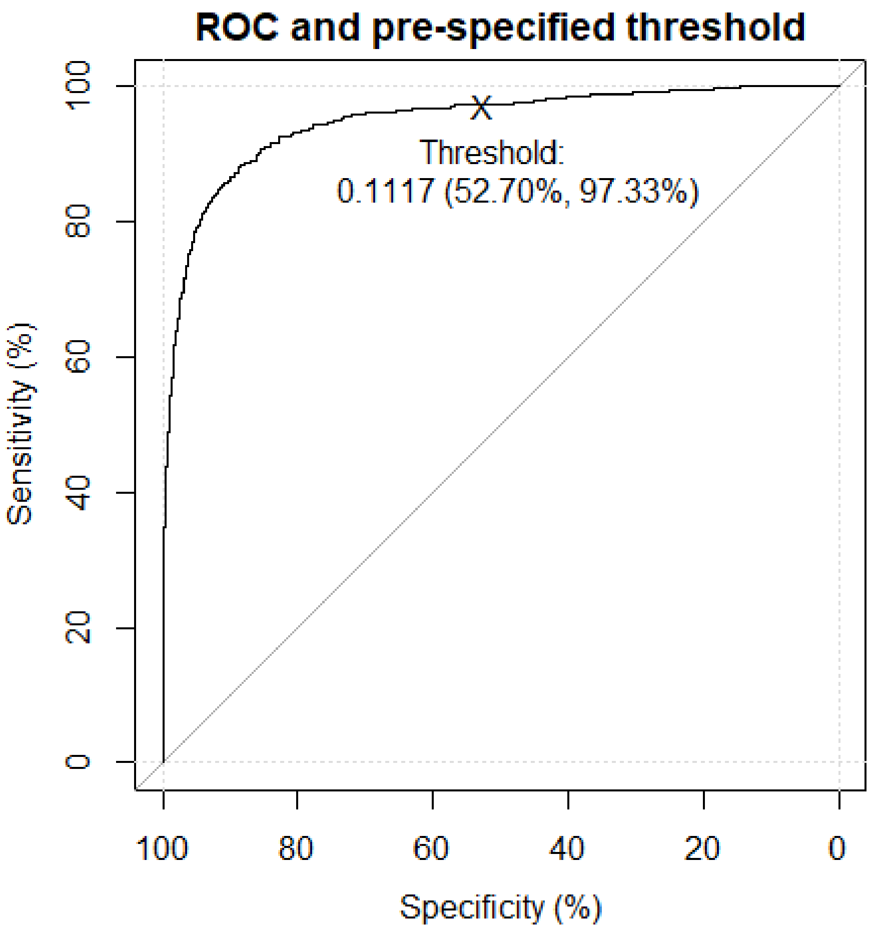
2 Original dataset

Number of datapoints: 55916

Number of cancers: 450

2.1 ROC curve

```
# ROC plot on all data with pre-specified threshold indicated
ROC <- roc(response = data$cancer, predictor = data$MIA, percent=T)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot.roc(ROC, grid.v = c(100,0), grid.h = c(100,0), percent=T, type='l', lty = 1, lwd = 1,
         print.thres = threshold_OOB,
         print.thres.pattern = "Threshold: \n%.4f (%.2f%%, %.2f%%)",
         print.thres.pch = 'X',
         print.thres.adj = c(0.4,1.5))
title(main='ROC and pre-specified threshold', line = 2.5)
```



```
print(auc(ROC), digits = 4)

## Area under the curve: 94.76%

print(ci.auc(ROC), digits = 4)

## 95% CI: 93.64%-95.89% (DeLong)
```

2.2 Evaluation of pre-specified threshold on screen detected cancers

Pre-specified threshold: 0.111697

```
data$Mia_recall_OOB <- ifelse(data$MIA >= threshold_OOB, 1, 0)

RR <- calc_RR(data$Mia_recall_OOB)
```

```
CDR <- calc_CDR(data$Mia_recall_OOB, data$cancer)

sensitivity <- calc_sens(data$Mia_recall_OOB, data$cancer)
specificity <- calc_spec(data$Mia_recall_OOB, data$cancer)

PPV <- calc_PPV(data$Mia_recall_OOB, data$cancer)
NPV <- calc_NPV(data$Mia_recall_OOB, data$cancer)

perf_print = function(value, scale){

  if(missing(scale)){
    scale = "percent"
  }

  if(scale == "percent"){
    sprintf("%.2f%% [%i / %i], CI: %.2f-%.2f",
            round(value$estimate*100, 2),
            value$statistic,
            value$parameter,
            round(value$conf.int[1]*100, 2),
            round(value$conf.int[2]*100, 2)
            )
  } else if(scale == "permil"){
    sprintf("%.2f per thousand [%i / %i], CI: %.2f-%.2f",
            round(value$estimate*1000, 2),
            value$statistic,
            value$parameter,
            round(value$conf.int[1]*1000, 2),
            round(value$conf.int[2]*1000, 2)
            )
  }
}
```

Numbers in square brackets indicate the numerator & denominator. CI refers to the 95% confidence interval

- Sensitivity: 97.33% [438 / 450], CI: 95.39-98.61
- Specificity: 52.70% [29233 / 55466], CI: 52.29-53.12
- Positive predictive value (PPV): 1.64% [438 / 26671], CI: 1.49-1.80
- Negative predictive value (NPV): 99.96% [29233 / 29245], CI: 99.93-99.98
- Recall rate: 47.70% [26671 / 55916], CI: 47.28-48.11
- Cancer detection rate: 7.83 per thousand [438 / 55916], CI: 7.12-8.60

3 Test dataset

```
# create subset, which excludes data used for optimising the t
hreshold
subset <- filter(data, dataset != 'validation_v1', dataset !=
'validation_v2')

sensitivity <- calc_sens(subset$Mia_recall_OOB, subset$cancer)
specificity <- calc_spec(subset$Mia_recall_OOB, subset$cancer)

PPV <- calc_PPV(subset$Mia_recall_OOB, subset$cancer)
NPV <- calc_NPV(subset$Mia_recall_OOB, subset$cancer)

RR <- calc_RR(subset$Mia_recall_OOB)
CDR <- calc_CDR(subset$Mia_recall_OOB, subset$cancer)
```

Test dataset: subset of data not used for threshold optimisation

- Number of datapoints: 45444
- Number of cancers: 303

3.1 Evaluation of pre-specified threshold on screen detected cancers

Pre-specified threshold: 0.271178

- Sensitivity: 98.35% [298 / 303], CI: 96.19-99.46
- Specificity: 52.08% [23510 / 45141], CI: 51.62-52.54
- Positive predictive value (PPV): 1.36% [298 / 21929], CI: 1.21-1.52
- Negative predictive value (NPV): 99.98% [23510 / 23515], CI: 99.95-99.99
- Recall rate: 48.25% [21929 / 45444], CI: 47.79-48.72
- Cancer detection rate: 6.56 per thousand [298 / 45444], CI: 5.84-7.34

3.2 Evaluation of optimised thresholds

Threshold pre-software upgrade: 0.271178

Threshold post-software upgrade: 0.431894

3.2.1 Screen detected cancers

```
# determine threshold for each case
subset <- subset %>%
  mutate(Mia_threshold = determine_threshold(Unit, StudyDate))
%>%
  mutate(Mia_recall_new = ifelse(MIA >= Mia_threshold, 1, 0))

# calculate Mia performance on subset using optimised threshol
ds

sensitivity <- calc_sens(subset$Mia_recall_new, subset$cancer)
specificity <- calc_spec(subset$Mia_recall_new, subset$cancer)

PPV <- calc_PPV(subset$Mia_recall_new, subset$cancer)
NPV <- calc_NPV(subset$Mia_recall_new, subset$cancer)

RR <- calc_RR(subset$Mia_recall_new)
CDR <- calc_CDR(subset$Mia_recall_new, subset$cancer)
```

- Sensitivity: 91.42% [277 / 303], CI: 87.68-94.32
- Specificity: 87.55% [39522 / 45141], CI: 87.24-87.86
- Positive predictive value (PPV): 4.70% [277 / 5896], CI: 4.17-5.27
- Negative predictive value (NPV): 99.93% [39522 / 39548], CI: 99.90-99.96
- Recall rate: 12.97% [5896 / 45444], CI: 12.67-13.29
- Cancer detection rate: 6.10 per thousand [277 / 45444], CI: 5.40-6.85

3.2.2 Interval cancers

```
# create interval cancers data frame (IC) by filtering the tes
t dataset (subset)
IC <- subset %>%
  filter(IC == 1)

# how many would Mia correctly classify?
perc_Mia_correct <- sum(IC$Mia_recall_new == 1)/nrow(IC)*100
```

Number of interval cancers: 138

Tumour size:

```
size <- IC$NISize + IC$InvSize
size_cat <- if_else(size >= 15, ">=15mm", "<15mm")
table(size_cat, useNA = "ifany")
```

```
## size_cat
##  <15mm >=15mm
##     81     57
```

Invasive?

```
table(IC$Invasive, useNA = "ifany")
```

```
## 
##  No Yes 
##  14 124
```

Group Opinion [Category 1: could not be detected on mammogram. Category 3: should have been detected on mammogram. Occult: could not be detected on mammogram or follow-up mammogram]:

```
table(IC$GroupOpinion, useNA = "ifany")
```

```
## 
##                       Category 1 - Satisfactory 
##                                              56 
## Category 2 - Satisfactory with learning points 
##                                              14 
##                     Category 3 - Unsatisfactory 
##                                               3 
##                                          Occult 
##                                               9 
##                                            <NA> 
##                                              56
```

Invasive Grade (NA indicates DCIS):

```
table(IC$InvasiveGrade, useNA = "ifany")
```

```
## 
##    I   II  III <NA> 
##    5   64   58   11
```

Number of interval cancers recalled by Mia using new threshold: 47 (34.06%)

Mia recalls divided by Group Opinion ('1' indicates recall, '0' indicates don't recall):

```
## 
##                                                    0  1
##   Category 1 - Satisfactory                      41 15
##   Category 2 - Satisfactory with learning points 10  4
##   Category 3 - Unsatisfactory                     0  3
##   Occult                                                    7  2
##   <NA>                                                     33 23
```

3.3 Reader 1 performance on screen detected cancers

```r
# Combine the first reader's opinion on left and right breast to determine overall Reader 1 opinion
subset <- subset %>%
  mutate(Reader1Opinion = case_when(
  OpinionLeft_1 == OpinionRight_1 ~ OpinionLeft_1,
  OpinionLeft_1 == 'Review Required' | OpinionRight_1 == 'Review Required' ~ 'Review Required',
  OpinionLeft_1 == 'Review (Symptoms)' | OpinionRight_1 == 'Review (Symptoms)' ~ 'Review (Symptoms)',
  OpinionLeft_1 == 'Routine Recall' | OpinionRight_1 == 'Routine Recall' ~ 'Routine Recall'
))

subset$Reader1Opinion[subset$Reader1Opinion == 'Review Required' | subset$Reader1Opinion == 'Review (Symptoms)'] <- 1
subset$Reader1Opinion[subset$Reader1Opinion == 'Routine Recall'] <- 0

R1_sens <- calc_sens(subset$Reader1Opinion, subset$cancer)
R1_spec <- calc_spec(subset$Reader1Opinion, subset$cancer)
R1_PPV <- calc_PPV(subset$Reader1Opinion, subset$cancer)
R1_NPV <- calc_NPV(subset$Reader1Opinion, subset$cancer)
R1_RR <- calc_RR(subset$Reader1Opinion)
R1_CDR <- calc_CDR(subset$Reader1Opinion, subset$cancer)
```

- Reader 1 sensitivity: 86.14% [261 / 303], CI: 81.73-89.82
- Reader 1 specificity: 95.16% [42956 / 45141], CI: 94.96-95.36
- Reader 1 positive predictive value (PPV): 10.67% [261 / 2446], CI: 9.47-11.96
- Reader 1 negative predictive value (NPV): 99.90% [42956 / 42998], CI: 99.87-99.93
- Reader 1 recall rate: 5.38% [2446 / 45444], CI: 5.18-5.59
- Reader 1 cancer detection rate: 5.74 per thousand [261 / 45444], CI: 5.07-6.48

Reader 1 missed 42 cancers, while Mia missed 26 cancers

3.4 Per-machine evaluation on screen detected cancers

```r
# Calculate performance for each X-ray Unit
subset %>%
  group_by(Unit) %>%
  summarise(N = n(),
            N_cancers = sum(cancer == 1),
            sens = perf_print(calc_sens(Mia_recall_new, cancer
)),
            spec = perf_print(calc_spec(Mia_recall_new, cancer
)),
            PPV = perf_print(calc_PPV(Mia_recall_new, cancer))
,
            NPV = perf_print(calc_NPV(Mia_recall_new, cancer))
,
            RR = perf_print(calc_RR(Mia_recall_new)),
            CDR = perf_print(calc_CDR(Mia_recall_new, cancer),
scale = "permil")
            ) %>%
  knitr::kable(.)
```

Unit	N	N_cancers	sens	spec	PPV	NPV	RR	CDR
1	13104	94	93.62% [88 / 94], CI: 86.62-97.62	87.79% [11421 / 13010], CI: 87.21-88.34	5.25% [88 / 1677], CI: 4.23-6.43	99.95% [11421 / 11427], CI: 99.89-99.98	12.80% [1677 / 13104], CI: 12.23-13.38	6.72 per thousand [88 / 13104], CI: 5.39-8.27
2	9960	78	92.31% [72 / 78], CI: 84.01-97.12	86.16% [8514 / 9882], CI: 85.46-86.83	5.00% [72 / 1440], CI: 3.93-6.26	99.93% [8514 / 8520], CI: 99.85-99.97	14.46% [1440 / 9960], CI: 13.77-15.16	7.23 per thousand [72 / 9960], CI: 5.66-9.10
3	13000	95	90.53% [86 / 95], CI: 82.78-95.58	88.69% [11445 / 12905], CI: 88.13-89.23	5.56% [86 / 1546], CI: 4.47-6.82	99.92% [11445 / 11454], CI: 99.85-99.96	11.89% [1546 / 13000], CI: 11.34-12.46	6.62 per thousand [86 / 13000], CI: 5.29-8.16
4	8541	31	83.87% [26 /	87.34%	2.36%	99.93% [7433 /	12.91%	3.04 per

Unit	N	N_cancers	sens	spec	PPV	NPV	RR	CDR
			31], CI: 66.27-94.55	[7433 / 8510], CI: 86.62-88.04	[26 / 1103], CI: 1.55-3.43	7438], CI: 99.84-99.98	[1103 / 8541], CI: 12.21-13.64	thousand [26 / 8541], CI: 1.99-4.46
5	839	5	100.00% [5 / 5], CI: 47.82-100.00	85.01% [709 / 834], CI: 82.41-87.37	3.85% [5 / 130], CI: 1.26-8.75	100.00% [709 / 709], CI: 99.48-100.00	15.49% [130 / 839], CI: 13.11-18.12	5.96 per thousand [5 / 839], CI: 1.94-13.85

RR, recall rate; CDR, cancer detection rate; sens, sensitivity; spec, specificity.

3.4.1 Are there differences in performance across mammography X-ray units?

3.4.1.1 Sensitivity

```
sens_subset <- filter(subset, cancer == 1)
#Exact Fisher's test due to small counts in contingueny table
fisher.test(table(sens_subset$Unit, sens_subset$Mia_recall_new
))

##
##  Fisher's Exact Test for Count Data
##
## data:  table(sens_subset$Unit, sens_subset$Mia_recall_new)
## p-value = 0.51
## alternative hypothesis: two.sided
```

3.4.1.2 Specificity

```
spec_subset <- filter(subset, cancer == 0)
#Chi-squared test
chisq.test(table(spec_subset$Unit, spec_subset$Mia_recall_new)
)

##
##  Pearson's Chi-squared test
##
```

```
## data:  table(spec_subset$Unit, spec_subset$Mia_recall_new)
## X-squared = 38.83, df = 4, p-value = 7.57e-08
```

3.4.1.3 Recall rate

```
RR_subset <- subset
#Chi-squared test
chisq.test(table(RR_subset$Unit, RR_subset$Mia_recall_new))

## 
##  Pearson's Chi-squared test
## 
## data:  table(RR_subset$Unit, RR_subset$Mia_recall_new)
## X-squared = 38, df = 4, p-value = 1.12e-07
```

3.4.1.4 Cancer detection rate

Unit 4 was excluded from the per-unit comparison of cancer detection rate. 35 additional positive studies from this unit were provided to the vendor. Therefore, the cancer detection rate reported for this unit (for the test dataset) was artificially low.

```
CDR_subset <- filter(subset, Unit != "4")

# cancers detected by MIA
CDR_subset$cancer_MIA <- if_else(CDR_subset$cancer == 1 & CDR_
subset$Mia_recall_new == 1, 1, 0)

#Chi-squared test
chisq.test(table(CDR_subset$Unit, CDR_subset$cancer_MIA))

## 
##  Pearson's Chi-squared test
## 
## data:  table(CDR_subset$Unit, CDR_subset$cancer_MIA)
## X-squared = 0.4384, df = 3, p-value = 0.932
```

3.5 Performance on screen detected cancers stratified by lesion size

```
subset_small <- subset %>%
  filter(is.na(size_tumour) | size_tumour ==  "<15mm")

Mia_sens_small <- calc_sens(subset_small$Mia_recall_new, subse
t_small$cancer)

subset_big <- subset %>%
  filter(is.na(size_tumour) | size_tumour == ">=15mm")

Mia_sens_big <- calc_sens(subset_big$Mia_recall_new, subset_bi
g$cancer)

sens_subset <- filter(subset, size_tumour == "<15mm" | size_tu
mour == ">=15mm")
```

- Mia performance on lesions <15mm: 91.01% [162 / 178], CI: 85.81-94.77
- Mia performance on lesions >=15mm: 93.69% [104 / 111], CI: 87.44-97.43

3.5.1 Is there a difference in performance between small and large tumours?

```
chisq.test(table(sens_subset$size_tumour, sens_subset$Mia_reca
ll_new))

## 
##  Pearson's Chi-squared test with Yates' continuity correcti
on
## 
## data:  table(sens_subset$size_tumour, sens_subset$Mia_recal
l_new)
## X-squared = 0.3553, df = 1, p-value = 0.551
```

4 Mia recall rate range pre and post software update

Monthly recall rate range pre and post software update with the initial recalibrated site-specific threshold (0.293801). Minimum and maximum recall rate shown

```
data_full$Mia_recall_overall <- ifelse(data_full$MIA >= new_th
reshold_overall, 1, 0)

# mammography unit upgrade dates
unit1_upgrade <- as.Date("2019/02/01")
unit2_upgrade <- as.Date("2019/03/01")
unit3_upgrade <- as.Date("2019/05/01")
unit4_upgrade <- as.Date("2016/09/01")

# exclude validation set
# exclude Feb 2016 as only partial data is available for that
month
```

```
# determine software version

data_subset <- data_full %>%
  filter(dataset != 'validation_v1', dataset != 'validation_v2
') %>%
  filter(StudyDate >= as.POSIXct("2016/03/01")) %>%
  mutate(software = case_when(
    Unit == '1' & StudyDate > unit1_upgrade ~ '1.8',
    Unit == '2' & StudyDate > unit2_upgrade ~ '1.8',
    Unit == '3' & StudyDate > unit3_upgrade ~ '1.8',
    Unit == '4' & StudyDate > unit4_upgrade ~ '1.8',
    TRUE ~ '1.7'
  ))

data_subset %>%
  group_by(software, month = floor_date(StudyDate, 'month')) %
>%
  summarise(recalls = sum(Mia_recall_overall), N = n(), RR = r
ecalls/N*100) %>%
  select(-month) %>%
  ungroup() %>%
  group_by(software) %>%
  arrange(RR) %>%
  slice(c(1, n())) %>%
  knitr::kable(.)

## `summarise()` has grouped output by 'software'. You can ove
rride using the
## `.groups` argument.
```

software	recalls	N	RR
1.7	63	760	8.28947
1.7	183	1382	13.24168
1.8	79	332	23.79518
1.8	86	223	38.56502

5 Reader 1 recall rate range pre and post software update

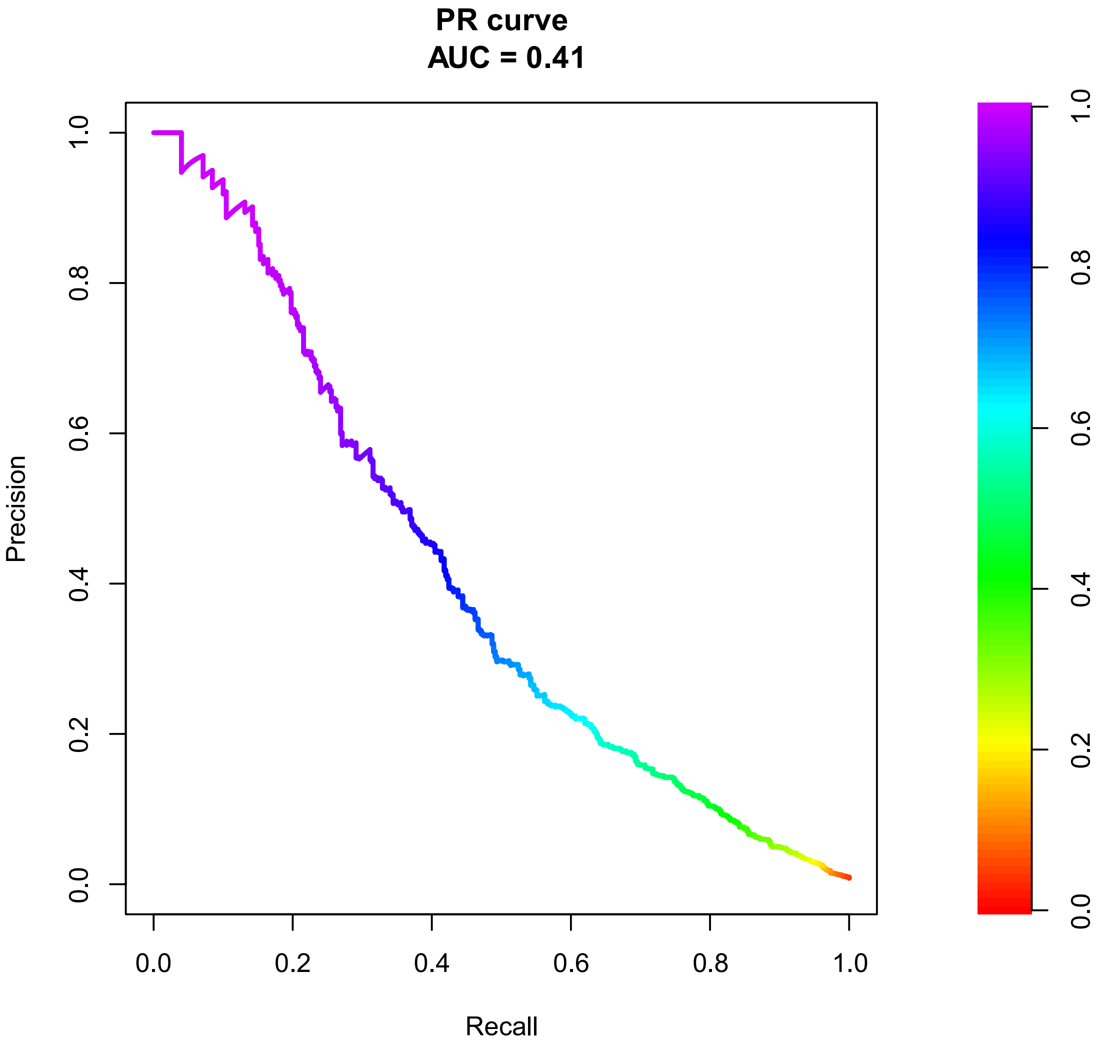
Monthly recall rate range pre and post software update. Minimum and maximum recall rate shown

```
data_full$Mia_recall_overall <- ifelse(data_full$MIA >= new_th
reshold_overall, 1, 0)
```

```
# Convert reader opinions to 1 (recall) & 0 (no recall)

data_subset <- data_subset %>%
  mutate(Reader1Opinion = case_when(
  OpinionLeft_1 == OpinionRight_1 ~ OpinionLeft_1,
  OpinionLeft_1 == 'Review Required' | OpinionRight_1 == 'Revi
ew Required' ~ 'Review Required',
  OpinionLeft_1 == 'Review (Symptoms)' | OpinionRight_1 == 'Re
view (Symptoms)' ~ 'Review (Symptoms)',
  OpinionLeft_1 == 'Routine Recall' | OpinionRight_1 == 'Routi
ne Recall' ~ 'Routine Recall'
))

data_subset$Reader1Opinion[data_subset$Reader1Opinion == 'Revi
ew Required' | data_subset$Reader1Opinion == 'Review (Symptoms
)'] <- 1
data_subset$Reader1Opinion[data_subset$Reader1Opinion == 'Rout
ine Recall'] <- 0

data_subset %>%
  group_by(software, month = floor_date(StudyDate, 'month')) %
>%
  summarise(recalls = sum(Reader1Opinion == "1"), N = n(), RR
= recalls/N*100) %>%
  select(-month) %>%
  ungroup() %>%
  group_by(software) %>%
  arrange(RR) %>%
  slice(c(1, n())) %>%
  knitr::kable(.)

## `summarise()` has grouped output by 'software'. You can ove
rride using the
## `.groups` argument.
```

software	recalls	N	RR
1.7	37	966	3.83023
1.7	84	1218	6.89655
1.8	7	282	2.48227
1.8	13	164	7.92683

Appendix E4 – Precision-recall curve



PR stands for precision recall. The y-axis shows positive predictive value (“precision”); the x-axis shows sensitivity (“recall”).

AUC: area under the curve.