Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection

M. Luke Marinovich,[a,b,*] Elizabeth Wylie,[c] William Lotter,[d,e] Helen Lund,[c] Andrew Waddell,[c] Carolyn Madeley,[c] Gavin Pereira,[b] and Nehmat Houssami[a,f]

[a]The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, New South Wales, Australia
[b]Curtin School of Population Health, Curtin University, Perth, Western Australia, Australia
[c]BreastScreen WA, Perth, Western Australia, Australia
[d]Dana-Farber Cancer Institute, Boston, MA, USA
[e]Harvard Medical School, Boston, MA, USA
[f]Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Camperdown, New South Wales, Australia

eBioMedicine
2023;90: 104498
Published Online 28 February 2023
https://doi.org/10.1016/j.ebiom.2023.104498

Summary

Background Artificial intelligence (AI) has been proposed to reduce false-positive screens, increase cancer detection rates (CDRs), and address resourcing challenges faced by breast screening programs. We compared the accuracy of AI versus radiologists in real-world population breast cancer screening, and estimated potential impacts on CDR, recall and workload for simulated AI-radiologist reading.

Methods External validation of a commercially-available AI algorithm in a retrospective cohort of 108,970 consecutive mammograms from a population-based screening program, with ascertained outcomes (including interval cancers by registry linkage). Area under the ROC curve (AUC), sensitivity and specificity for AI were compared with radiologists who interpreted the screens in practice. CDR and recall were estimated for simulated AI-radiologist reading (with arbitration) and compared with program metrics.

Findings The AUC for AI was 0.83 compared with 0.93 for radiologists. At a prospective threshold, sensitivity for AI (0.67; 95% CI: 0.64–0.70) was comparable to radiologists (0.68; 95% CI: 0.66–0.71) with lower specificity (0.81 [95% CI: 0.81–0.81] versus 0.97 [95% CI: 0.97–0.97]). Recall rate for AI-radiologist reading (3.14%) was significantly lower than for the BSWA program (3.38%) (–0.25%; 95% CI: –0.31 to –0.18; $P < 0.001$). CDR was also lower (6.37 versus 6.97 per 1000) (–0.61; 95% CI: –0.77 to –0.44; $P < 0.001$); however, AI detected interval cancers that were not found by radiologists (0.72 per 1000; 95% CI: 0.57–0.90). AI-radiologist reading increased arbitration but decreased overall screen-reading volume by 41.4% (95% CI: 41.2–41.6).

Interpretation Replacement of one radiologist by AI (with arbitration) resulted in lower recall and overall screen-reading volume. There was a small reduction in CDR for AI-radiologist reading. AI detected interval cases that were not identified by radiologists, suggesting potentially higher CDR if radiologists were unblinded to AI findings. These results indicate AI's potential role as a screen-reader of mammograms, but prospective trials are required to determine whether CDR could improve if AI detection was actioned in double-reading with arbitration.

Funding National Breast Cancer Foundation (NBCF), National Health and Medical Research Council (NHMRC).

Keywords: Breast neoplasms; Artificial intelligence; Diagnostic screening programs; Sensitivity and specificity

Introduction

Population-based breast cancer screening programs have been shown to reduce breast cancer-specific mortality through early cancer detection.[1,2] Breast screening involves the interpretation of digital mammograms to identify suspicious abnormalities that warrant further investigation ("recall to assessment"). However, screen-reading is a subjective process that can not only detect

*Corresponding author. The Daffodil Centre, The University of Sydney, A Joint Venture with Cancer Council NSW, Sydney, NSW, Australia.
E-mail address: Luke.Marinovich@sydney.edu.au (M.L. Marinovich).

Research in context

Evidence before this study

We searched Medline between January 1, 2010, and October 31, 2018 using the MeSH term 'breast neoplasms', combined with a title search for 'artificial intelligence', 'deep learning', 'machine' or 'neural', to inform our research plan. We found 23 relevant papers reporting a wide range of areas under the ROC curve for AI interpretation of mammograms (0.69–98, median 0.88). Studies were predominantly small, retrospective, and used cancer-enriched datasets. There were few studies comparing the accuracy of AI with radiologists and limited validation of AI using external datasets. A methodological review of independent validation studies to December 10, 2020 found a high risk of bias through selection of non-consecutive patient cohorts, and a lack of linkage to cancer registry data to identify interval cancers.

Added value of this study

This is an external validation study of a commercially available AI algorithm in a retrospective, consecutive cohort of screening mammograms from a population breast screening program, with linkage to cancer registry data for ascertainment of interval cancers. Standalone accuracy of the AI was lower than that of radiologists who interpreted the mammograms in practice. At the summary sensitivity for radiologists, the AI had lower specificity. However, in simulated AI-radiologist reading where discordance between AI and radiologist results was arbitrated by a second radiologists' read, lower specificity did not translate to an increase in recall compared with double-reading in practice. Although there was a small reduction in cancer detection rate for simulated AI-radiologist reading, the AI detected some interval cases that were not identified by radiologists. The interval cancer detection rate in simulated AI-radiologist reading may have been underestimated by our retrospective design where radiologists were blinded to AI findings.

Implications of all the available evidence

We have shown that the high accuracy of AI observed in selected, cancer-enriched datasets may have limited applicability to real-world breast cancer population screening. However, lower AI specificity does not necessarily translate to increased recall when integrating AI into a reading work-flow including arbitration. Prospective trials are required to determine whether slightly lower cancer detection rates from simulated AI-radiologist reading could potentially improve if AI detection was actioned in a reading strategy including arbitration to maximise cancer detection while minimising unnecessary recall.

cancer but also yield false-positive results or miss cancers that are not perceptible to the screen-reader. False-positive recall is a downside of screening, potentially causing anxiety and unnecessary investigations.[2] Cancers that are not detected at screening often present symptomatically in the interval between screening rounds ("interval cancers"), and may be more fast-growing and aggressive than screen-detected cancer.[3,4] Automated reading of mammograms by artificial intelligence (AI) algorithms has been proposed to reduce false-positive recall, increase cancer detection through earlier identification of interval cancers,[5–8] and reduce workforce challenges faced by screening programs.[9–11]

Screening programs in Australia, Europe and the United Kingdom (UK) use double-reading, implemented as two independent screen-readings, with discordance typically resolved by arbitration or additional reads. Replacing one of the initial two human readers with an accurate AI algorithm has the potential to improve cancer detection and recall metrics. However, studies that have evaluated AI for breast cancer screening have commonly employed cancer-enriched datasets that are likely to be unrepresentative of disease spectrum in screening populations, and may lead to estimates of accuracy that are not generalisable to real-world screening.[12–14] External validation studies, in which AI is evaluated in datasets that are independent of those used to train the algorithm, are uncommon and have suffered from methodological shortcomings in cohort selection and outcome ascertainment.[15,16] Currently, the European Commission Initiatives on Breast and Colorectal Cancer recommends against the use of AI as second reader due to very low certainty of the evidence on test accuracy.[17] There is therefore a need to generate robust evidence of AI performance that is generalisable to routine screening practice to inform decisions about adopting the technology.[12,18]

In this cohort study, we compare AI reading of digital mammograms with human reading in a real-world, population breast cancer screening setting using consecutive screening mammograms. We aim firstly to compare the accuracy of AI with single human reading, and secondly to compare cancer detection and recall rates for simulated AI-human screen-reading with human double-reading (standard practice in most breast screening programs).

Methods

We conducted a retrospective independent validation of a commercially-available AI algorithm in a consecutive cohort of screening participants from the population breast screening program in Western Australia (WA), BreastScreen WA (BSWA), for whom screening data and outcomes were prospectively collected. Detailed methods are described in the previously-published protocol.[19]

Ethics
The study had ethical approval from the Women and Newborn Health Service Ethics Committee (EC00350). The committee provided a waiver of consent for this study. Participants in the BSWA programme provide written consent for their data to be used for research purposes each time they screen.

Study cohort
Consecutive mammography screens between 1 November 2015 and 31 December 2016 were identified from BSWA. Full-field digital mammograms were acquired with Siemens systems (MAMMOMAT Inspiration). Women were eligible if aged 50–74 years, aligning with the invited age range for breast cancer screening in Australia.[20] For women with multiple screening episodes, only the most recent was included. Deaths within 24 months of screening and out-of-state relocations were excluded to ensure a minimum follow-up period of 24 months for ascertaining interval cancers. Women with a previous mastectomy, implants, or an incomplete screening examination were excluded to ensure completeness of images for interpretation by the AI algorithm.

Measurement
For each woman, demographic characteristics and risk factors (age; screening round; personal history of breast and ovarian cancers; first degree family history of breast cancer; hormone replacement therapy in previous six months) were extracted from the Mammographic Screening Registry. Breast density was not available for the full the cohort because BSWA records density only for women not recalled to assessment. The final screening outcome (recall or not recall) and findings from each radiologist were also extracted, along with data on screen-detected cancers (date of diagnosis; histological type; tumour size). Screen-detected cancers were defined as invasive cancer or ductal carcinoma in situ (DCIS) detected at the index screening episode.[21] Interval cancers, defined as invasive cancers diagnosed after a negative index screen and before the next scheduled screening episode (i.e. within 24 months for biennial screeners, and 12 months for a minority of women scheduled for annual screening)[21] were identified through data linkage to the WA Cancer Registry.

Interpretation of mammograms by AI algorithm
The DeepHealth AI model used in this study underlies a triage product (Saige-Q v2.0.0) that is Food and Drug Administration (FDA)-cleared and commercially available in the United States (US). Development of the AI model has been described previously.[5] Training data sets from the US and UK (with images acquired using General Electric and Hologic systems) were independent of the Australian data set used for this external validation study. The algorithm evaluates each image in a mammographic study independently and aggregates all potential regions of interest to compute a single study-level score ranging from 0 to 1.

In processing a mammographic study, the commercial algorithm consists of the following steps: 1) checking each image and the entire study for acceptability for processing; 2) pre-processing the pixel data for input into the AI model; and 3) evaluation of the AI model on the pre-processed pixels. To enable processing of Siemens images for which the algorithm had not been previously validated or FDA-cleared, modifications were performed for steps 1 and 2 (majority of acceptability checks were removed, as well as a pre-processing step that crops out the image background), whereas the AI model (step 3) remained fixed.

Integration of AI and radiologist findings
The BSWA program uses double-reading, implemented as independent screen-readings by two radiologists with arbitration for discordance.[20] The integration of AI into double-reading was simulated by analytically pairing the first radiologists' read (Reader 1) per screen with the AI result. Recall to assessment was based on results of AI-radiologist reading, where agreement between Reader 1 and AI resulted in a decision to recall or not recall. In the case of disagreement, a new arbitrating read was not performed; rather, disagreement was arbitrated by the second radiologist read (Reader 2) that occurred in practice ("simulated AI-radiologist reading"). A sensitivity analysis resolved disagreement by Reader 3 when arbitration occurred in practice (and by Reader 2 when it did not). Strategies that did not include arbitration (recall when either Reader 1 or AI was positive for suspicious abnormality; recall when both Reader 1 and AI were positive) were also investigated.

Statistics
Detailed statistical methods are described in Supplementary Method S1. Sample size calculations are available in the published protocol.[19] Descriptive characteristics of the cohort were summarised by the mean and standard deviation (SD) for age and percentages for categorical variables. All tests of statistical significance were two-sided. The level chosen for statistical significance was $P < 0.05$; $P < 0.10$ was considered to indicate weak evidence of a difference. Analyses were undertaken in R 4.0.4 and SAS 9.4.

Accuracy measures
For each radiologist, sensitivity and specificity for detection of cancer were computed and the summary receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), and summary sensitivity and specificity were estimated.[22] For the AI algorithm, the empirical ROC curve and AUC were derived. Partial AUCs were also estimated.[23] Sensitivity and specificity for AI were computed at a prospectively-defined

threshold (Threshold 1) predicted to generate a 4% positivity rate (*a priori* expected recall rate for double-reading in BSWA[20]) based on DeepHealth's independent US validation data.

Screening detection measures

The cancer detection rate (CDR, per 1000 screens) and recall rate (percentage) were computed for double-reading by radiologists and compared with simulated AI-radiologist reading using Threshold 1 (McNemar's test). Retrospective thresholds generating a positivity rate for the *AI algorithm alone* that equalled the BSWA program recall rate observed in our study data (Threshold 2), and a recall rate for *simulated AI-radiologist reading (with arbitration)* that equalled the observed BSWA program recall rate (Threshold 3) were also explored. The proportion of AI-positive screen-detected cancers was stratified by age, screening round, pathologic type and tumour size. The proportion of AI-positive interval cancers was stratified by age, screening round, and time-to-diagnosis. Strata were compared with Chi-squared or Fisher's exact tests.

Role of funders

The study sponsors had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

Results

Cohort characteristics

A total of 113,818 women contributing unique, consecutive screening examinations were identified. After applying exclusion criteria, 109,000 women (95.8%) were eligible. A further 30 (<0.1%) were excluded due to mammograms that could not be retrieved, resulting in 108,970 women included in the analytic cohort (Supplementary Figure S1).

Descriptive characteristics and screening metrics are presented in Table 1. The mean age of participants was 61.0 (SD 6.9) years. There were 9071 baseline (incident) screens (8.3%); the remainder were subsequent (repeat) screens. A majority of women (87.2%) had a biennial screening interval (n = 95,017). There were 760 screen-detected breast cancers (605 invasive, 155 DCIS) and 235 interval cancers.

Accuracy of AI algorithm versus radiologists

During the study period, 27 radiologists interpreted mammograms as first or second readers. Each radiologist interpreted a subset of the mammograms comprising the study cohort (mean 8072, range 3066–15,938). Pairs of (sensitivity, 1-specificity) for each radiologist are plotted in Fig. 1A, along with the summary ROC curve and empirical ROC curve for the AI algorithm. Sensitivities and 1-specificities for radiologists lie above the algorithm ROC curve. The AUC for radiologists was 0.93 compared with 0.83 for the AI; pAUCs over the 1-specificity range for radiologists were 0.86 and 0.71, respectively. The AI algorithm AUC was 0.87 for screen-detected cancers and 0.67 for interval cancers (Supplementary Figure S2).

Fig. 1B plots the same ROC curves overlaid with summary sensitivity and specificity for radiologists and the AI. For radiologists, summary sensitivity was 0.68 (95% CI 0.66–0.71) and specificity was 0.97 (95% CI 0.97–0.97). At Threshold 1, the AI algorithm sensitivity (0.67; 95% CI 0.64–0.70) was comparable to summary sensitivity for radiologists but with lower specificity (0.81; 95% CI 0.81–0.81). The AI positivity rate at Threshold 1 was 19.5%. At AI sensitivity of 0.68 (95% CI 0.65–0.71) (i.e. equal to radiologist sensitivity), specificity was 0.80 (95% CI 0.80–0.80; estimates not shown in Fig. 1B). At AI specificity of 0.97 (95% CI 0.97–0.97) (i.e. equal to radiologist specificity), sensitivity was 0.40 (95% CI 0.37–0.43).

In a sensitivity analysis applying 12-month follow-up for ascertaining interval cancers (Supplementary Figure S3), AUCs were higher for both radiologists (0.96) and the algorithm (0.85), but the absolute difference (–0.11) was consistent with the primary analysis.

Cancer detection rates

Table 2 reports CDRs for double-reading in the BSWA program, and for simulated AI-radiologist reading with arbitration (AI-radiologist reading without arbitration shown in Supplementary Table S1). CDR for the BSWA program was 6.97 per 1000 (95% CI 6.49–7.49). CDRs at each AI threshold were statistically significantly lower than the BSWA CDR. Absolute differences between CDRs were –0.61 per 1000 (95% CI –0.77 to –0.44; $P < 0.001$ McNemar's) at Threshold 1; –0.97 per 1000 (95% CI –1.16 to –0.78; $P < 0.001$ McNemar's) at Threshold 2; and –0.51 per 1000 (95% CI –0.67 to –0.36; $P < 0.001$ McNemar's) at Threshold 3. In a sensitivity analysis incorporating Reader 3 in arbitration (when available), CDRs were generally comparable to the main analysis (Supplementary Table S2). Slightly higher CDRs are attributable to a small increase in the number of screen-detected cancers.

Simulated AI-radiologist reading detected 90.4% (95% CI 88.1–92.4) of screen-detected cancers at Threshold 1; 85.8% (95% CI 83.1–88.2) at Threshold 2; and 91.6% (95% CI 89.4–93.5) at Threshold 3 (Table 2). The proportions of screen-detected cancers that were AI true-positive (i.e. algorithm sensitivity for screen-detected cancers) were 76.8% (95% CI 73.7–79.8) at Threshold 1; 49.6% (95% CI 46.0–53.2) at Threshold 2; and 83.0% (95% CI 80.2–85.6) at Threshold 3. AI sensitivity was consistently higher for younger women, with a statistically significant trend for decreasing sensitivity with increasing age (Table 3). AI sensitivity was higher for incident than repeat screens. At

	N	% from all screened women
Cohort characteristics		
Age-group, years		
50–59	48,389	44.4%
60–69	45,616	41.9%
70–74	14,965	13.7%
Personal history of breast cancer		
Yes	3351	3.1%
No	105,618	96.9%
No response	1	<0.1%
Personal history of ovarian cancer		
Yes	631	0.6%
No	108,339	99.4%
First degree family history of breast cancer		
Yes	10,197	9.4%
No	98,773	90.6%
Hormone replacement therapy (past 6 months)		
Yes	12,497	11.5%
No	96,465	88.5%
No response	7	<0.1%
Screening round		
First	9071	8.3%
Repeat	99,899	91.7%
Recommended screening interval		
Annual	13,951	12.8%
Biennial	95,019	87.2%
Screening metrics		
Recalled to assessment	3684	3.4%
Screen-detected cancer	760	0.7%
Pathologic type		
Invasive	605	0.6%
DCIS	155	0.1%
Size		
≤15 mm	419	0.4%
>15 mm	340	0.3%
Missing	1	<0.1%
Interval cancer	235	0.2%
≤12 months post-screen	94	<0.1%
>12–24 months post-screen	141	0.1%

Table 1: **Cohort characteristics and screening metrics for 108,970 women screened by BreastScreen WA, November 2015 to December 2016.**

Threshold 2, there was weak evidence for higher sensitivity for invasive cancer than DCIS, but no evidence of this difference at other thresholds. The proportion of invasive screen-detected cancers detected by the AI was greater for large (>15 mm) than small (≤15 mm) cancers at all thresholds. There was weak evidence for this difference in screen-detected DCIS for Threshold 2 only, but these comparisons have reduced statistical power due to lower numbers of DCIS (Table 3).

The proportion of interval cancers detected by the AI (i.e. algorithm sensitivity for interval cancers) was 36.6% (95% CI 30.4–43.1) at Threshold 1; 10.2% (95% CI 6.7–14.8) at Threshold 2; and 45.5% (95% CI 39.0–52.1) at Threshold 3. There were no statistically significant differences by age, screening round and time-to-diagnosis (Table 3). Relatively small numbers of interval cancers were identified by simulated AI-radiologist reading (range 0.02–0.07 per 1000; Table 2); those strategies required positive findings by AI and one radiologist. A larger number of interval cancers were detected by the AI algorithm but not by either radiologist (range 0.20–0.91 per 1000), and hence were not recalled in simulated arbitration. To simulate a scenario in which the AI result was made available to aid radiologist interpretation, the maximum possible CDR (including interval cancers detected only by AI) was estimated (Table 2). At Threshold 1, the maximum CDR was not different to the BSWA CDR (difference 0.12 per 1000; 95% CI –0.11 to 0.35; $P = 0.30$ McNemar's). The maximum CDR was statistically significantly less than the BSWA CDR at Threshold 2 (–0.77 per 1000; 95% CI –0.98 to –0.56; $P < 0.001$ McNemar's), and statistically significantly greater at Threshold 3 (0.39 per 1000; 95% CI 0.16–0.63; $P = 0.001$ McNemar's).

Recall rates

Table 2 presents recall rates for simulated AI-radiologist reading (with arbitration) compared with the BSWA recall rate (AI-radiologist reading without arbitration is described in Supplementary Table S1). The BSWA program recall rate was 3.38% (95% CI 3.27–3.49). The simulated AI-radiologist reading recall rate was significantly lower than the BSWA recall rate for both Threshold 1 (difference –0.25%; 95% CI –0.31 to –0.18; $P < 0.001$ McNemar's) and Threshold 2 (–0.81%; 95% CI –0.87 to –0.75; $P < 0.001$ McNemar's). By definition, recall rates were equal for Threshold 3. In a sensitivity analysis incorporating Reader 3 in arbitration (when available), recall rates for Thresholds 1 and 2 increased slightly but were generally comparable to the main analysis (Supplementary Table S2).

Recall rates associated with potential maximum CDRs were estimated as a range of values. "Low" estimates assumed that only AI true positives for interval cancers were recalled; "high" estimates assumed that all AI positives were recalled. "Low" estimates were significantly lower than the BSWA recall rate except at Threshold 3, which resulted in a small increase in recall (difference 0.09%; 95% CI 0.03–0.16; $P = 0.006$ McNemar's). Large increases in recall rates were evident in all "high" estimate scenarios.

Screen-reading volume

Table 4 reports the number of first, second and arbitrating radiologist reads in the BSWA program compared with simulated AI-radiologist reading. For

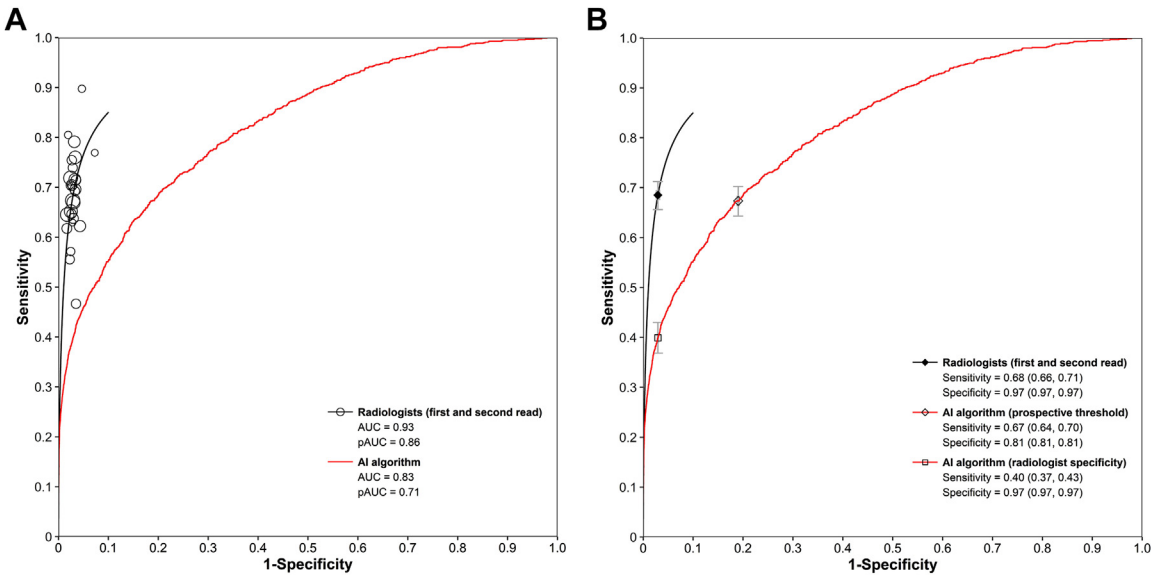

Fig. 1: Summary ROC curve for radiologists (N = 27) and empirical ROC curve for AI algorithm showing A) AUCs and partial AUCs, and B) sensitivity and specificity for radiologists (summary) and AI algorithm (at prospective threshold and at radiologist specificity).

simulated AI-radiologist reading, the number of arbitrating reads increased at all AI thresholds, but replacement of the second radiologist read by the AI algorithm resulted in a reduction in the total number of reads of between 37.7% and 48.4%. For strategies combining radiologists and AI with no arbitration, the estimated reduction in the number of radiologist reads was 51.2% (Supplementary Table S3).

Discussion

Population mammography screening is the main cancer control strategy for female breast cancer. The development and validation of AI algorithms for automated reading of mammograms has been the subject of intensive research aiming to improve screening metrics and address resourcing of screen-readings in population screening programs.[5–8] Despite this effort, there remains substantial uncertainty about the accuracy of AI in real-world screening and the transferability of evidence between settings due to limited external validation, the use of selected (non-consecutive) subjects and incomplete ascertainment of interval cancers leading to a high risk of bias.[12,15,16] In our study, we assembled a well-defined cohort of consecutive mammograms from a population-based screening program with ascertained outcomes (ensuring all interval cancers were included). Unlike the selected, cancer-enriched datasets commonly reported in the literature, our cohort was highly representative of screening participants in Australia. Our methods also represent a stringent test of transferability by applying an AI algorithm that was trained and validated using independent, non-Australian data, and with images acquired by different hardware vendors to that used by the screening program.[24] The accuracy of the algorithm observed in our study (AUC 0.83 for all cancers; AUC 0.87 for screen-detected cancers) was lower than previously found when tested using a cancer-enriched, US dataset (AUC 0.94 for screen-detected cancers).[5]

Several factors may have contributed to this performance difference, including the consecutive cohort versus cancer-enriched selection, and differences in screening populations, mammography vendors, and criteria for defining ground truth (confirmation of non-cancer status; inclusion of interval cancers; consideration of "misses" that are identified on subsequent screens). Notably, an earlier study reported lower performance for a UK-trained and validated algorithm when tested in a (cancer-enriched) US setting.[6] Although the difference in performance between countries is consistent with our results (albeit with a smaller magnitude of difference in our study), that earlier comparison may itself be confounded by a difference in cohort selection. Recognising and further exploring these factors and how they contribute to transferability of AI are important lines of research to inform translation into screening practice.

Our study reports various comparisons of AI relative to human readers to judge its potential for further evaluation or application in breast cancer screening. In contrast to a previous report in which AI accuracy was non-inferior to radiologists in a selected, cancer-

enriched dataset assembled from multiple screening settings in different countries,[7] our study using a well-defined, consecutive cohort found that the algorithm exhibited lower overall accuracy (AUC 0.83) than radiologists who interpreted the screens in practice (AUC 0.93). However, at a prospectively-defined positivity threshold, the algorithm's sensitivity (0.67) was comparable to the summary sensitivity for radiologists (0.68), but with lower specificity (0.81 versus 0.97). This threshold was selected *a priori* to produce a relative low algorithm positivity rate of 4% (based on independent, non-Australian validation data), and hence an expected specificity of approximately 0.96. This difference between expected and observed AI positivity rates and resulting specificity highlights the inherent difficulties of threshold selection using independent datasets. Applying an optimal threshold is likely to remain challenging where calibration of algorithms to local conditions is not feasible, and emphasises the importance of better understanding factors potentially affecting transferability.

In simulated AI-radiologist reading, where the AI and human reads were analytically integrated with arbitration for discordance, the AI's lower specificity increased the arbitration rate but did not translate to an increase in the recall rate. This suggests that arbitration by a radiologist can mitigate additional AI false-positives. The increase in arbitrating reads partially offset a reduction in reading workload from using AI as one reader, but this strategy still resulted in an overall reduction in reading volume of between 37.7% and 48.4%. A similar increase in arbitration coupled with a reduction in overall screen-reading volume was observed in previous studies simulating integrated AI-radiologist reading,[25] preprint,[26] highlighting the potential efficiencies from using AI in screen-reading in the context of arbitrated double-reading.

At the prospective threshold, the AI algorithm was positive for 76.8% of cancers that were screen-detected in BSWA practice and 36.6% of interval cancers; at a lower (retrospective) threshold, sensitivity (83.0% for screen detected, 45.5% for interval cancers) was comparable to that found in external validation of a different AI system in a Norwegian screening program (86.8% and 44.9%; small differences in estimates may reflect age differences between cohorts).[27] Consistent with expectations, the AI sensitivity for screen-detected cancers was higher for incident screens[20] and larger cancers.[28] Our data did not allow stratification of results by breast density classification; future work should address this limitation by providing AI accuracy by density. However, we presented comparisons by age as a potential proxy for density,[29] and found evidence of a decreasing proportion of detection of screen-detected cancers with increasing age (i.e. the AI was more sensitive in younger age groups). Given that mammography sensitivity increases with age,[20] a relationship also

Screen reading strategy	Threshold for AI algorithm	Recall rate		CDR						Potential maximum CDR and recall rate					
				All cancers		Screen-detected		Interval		Interval cancers detected by AI but not readers 1 and 2			Potential recall % range		
		N recalls	Recall % (95% CI)	N cancers	CDR per 1000 (95% CI)	N cancers	CDR per 1000 (95% CI)	N cancers[a]	CDR per 1000 (95% CI)	N cancers	Additional CDR per 1000 (95% CI)	Potential maximum CDR per 1000 (95% CI)	Low	High
BSWA screening practice: Radiologist double-reading, disagreement arbitrated by additional read/s	N/A	3684	3.38 (3.27–3.49)	760	6.97 (6.49–7.49)	760	6.97 (6.49–7.49)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Simulated AI-radiologist reading: Reader 1 paired with AI algorithm, disagreement arbitrated by Reader 2	Threshold 1 (prospective)	3417	3.14** (3.03–3.24)	694	6.37** (5.91–6.86)	687	6.30 (5.84–6.79)	+7	+0.06 (0.03–0.13)	+79	0.72 (0.57–0.90)	7.09 (6.60–7.61)	3.21**	20.79**
	Threshold 2	2799	2.57** (2.48–2.66)	654	6.00** (5.55–6.48)	652	5.98 (5.53–6.46)	+2	+0.02 (0.00–0.07)	+22	0.20 (0.13–0.31)	6.20** (5.75–6.69)	2.59**	5.29**
	Threshold 3	3684	3.38† (3.27–3.49)	704	6.46** (5.99–6.95)	696	6.39 (5.92–6.88)	+8	+0.07 (0.03–0.14)	+99	0.91 (0.74–1.11)	7.37** (6.87–7.89)	3.47*	28.73**

P-value for comparison with BSWA practice (McNemar's test): *P < 0.01; **P < 0.001; † not applicable. Abbreviations: AI, artificial intelligence; BSWA, BreastScreen WA; CDR, cancer detection rate; CI, confidence interval; N/A, not applicable.
[a]Number of interval cancers not detected by Reader 1 but detected by the AI and Reader 2 (arbitrating read in our study). In BSWA practice, the disagreement between Readers 1 and 2 was arbitrated in favour of Reader 1 and these cases were not recalled.

Table 2: **Recall and CDR for BSWA program and simulated AI-radiologist reading (N = 108,970 screens).**

Stratification variable	Category	N cancers	Threshold 1 (prospective)		Threshold 2		Threshold 3	
			AI true positive proportion (95% CI)	*P*-value[a]	AI true positive proportion (95% CI)	*P*-value[a]	AI true positive proportion (95% CI)	*P*-value[a]
Screen-detected cancers	N/A	760	76.8% (73.7–79.8)	N/A	49.6% (46.0–53.2)	N/A	83.0% (80.2–85.6)	N/A
Age	50–59	267	81.7% (76.5–86.1)	0.07 (0.03 for trend)	56.9% (50.7–62.9)	0.01 (0.002 for trend)	86.1% (81.4–90.1)	0.06 (0.02 for trend)
	60–69	335	74.6% (69.6–79.2)		46.3% (40.8–51.8)		83.3% (78.9–87.1)	
	70–74	158	73.4% (65.8–80.1)		44.3% (36.4–52.4)		77.2% (69.9–83.5)	
Screening round	First	113	85.8% (78.0–91.7)	0.01	62.8% (53.2–71.7)	0.002	88.5% (81.1–93.7)	0.09
	Repeat	647	75.3% (71.8–78.5)		47.3% (43.4–51.2)		82.1% (78.9–84.9)	
Pathologic type	DCIS	155	76.8% (69.3–83.2)	0.98	42.6% (34.7–50.8)	0.05	82.6% (75.7–88.2)	0.87
	Invasive	605	76.9% (73.3–80.2)		51.4% (47.3–55.5)		83.1% (79.9–86.0)	
Tumor size								
Invasive[c]	≤15 mm	350	72.9% (67.9–77.5)	0.005	43.7% (38.5–49.1)	<0.001	80.9% (76.3–84.9)	0.06
	>15 mm	254	82.7% (77.5–87.1)		62.2% (55.9–68.2)		86.6% (81.8–90.5)	
DCIS	≤15 mm	69	71.0% (58.8–81.3)	0.13	34.8% (23.7–47.2)	0.08	79.7% (68.3–88.4)	0.40
	>15 mm	86	81.4% (71.5–89.0)		48.8% (37.9–59.9)		84.9% (75.5–91.7)	
Interval cancers[d]	N/A	235	36.6% (30.4–43.1)	N/A	10.2% (6.7–14.8)	N/A	45.5% (39.0–52.1)	N/A
Age	50–59	99	40.4% (30.7–50.7)	0.45	8.1% (3.6–15.3)	0.63[b]	49.5% (39.3–59.7)	0.51
	60–69	97	32.0% (22.9–42.2)		12.4% (6.6–20.6)		41.2% (31.3–51.7)	
	70–74	39	38.5% (23.4–55.4)		10.3% (2.9–24.4)		46.1% (30.1–62.8)	
Screening round	First	27	37.0% (19.4–57.6)	0.96	3.7% (0.1–19.0)	0.33[b]	44.4% (25.5–64.7)	0.90
	Repeat	208	36.5% (30.0–43.5)		11.1% (7.1–16.1)		45.7% (38.8–52.7)	
Time to diagnosis	≤12 months	94	39.0% (30.9–47.6)	0.35	9.9% (5.5–16.1)	0.86	45.4% (37.2–53.6)	0.96
	>12–24 months	141	33.0% (23.6–43.4)		10.6% (5.2–18.7)		45.7% (35.4–56.3)	

Abbreviations: AI, artificial intelligence; CI, confidence interval; DCIS, ductal carcinoma in situ; N/A, not applicable. [a]Chi-squared test; Wald Chi-squared test for trend in parentheses. [b]Fisher's exact test. [c]N = 1 invasive screen-detected cancer with missing size data. [d]Invasive only, consistent with the BreastScreen Australia definition.[21] Tumour size was not recorded for interval cancers.

Table 3: **Proportion of screen-detected and interval cancers that were positive on AI, stratified by age, screening round, pathologic type and tumour size.**

observed in previous reports of AI sensitivity,[30] this new finding was unexpected and warrants further exploration, noting however that this was not observed for the interval cancers.

Cancer screening programs may be reluctant to adopt new technologies if cancer detection is not at least equivalent to current practice, regardless of potential benefits to recall and workflow. A study of simulated AI-radiologist reading in the Norwegian screening program found a small reduction in CDR relative to program metrics; however, the estimates do not include detection of interval cancers as arbitration was not applied to cases that were not recalled in practice.[26] Similarly, a previous study with incomplete interval cancer ascertainment found that the CDR for integrated AI-radiologist reading did not exceed that of double-reading by radiologists.[25] [preprint] In our study including all interval cancers, the CDR for simulated AI-radiologist reading with arbitration was lower compared with the BSWA program, although the difference was relatively small (–0.61 per 1000). However, it is possible that the CDRs observed for simulated AI-radiologist reading are an underestimation of what might be observed in a prospective study, where all AI-radiologist recalled cases would undergo assessment and additional (non-interval) cancers may be detected.[6] Also, the use of Reader 2 reads for arbitration in our study may have underestimated CDR due to likely performance differences of real-world arbitrating readers and initial readers. Our sensitivity analyses suggested that both CDR and recall may increase with real-world arbitration, although the observed improvement in CDR was attributable to recall of screen-detected (not interval) cancers.

Furthermore, we found that radiologist arbitration corrected AI false-positives but also "arbitrated out" most interval cancers detected by AI. Radiologists were blinded to AI findings, but it is possible that some Reader 2 (arbitrating) reads were not conducted independently from Reader 1. This may bias arbitration in favour of agreement with Reader 1 rather than with AI. It is therefore possible that a larger number of interval cancers may be detected in a prospective setting where arbitrating radiologists are not blinded to AI findings, and our simulation showed the potential for unblinding to result in increases in CDR. An earlier study, using the same algorithm with a cancer-enriched dataset, also

Screen reading strategy	Threshold for AI algorithm	Number of radiologist reads				Reduction relative to BSWA practice (%) (95% CI)
		First read	Second read	Arbitrating reads	Total	
BSWA screening practice: Radiologist double-reading, disagreement arbitrated by additional read/s	N/A	108,970	108,970	5415	223,355	N/A
Simulated double-reading: Reader 1 paired with AI algorithm, disagreement arbitrated by Reader 2	Threshold 1 (prospective)	108,970	0	21,960	130,930	↓ 41.4% (41.2–41.6)
	Threshold 2	108,970	0	6303	115,273	↓ 48.4% (48.2–48.6)
	Threshold 3	108,970	0	30,100	139,070	↓ 37.7% (37.5–37.9)

Abbreviations: AI, artificial intelligence; BSWA, BreastScreen WA; CI, confidence interval; N/A, not applicable.

Table 4: **Number of radiologist reads for BSWA program and simulated AI-radiologist reading (N = 108,970 screens).**

showed that the AI could detect cancers that had not been detected by radiologists in real-world screening,[5] and the potential for interval cancer detection has been highlighted in a previous retrospective study without arbitration of additional AI findings.[26] However, our analysis was case-based and correlation of AI findings with subsequently diagnosed interval cancer (lesion) location was not undertaken. Our simulation showed probable increases in recall associated with any increase in CDR from additional interval cancer detection. Prospective trials would provide the strongest evidence about the true magnitude of changes in CDR and recall; however, extensions of retrospective studies in which all AI-radiologist discordant cases are subject to new third reader arbitration are possible, and could elucidate this impact in shorter timeframes.

In this external validation cohort study, using consecutive screening mammograms, we found that the accuracy of AI observed in selected, cancer-enriched datasets may not be directly applicable to real-world breast cancer population screening. Differences in screening participants, mammography vendors and ground truth criteria may have also impacted the transferability of the AI's performance. However, the lower AI specificity compared with radiologists observed in our study does not necessarily translate to increased recall when integrating AI into double-reading, where additional arbitration may mitigate false positives while still resulting in reductions in overall screen-reading volume. The evidence from our study suggests a potential role for AI in automated reading of breast cancer screening mammograms; however, prospective trials are required to determine whether the slightly lower CDR from simulated AI-radiologist reading could improve if AI detection was actioned in a reading workflow including arbitration to minimise unnecessary recall while maximising cancer detection to improve screening outcomes for women.

Contributors

MLM and NH conceptualised the study. MLM, EW, WL, HL, AW, and NH planned and designed the study. GP contributed to the study methods and analysis plan. MLM, HL, and AW accessed and verified the underlying data. WL oversaw validity checks of the algorithm output but did not access the underlying data. MLM did the formal data and statistical analyses and wrote the first draft. EW, WL, HL, AW, CM, GP, and NH critically revised the draft for important intellectual content. All authors have approved the final written manuscript.

Data sharing statement

Datasets generated and/or analysed during the current study are not publicly available due to data confidentiality agreements with data custodians.

Declaration of interests

WL is a consultant for DeepHealth, RadNet AI Solutions, was a full-time employee during conduct of this study, and owns stock in that entity. Other authors have no competing interest to declare.

Acknowledgments

This work was supported by funding from a National Breast Cancer Foundation (NBCF) Investigator Initiated Research Scheme grant (IIRS-20-011 to MLM, NH, EW and WL). GP was supported with funding from National Health and Medical Research Council (NHMRC) Project and Investigator Grants #1099655 and #1173991. NH was supported by the National Breast Cancer Foundation (NBCF) Chair in Breast Cancer Prevention program (EC-21-001) and by a NHMRC Investigator (Leader) grant (1194410).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104498.

References

1 Hanley JA, Hannigan A, O'Brien KM. Mortality reductions due to mammography screening: contemporary population-based data. *PLoS One*. 2017;12(12):e0188947.
2 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.
3 Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer*. 2017;3(1):12.
4 Baré M, Torà N, Salas D, et al. Mammographic and clinical characteristics of different phenotypes of screen-detected and interval breast cancers in a nationwide screening program. *Breast Cancer Res Treat*. 2015;154(2):403–415.
5 Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med*. 2021;27(2):244–249.
6 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89–94.
7 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst*. 2019;111(9):djy222.
8 Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent

assessment of screening mammograms. *JAMA Oncol*. 2020;6(10):1581–1588.
9 Crouch B. *Shortage of radiologists pushing out breast scan result times for patients*. The Advertiser; 2018.
10 Harvey H, Karpati E, Khara G, et al. The role of deep learning in breast screening. *Curr Breast Cancer Rep*. 2019;11(1):17–22.
11 The Royal Australian and New Zealand College of Radiologists. *2016 RANZCR clinical radiology workforce census report: Australia*. Sydney, NSW: The Royal Australian and New Zealand College of Radiologists; 2018.
12 Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices*. 2019;16(5):351–362.
13 Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. *Can Med Assoc J*. 2013;185(11):E537–E544.
14 Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. *Radiology*. 2019;290(1):272–273.
15 Anderson AW, Marinovich ML, Houssami N, et al. Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: a systematic review. *J Am Coll Radiol*. 2022;19(2):259–273.
16 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.
17 European Commission Initiatives on Breast and Colorectal Cancer. Use of artificial intelligence 2022 [updated 08/11/2022]. Available from: https://healthcare-quality.jrc.ec.europa.eu/ecibc/european-breast-cancer-guidelines/artificial-intelligence.
18 Elmore JG, Lee CI. Artificial intelligence for breast cancer imaging: the new frontier? *J Natl Cancer Inst*. 2019;111(9):djy223.
19 Marinovich ML, Wylie E, Lotter W, et al. Artificial intelligence (AI) to enhance breast cancer screening: protocol for population-based cohort study of cancer detection. *BMJ Open*. 2022;12(1):e054005.
20 Australian Institute of Health and Welfare. *BreastScreen Australia monitoring report 2021. Cat. no. CAN 140*. Canberra: AIHW; 2021. Contract No.: Cat. no. CAN 140.
21 Australian Institute of Health and Welfare. *BreastScreen Australia data dictionary version 1.2*. 2019.
22 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20(19):2865–2884.
23 McClish DK. Analyzing a Portion of the ROC curve. *Med Decis Making*. 1989;9(3):190–195.
24 Grimm LJ, Plichta JK, Hwang ES. More than incremental: harnessing machine learning to predict breast cancer risk. *J Clin Oncol*. 2023;0(0). JCO.21.02733.
25 Sharma N, Ng AY, James JJ, et al. Retrospective large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. *medRxiv*. 2022. https://doi.org/10.1101/2021.02.26.21252537.
26 Larsen M, Aglen CF, Hoff SR, Lund-Hanssen H, Hofvind S. Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122,969 screening examinations. *Eur Radiol*. 2022;32(12):8238–8246.
27 Larsen M, Aglen CF, Lee CI, et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. *Radiology*. 2022;303(3):502–511.
28 Wang J, Gottschal P, Ding L, et al. Mammographic sensitivity as a function of tumor size: a novel estimation based on population-based screening data. *Breast*. 2021;55:69–74.
29 Checka CM, Chun JE, Schnabel FR, Lee J, Toth H. The relationship of mammographic density and age: implications for breast cancer screening. *Am J Roentgenol*. 2012;198(3):W292–W295.
30 Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open*. 2020;3(3):e200265.