Integrated AI reader development and evaluation provides clinically-relevant guidance for human-AI collaboration in population mammographic screening

Helen ML Frazer [1 2 †] ✉, Carlos A Peña-Solorzano [3 4 †], Chun Fung Kwok [3 4 †], Michael S Elliott [3 4 †], Yuanhong Chen[5], Chong Wang[5], the BRAIx team[9], Jocelyn Lippey[1 6], John Hopper[7], Peter Brotchie[8], Gustavo Carneiro[5], Davis J McCarthy [3 4]

[1]St Vincent's BreastScreen, St Vincent's Hospital Melbourne, Victoria, Australia; [2]BreastScreen Victoria, Victoria, Australia; [3]Bioinformatics and Cellular Genomics Unit, St Vincent's Institute of Medical Research, Victoria, Australia; [4]Melbourne Integrative Genomics, School of Mathematics and Statistics/School of BioSciences, Faculty of Science, University of Melbourne, Victoria, Australia; [5]School of Computer Science, Australian Institute for Machine Learning, University of Adelaide, South Australia, Australia; [6]Department of Surgery, St Vincents Hospital Melbourne, Victoria, Australia; [7]Centre for Epidemiology & Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Victoria, Australia; [8]Department of Radiology, St Vincent's Hospital Melbourne, Victoria, Australia; [9]See Contributors section for full list of BRAIx team members.

✉ For correspondence: helen.frazer@svha.org.au (HMLF)

[†]These authors contributed equally to the work.

Present address: St Vincent's BreastScreen Clinic, St Vincent's Hospital Melbourne, 41 Victoria Parade, Fitzroy VIC 3065 Australia

Funding: This work is supported by funding from the Australian Government under the Medical Research Future Fund Grant (MRFAI000090) for the Transforming Breast Cancer Screening with Artificial Intelligence (BRAIx) Project awarded to HMLF, DJM, PB, Jl, JH, GC and a National Health and Medical Research Council Investigator Grant (GNT1195595) awarded to DJM. This work is also supported by a Ramaciotti Health Investment Grant awarded to DJM and funding from a Royal Australian and New Zealand College of Radiologists Clinical Research Grant and the St Vincent's Hospital Melbourne Research Endowment Fund awarded to HMLF. The funders had no role in the work or decision to publish.

Competing interests: Peter Brotchie is an employee of annalise.ai.

Abstract

Artificial intelligence (AI) holds promise for improving breast cancer screening, but many challenges remain in implementing AI tools in clinical screening services. We developed an ensemble AI reader using a high-quality retrospective mammography dataset from Victoria, Australia. The dataset comprises 629,853 unique clients with 1,039,728 screening episodes and 4,374,956 images from 2013–19, randomly split by client into training, development and testing sets. We evaluated our AI reader's performance on the testing set of our retrospective cohort comprising 92,839 clients with 149,105 screening episodes, a prospectively collected cohort of 25,848 clients and screening episodes, and three public external datasets. The AI reader achieved high accuracy (AUC 0.932, 95% CI 0.923, 0.940) on our testing set and outperformed published state-of-the-art models on external datasets. Successful integration of AI readers into population breast cancer screening demands a better understanding of possible models of human-AI collaboration. To address this need, we conducted detailed simulations of three plausible AI-integrated screening pathways: AI reader replacement, AI band-pass filtering, and AI triage. The reader replacement simulation, with AI replacing the second reader, achieved higher sensitivity (82.3%, $P < .0025$) and higher specificity (96.3%, $P < 4.21 \times 10^{-6}$) than the current two reader with arbitration system (96.0%, 79.6%) at a reduced workload (−48.0%). The AI band-pass filter approach also achieved higher sensitivity (81.7%, $P < .0058$) and higher specificity (96.6%, $P < 2.12 \times 10^{-25}$) with drastically reduced workload (−80.7%), while AI triage performed less strongly. This study provides insight into feasible approaches for implementing human-AI collaboration in population mammographic screening, and evidence to support their urgent assessment with prospective studies to validate real-world performance and open pathways to clinical adoption.

Introduction

Breast cancer is the most common cancer globally and the most common cause of cancer-related death in women [1]. The BreastScreen Australia program has been operating for 30 years offering free biennial mammographic screening targeted to women aged 50-74 years (available from age 40). Approximately one million Australian women are screened each year [2]. The program has successfully contributed to a 41-52% reduction in mortality for screening participants and a 21% reduction in population-level breast cancer mortality in Australia [3]. In 2020, approximately 59 per 10,000 participants were diagnosed with breast cancer and 16 per 10,000 participants were diagnosed with Ductal Carcinoma In Situ (DCIS) [2].

Largely thanks to its success there has been little change in the manner of breast cancer screening in the 30 years since BreastScreen Australia was established, but there are increasing challenges in accuracy, service experience, and efficiency in both time and cost. Despite a process of independent double reading of all mammograms by radiologists, and a third arbitration read when there is discordance (henceforth called "two reader with arbitration system"), in 2020 approximately 368 per 10,000 participants were recalled for assessment and later determined not to have breast cancer (false positive). Also, approximately 9 per 10,000 participants aged 50-74 years (2015-2017) subsequently discovered they had an "interval" breast cancer before their next scheduled screen after receiving an "all clear" result (false negative). Only 49.9% of eligible women participated in screening in 2019-2020, down from 54% in 2018-2019, due to the COVID 19 pandemic. Additionally, it takes up to 14 days to receive an "all clear" result, and there is little personalisation for individual risk [2].

Using artificial intelligence (AI), based on deep learning (DL) and neural networks to help read mammograms has the potential to transform breast cancer screening by addressing the three key challenges of accuracy, client experience, and efficiency. The evidence base for AI readers in breast cancer screening has been growing rapidly in recent years, with studies demonstrating the potential of AI to detect breast cancer on mammographic images with a similar accuracy to radiologists [4–12] and addressing key limitations of earlier concern [13]. There remain few studies published in breast cancer screening pathway applications with large population datasets, but important contributions have come from: (1) using a retrospective UK screening population testing dataset of 25,856 episodes, where an AI reader retrospectively gave an accuracy similar to an individual radiologist but was not able to match the accuracy of the double-reading consensus model [6]; (2) using a retrospective Danish screening population testing dataset of 114,421 episodes to filter high confidence non-cancer and cancer classifications with an AI reader, with all other episodes proceeding to a double-reading consensus model, showing potential for reducing radiologist workload and false positives [14]; (3) using a retrospective testing dataset of 213,694 German population episodes to show a decision referral approach applied to the second reader was able to improve sensitivity and specificity relative to individual radiologists [15]; (4) using a retrospective dataset of 122,969 episodes from Norway to assess performance of an AI reader [16]. Several papers have included accurate ground truths for episodes including false negatives (interval cancers) from appropriate follow-up confirmations [6, 10, 14, 17]. Similarly, various aspects of generalisability have been addressed by a handful of studies [6, 9, 10, 18]. Finally, some studies have explored AI-integration with human readers in screening pathways with simulation studies [15, 18–21].

The final frontier for the implementation of AI in breast cancer screening is building evidence of its efficacy through prospective studies. At the time of writing, the only published prospective studies are a recent clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study [22] and a prospective, population-based, paired-reader, non-inferiority study [23], both from Sweden. Showing promising results for AI-supported mammography screening in prospective studies is a major milestone, but crucial assessments of interval cancer rate can only occur after a further two-year follow-up period, leading to calls for caution in the interpretation of these results [24].

Here, we build on insights from the papers discussed above and present a new AI reader and detailed simulation studies that provide clinically relevant guidance for human-AI collaboration in

population mammographic screening, specifically targeted to implementation within population breast cancer screening programs in Australia.

Our dataset is derived from the ADMANI datasets from screening services across Victoria (Australia), totalling 629,853 clients, with 1,039,728 screening episodes and 4,374,956 mammographic images [25]. The dataset spans 2013–19 (with 2016–19 complete and sequential), comprise the screening population of women aged 40 years and above (target recruitment is age 50–74), include screening sites across eight regions, and contain six mammographic machine vendors. The dataset features a high degree of diversity in client ethnicity/region of birth, age, mammography machine manufacturer, screening round, and cancer type and grade. Our dataset has highly accurate ground truth labels, as all cancer episodes (screen-detected and interval) are histopathologically confirmed from biopsies and all non-cancer episodes are confirmed by observation of a two-year interval without cancer following an all-clear, benign, or no significant abnormality (NSA) screening result. All episodes from clients who developed an interval cancer are included.

We randomly split the full cohort by unique screening client into training (506,068 clients), development (30,946 clients), and testing (92,839 clients) sets. Our testing set of 92,839 clients comprises 149,105 screening episodes and 627,647 images. We also evaluate AI reader performance on a prospectively-collected cohort of 25,848 episodes (all unique clients) and 108,654 images from December 2021 to May 2022 from a single breast screening site in Victoria. To further evaluate the performance and robustness of our AI reader we include results on three public external datasets. We focus on clinically meaningful outcomes such as improvements in interval cancer detection and reductions in unnecessary recalls, alongside reduced radiologist workload.

To gain insight into the performance of AI readers in collaboration with human readers and gain clinically-relevant guidance for the implementation of AI readers in screening programs, we conducted detailed simulation studies of AI integrated screening scenarios using our retrospective study cohort. We use the richness and diversity of our cohort combined with careful simulation of reader behaviour (especially that of the third reader in arbitration situations) to substantially improve on previous efforts [15, 18–21] and provide actionable information relevant to current directions for AI implementation [22]. We report on three promising pathway scenarios: (1) the AI reader-replacement scenario where one of the first two readers is replaced by an AI reader in the current two reader with arbitration system; (2) the AI band-pass scenario where the AI reader is used to filter episodes with high-confidence cancer detection scores straight to recall and high-confidence normal scores to no-recall outcomes without human involvement, while the remaining episodes with intermediate scores from the AI reader are passed to the two reader with arbitration system; and (3) the AI triage scenario in which an AI reader sends high detection score episodes (e.g., top decile) for assessment in a two reader with arbitration system and all other episodes to a single reader. The AI triage scenario is currently being assessed in a large-scale randomised controlled study in Sweden that has recently reported on a clinical safety milestone [22].

Our goal of developing an AI reader for implementation in the Australian context is aided by our ability to train the AI reader on the target population. Our BRAIx AI reader performs with high accuracy on the retrospective test data (AUC 0.932, 95% CI 0.923, 0.940). It outperforms 91% of individual radiologists and achieves significant superiority over human readers in both sensitivity and specificity. Our AI reader shows stable performance and high consistency across the Victorian population and various sub-groups, including machine manufacturers, age groups, number of screening rounds, First Nations Australians or region of birth, and cancer type and grade. Testing on the prospectively-collected data from 2021-2022 shows that there is no evidence of algorithmic drift or diminished performance when the AI reader is applied to episodes acquired more than two years after the time period of the training data. Testing on completely external datasets demonstrates that our AI reader outperforms previously published state-of-the-art AI reader models. Our simulation studies show that both AI reader-replacement and AI band-pass scenarios can significantly improve system sensitivity and specificity relative to the current two reader with arbitration system. Our investigation of the possible positive effects of human-AI collaboration in the AI triage simulations define required

conditions for that approach to achieve non-inferiority relative to the current human-reader system.

Taken together, our results show that we have created a new, high-performing AI model for cancer detection from mammograms that is particularly relevant for application in the Victorian and wider Australian population from which and for which it was developed. Our simulation results present the AI reader-replacement scenario as the top candidate approach for enabling human-AI collaboration to improve population mammographic screening. An AI reader-replacement system can achieve higher accuracy, at higher speed, with lower workload and reduced costs, all while keeping humans "in the loop" throughout a client's screening journey.

Results

AI reader demonstrates strong performance compared to human readers

The AI reader was trained on our retrospective training set, AI reader operating thresholds were determined using the development set, and then performance was evaluated on all episodes from the retrospective test set (Methods). The AI reader achieved an area under the ROC curve of 0.932 (95% CI 0.923, 0.940), superior to the weighted mean performance of the individual first and second readers (see Methods), but lower than the performance of the two reader with arbitration system (Figure 1A; also referred to as "reader consensus" for brevity). When compared against individual readers in the first or second reader position with at least 1,000 reads the AI reader equalled or improved upon 74 out of 81 (91.4%). Using an operating point to match the weighted mean individual reader's specificity (95.6%; Figure 1B and C), the AI reader achieved significant superiority over human readers in both sensitivity (75.0%; +8.3 percentage points; $P < 4.11 \times 10^{-8}$) and specificity (96.0%; $P < 1.11 \times 10^{-9}$; Supplementary Table 1; Supplementary Figure 1).

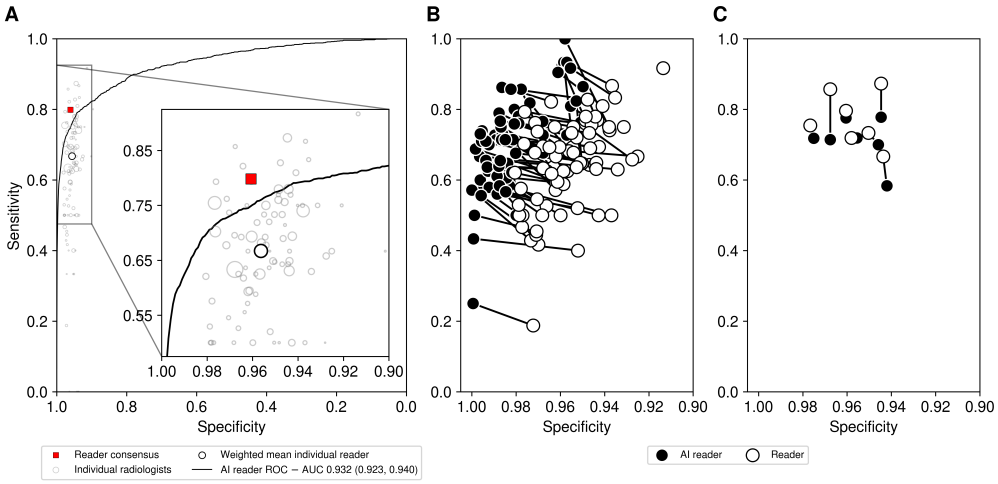

Figure 1. Performance of the AI reader on retrospective cohort. (A) The AI reader ROC curve compared with the weighted mean individual reader and reader consensus for 149,105 screening episodes. The AI reader achieved an AUC of 0.932 (95% CI 0.923, 0.940) above the weighted mean individual reader performance (95.6% specificity, 66.7% sensitivity) but below the reader consensus performance (96.1% specificity, 79.8% sensitivity; standard of care). The weighted mean individual reader (black circle) is the mean sensitivity and specificity of all the individual readers (grey circles; 125 readers) weighted by their respective total number of reads. (B) and (C) AI reader compared against 81 individual readers (min. 1,000 reads). A single point from each AI reader ROC curve is shown for each comparison. The individual AI reader ROC curve passes above or inline with 74 readers (253,328 reads, 88.3%; B) and below 7 readers (33,525, 11.7%; C).

The AI reader performed well on both screen-detected and interval cancers with the top 10% of AI reader scores containing 92.7% of screen-detected cancers, and 42.2% of interval cancers (81.8% of all cancers; Table 1). The top 50% contained 99.2% of screen-detected cancers and 84.0% of interval cancers (95.9% of all cancers), while the bottom 50% contained 25.3% of all false positives (Benign

AI reader score bin	Screen-detected cancer	Interval cancer	Normal	Benign	NSA
1	0 (0.0)	4 (1.4)	14,793 (10.4)	30 (1.0)	84 (3.0)
2	1 (0.1)	5 (1.7)	14,688 (10.3)	59 (2.0)	157 (5.6)
3	0 (0.0)	11 (3.8)	14,606 (10.3)	95 (3.1)	199 (7.1)
4	2 (0.2)	13 (4.5)	14,513 (10.2)	156 (5.2)	226 (8.1)
5	5 (0.5)	13 (4.5)	14,431 (10.2)	189 (6.3)	273 (9.8)
6	6 (0.6)	17 (5.9)	14,369 (10.1)	223 (7.4)	295 (10.6)
7	5 (0.5)	26 (9.1)	14,269 (10.1)	293 (9.7)	317 (11.4)
8	21 (2.0)	31 (10.8)	14,116 (9.9)	369 (12.2)	374 (13.4)
9	36 (3.5)	46 (16.0)	13,844 (9.8)	565 (18.7)	419 (15.0)
10	966 (92.7)	121 (42.2)	12,339 (8.7)	1039 (34.4)	446 (16.0)

Table 1. AI reader scores stratified by ground truth label. The AI reader scores binned by deciles. Bin 1 represents the episodes with the lowest 10% of AI reader scores and bin 10 the episodes with the highest 10%. The number of episodes for each category are shown for each bin, with the percentage of all episodes in the category falling into that bin shown in parentheses.

and NSA), 0.8% of screen-detected cancers, and 15.9% of interval cancers. The AI reader had excellent performance distinguishing screen-detected cancers from non-cancers and screen-detected cancers from normal (AUC 0.975; 95% CI 0.970, 0.979 and AUC 0.976; 95% CI 0.972, 0.981, respectively). On the interval cancer class (cancer cases that were not detected at screening) the AI reader achieved an AUC of 0.774 (95% CI 0.746, 0.801), similar to or better than previously published deep learning approaches [17, 26, 27]. On the false positive classes, no significant abnormality and benign, the AI reader performed well at detecting screen-detected cancer versus NSA (AUC 0.962; 95% CI 0.954, 0.969) and benign (AUC 0.920; 95% CI 0.910, 0.931). The lower performance on the benign class was consistent with expectations that they are generally more difficult to classify [17].

The AI reader generalised well across screening round, client demographics, machine manufacturer, and cancer type in our retrospective cohort, with the AI reader-replacement operating point used for all comparisons. Specifically, we compared AI reader performance with human individual and reader consensus performance on cohorts split by screening round, age, device manufacturer, self-reported ethnicity (First Nations Australians) or region of birth, and different groupings of cancer invasive status and grade (Supplementary Figure 2). For all categories, the AI reader showed very consistent performance, often exhibiting less variability than the single and reader consensus decisions. The sensitivity and specificity of the AI reader was generally higher than the weighted mean individual reader and lower than the reader consensus. The strongest result is that of the AI reader on first round screening clients and younger women (40–52 years of age) where the AI reader achieves much higher specificity at a sensitivity equivalent to the reader consensus. The first round screening clients were, on average, younger and proportionally more cancers were detected. The AI reader also performed best on high grade invasive cancers relative to the weighted mean individual and reader consensus, with low and medium grade in-situ the most challenging. Conversely, we observed that the AI reader had a general decline in specificity with client age and a consistent or slightly increasing sensitivity across different age groupings, whereas the radiologists' performance improved with client age. The AI reader provided consistent performance across different mammography machines and associated processing algorithms. The AI reader performance was in line with the weighted mean individual reader variability across manufacturers and better in some cases (manufacturer E). We examined the AI reader performance in the small cohort of First Nations Australians (n=579) and the rest of the testing set by region of birth and found no evidence of underperformance.

Taken together, evaluation of our AI reader on the testing set from our retrospective cohort shows it performs strongly in comparison to human readers and is highly consistent across different categories of screening episode.

AI reader achieves state-of-the-art performance on validation datasets

To better understand and assess the generalisability of our AI reader we assessed its performance on one local prospectively-collected dataset and three external validation datasets.

First, we collected a small prospective dataset (ADMANI Prospective data; see Methods). We collected data from December 2021 to May 2022 in real time (daily) from a single reading and assessment unit (St Vincent's BreastScreen, Melbourne, Australia) using two mammography machine manufacturers. This prospective dataset is temporally distinct from the retrospective dataset used for training our AI models (data from 2013–2019) and thus offers the opportunity to assess possible algorithmic drift or diminished performance on recently acquired screening episodes. As insufficient time has elapsed to enable a full two-year follow-up period, reporting of interval cancer cases is currently incomplete and so we restrict our analyses to screen-detected cancers in this cohort. The BRAIx AI reader performs strongly when tested on the prospective cohort, achieving an area under the ROC curve for screen-detected cancer of 0.970 (95% CI 0.953, 0.983). At the retrospective operating point and looking just at screen-detected cancers, the BRAIx AI reader achieves non-inferior sensitivity and superior specificity ($\boldsymbol{P} < 1.11 \times 10^{-9}$) compared to the radiologists on the retrospective dataset and superior sensitivity ($\boldsymbol{P} < 0.020$) and non-inferior specificity compared to the radiologists on the prospective cohort (Supplementary Table 2). Thus, we find no evidence of algorithmic drift or diminished performance of the BRAIx AI reader on episodes acquired more than two years after the most recent episodes in the training dataset. On the contrary, our results show superior AI performance on the retrospective dataset and testing on the prospective cohort provides further evidence of the superiority (or at least non-inferiority) of the AI reader relative to human readers.

Finally, we tested the AI reader on three external datasets (CMMD, CSAW-CC and BREAST; see Methods) that were not used in model development or training and derive from very different cohorts than our retrospective and prospective studies from Victoria, Australia. The BRAIx AI reader performed very strongly on the three external datasets, with AUC results ranging from 0.906 to 0.994 and excellent balancing of precision and recall (Table 2). Notably, the BRAIx AI reader achieved an AUC result of 0.994 (95% CI 0.990, 0.997) on the CSAW-CC screen-detected cancer only dataset and an AUC of 0.906 (95% CI 0.895, 0.918) on CMMD. These results improved significantly on previously reported results, represented here by the GMIC reader, which was the best-performing of six purported state-of-the-art AI readers in recent independent evaluations [28, 29]. The strong results for the BRAIx AI reader show the relative strength of our model compared to previous state-of-the-art AI readers, as well as the quality and relative difficulty of our dataset. Published performance evaluations of AI readers on publicly-available external test datasets such as the CSAW-CC and CMMD datasets remain rare, precluding further direct comparisons between the BRAIx AI reader and alternative readers from academic and commercial groups.

AI-integrated screening scenarios

To gain insight into the performance of AI readers in collaboration with human readers, we conducted detailed simulation studies of AI integrated screening scenarios using our retrospective study cohort. Specifically, we simulated three AI-integrated pathways: the "AI reader-replacement", "AI band-pass" and "AI triage" scenarios (Figure 2; Methods). Our baseline for comparisons between possible scenarios is the current reader system implemented in population breast screening in Australia, namely two readers with arbitration (Figure 2A). The AI reader-replacement scenario simply replaces one of the initial two (human) readers in the current reader system with an AI reader (Figure 2B). In the AI band-pass scenario, all episodes are first read by the AI reader, which sends episodes with a sufficiently high score straight for recall, determines episodes with a sufficiently low score to be "no recall", and passes all other episodes (with intermediate scores) onto the current reader system (Figure 2C). Finally, the AI triage scenario uses an AI reader to make an initial read of all episodes, after which episodes with a high AI reader score (e.g., top decile) are passed onto the current reader system, while the rest are passed to a single human reader for the final recall/no recall decision (Figure 2D). The triage scenario provides challenges for simulation as the AI simply makes a decision on whether each episode

Dataset	Episodes			AUC	BRAIx	GMIC
ADMANI Prospective	Normal	24990	Breast	ROC	0.976 (0.962, 0.989)	-
	Benign	663		PR	0.667 (0.601, 0.725)	-
	Screen-detected	195	Episode	ROC	0.970 (0.953, 0.983)	-
	Interval	0		PR	0.670 (0.605, 0.728)	-
CSAW-CC[1]	Normal	22868	Breast	ROC	0.994 (0.990, 0.997)	0.943 (0.931, 0.954)
	Benign	0		PR	0.860 (0.832, 0.886)	0.495 (0.447, 0.543)
	Screen-detected	524	Episode	ROC	0.993 (0.991, 0.995)	-
	Interval	0		PR	0.889 (0.867, 0.909)	-
CSAW-CC	Normal	23903	Breast	ROC	0.943 (0.933, 0.953)	-
	Benign	0		PR	0.651 (0.617, 0.682)	-
	Screen-detected	524	Episode	ROC	0.934 (0.923, 0.944)	-
	Interval	267		PR	0.685 (0.655, 0.715)	-
CMMD	Normal	0	Breast	ROC	0.906 (0.895, 0.918)	0.831 (0.815, 0.846)
	Benign	465		PR	0.922 (0.911, 0.932)	0.859 (0.842, 0.875)
	Screen-detected	1310	Episode	ROC	-	-
	Interval	0		PR	-	-
BREAST	Normal	361	Breast	ROC	0.972 (0.960, 0.982)	-
	Benign	0		PR	0.910 (0.876, 0.939)	-
	Screen-detected	179	Episode	ROC	0.962 (0.947, 0.976)	-
	Interval	0		PR	0.939 (0.913, 0.961)	-

Table 2. Results of testing the BRAIx and GMIC AI readers on prospective and external datasets. 95% confidence interval were calculated using 2,000 bootstrap replicates. "Screen-detected cancer" and "Interval" cancer are abbreviated in the table to "Screen-detected" and "Interval", respectively.
[1] screen-detected cancers only

is ultimately seen by the two reader with arbitration system or a single human reader. With this setup, the triage scenario cannot perform as well as or better than the baseline scenario, in which all episodes go through the two reader with arbitration system, unless there is positive human-AI interaction in the sense that single human reader performance improves with the knowledge of the AI output for the episode. We explore this effect in our simulation studies. It is a highly relevant clinical question as the AI triage scenario is currently being assessed in a large-scale randomised clinical trial [22], which has reported intermediate results of non-inferiority of the AI triage scenario compared to their current reader system with respect to cancer detection rates. We analysed each screening episode of the testing set in the different AI-integrated screening scenarios and compared the ground truth screening outcomes, workload implications, and costs (Table 3; see also Supplementary Table 3, 4, 5, and 6 and Supplementary Figure 3 and 4 for the detailed breakdown). These AI-integrated scenarios were ultimately compared with the existing (baseline) reader system, which has a sensitivity of 79.8% and specificity of 96.0% (Figure 3).

The AI reader-replacement scenario produced higher system sensitivity (82.3%: 95% CI 81.5, 83.1, $P < 0.0025$) and higher specificity (96.3%: 95% CI 96.2, 96.3, $P < 4.3 \times 10^{-6}$) when compared with the current two reader with arbitration system. Across 149,105 screening episodes in our retrospective testing set, the AI reader-replacement scenario had 354 fewer unnecessary recalls (-6%, false positives), detected 33 more cancers (+3.1%, true positives) with a reduction of 147,959 human reads (-48%) but requiring 11.6% more third (arbitration) reads (Table 3). The modelled reading and assessment cost reduction was 15%.

The AI band-pass screening scenario also achieved both higher system sensitivity (81.7%, $P < 0.0058$) and specificity (96.6%, $P < 2.2 \times 10^{-25}$) when compared with the current reader system. The AI band-pass screening scenario had 779 fewer unnecessary recalls (-13.3%), detected 25 more cancers (+2.4%) with a reduction of 248,638 human reads (-80.7%) including a 67.9% reduction in third reads (Table 3). The modelled reading and assessment cost reduction was 28.3%.

The AI triage screening scenario, without modelled AI-human interaction, produced lower sensitivity (77.2%) and specificity (95.7%) than the current reader system. This AI triage screening scenario had 425 more unnecessary recalls (+7.3%) and detected 35 fewer cancers (-3.3%) with a reduction of 141,799 human reads (-46%), including a 77.0% reduction in third reads (Table 3). The modelled reading and assessment cost reduction was 7.5%. However, as indicated above, the AI triage approach of allocating a large proportion of episodes to be read by only a single human reader will not outperform the current reader system unless AI-human interaction (through human readers having access to the output of the AI reader) improves human reader performance. Therefore we ran multiple versions of the AI triage simulation in which we varied the positive effect of AI-human interaction for the single human reader from a 0% improvement in human performance to 100% improvement (i.e., all mistakes by the single human reader are avoided when they have access to the AI reader output; Figure 3). With a 100% improvement in human performance thanks to AI-human interaction, the AI triage scenario can achieve sensitivity of 91.8% and specificity of 97.9%. To match or surpass the performance of the current reader system, AI-human interaction needs to avoid or rectify at least 14.1% of the single human reader's errors (Supplementary Table 7).

Variable	Current reader system	AI reader-replacement	AI band-pass	AI triage
Sensitivity (%)	79.8	82.3 (81.5, 83.1)	81.7	77.2
Specificity (%)	96.0	96.3 (96.2, 96.3)	96.6	95.7
No. of episodes	149105	149105	149105	149105
True positive	1061	1094 (+3.1%)	1086 (+2.4%)	1026 (-3.3%)
True negative	141915	142269 (+0.2%)	142694 (+0.5%)	141490 (-0.3%)
False positive	5861	5507 (-6%)	5082 (-13.3%)	6286 (+7.3%)
False negative	268	235 (-12.4%)	243 (-9.3%)	303 (+13.1%)
Workload				
Assessments	6922	6602 (-4.6%)	6168 (-10.9%)	7312 (+5.6%)
Human reads	308091	160132 (-48%)	59453 (-80.7%)	166292 (-46%)
*Third reads	9881	11027 (+11.6%)	3173 (-67.9%)	2276 (-77%)

*The third reads are part of the human readings. The separate entry is presented to show the workload impact on the third reader.

Table 3. Comparison of the current reader system and the AI-integrated scenarios by the screening outcomes, workload and cost. For the AI reader replacement scenario, mean values based on 1000 simulations are presented (rounded to the nearest integer, and the 95% confidence intervals (CI) are presented for the sensitivity and specificity). The AI band-pass and AI triage scenarios do not require simulating the third reader, and the results are entirely based on the real data (i.e., no CI). Percentages in the brackets with the plus / minus sign indicate the percentage change compared with the current reader system.

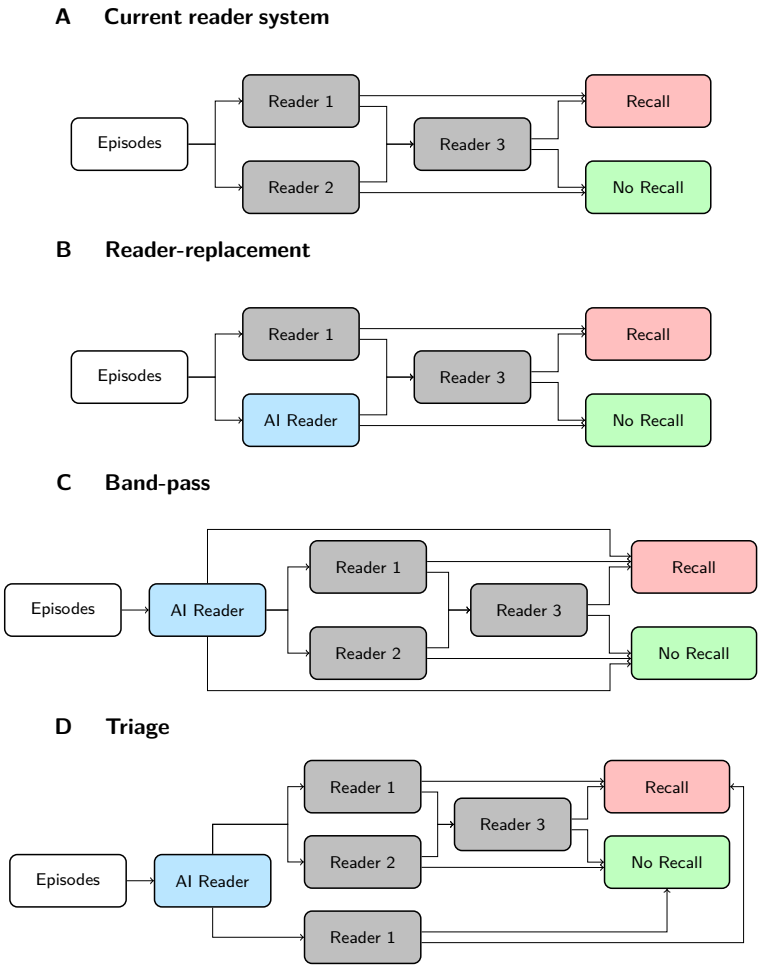

Figure 2. Screening episode flows for the current reader system and AI integration scenarios. (A) The current two reader with arbitration system: Readers 1 and 2 see the same episode and opt to recall or not-recall, if they disagree Reader 3 arbitrates. (B) AI reader-replacement: same as (A) but with AI Reader replacing Reader 2. (C) AI band-pass scenario: AI Reader screens out episodes before Readers 1 and 2. Episodes with high scores trigger the recall decision directly, and episodes with low scores trigger the no-recall decision directly. The other episodes continue to the usual reader system. (D) AI triage scenario: AI reader triages the episodes before Readers 1 and 2. Episodes with high scores continue to the usual system, and episodes with low scores go through the path with only 1 reader.

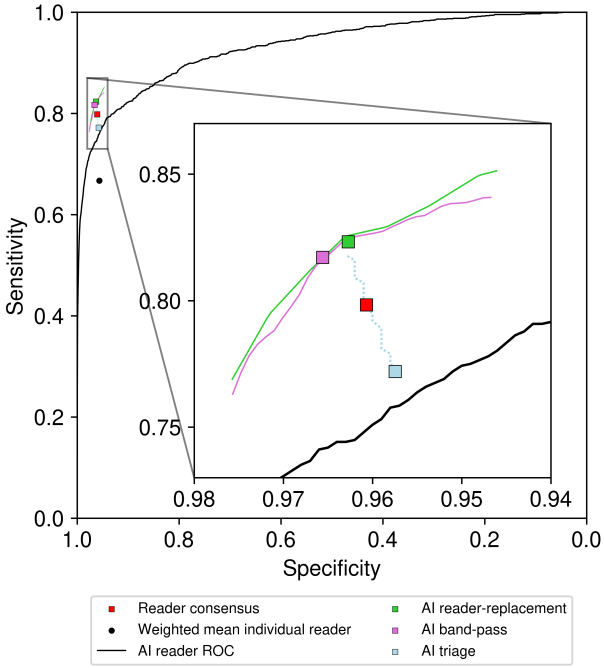

Figure 3. Comparison of AI integrated scenarios. Human reader consensus performance compared with AI reader-replacement, AI band-pass, and AI triage on the retrospective cohort (149,105 screening episodes). Two representative points are shown for AI reader-replacement (96.3% specificity, 82.3% sensitivity) and AI band-pass (96.6%, 81.7%); other operating points are shown as a continuous line. AI triage (95.7% specificity, 77.2% sensitivity) is shown (blue square) as well as AI triage modelled with 0% to 25% randomly corrected individual reader decisions (dotted blue line). Both AI reader-replacement and AI band-pass improved performance over the human reader consensus (96.1% specificity, 79.8% sensitivity); AI triage equals the human reader consensus performance with 15% of individual reader decisions corrected.

Discussion

In this study we sought to develop an AI reader for breast cancer detection in population mammographic screening that could be implemented in the Australian context within the well-established two reader with arbitration system for breast screening. Our results demonstrate the potential for an AI reader integrated into a population breast screening program to deliver improved performance compared to the current system with respect to sensitivity, specificity, and overall accuracy, while also achieving lower radiologist workload and reduced program costs, in combination leading to a better client experience.

Our AI reader, an ensemble of deep learning models trained on our retrospective training dataset of 840,827 screening episodes and 3,537,758 images, demonstrated itself to be a strong individual reader with significantly higher sensitivity and specificity than the weighted mean individual human reader and better performance than 91% of individual human readers on the testing dataset. Of note was the AI reader's particularly good performance on first-round episodes (i.e., the first screening episode for a particular client) compared to human readers. First round screening clients had a higher proportion of breast cancer (1.15% vs 0.88%) and are more challenging for human readers as there are no prior mammograms. The AI reader had higher specificity than individual readers and the reader consensus for first round episodes. We hypothesise that radiologists take good advantage of the opportunity to review previous screening rounds for a client, whereas the AI reader in its current form is limited to assessing each client episode in isolation. The AI reader also showed higher sensitivity on medium and high-grade invasive cancers and lower sensitivity on low and medium grade ductal carcinoma in situ compared with human readers. This result indicates AI can alter the spectrum of disease detected in

a positive way, which was a concern raised in reviews of AI systems in screening programs [13]. When tested on a Victorian cohort collected more than two years after the end of the acquisition period for the training dataset, our AI reader showed no evidence of algorithmic drift or diminished performance. When tested on external datasets, our AI reader outperformed previously published state-of-the-art AI readers.

Despite the AI reader's impressive performance as a single reader, however, as a standalone system the AI reader cannot match the performance of the two reader with arbitration system used in Australia and other countries with national, population breast cancer screening programs. In Australia, the first and second readers, whose function our AI reader was designed to emulate, operate independently to support a high system specificity (target recall rate of less than 10% for first round episodes and less than 5% for second and subsequent rounds). Third readers (needed when there is not a consensus) view the first and second reader decisions and arbitrate with high sensitivity as they assess a cohort of episodes containing a higher proportion of cancers. The resulting two reader with arbitration system lifts overall sensitivity without loss of specificity, and above that of an individual radiologist and our AI reader. The performance gap between a standalone AI reader and the current system necessitates some human-AI collaboration if we are to improve screening outcomes [30]. This AI collaboration could conceivably be done in many ways, such as in the form of algorithmic triage, where some cases are deferred to the human expert [31, 32], or by presenting the AI prediction and model outputs with an associated indication of confidence to improve human performance [33].

We generated detailed simulations to study three timely, feasible AI-integrated scenarios for the screening pathway: AI reader-replacement, AI band-pass and AI triage. Both AI reader-replacement and AI band-pass demonstrated significantly superior sensitivity and specificity compared to the two reader with arbitration system. As a consequence of its inherent structure, the AI triage scenario relies on single (human) reader performance improving with the availability of AI reader output to achieve performance equivalent to the current system. Under an assumption of no increase in single human reader performance the AI triage scenario significantly underperforms relative to the human-only, AI reader-replacement, and AI band-pass options. However, if AI-human interaction improves human reader performance by at least 14% then the AI triage scenario can achieve equal or better performance than the current system.

As well as exhibiting performance differences, the three AI-integrated scenarios also present differing considerations when it comes to implementation and clinical application. AI reader-replacement is conceptually straight-forward and would be the least disruptive AI-integrated reader system to implement as it does not change the overall structure of the reader system—each episode is read by at least one human reader, and humans are always "in the loop" with respect to decision making. With great flexibility in selecting "recall" and "no recall" thresholds for the AI reader, the band-pass scenario offers perhaps the greatest potential for tuning the pathway to maximise sensitivity or workload and cost reduction. However, the band-pass scenario also allows (or requires) the AI reader to make final decisions about the outcome of many episodes; with humans "out of the loop" for many decisions, the band-pass approach is likely to face the greatest resistance to clinical implementation on grounds of ethics and trust in AI for independent high-consequence decision-making. The AI triage scenario is clinically highly relevant as it is currently being assessed in a large-scale randomised clinical trial [22], which has reported intermediate results of non-inferiority of the AI triage scenario compared to the two reader with arbitration system with respect to cancer detection rates.

Interestingly, the empirical observation of non-inferiority of the AI triage scenario in that study implies that there is indeed a positive interaction between the human and AI readers that improves human reader performance. From our simulation studies we would roughly estimate the improvement thanks to human-AI interaction to be around 15%. If so, the AI triage scenario is a legitimate option for successful clinical use. The AI reader-replacement and AI band-pass scenarios offer potentially greater opportunities for improved outcomes and reduced costs (especially if human-AI interaction in those scenarios also improves human reader performance), but they have not yet been evaluated prospectively. In our judgement, considering both performance and other factors relating to implementation, the AI

reader-replacement scenario emerges as the top candidate for implementing an optimal AI-integrated pathway for a population mammographic screening service.

A key limitation of our study, as in any retrospective study, is that we do not know the impact of AI on human performance. While we have investigated some aspects of positive human-AI collaboration through our simulation studies, we cannot yet truly know to what extent and in what circumstances radiologists, with knowledge of the involvement of AI reading, access to AI reader outputs, and facing a different mix of episodes, may deliver different reading outcomes. Further prospective studies are now required to examine the impact of integration of an AI reader on human behaviour. Alongside prospective studies and further development of the datasets, translation into screening programs will require, or be aided by, many programs of work not covered in this study, including: Development of approaches to algorithm quality assurance to assess bias and drift on an ongoing basis; Deeper examination of AI explainability for client and clinician, and their acceptance of AI reader false negatives and positives; Further development of algorithms to focus on interval cancers; Utilising AI to predict risk and support personalized screening pathways.

Necessary future work building on the present study includes exploring further options for human-AI collaboration, including AI reader decision referral, and developing a detailed understanding of the implications of AI reader scores for future breast cancer risk. There is a need for improved annotation and reporting capabilities of AI readers beyond a cancer detection score when integrating into population breast cancer screening services. It looks promising that AI reader performance may be improved by integrating a longitudinal view of episodes for a given client over multiple screening rounds, taking advantage of information from prior screens as human readers do. Real-world performance of all of these ideas needs, eventually, to be assessed with prospective studies, and important questions face the field in that respect, including consideration of when or whether full randomised controlled trials are necessary to provide sufficient clinical evidence of efficacy or superiority [34].

In summary, using a large retrospective dataset, we developed the BRAIx AI reader, which shows strong performance in detecting breast cancer from mammograms. As the AI reader in an AI reader-replacement approach to integrating an AI reader with human readers, the BRAIx AI reader could plausibly be implemented in a population screening service in Victoria or Australia or beyond to transform breast cancer screening with higher accuracy, lower costs, and better client experience.

Contributors

Conceptualization: HMLF, DJM, JL, JH, GC; Methodology: HMLF, CAPS, CFK, ME, YC, CW, PB, GC, DJM; AI model and code development: CAPS, CFK, ME, YC, CW; Dataset development: HMLF, ME, CAPS, CFK; Data analysis: CAPS, CFK, ME; Resources: DJM, HMLF, PB, JL, JH, GC; Writing - original draft, review & editing: CAPS, CFK, ME, HMLF, DJM, GC; Supervision: HMLF, DJM, GC; Project administration: HMLF, DJM, GC; Funding acquisition: HMLF, DJM, PB, JL, JH, GC.

The BRAIx team comprises: Helen ML Frazer[1,2], Carlos Andres Peña-Solorzano[3,4], Chun Fung Kwok[3,4], Michael S Elliott[3,4], Yuanhong Chen[5], Chong Wang[5], Prabhathi Basnayake Ralalage[1], Ravishankar Karthik[2], Katrina Kunicki[3], Daniel Schmidt[7], Enes Makalic[7], Osamah Al-Qershi[7], Prue C Weideman[7], Samantha K Fox[7], Shuai Li[7], Tuong L Nguyen[7], Brendan Hill, Jocelyn F Lippey[1,6], John L Hopper[7], Peter Brotchie[8], Gustavo Carneiro[5], Davis J McCarthy[3,4]. [1]St Vincents BreastScreen, St Vincents Hospital Melbourne, Victoria, Australia; [2]BreastScreen Victoria, Victoria, Australia; [3]Bioinformatics and Cellular Genomics Unit, St Vincent's Institute of Medical Research, Victoria, Australia; [4]Melbourne Integrative Genomics, School of Mathematics and Statistics/School of BioSciences, Faculty of Science, University of Melbourne, Victoria, Australia; [5]School of Computer Science, Australian Institute for Machine Learning, University of Adelaide, South Australia, Australia; [6]Department of Surgery, St Vincents Hospital Melbourne, Victoria, Australia; [7]Centre for Epidemiology & Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Victoria, Australia; [8]Department of Radiology, St Vincents Hospital Melbourne, Victoria, Australia.

Declaration of interests

Peter Brotchie is an employee of annalise.ai.

Data availability

The non-transformed image and non-image data that established the ADMANI datasets were accessed under license agreement with BreastScreen Victoria. Further details about the ADMANI datasets are available

in the data descriptor paper [25]. The three datasets used as external validation are publicly available or available via request. The Chinese Mammography Dataset (CMMD) is publicly available from the website (https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70230508). The Cohort of Screen-age Women - Case control (CSAW-CC) dataset is available via request from the website (https://snd.gu.se/en/catalogue/study/2021-204). The Breast Screen Reader Assessment Strategy Australia (BREAST Australia) is available via request from the website (https://breast-australia.sydney.edu.au/research/).

Code availability

The code used for training the BRAIx AI reader is largely based on open-source algorithms and training techniques, the model names and training procedures have been provided in detail in the Methods. We are required to protect potentially commercially valuable project intellectual property, which the source code constitutes, as part of our multi-institution agreement and grant obligations. The description of the model training procedure and models used are provided in the methods and can be implemented with open-source frameworks. The main conclusions drawn in our work relate to our AI reader simulation experiments. The code used to simulate the AI reader operating within the screening program and validate the external results are available publicly at https://github.com/BRAIx-project/retrospective-cohort-study.

Acknowledgments

The authors would like to acknowledge Rita Butera, Luke Neill and Georgina Marr from BreastScreen Victoria and the leadership and staff of St Vincent's Hospital Melbourne for their support of the project. The authors would like to acknowledge Katrina Kunicki, Anne Johnston, Colleen Elso, Liz Campbell and Tom Kay of St Vincents Institute of Medical Research (SVI) for their extensive help and support with many aspects of this project. The authors thank Wayne Crismani for thoughtful comments on the manuscript. SVI provided IT support, infrastructure, and GPUs used for numerical calculations in this paper enabled via funding from the Victorian State Government Operational Infrastructure Support Program to St Vincents Institute of Medical Research. HMLF acknowledges the generous support of the Royal Australian and New Zealand College of Radiologists (Clinical Research Grant), St Vincent's Hospital Melbourne (Research Endowment Fund) and The Medical Device Partnering Program for enabling the prior foundational research to be undertaken. DJM acknowledges generous support from Paul Holyoake and Marg Downey.

References

1. World Cancer Research Fund. American Institute for Cancer Research. Breast cancer https://www.wcrf.org/dietandcancer/breast-cancer/. Accessed: 2021-08-11. 2021.
2. Australian Institute of Health and Welfare. BreastScreen Australia Monitoring Report 2022 Accessed: 2022-11-07 (Australian Institute of Health and Welfare, Australian Government, 2022).
3. Morrell, S., Taylor, R., Roder, D. & Dobson, A. Mammography screening and breast cancer mortality in Australia: an aggregate cohort study. Journal of medical screening 19, 26–34 (2012).
4. Rodríguez-Ruiz, A. et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologistsdembrower. JNCI: Journal of the National Cancer Institute 111, 916–922 (2019).
5. Rodríguez-Ruiz, A. et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 290, 305–314 (2019).
6. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. Nature 577, 89–94 (2020).
7. Kim, H.-E. et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. The Lancet Digital Health 2, e138–e148. ISSN: 2589-7500. https://www.sciencedirect.com/science/article/pii/S2589750020300030 (2020).
8. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with deep learning. Scientific Reports 8, 1–7 (2018).
9. Salim, M. et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA oncology 6, 1581–1588 (2020).
10. Schaffter, T. et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA network open 3, e200265–e200265 (2020).

11. Wu, N. et al. Deep Neural Networks Improve Radiologists Performance in Breast Cancer Screening. IEEE Transactions on Medical Imaging 39, 1184–1194 (2020).

12. Frazer, H. M., Qin, A. K., Pan, H. & Brotchie, P. Evaluation of deep learning-based artificial intelligence techniques for breast cancer detection on mammograms: Results from a retrospective study using a BreastScreen Victoria dataset. Journal of medical imaging and radiation oncology 65, 529–537 (2021).

13. Freeman, K. et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. bmj 374 (2021).

14. Lauritzen, A. D. et al. An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. Radiology, 210948 (2022).

15. Leibig, C. et al. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. The Lancet Digital Health 4, e507–e519 (2022).

16. Larsen, M. et al. Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program. en. Radiology 303, 502–511. ISSN: 0033-8419, 1527-1315. http://dx.doi.org/10.1148/radiol.212381 (2022).

17. Lotter, W. et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nature Medicine 27, 244–249 (2021).

18. Sharma, N. et al. Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. en. BMC Cancer 23, 460 (2023).

19. Yala, A., Schuster, T., Miles, R., Barzilay, R. & Lehman, C. A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. en. Radiology 293, 38–46. ISSN: 0033-8419, 1527-1315. http://dx.doi.org/10.1148/radiol.2019182908 (2019).

20. Dembrower, K. et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. The Lancet Digital Health 2, e468–e474 (2020).

21. Shoshan, Y. et al. Artificial Intelligence for Reducing Workload in Breast Cancer Screening with Digital Breast Tomosynthesis. en. Radiology 303, 69–77. ISSN: 0033-8419, 1527-1315. http://dx.doi.org/10.1148/radiol.211105 (2022).

22. Lång, K. et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. The Lancet Oncology 24, 936–944. ISSN: 1470-2045. https://www.sciencedirect.com/science/article/pii/S147020452300298X (2023).

23. Dembrower, K. et al. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. en. The Lancet. Digital health. ISSN: 2589-7500. http://dx.doi.org/10.1016/S2589-7500(23)00153-X (2023).

24. Segnan, N. & Ponti, A. Artificial intelligence for breast cancer screening: breathtaking results and a word of caution. en. The Lancet Oncology 24, 830–832. ISSN: 1470-2045, 1474-5488. http://dx.doi.org/10.1016/S1470-2045(23)00336-4 (2023).

25. Frazer, H. M. et al. ADMANI: Annotated Digital Mammograms and Associated Non-Image Datasets. Radiology: Artificial Intelligence 5, e220072 (2022).

26. Zhu, X. et al. Deep Learning Predicts Interval and Screening-detected Cancer from Screening Mammograms: A Case-Case-Control Study in 6369 Women. Radiology 301. PMID: 34491131, 550–558. eprint: https://doi.org/10.1148/radiol.2021203758. https://doi.org/10.1148/radiol.2021203758 (2021).

27. Damiani, C. et al. Evaluation of an AI Model to Assess Future Breast Cancer Risk. Radiology 307. PMID: 37310244, e222679. eprint: https://doi.org/10.1148/radiol.222679. https://doi.org/10.1148/radiol.222679 (2023).

28. Stadnick, B. et al. Meta-repository of screening mammography classifiers. arxiv:2108.04800 (2021).

29. Velarde, O. M. & Parra, L. Independent evaluation of state-of-the-art deep networks for mammography. arXiv: 2206.12407 [eess.IV]. http://arxiv.org/abs/2206.12407 (2022).

30. Wilder, B., Horvitz, E. & Kamar, E. Learning to Complement Humans 2020. arXiv: 2005.00582 [cs.AI].

31. Okati, N., De, A. & Gomez-Rodriguez, M. Differentiable Learning Under Triage 2021. arXiv: 2103.08902 [stat.ML].

32. Dvijotham, K. et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. Nature Medicine 29, 1814–1820. https://doi.org/10.1038/s41591-023-02437-x (2023).

33. Vodrahalli, K., Gerstenberg, T. & Zou, J. Uncalibrated Models Can Improve Human-AI Collaboration 2022. arXiv: 2202.05983 [cs.AI].

34. Van Nijnatten, T. J. A., Payne, N. R., Hickman, S. E., Ashrafian, H. & Gilbert, F. J. Overview of trials on artificial intelligence algorithms in breast cancer screening - A roadmap for international evaluation and implementation. en. European journal of radiology 167, 111087. ISSN: 0720-048X, 1872-7727. http://dx.doi.org/10.1016/j.ejrad.2023.111087 (2023).

35. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library in Advances in Neural Information Processing Systems 32 (eds Wallach, H. et al.) (Curran Associates, Inc., 2019), 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

36. Wightman, R. PyTorch Image Models https://github.com/rwightman/pytorch-image-models. 2019.

37. Strand, F. CSAW-CC (mammography) en. 2022. https://doi.org/10.5878/45vm-t798.

38. Cui, C. et al. The Chinese Mammography Database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast. The Cancer Imaging Archive 1 (2021).

39. Brennan, P., Tapia, K., Ryan, J. & Lee, W. BREAST: a novel method to improve the diagnostic efficacy of mammography in Medical imaging 2013: image perception, observer performance, and technology assessment 8673 (2013), 28–32.

40. Youden, W. J. Index for rating diagnostic tests. Cancer 3, 32–35 (1950).

41. With contributions from Bernardo Sousa-Pinto, P. D. mada: Meta-Analysis of Diagnostic Accuracy R package version 0.5.11 (2022). https://CRAN.R-project.org/package=mada.

42. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157 (1947).

43. Trajman, A. & Luiz, R. McNemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations. Scandinavian journal of clinical and laboratory investigation 68, 77–80 (2008).

44. Kim, S. & Lee, W. Does McNemars test compare the sensitivities and specificities of two diagnostic tests? Statistical methods in medical research 26, 142–154 (2017).

45. Marinovich, M. L. et al. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. en. EBioMedicine 90, 104498 (2023).

Methods

Inclusion and ethics

The conceptualisation, design, and implementation of this study was conducted with close collaboration between clinical staff working in the organised population breast screening service in Victoria, Australia and local academic researchers. Ethics approval for this study was provided by the St Vincent's Hospital, Melbourne human research ethics committee (approval no. LNR/18/SVHM/162). All women sign a consent form at screening registration that provides for the use of the de-identified data for research purposes. A unique identifier is used for the purposes of the ADMANI datasets, with all image and non-image data de-identified.

Screening program

The BreastScreen Victoria screening program is a population screening program open to women aged 40+ with those between the ages 50–74 actively recruited. A typical BreastScreen Victoria client has a mammogram taken with a minimum of four standard mammographic views (left and right mediolateral oblique, MLO, and craniocaudal, CC) every two years. Annual screening is offered to a small proportion of high-risk clients (<2%).

Every client undergoing screening through BreastScreen Victoria experiences a standardised screening pathway and data generation process (Supplementary Figure 5). Each mammogram is read independently by two breast imaging radiologists who indicate suspicion of cancer, all clear, or technical rescreen. If there is disagreement a third reader, with visibility of the original two readers' decisions, determines the final reading outcome. Clients with a suspicion of cancer are recalled for assessment. At assessment, further clinical workup and imaging is performed. Any client who has a biopsy-confirmed cancer at assessment (within six months of screening) is classified as a screen-detected cancer (true positive). Any clients who are recalled but confirmed with no cancer after follow-up assessments are classified as either benign or no significant abnormality (false positive). Clients who were not recalled at reading and do not develop breast cancer within the next screening interval are classified as normal (true negative). Clients who develop breast cancer between six months after a screen and the date of their next screen (12 or 24 months) are classified as interval cancers (false negative). The datasets we use are structured around individual screening episodes of clients attending BreastScreen Victoria. A screening episode is defined as a single screening round that includes mammography, reading, assessment, and the subsequent screening interval.

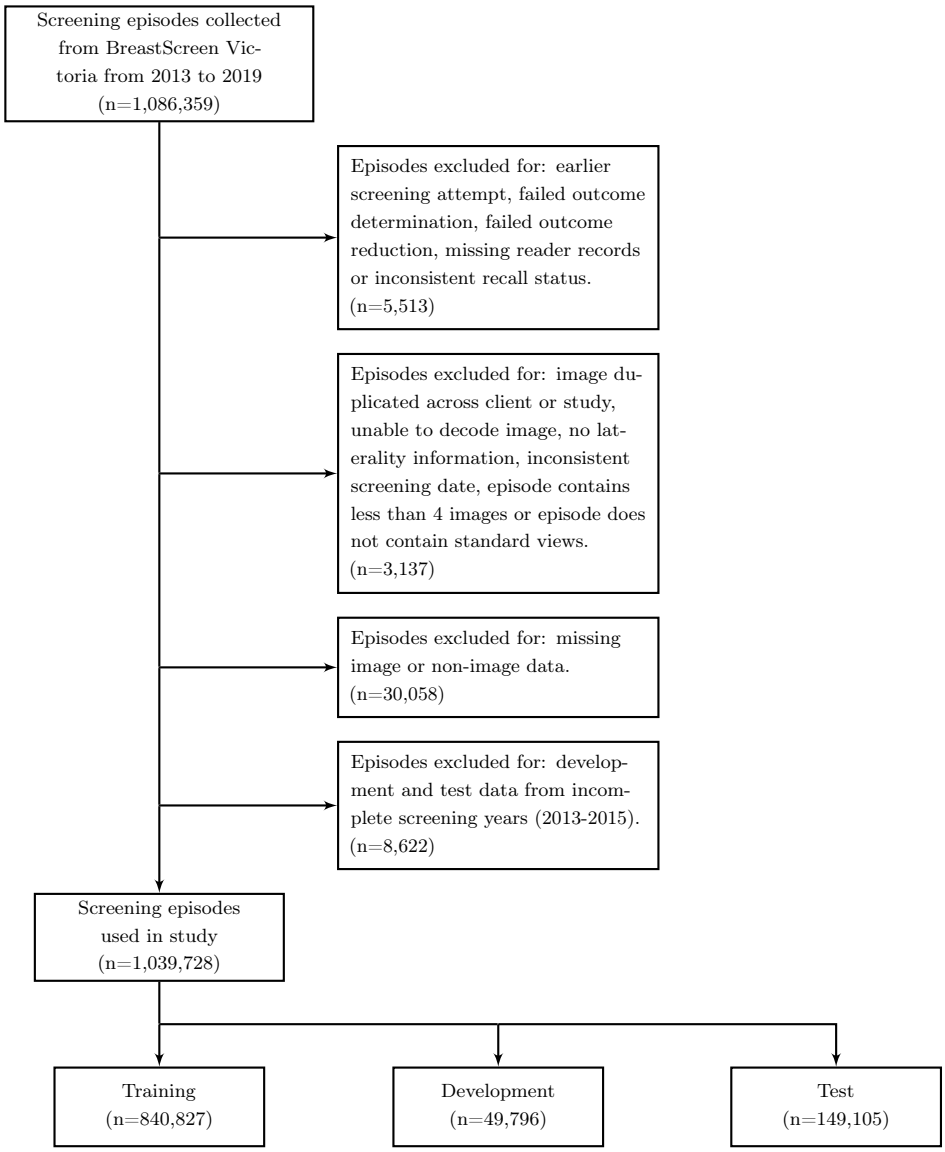

Figure 4. Screening episode exclusion criteria. Flow diagram of study exclusion criteria for screening episodes from the standardised screening pathway at BreastScreen Victoria.

Study Datasets

A large, population breast cancer screening dataset was curated from clients across 46 permanent screening clinics and two mobile services across the state of Victoria, Australia from 2013 to 2019 (Table 4). Screening episodes that were missing any of the standard mammographic views, had incomplete image or non-image data, or did not have interval cancer follow-up were excluded (Figure 4). If a screening episode had multiple screening attempts only the final attempt was used. Clients with breast implants or other medical devices were included.

The final dataset contained a total of 1,039,728 episodes and 4,374,956 images from 629,853 clients. Screening episodes from 2013, 2014, and 2015 were cancer enriched and contained 44,605 clients, 45,650 episodes, and 192,670 images. Screening episodes from 2016–2019 were sequential population cohorts containing 618,973 clients, 994,078 episodes, and 4,182,286 images.

The dataset was split by client into training, development, and testing sets (Table 4). For the years with complete data (2016–2019), 15% of the clients were allocated to the testing set, 5% of the clients allocated to the development set, and all remaining clients to the training set, including those from

2013–2015 that did not appear in later years. All screening episodes for each client were allocated to their respective sets. The testing, development, and training sets contain 92,839 clients, 30,946 clients, and 506,068 clients respectively.

The dataset has strong ground truths for all cancer (screen detected and interval) and non-cancer (normal, benign, or no significant abnormality (NSA) with no interval cancer). Cancer was confirmed by histopathology or when recorded in the screening interval. The histopathological proof was predominantly from an assessment biopsy confirmed with subsequent surgery. The ground truth for clients without cancer was a non-cancer outcome after reading and no interval cancer (normal) or non-cancer outcome after assessment and no interval cancer (benign or NSA). Information on country of birth, whether or not the client identifies as Aboriginal and/or Torres Strait Islander, and age was collected at the time of screening. Responses for country of birth and Aboriginal and/or Torres Strait Islander identification were aggregated into categories of First Nations Australians and regions.

Separately to the retrospective analysis a prospective dataset was collected from December 2021 to May 2022. Data was collected in real-time (daily) from a single reading and assessment unit (St Vincent's Breastscreen, Melbourne, Australia) using two mammography machine manufacturers from December 2021 to May 2022. The prospective dataset contains the same ground truth and demographic information with the exception of interval cancer data as it was not yet available at the time of publication. The prospective dataset consisted of a total of 25,848 episodes and 108,654 images from 25,848 clients with a total of 195 screen-detected cancers (Supplementary Table 8).

AI Reader System

The BRAIx AI reader (v3.0.7) system consists of multiple deep learning models trained on different subsets of the training dataset. The models are all variants of modern convolutional neural networks (CNNs) implemented using the PyTorch [35] and the PyTorch Image Models [36] libraries. The models are optimised for performance on the manufacturer and population characteristics of the ADMANI datasets, specifically operating within the first or second reader position which demands high specificity with good sensitivity. The AI reader provides a score for each breast and the maximum breast score is the episode score.

Simulation design, operating points and evaluation metrics

To provide insights into the AI reader and its potential in clinical application, we performed retrospective simulation studies, where we evaluated the AI reader performance as a standalone reader and in three AI-integrated screening scenarios.

AI as a standalone reader

In the standalone reader case the AI reader was evaluated against the radiologist performance and the two reader with arbitration system. The AI reader episode predictions were compared with the real episode outcomes and the original reader decisions. The overall performance was summarised using the area under the curve (AUC) score obtained from the receiver operating characteristic (ROC) curve as well as at binary-decision operating points where sensitivity and specificity can be calculated.

External datasets

We accessed three external datasets for the purpose of validating the AI reader performance: the Cohort of Screen-age Women - Case control (CSAW-CC) [37], the Chinese Mammography Dataset (CMMD) [38], and the BreastScreen Reader Assessment Strategy Australia (BREAST Australia) [39]. CSAW-CC is a subset of a large screening population dataset from Sweden containing 23,903 normal, 524 screen-detected cancer and 267 interval cancer episodes. We report breast and episode level results for the entire dataset and also for normal versus screen-detected only. CMMD is a publicly available dataset containing 465 benign and 1,310 malignant mammograms. As not all clients in the CMMD dataset have four mammographic images we report breast level results only. BREAST is a collection of curated test sets designed to allow radiologists to assess their performance, containing 361 normal and 179 screen-detected cancer episodes. We present the breast and episode results excluding any

prior episodes. These datasets contain mammography vendors not used in the training of our models and no fine-tuning of the AI reader was done on these datasets. They were used purely as external test sets for the AI reader. Other datasets commonly used for comparisons - INbreast, DDSM, OPTIMAM - were considered, but omitted for reasons of sample size or licensing/access restrictions.

AI-integrated screening scenarios

Three scenarios were considered to evaluate the AI reader integrated in the screening pathway, AI reader-replacement, AI band-pass, and AI triage. In the current screening pathway the first two readers individually make a decision on whether or not to recall the client for further assessment. If the two readers agree, that is the final reading outcome. If they disagree, a third reader, who has access to the first two readers' decisions and image annotations, arbitrates the decision.

In the AI reader-replacement scenario, the AI reader replaces one of the first two readers in the screening pathway and provides the same binary recall or no-recall outcome as the human readers. The first and second readers are replaced at random (with equal probability) for each episode. As the AI reader could trigger a third read that did not exist in the original dataset, the third reader was simulated for all episodes even if an original third read was present. This approach is to prevent unduly tying the result to the dataset and to obtain better variability estimates. Sensitivity analysis where the third reader uses the real data when possible and where the replaced second reader is used as the third reader was also performed (Supplementary Table 9). The third reader in our retrospective cohort operated with a sensitivity of 97.5% and a specificity of 56.3% and we simulated the third reader with respect to this performance. Confidence intervals were generated through 1000 repetitions of each simulation.

In the AI band-pass scenario, the AI reader was used analogously to a band-pass filter. The AI reader provided one of three outcomes: recall, pass, and no-recall. All episodes with recall outcome were automatically recalled. All episodes with the no-recall outcome were not recalled. All episodes with the pass outcome were sent to the usual human screening pathway. The AI reader made the final decision on the recall and no-recall episodes with no human reader involvement, and for all episodes that passed to the human screening pathway the original reader decisions were used.

In the AI triage scenario, the AI reader triages the episodes before the human readers. Episodes with high scores continue to the standard pathway, and episodes with low scores go through the pathway with only 1 reader. For episodes sent to the standard pathway, the original reader decisions were used, and for episodes sent to the single-reader path, the reader decision is sampled randomly (with equal probability) from the first and second readers. The AI reader made no final decision on any of the episodes.

AI operating points

Two sets of operating points were used as part of the study: the AI reader-replacement reader and the AI band-pass reader. All operating points were set on the development set (Supplementary Figure 1).

The AI reader-replacement operating point used a set of manufacturer-specific thresholds to convert the prediction scores into a binary outcome: recall or no-recall. The operating point was chosen to improve on the weighted mean individual reader's sensitivity and specificity. The weighted mean individual reader was the weighted (by number of reads) mean of the sensitivity and specificity of the individual (first and second) radiologists when they were operating as a first or second reader. For all operating points that improved on the weighted mean individual readers sensitivity and specificity the point with the maximum Youden's index [40] was chosen.

The AI band-pass reader used two sets of manufacturer-specific thresholds to convert the prediction scores into three outcomes: recall, pass and no-recall. The AI band-pass simulation was evaluated at different AI reader thresholds via a grid search. At each evaluation, two thresholds for each manufacturer were set to a target high specificity and high sensitivity point. All episodes with a score above the high sensitivity point were given the recall outcome, all episodes below the high specificity point the no-recall outcome, and all episodes in between points were given the pass outcome.

The final AI band-pass reader thresholds were chosen from the simulation result with the maximum Youden's index of the points with non-inferior sensitivity and specificity than that of the two reader with arbitration system.

The AI triage reader used the 90% quantile of the prediction scores as the threshold to convert the prediction scores into the triage outcome: standard pathway or one-reader pathway. Episodes with prediction scores less than the threshold are assigned to the one-reader pathway; otherwise, they are assigned to the standard pathway.

Evaluation metrics

The AUC, based on the ROC curve, is used to summarise the AI reader's standalone performance. The ROC curve is a curve that measures a classifier's ability to distinguish between two classes in a binary classification problem. It is created by plotting the true positive rate (TPR or sensitivity) against the false positive rate (FPR) at various threshold settings (FPR can be calculated as $1 - \text{specificity}$). The area under the ROC curve, denoted by AUROC or AUC, summarizes the ROC curve into a single value between 0 and 1 that describes the performance of a model for multiple thresholds at the same time. Higher values indicate better performance, with a score of exactly 1 only obtained for a perfect classifier. The reader summary ROC curve was fitted using a bivariate model using the individual radiologist sensitivities and specificities as provided by the `mada` R package [41].

Sensitivity and specificity are used to compare the AI reader with the radiologists and the AI-integrated screening scenarios with the current screening pathway. Sensitivity, or the true positive rate (TPR), is computed by dividing the number of correctly identified cancers (the "true positives") by the total number of the observed cancers ("all positives", i.e., including both screen-detected cancers and interval cancers). It measures the success rate of the classifier in detecting the cancer. This is a key performance metric because early detection of cancer leads to more effective treatment. Specificity, or the true negative rate (TNR), is computed by dividing the number of correctly identified non-cancer episodes ("true negatives") by the total number of the observed non-cancer episodes ("all negatives"). It measures the success rate of the classifier in correctly not recalling a client when cancer is absent in the client. This is a key performance metric because unnecessarily recalling clients to assessment center is costly and induces significant stress on clients and their families.

To compare the sensitivity and specificity of the AI-integrated screening scenarios with the current screening pathway, McNemar's test (with continuity correction) is used to test for differences [42], and the binomial exact test (one-sided) is used to test for superiority. Both tests adhere to the correct design for McNemar's test, in which a 2-by-2 contingency table is constructed based on the paired samples from the two comparison scenarios. The tests are conducted once for each of the sensitivity and specificity [43, 44] at a significance level of 5%.

No. of episodes	Test	Development	Train	All
Age group				
40 to 49	11,873	3,874	67,207	82,954
50 to 59	58,835	19,596	332,616	411,047
60 to 69	54,491	18,284	309,442	382,217
70 to 74	19,482	6,574	107,508	133,564
75+	4,424	1,468	24,053	29,945
Screening round				
1	19,374	6,474	111,181	137,029
2+	129,731	43,322	729,645	902,698
Country of birth				
Australia	97,880	32,841	555,276	685,997
United Kingdom	9,666	3,133	53,841	66,640
Vietnam	3,579	1,182	19,515	24,276
Italy	3,340	1,145	18,827	23,312
Other	34,640	11,495	193,367	239,502
Risk category				
None	34,325	11,499	227,646	273,470
Average	102,098	34,077	545,447	681,622
Moderate	7,588	2,532	40,539	50,659
High	5,094	1,688	27,194	33,976
Personal history of breast cancer				
Yes	683	250	3,577	4,510
No	148,422	49,546	837,249	1,035,217
Recalled for assessment	6,871	2,253	52,119	61,243
Screening outcome				
Screen-detected cancer	1,042	316	8,888	10,246
Interval cancer	287	100	2,745	3,132
Normal	141,968	47,449	786,191	975,608
Benign	3,018	983	22,268	26,269
No significant abnormality	2,790	948	20,734	24,472
Total	149,105	49,796	840,826	1,039,727

Table 4. Summary and characteristics of data used in the study.