AI integration improves breast cancer screening in a real-world, retrospective cohort study

Helen ML Frazer [1 2 †✉], **Carlos A Peña-Solorzano** [3 4 †], **Chun Fung Kwok** [3 4 †], **Michael 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 Vincent's 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, JI, 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

Background Artificial intelligence (AI) readers, derived from applying deep learning models to medical image analysis, hold great promise for improving population breast cancer screening. However, previous evaluations of AI readers for breast cancer screening have mostly been conducted using cancer-enriched cohorts and have lacked assessment of the potential use of AI readers alongside radiologists in multi-reader screening programs. Here, we present a new AI reader for detecting breast cancer from mammograms in a large-scale population screening setting, and a novel analysis of the potential for human-AI reader collaboration in a well-established, high-performing population screening program. We evaluated the performance of our AI reader and AI-integrated screening scenarios using a two-year, real-world, population dataset from Victoria, Australia, a screening program in which two radiologists independently assess each episode and disagreements are arbitrated by a third radiologist.

Methods We used a retrospective full-field digital mammography image and non-image dataset comprising 808,318 episodes, 577,576 clients and 3,404,326 images in the period 2013 to 2019. Screening episodes from 2016, 2017 and 2018 were sequential population cohorts containing 752,609 episodes, 565,087 clients and 3,169,322 images. The dataset was split by screening date into training, development, and testing sets. All episodes from 2017 and 2018 were allocated to the testing set (509,109 episodes; 3,651 screen-detected cancer episodes). Eight distinct AI models were trained on subsets of the training set (which included a validation set) and combined into our ensemble AI reader. Operating points were selected using the development set. We evaluated our AI reader on our testing set and on external datasets previously unseen by our models.

Findings The AI reader outperformed the mean individual radiologist on this large retrospective testing dataset with an area under the receiver operator characteristic curve of 0.92 (95% CI 0.91, 0.92). The AI reader generalised well across screening round, client demographics, device manufacturer and cancer type, and achieved state-of-the-art performance on external datasets compared to recently published AI readers. Our simulations of AI-integrated screening scenarios demonstrated that a reader-replacement human-AI collaborative system could achieve better sensitivity and specificity (82.6%, 96.1%) compared to the current two-reader consensus system (79.9%, 96.0%), with reduced human reading workload and cost. Our band-pass AI-integrated scenario also enabled both higher sensitivity and specificity (80.6%, 96.2%) with larger reductions in human reading workload and cost.

Interpretation This study demonstrated that human-AI collaboration in a population breast cancer screening program has potential to improve accuracy and lower radiologist workload and costs in real world screening programs. The next stage of validation is to undertake prospective studies that can also assess the effects of the AI systems on human performance and behaviour.

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) with approximately one million Australian women 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].

However, there has been little change in the manner of cancer screening in the 30 years since BreastScreen Australia was established, and there are challenges in accuracy, experience and efficiency. Despite a process of independent double reading of all mammograms by radiologists (and a third arbitration read when there is discordance), 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 after receiving an “all clear” result and before their next scheduled screen. Only 49.9% of eligible women participated in screening in 2019-2020, down from 54% participation 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 improve the accuracy, experience, and efficiency of breast cancer screening. Recent studies have demonstrated the potential of AI to detect breast cancer on mammographic images with a similar accuracy to radiologists [4–12]. However, current evidence is insufficient for clinical translation [13]. Limitations include:

- the use of small, retrospective, often cancer-enriched datasets;
- incomplete ground truths, particularly with respect to interval cancers diagnosed following an “all clear” screening mammogram and prior to the next screening round;
- lack of generalisability and assessment of performance across different population sub-groups of client, machine manufacturers, age groups, number of rounds, ethnicity or region of birth, and cancer type and grade; and
- limited/no validation with screening pathway application, understanding of impact on radiologist behaviour and prospective studies.

There are few studies published in breast cancer screening pathway applications with large population datasets. Using a UK screening population testing dataset of 25,856 episodes, 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]. Using a Danish screening population testing dataset of 114,421 screening episodes, an AI reader was used to filter high confidence no-cancer and cancer classifications, with all other episodes proceeding to a double reading consensus model. A retrospective study showed potential for reduced radiologist workload and lower false positives [14]. Using a testing dataset of 213,694 German population screening episodes, a decision referral approach applied to the second reader was found to improve sensitivity and specificity relative to individual radiologists [15].

Our paper builds on these insights and seeks to address many of the limitations by using a large population testing dataset, strong ground truths, examination of generalisability, and complete screening pathway simulation with a unique digital twin application.

Our dataset derives from the ADMANI datasets [16] from screening services across Victoria (Australia) and comprises 808,328 screening episodes with 3,404,367 mammographic images. Our work advances previous studies with a large, sequential screening population testing set of 509,109 screening episodes and retrospective simulation of population screening outcomes against the actual two reader (third arbitration read if required) consensus model. We focus on clinically meaningful outcomes such as improvements in interval cancer detection and reductions in unnecessary recalls, alongside reduced workload. Our dataset has exceptionally accurate ground truth labels, as all cancer episodes (screen-detected and interval) are histopathologically confirmed from biopsies and all no-cancer episodes are confirmed by observation of a two-year interval without cancer following an all-clear, benign, or no significant abnormality screening result. All episodes from clients who developed an interval cancer are included. The datasets comprise the screening population of women aged 40 years and above (target recruitment is age 50-74), with multiple screening sites and mammographic machine vendors. A digital twin screening application enabled simulation of different AI uses in the screening pathway. Results from two screening pathway scenarios are presented:

- The *band-pass scenario* where the AI reader is used to filter high-confidence episodes without human involvement. Those episodes for which the AI reader has a high confidence indication of cancer are immediately recalled for further assessment and those with high confidence indication of no cancer are given an "all clear" and invited for routine re-screening in two years' time. Only those ambiguous episodes for which the AI reader does not have high confidence in a cancer or no cancer decision are managed by the current two-reader consensus model.
- The *reader replacement scenario* where one of the first two readers is replaced by an AI reader in the current consensus model.

The reader replacement scenario has the advantage that a radiologist remains in charge for all episodes, so that the mammogram analysis is not fully automated. The band-pass scenario requires care with threshold settings to enable high confidence in both cancer and no cancer filters.

We investigate classification accuracy of benign and no significant abnormality findings, as these cases are integral parts of real-world screening. We examine differences between radiologist and AI readers acknowledging that even when the AI reader performs at the same accuracy level as a radiologist, its predictions sometimes differ. This feature has potentially impactful clinical implications because in an independent double read (with third read if discordant) system, these types of disagreement lead to arbitration, which may increase the overall arbitration workload. We present analyses across the population and various sub-groups, including machine manufacturers, age groups, number of screening rounds, ethnicity or region of birth, and cancer type and grade. We evaluate performance using sensitivity, specificity, present workload, and economic analyses.

Box 1. Research in context

Evidence before this study

We searched PubMed for English-language studies of deep learning-based breast cancer detection with mammography published from Jan 1, 2012 to July 31, 2022. We used the terms "deep learning", "breast cancer screening", and "mammography" to identify papers that triage screening mammograms between normal and suspicious, or that improved the screening efficacy by replacing one or more of the radiologist positions (readers) by artificial intelligence (AI) models in AI-in-the-loop approaches. We found studies using AI algorithms for the classification of breast cancer mammograms, completely replacing the human diagnosis. One study focused on the decrease of false positives and false negatives obtained by the AI system when compared to human readers, while maintaining a non-inferior performance when used as part of a double-reading process. The comparison of AI models against individual readers helps to contextualize the performance of the AI systems in the clinical setting, to define useful metrics, and to validate using external datasets, but results cannot be translated directly to the clinic due to the regular use in breast cancer screening programs of decision pathways that use two or more readers. These decision pathways present, in general, a better performance than achieved by individual readers, and although replacing one of the readers by an AI model can reduce the total workload, in-depth analyses of the whole-system performance have not been presented. Major limitations of prior studies include the use of small, retrospective, often cancer-enriched datasets, incomplete ground truths, exclusion of interval cancers, lack of generalisability and performance assessment across differing population cohorts, and minimal or no validation with screening pathway applications.

Added value of this study

We trained and validated eight deep learning algorithms and then combined them to create an ensemble AI reader to diagnose breast cancer on mammograms from 808,320 episodes and 577,581 clients from multiple screening locations, including multiple mammography machine vendors. To the best of our knowledge, our study utilizes the largest dataset with the most comprehensive demographic data of clients screened for breast cancer for the development of deep learning models. This study presents a pathway towards the inclusion of a high-performance AI-based reader in AI-integrated screening scenarios in which human and AI readers collaborate to improve screening outcomes. Specifically, we use detailed simulations to assess the performance of AI-integrated scenarios in which an AI model triages cases for human readers ("band-pass" AI-integrated screening scenario) or the AI model replaces one of the initial human readers in a two-reader with third reader arbitration consensus system ("reader replacement" AI-integrated screening scenario). We proposed and studied different optimization points for these AI-integrated pathways to combine the strengths of both the human and the AI readers. We used our operational experience in breast cancer screening in Victoria along with the detailed non-image data in our dataset to conduct in-depth economic and cohort-driven analyses of the effects caused by different screening scenarios and AI optimization goals. We show that adding an AI reader into the already existing decision pathways, not necessarily operating where expected, can lead to improved results, and also that our stand-alone AI reader represents a new state of the art when compared against radiologist-made decisions on external validation datasets.

Implications of all the available evidence

Both the band-pass and reader replacement AI-integrated pathways achieve better sensitivity and specificity than the original screening pathway, while significantly reducing the workload of human readers, and in turn the associated costs of screening. The band-pass system performs strongly and provides greater cost savings than the reader replacement pathway, but it allows the AI model to make recall/no-recall decisions on episodes never seen by a human reader. This removal of human intervention potentially increases the operational risk of implementing the model and may hinder acceptance by consumers and clinicians, affecting its adoption in the clinic. In contrast, the reader replacement AI-integrated scenario achieves superior performance and reduced costs while retaining human involvement in the decision for every episode. This model for human-AI collaboration in screening may aid the acceptance and adoption of AI-integrated screening pathways. In summary, this study provides evidence for the application of AI algorithms in AI-in-the-loop breast cancer screening pathways, leading to improvements in breast cancer detection, reduction in unnecessary recall to assessment, and significant reduction to human reader workload and screening costs.

Materials and Methods

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 1). 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 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 an individual screening episode of a client attending BreastScreen Victoria. A screening episode is defined as a single screening round that includes mammography, reading, assessment and the subsequent screening interval.

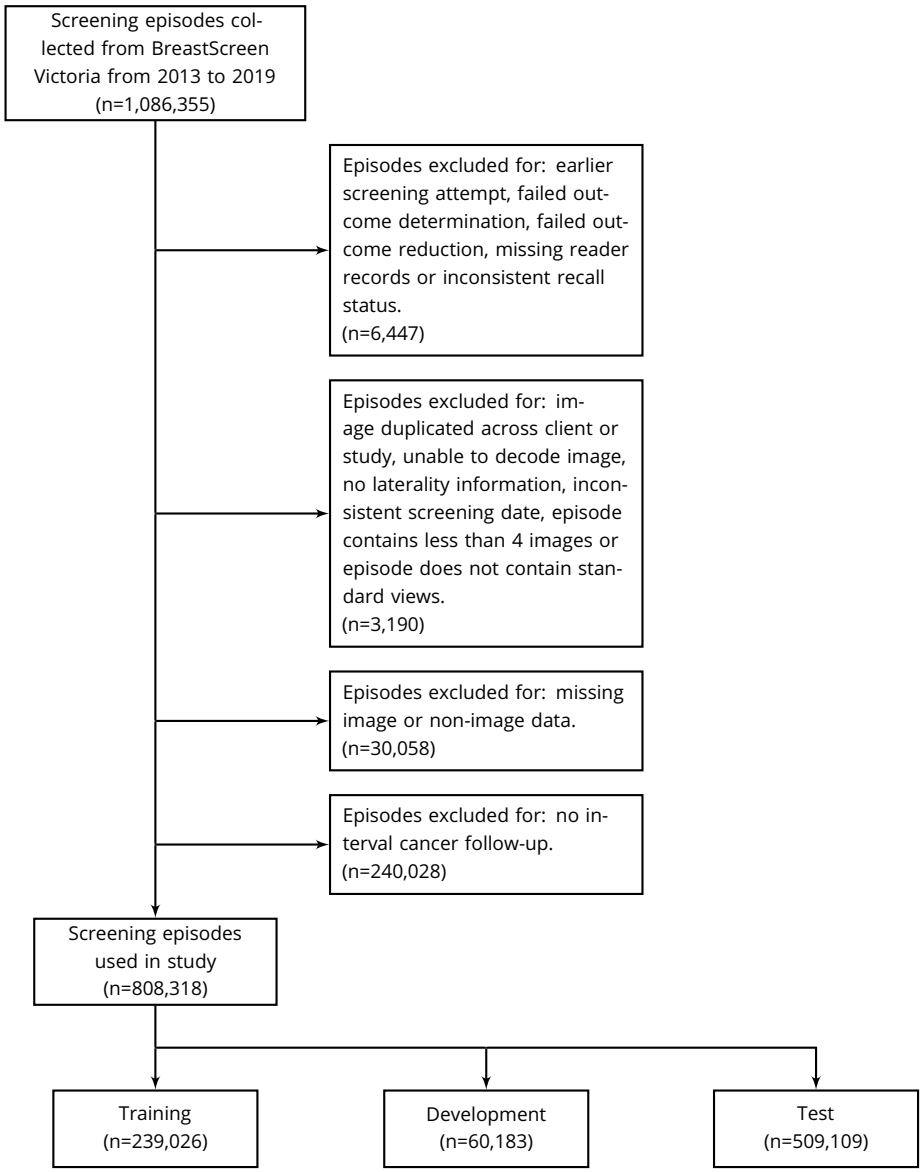

Figure 1. Screening episode exclusion criteria. Flow diagram of study exclusion criteria for screening episodes from the standardised screening pathway at BreastScreen Victoria.

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 1). 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 1). 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 808,318 episodes and 3,404,326 images from 577,576 clients. Screening episodes from 2013, 2014, 2015 and 2019 were cancer enriched and contained 54,395 clients, 55,709 episodes, and 235,004 images. Screening episodes from 2016, 2017 and 2018 were sequential population cohorts containing 565,087 clients, 752,609 episodes, and 3,169,322 images. This three year population cohort contained 98.35% of the published total episodes and 98.49% of the screen-detected cancer episodes.

The dataset was split by screening date into training, development, and testing sets (Table 2). Model training and validation sets were drawn from the training set. The development set was used to set model operating points (e.g., classification threshold). Reported performance was calculated on the testing set. All screening episodes from 2013, 2014, 2015, 2019 and January to September of 2016 were allocated to the training set. The 2019 data consisted only of screen detected cancers. All screening episodes from October to December 2016 were allocated to the development set. All episodes from 2017 and 2018 were allocated to the testing set, which established the sequential population testing set.

Our dataset had 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 on 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 no-cancer outcome after reading and no interval cancer (normal) or 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 were aggregated into categories of First Nations Australians and regions.

Outcome	No. of episodes							
	2013	2014	2015	2016	2017	2018	2019[1]	All
Screen-detected cancer	6	1,655	1,730	1,747	1,795	1,856	1,458	10,247
Interval cancer	397	442	440	526	495	496	0	2,796
Normal	62	22,534	12,660	231,610	241,465	243,101	0	751,432
Benign	4,378	2,131	784	4,965	5,258	5,092	0	22,608
No significant abnormality	4,596	1,816	620	4,652	4,926	4,625	0	21,235
All	9,439	28,578	16,234	243,500	253,939	255,170	1,458	808,318

Table 1. Number of episodes in entire dataset, split by year.
[1] awaiting interval cancer data extract

Outcome	No. of episodes			
	Training	Development	Test	All
Screen-detected cancer	6,165	431	3,651	10,247
Interval cancer	1,668	137	991	2,796
Normal	209,477	57,389	484,566	751,432
Benign	11,165	1,093	10,350	22,608
No significant abnormality	10,551	1,133	9,551	21,235
All	239,026	60,183	509,109	808,318

Table 2. Number of episodes in training, testing and development splits.

AI Reader System

An initial investigation was performed over a large number of AI readers based on Convolutional Neural Network models (CNNs). They were trained and tested on a small subset (~4,000 cancer and ~4,000 non-cancer images) of the dataset; the subset was used only on this stage to provide an initial filter of possible AI models found in the literature. The models were shortlisted based on their documented performance on the ImageNet dataset [17], with the requirement that they were open source and had pretrained weights and model variants available and implemented in the PyTorch

Image Models (timm) library [18]. The CNN families studied include: ResNet [19], DenseNet [20], ECA-Net [21], EfficientNet [22], Inception [23], and Xception [24].

The performance of the models on the chosen subset allowed for a selection of the four best models for our application, while initially limiting the complexity of the system by only allowing the use of single input models, that is, models with which only one image is analyzed and scored at a time. More complex models were developed, specifically tailored to be used with our dataset, that are able to use multiple mammographic views of the breast to generate a cancer prediction score, or are able to be used in conjunction with single input models to provide interpretable results.

The final AI reader system comprises an ensemble of eight individual CNN models, four off-the-shelf CNN models and four model architectures developed specifically for this problem. More specifically, the ensemble combines four standard single image input models (versions of ResNet and ECA-Net), two multi-view models that use two mammographic views of the same breast concurrently [25], and two single-image interpretable models that provide improved prediction localisation and interpretability [26].

ResNets, or Residual Networks, allow information that comes directly from previous layers to be added on subsequent layers, ensuring that knowledge of the original image is not lost on deeper layers. ResNets allow for the training of deeper networks that are easier to optimise and present improved performance [11, 27, 28]. ECA-Nets or Efficient Channel Attention networks use architectural units (ECA modules) capable of capturing interaction between channels, improving performance and task generalization. The network used in this work is ECA-ResNet, a ResNet variant that uses the ECA module, improving its baseline performance without increasing the computational burden [29, 30]. The multi-view network, BRAIxMVCCL, uses two image-views of a single breast as inputs: a bilateral craniocaudal (CC) view, and a mediolateral oblique (MLO) view. The use of multiple views allows the network to learn both whole-image representations and the correspondence in local regions between the views, merging the information and improving the decision-making process before producing a prediction [27, 31]. For the interpretable model, the BRAIxProtoPNet++ framework was used, which allows for the integration of an interpretable model (in this case prototype-based [32, 33]) into a breast cancer classifier network that uses mammograms as inputs. This approach generates patches for each class from the training images (prototypes), with patches similar to the characteristics of the test image being loaded along a similarity map presenting evidence for the prediction chosen.

Each model from the ensemble was implemented in PyTorch [34] and trained on data splits from the training set. The four standard single image input models from the ensemble were trained using the Adam optimiser [35] with an initial learning rate of 0.005, without weight decay, and with the AMSGrad variant enabled [36]. During training, the learning rate was reduced by a factor of 0.1 based on the binary cross-entropy loss (BCE) during validation if the loss did not improve for three sequential epochs, and the training was stopped if the validation loss did not improve for five sequential epochs. Similarly, the BRAIxMVCCL model and the BRAIxProtoPNet++ framework were trained for 20 epochs using the Adam optimiser, but with learning rate of 0.0001, weight decay of 10^{-6}, and with the same learning rate reduction functionality.

The training set was selected to have a 10:1 ratio for non-cancers (benign, no significant abnormality and normal) and screen-detected cancers, respectively, as we found this ratio to represent an efficient training regime. To enforce the chosen ratio, not all the available non-cancer images in the dataset are used during training of the models. This training choice has a direct impact on the performance of the model, improving its specificity, a measure of how well the model is able to identify true negatives (non-cancer episodes). Images were pre-processed to remove text and background annotations outside the breast region, then cropped and padded to keep the same ratio as the final resized image: 1536×768 pixels. Data augmentation consisted of random affine transformations [37].

Individual model predictions are obtained for each episode by taking the maximum prediction from each breast on the episode, with each breast calculated as the average of the predictions on

the individual views (CC, MLO). The model predictions of the eight deep learning models are averaged together to produce an output ranging from 0 to 1, with larger numbers indicating increased probabilities of cancer present on the episode.

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 two 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 consensus reader 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.

AI-integrated screening scenarios

Two AI-in-the-loop scenarios were considered to evaluate the AI reader while integrated within the current screening pathway. 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 annotations, arbitrates the decision.

In the first AI-integrated scenario ("reader-replacement"), the AI reader randomly replaced one of the first two readers in the screening pathway and provided the same binary recall or no recall outcome as the readers. The readers were replaced with equal probability using a (pseudo) random number generator from the statistical software R. For each episode the AI reader outcome was combined with an original reader outcome. If a third read was required and the original third read was available then the original third read was used. In the event the AI reader triggered a third read which did not exist in the original dataset, the third reader was simulated. The third reader was simulated such that the simulated third-reading performance matched the retrospective third-reading performance from the dataset. In the reader-replacement scenario, all episodes were inspected by at least one human reader.

In the second AI-integrated scenario ("band-pass"), the AI reader was used analogously to a band-pass filter. The AI reader provided a ternary outcome based on two thresholds: recall, pass and no-recall. All episodes above the recall threshold were recalled. All episodes below the no-recall threshold were not recalled. All the episodes between the two thresholds were passed to the usual 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.

AI operating points

Three operating points were used as part of the study: the AI improved mean reader, AI improved system reader and the AI band-pass reader. All operating points were set on the development set (see Supplementary Figure 2 for the operating points on the test sets).

The AI improved mean reader used a single threshold to convert the prediction scores into binary outcomes. The operating point was designed to improve on the actual mean reader sensitivity and specificity while controlling for manufacturer differences. The threshold was chosen by sampling points on the development set ROC curve set with sensitivity and specificity greater than the mean individual reader. For each sampled point, a threshold for each manufacturer was set to

match the target specificity and recalculate the overall performance using the individual manufacturer thresholds. The final AI improved mean reader operating point was the point that achieved the maximum Youden's index [38] of all sampled points.

The AI improved system reader used a single threshold to convert the prediction scores into a binary outcome. The operating point was designed to improve on the consensus system (as opposed to the individual reader) sensitivity and specificity while controlling for manufacturer differences. The AI improved system reader operating point was an output of the simulation study. The threshold was chosen by performing a grid search across the development set ROC curve, adjusting for each manufacturer, within the AI-integrated scenario. The final AI improved system reader operating point was the point that achieved the maximum Youden's index of the AI-integrated reader replacement consensus performance where both the *system-wide* sensitivity and specificity of the AI-integrated scenarios were non-inferior to the consensus performance of the current screening pathway.

The AI band-pass reader used two thresholds to convert the prediction scores into a ternary outcome. The operating point was again designed to improve on the consensus system sensitivity and specificity while controlling for manufacturer differences. The AI band-pass reader was an output of the band-pass AI-integrated scenario. The two thresholds were set by performing a grid search across the development set ROC curve, setting both thresholds and evaluating the system performance. The final AI band-pass operating point was the point with the maximum Youden's index of the AI-integrated band-pass consensus performance where both the *system-wide* sensitivity and specificity of the AI-integrated scenarios were non-inferior to the consensus performance of the current screening 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 $1 -$ 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.

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.

Results

AI reader performance

AI reader predictive scores align with outcomes

We first characterised the distributions of predictive scores for the AI reader by episode outcome (Figure 2) by converting raw prediction scores to binned scores based on deciles (Supplementary

Figure 3). Comparing the deciles of the prediction scores—with 1 representing the 0th–10th percentile and 10 representing the 90th–100th percentile—relative to the proportion of episodes for each of the label classes provided a high-level summary of the predictive signal in the model outputs. In the case of screen-detected cancer (Figure 2E), we observed that the top 10% prediction scores contained about 90% of the screen-detected cancers. The same top 10% scores contained about 40% of the interval cancers and benign episodes (Figure 2C, D), 20% of the NSA episodes (Figure 2B), and less than 10% of the normal episodes (Figure 2A). The mean binned prediction scores of the outcomes were 5.39 for normal episodes, 6.76 for NSA episodes, 7.85 for benign episodes, 7.80 for interval cancer and 9.75 for screen-detected cancers. In general, screen-detected cancers had higher prediction scores than all the other outcomes. Interval cancers were harder to detect compared with screen-detected cancers, and both benign and NSA were harder to identify compared with normal, given their higher average prediction scores and greater overlap in scores with cancer episodes.

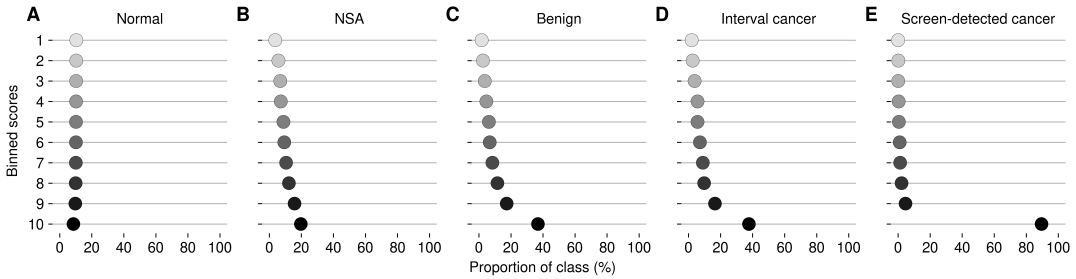

Figure 2. Distribution of episode outcomes by decile scores. Each of the five figures plots the decile scores against the proportion of cases in the respective outcome. The total number of episodes in each of the classes were 484,566 for normal, 9,551 for NSA, 10,350 for benign, 3,651 for screen-detected cancer, and 991 for interval cancer.

The distribution of the normal category was mostly uniform across the deciles, while the distribution of the screen-detected cancers was highly skewed, and the rest were in between. This result indicates there was a strong signal for the screen detected cancers and a weak signal for interval cancers. Considering the normal (Figure 2A) and the screen-detected cancer (Figure 2E), if the operating point threshold were set at bin 10, then the AI reader would recall 3,241 cancers correctly and 40,000 normal episodes unnecessarily. The cancer-to-normal trade-off would be 1 correctly identified cancer to 12 unnecessary recalls of normal. Moving the threshold down to bin 9 would recall an additional 168 cancers and 50,000 normal. The trade-off ratio of 1:298, about 25 times worse than bin 10. Similarly, bin 1 contained about 50,000 normals but 2 cancers only. The cancer-normal ratio was 1:25,000, about 2,000 times worse than bin 10. The final decile was used as a point of comparison since this decile was where the operating point would be if the sensitivity and specificity of the AI reader were tuned to match those of the average radiologist.

AI reader outperforms the mean individual reader

Overall, the AI reader performance was superior to the mean individual first and second reader performance, but lower than the consensus human performance after the third reader arbitration (Figure 3A). The AI reader was evaluated on all episodes from the retrospective test cohort using the grouped ground truth labels. The interval and screen-detected cancers are grouped into the cancer class, and normal, benign, and NSA are grouped into the no-cancer class. The AI reader achieved an area under the receiver operator characteristic of 0.92 (95% CI 0.91, 0.92) and outperformed the mean individual reader (67.6%, 95.7%). At the mean reader's sensitivity, the AI reader achieved an improved specificity of 96.9% (1.2% better), and at the mean reader's specificity, the AI reader achieved an improved sensitivity of 71.0% (3.4% better).

The AI reader had excellent performance distinguishing screen-detected cancers from non-

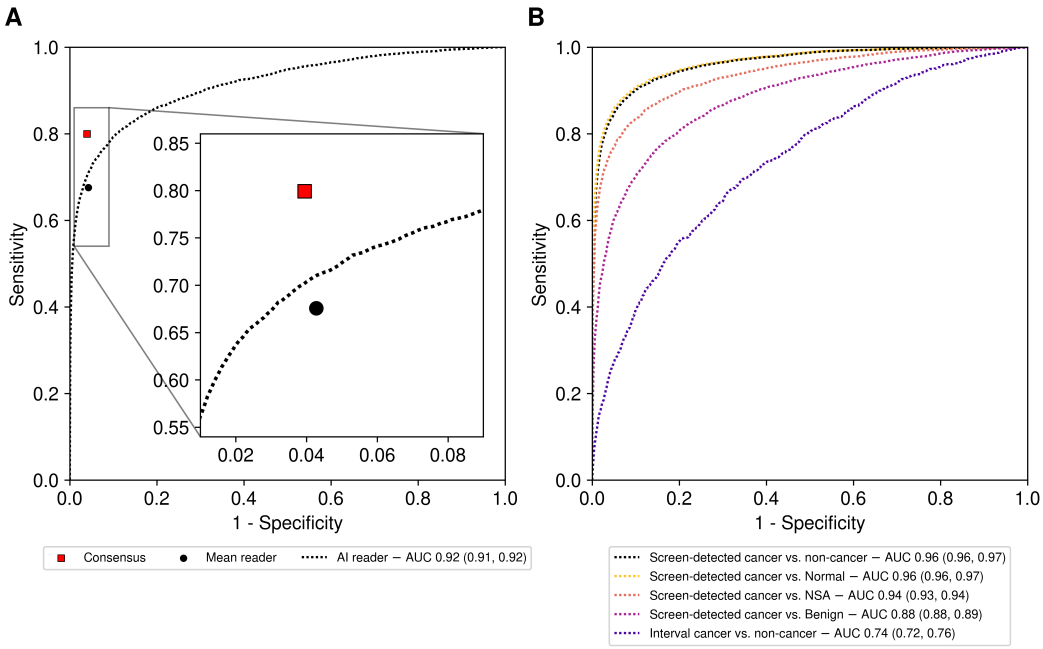

Figure 3. AI reader receiver operating characteristic curves. (A) AI reader performance compared with current system assessed on the testing set. The mean reader (black circle) is the mean sensitivity and specificity of all the first and second readers each weighted by their respective total number of reads. The consensus radiologist performance (red square) is the overall system performance after any third reader arbitration. **(B)** AI reader performance on each ground truth label assessed on the testing set.

cancers and screen-detected cancers from normal (Figure 3B, both AUC 0.96; 95%CI 0.96, 0.97). On the interval cancer class, cancer cases that were not detected at screening, the AI reader achieved an AUC of 0.74 (95%CI 0.72, 0.76). On the false positive classes, NSA and benign, the AI reader performed well detecting screen-detected cancer versus NSA (AUC 0.94; 95% CI 0.93, 0.94) and benign (AUC 0.88; 95%CI 0.88, 0.89). The lower performance on the benign class was consistent with expectations they are generally more difficult to classify.

AI reader generalises across screening round, client demographics, manufacturer and cancer type

Using the AI improved mean reader threshold setting, we compared AI reader performance with individual and consensus reader performance on cohorts split by screening round, age, device manufacturer, self-reported ethnicity or region of birth and different groupings of cancer invasive status and grade (Figure 4). The AI improved mean reader operating point was used for all comparisons. We found the AI reader performed with better specificity than individual readers and the consensus system in the first round cohort (Figure 4A). The first round screening clients were younger and proportionally more cancers were detected (Supplementary Figure 4 and 5).

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 (Figure 4B). The radiologists' performance improved with client age. Age groupings were selected to keep the proportions approximately similar; the 40-52 age group contains 69.23% of all first round episodes (45,812).

We adjusted the operating thresholds for distinct manufacturer variations and found that the AI reader provided consistent performance across different mammography machines and associated processing algorithms (Figure 4C). The AI reader performance was inline with the mean individual reader variability across manufacturers and improved upon in some cases (manufacturer E).

We found no indication of under-performance of the AI reader in any minority groupings of

ethnicity or country of birth (Figure 4D) absolutely or relative to the readers. Sensitivity was consistent across all groupings and specificity was non-inferior or superior to the dominant grouping (Australia and Pacific Islands).

The AI reader performed best on high grade invasive cancers relative to the mean and consensus readers, with low and medium grade in-situ the most challenging (Figure 4E and F).

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

We tested the AI reader on three external datasets that were not used in training: the Cohort of Screen-age Women - Case control (CSAW-CC) [39], the Chinese Mammography Dataset (CMMD) [40], and the BreastScreen Reader Assessment Strategy Australia (BREAST Australia) [41]. 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 mammogram 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.

The AI reader performed very strongly on the three external datasets, with AUC results ranging from 0.877 to 0.989 and excellent balancing of precision and recall (Table 3). The BRAIx AI reader AUC result of 0.989 (95% CI 0.983, 0.993) on the CSAW-CC screen-detected cancer only dataset and AUC of 0.877 (95% CI 0.863, 0.890) on CMMD improve significantly on previously reported results [42]. The strong results show the relative strength of our model as well as the quality and relative difficulty of our dataset.

Dataset	Episodes			AUC	Ours	GMIC
CSAW-CC[1]	Normal	22868	Breast	ROC	0.989 (0.983, 0.993)	0.943 (0.931, 0.954)
	Benign	0		PR	0.795 (0.757, 0.830)	0.495 (0.447, 0.543)
	Screen-detected cancer	524	Episode	ROC	0.986 (0.981, 0.991)	-
	Interval cancer	0		PR	0.825 (0.796, 0.853)	-
CSAW-CC	Normal	23903	Breast	ROC	0.926 (0.914, 0.937)	-
	Benign	0		PR	0.591 (0.555, 0.626)	-
	Screen-detected cancer	524	Episode	ROC	0.919 (0.907, 0.930)	-
	Interval cancer	267		PR	0.628 (0.594, 0.661)	-
CMMD	Normal	0	Breast	ROC	0.877 (0.863, 0.890)	0.831 (0.815, 0.846)
	Benign	465		PR	0.900 (0.888, 0.912)	0.859 (0.842, 0.875)
	Screen-detected cancer	1310	Episode	ROC	-	-
	Interval cancer	0		PR	-	-
BREAST	Normal	361	Breast	ROC	0.968 (0.956, 0.978)	-
	Benign	0		PR	0.891 (0.850, 0.925)	-
	Screen-detected cancer	179	Episode	ROC	0.957 (0.939, 0.972)	-
	Interval cancer	0		PR	0.931 (0.901, 0.956)	-

Table 3. Results of testing the AI reader on external datasets, 95% confidence interval calculated using 2,000 bootstrap replicates.
[1] screen-detected cancers only

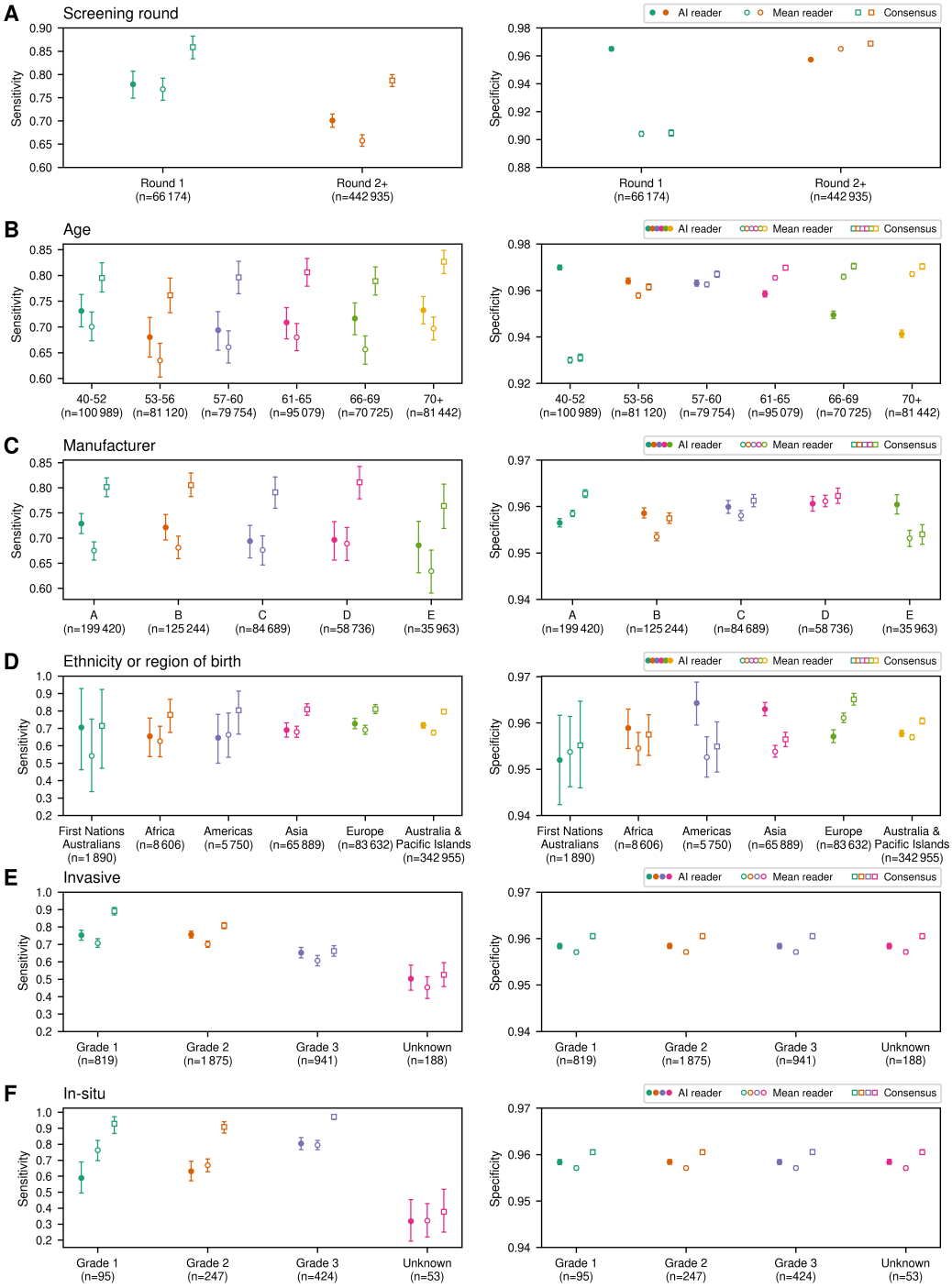

Figure 4. AI reader performance compared to radiologist performance on different episode groupings. The sensitivity and specificity with bootstrapped 95% confidence intervals of the AI reader at the AI improved mean reader operating points, the mean individual reader and the consensus reader are reported for 2,000 replicates of each grouping. The top three countries for each regional grouping were—Europe: England (23%), Italy (14%), Greece (10%); Asia: Vietnam (18%), China (16%), Philippines (9%); Americas: USA (30%), Chile (19%), Canada (14%); Africa: South Africa (35%), Egypt (20%), Mauritius (18%); and Australia & Pacific Islands: Australia (97%), New Zealand (2%), Fiji (< 1%).

Comparing the AI reader with individual radiologists

Individual radiologist comparisons demonstrate the AI reader's strength and consistency. We evaluated how the AI reader ranks among the human reader cohort by comparing the performance of the AI reader with each radiologist on their individual subset of reads (Figure 5). The comparison was limited to radiologists who had read more than 2,500 screening episodes as a first or second reader to ensure all readers had read a cancer episode. The comparison included 509,101 unique episodes and each episode was read by at least two readers. There were 109 readers in total and a subset of 83 had more than 2,500 reads. The operating point used was the AI improved mean reader.

In the 83 reader subset the median number of reads was 8,474 (IQR 5,144 - 14,414), the maximum number of reads by an individual was 68,225 and the minimum was 2,645. The AI reader had a higher sensitivity and specificity than 22 readers (234,187 reads by radiologists, 23.4%) and lower sensitivity and specificity than four readers (78,805 radiologist reads, 7.9%). For the remaining 57 readers (688,035 radiologist reads, 68.7%) we observed a trade-off between sensitivity and specificity, with the AI reader generally trading specificity for sensitivity due to the choice of operating point. The AI reader was more specific but less sensitive than 21 of the readers and more sensitive but less specific than 36 readers.

Overall, the AI reader had higher sensitivity with less variance than the radiologists, with a mean sensitivity of 71.23% (SD 6.22%) and mean specificity of 95.75% (SD 0.59%) compared with the radiologists who had a mean sensitivity of 65.94% (SD 10.94%) and specificity of 95.73% (SD 1.42%). With strong performance of the AI model in comparison to individual radiologists we can productively explore plausible screening models in which the AI model works together with human readers to improve the overall performance of the screening system.

AI-integrated screening scenarios

We conducted detailed simulation studies of the "reader-replacement" and the "band-pass" AI-integrated screening scenarios. We analysed the flow of the mammograms through the different AI-integrated screening scenarios (Figure 6 and Figure 7) and the cost implications when compared with the existing screening pathway (Table 4; see also Supplementary Table 1, 2 and 3 for the detailed breakdown). These AI-integrated scenarios were also compared with the existing screening pathway using sensitivity and specificity (Figure 8).

The AI reader-replacement (improved mean) scenario produced a trade-off between sensitivity and specificity when compared with the existing screening pathway. The consensus performance after integration with the AI improved mean reader had a lower sensitivity (77.4%, -2.5%) but higher specificity (97.1%, +1.1%) when compared with the existing consensus performance (79.9%, 96.0%). The AI reader-replacement (improved mean) scenario had 5,535 fewer unnecessary recalls with a cost reduction of 29.4% (A$12.3 million) but missed 114 more cancers (Figure 7A). The reduction in reading and assessment costs was the result of 509,109 fewer first or second reads, 3,749 more third reads, and 5,649 fewer recalls to assessment.

The AI reader-replacement (improved system) scenario achieved improved sensitivity and specificity when compared with the existing screening pathway. The consensus performance after integrating with the AI improved system reader had a higher sensitivity (82.6%, +2.7%) and specificity (96.1%, +0.1%) than the existing consensus performance (Figure 7B). The AI reader-replacement (improved system) had 445 fewer unnecessary recalls, 126 fewer missed cancers and an 8.7% cost reduction (A$3.67 million). The scenario had 509,109 fewer first and second reads and 86,771 more third reads for an overall reduction of 422,338 human reads. The increased third reads was a result of increased disagreement between the AI and human readers at the first and second position. The reduction in human reads alongside 319 fewer recalls to assessment led to the A$3.67 million in cost savings.

The AI band-pass screening also achieved improved system sensitivity and specificity (80.6%,

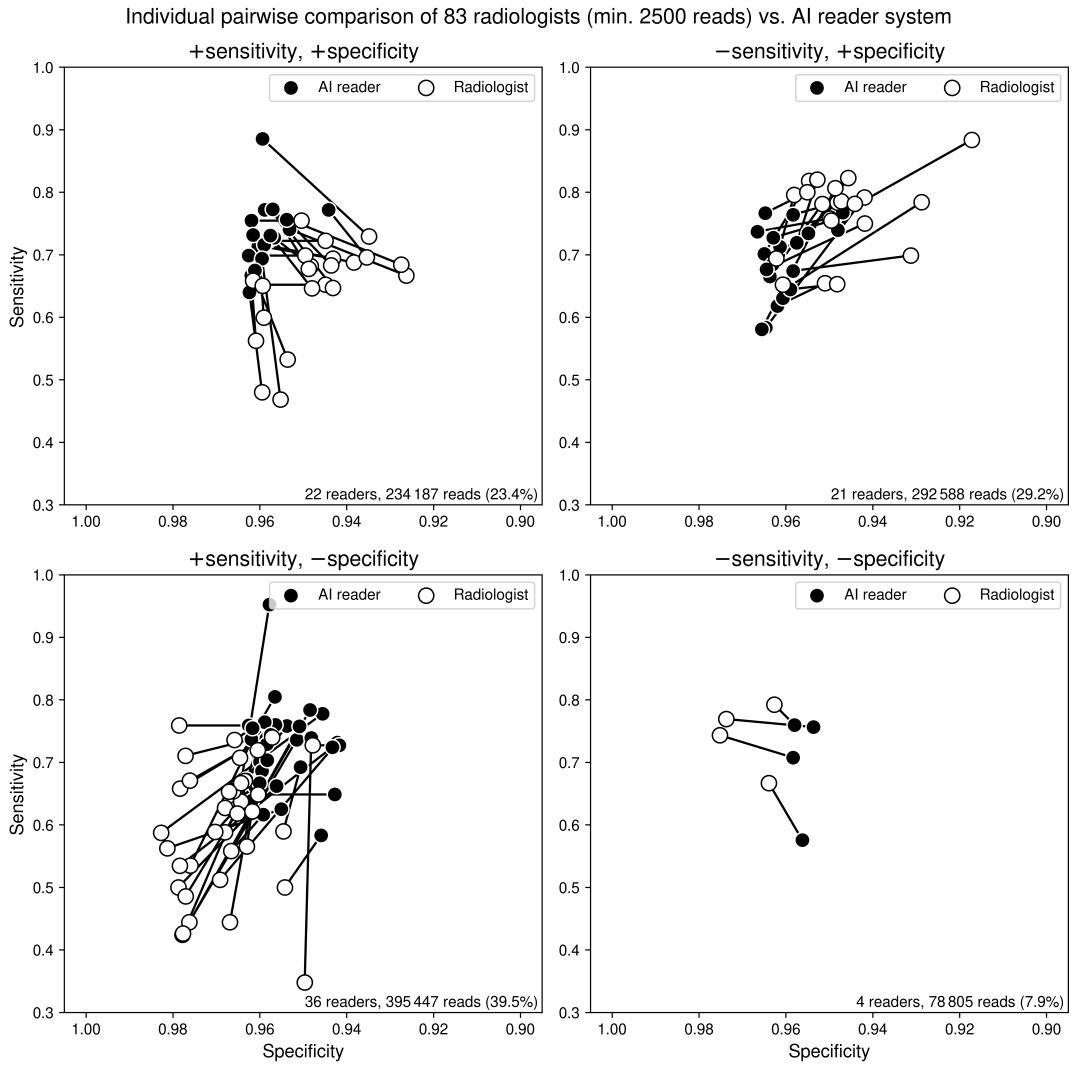

Figure 5. Individual pairwise comparison of 83 radiologists and the AI reader system. The AI reader and each radiologist (83 radiologists who read a minimum of 2,500 episodes) were compared on the same episodes. For 22 of 83 readers (23.4% of radiologist reads), the AI reader was strictly better (better sensitivity and specificity) and for 4 of 83 (7.9% of radiologist reads) the AI reader was strictly worse (lower sensitivity and specificity). For the remaining 57 of 83 readers (68.7% of radiologist reads) the AI reader had either higher sensitivity or higher specificity. There were 509,101 unique episodes and each episode was read more than once by the radiologists as a result of the two individual readers in the system. The AI reader operating point was the AI improved mean reader.

96.2%) when compared with the existing screening pathway. The AI band-pass screening had 1,135 fewer unnecessary recalls, a 19.6% cost reduction (A$8.3 million), and detected 34 more cancers (Figure 6B). As the AI band-pass reader operates before the first two human readers, when the AI reader decides on an episode directly two human reads are saved instead of one. In the AI band-pass screening scenario there was an overall reduction of 687,105 (65.3%) human reads with the number of third reads reduced by half (16,393, 49.1%).

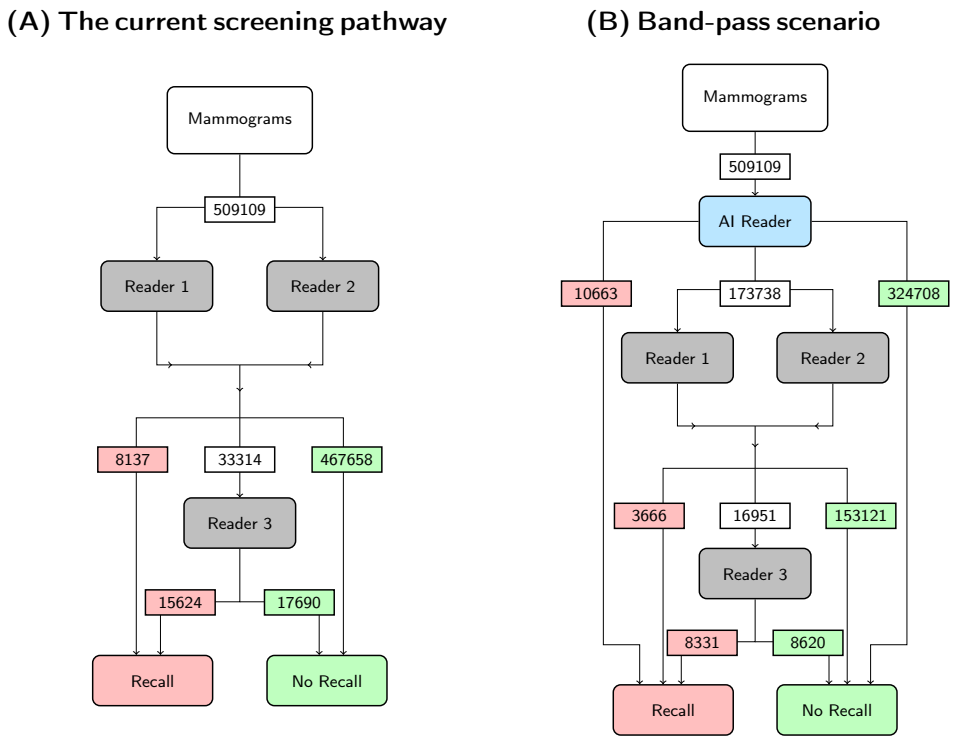

Figure 6. The current screening pathway and the band-pass AI-integrated screening scenario. (A) The current screening pathway: Readers 1 and 2 see the same episode and opt to recall or not-recall, if they disagree Reader 3 arbitrates. The sensitivity and specificity of the system is 79.9% and 96.0% respectively. (B) The band-pass AI-integrated screening scenario: An AI reader is set up to screen out the mammograms before Readers 1 and 2. Mammograms with high scores trigger the recall decision directly, and mammograms with low scores trigger the no-recall decision directly. The mid-score mammograms continue to the usual screening pathway. The sensitivity and specificity of the system are 80.6% and 96.2%. The numbers are in episodes. The band-pass scenario has higher sensitivity and specificity than the current screening pathway, with 687,105 savings of human readings.

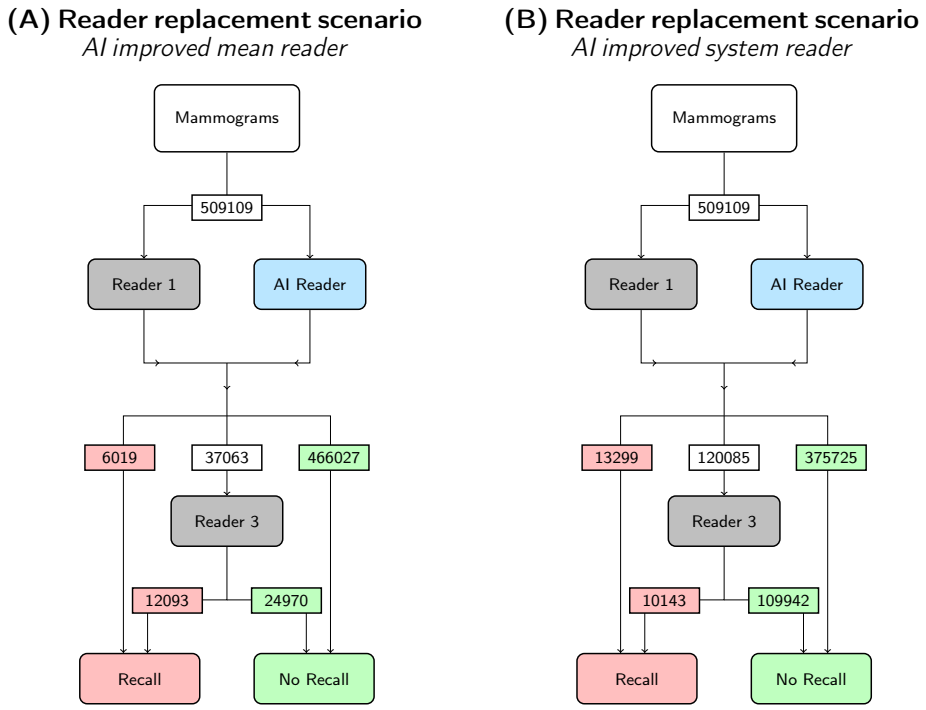

Figure 7. The reader replacement AI-integrated screening scenarios. (A) (A) Reader replacement AI-integrated screening scenario with the AI improved mean reader. (B) Reader replacement AI-integrated screening scenario with the AI improved system reader. The numbers are in episodes. The AI reader replaces one of the first two readers (randomly). Both readers see the same episode and opt to recall or not-recall, if they disagree Reader 3 arbitrates. The sensitivity and specificity of the system is 77.4% and 97.1% for the AI improved mean reader and 82.6% and 96.1% for the AI improved system reader respectively. The AI improved mean reader has lower sensitivity and higher specificity than the current screening pathway, with 509,109 savings of human readings at the beginning and 3,749 extra human readings at the Reader 3 position. The AI improved system reader has both higher sensitivity and specificity than the current screening pathway, also with 509,109 savings of human readings upfront but with 86,771 extra human readings at the Reader 3 position.

Variable	Current screening pathway	AI reader replacement (improved mean)	AI reader replacement (improved system)	AI band-pass screening
Sensitivity (%)	79.9%	77.4% (76.9%, 77.9%)	**82.6% (82.0%, 83.1%)**	80.6%
Specificity (%)	96.0%	**97.1% (97.1%, 97.1%)**	96.1% (96.1%, 96.1%)	96.2%
No. of episodes	509109	509109	509109	509109
True positive	3707	3593 (-3.1%)	**3833 (+3.4%)**	3741 (+0.9%)
True negative	484413	**489948 (+1.1%)**	484858 (+0.1%)	485548 (+0.2%)
False positive	20054	**14519 (-27.6%)**	19609 (-2.2%)	18919 (-5.7%)
False negative	935	1049 (+12.2%)	**809 (-13.5%)**	901 (-3.6%)
Workload				
Assessments	23761	**18112 (-23.8%)**	23442 (-1.3%)	22660 (-4.6%)
Human readings	1051532	546172 (-48.1%)	629194 (-40.2%)	**364427 (-65.3%)**
*Third reads	33314	37063 (+11.3%)	120085 (+260.5%)	**16951 (-49.1%)**
Reading and assessment cost	$42,174,722	**$29,788,970 (-29.4%)**	$38,501,280 (-8.7%)	$33,897,520 (-19.6%)

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

Table 4. Comparison of the current screening pathway 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 (the standard deviations are presented for the sensitivity and specificity). The AI band-pass scenario does not require simulating the third reader, and the result is entirely based on the real data. Percentages in the brackets with the plus / minus sign indicate the percentages change compared with the current screening pathway.

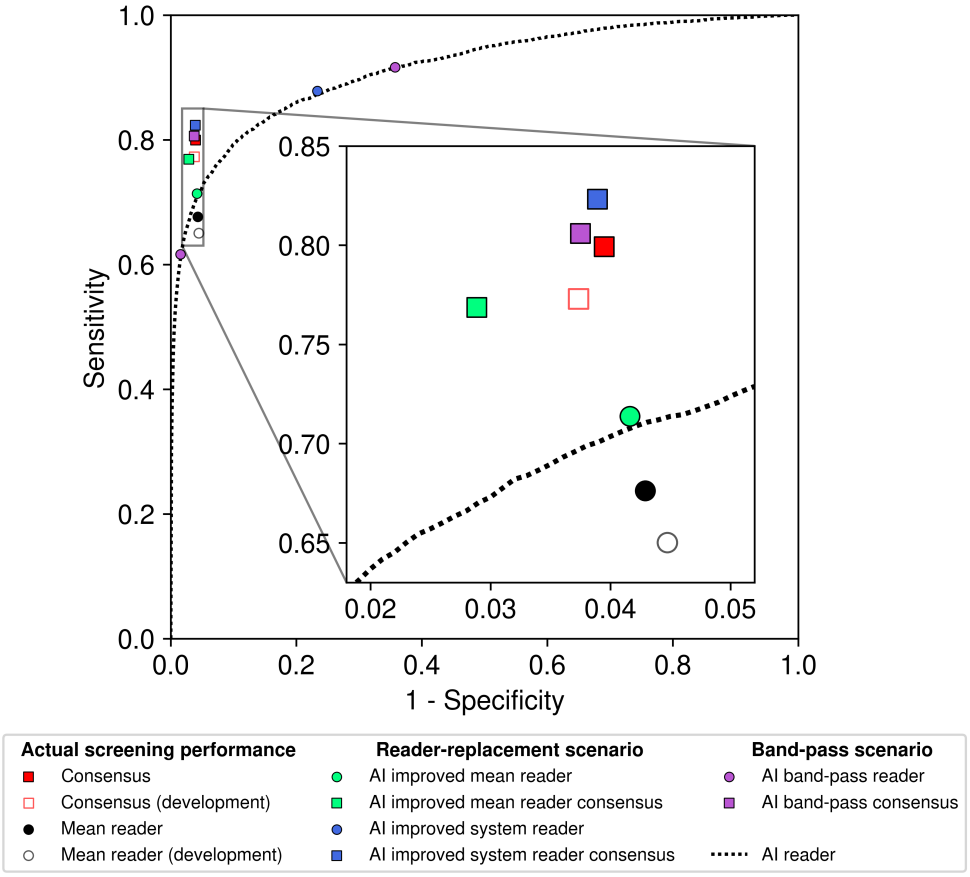

Figure 8. Screening performance for AI integrated scenario thresholds and consensus outcomes. The circles represent the human and AI reader operating points. The squares represent the current consensus performance and the system consensus performance of the AI-integrated scenarios. The mean reader is the weighted average (in sensitivity and specificity) of all the individual radiologists. The AI improved mean reader and consensus (green) are the operating point and resulting system performance of the AI reader replacement scenario when the operating point was set to improve on the mean reader. The AI improved system reader and consensus (blue) are the operating point and resulting system performance of the AI reader-replacement scenario when the operating point was set to improve the system performance. The AI band-pass recall, no-recall and consensus (purple) are the operating points and resulting threshold when the operating point thresholds were set to improve the system performance. The human mean reader and consensus are shown for both the development and test sets. Each operating point consists of a threshold for each manufacturer; the AI reader ROC curve (black) summarises the AI reader performance as a single reader across all possible thresholds.

Discussion

We demonstrated that our AI reader was a very strong individual reader on a large real-world, retrospective cohort of breast cancer screening episodes. The BRAIx AI reader with the improved mean reader threshold had better sensitivity (+3.77%) and specificity (+0.13%) than the mean radiologist reader on the testing dataset. When compared individually against radiologists the AI reader had a higher sensitivity and specificity than 22 of 83 readers and lower sensitivity and specificity than only four.

We also examined generalisability of the AI reader across a range of cohorts. Significant areas of deviation were observed in specificity related to first round screening clients and aging compared to current screening performance. 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 better specificity than individual readers and the consensus system for first round episodes but decreased in performance in subsequent rounds whereas the radiologists and the system improved their specificity in subsequent rounds. We hypothesise that radiologists take good advantage of the opportunity to inspect previous screening rounds for a client, whereas the AI reader in its current form is limited by looking at each episode for a client in isolation. Thus an opportunity for further improvement for the AI reader is to develop it such that it can take into account previous episodes for a client when calculating an episode score.

The AI reader showed higher sensitivity on medium and high grade invasive cancers and lower sensitivity on low and medium grade in situ compared with human readers. This result indicates that 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].

Despite the strong performance of our AI reader it remains clear from our study and previously reported results that AI readers as single readers cannot match the performance of the multiple reader consensus models used in Australia and other countries with national, population breast cancer screening programs. The first and second readers, whose function our AI reader was designed to emulate, are required to operate independently at a high specificity (96%). In contrast, the third readers must be highly sensitive, can view the decisions of the first and second readers, and the cohort of episodes assessed by the third reader contains a high proportion of cancers. In the Australian screening program those in the third reader pool are high performing and typically more experienced readers that also perform many of the first and second reads, but never more than one read per episode. The performance gap necessitates some human-AI collaboration to improve screening outcomes.

We used detailed simulations to study three AI-integrated scenarios for the screening pathway: reader replacement with an AI improved mean reader, reader replacement with an AI improved system reader and AI band-pass consensus.

In the reader replacement scenario with AI improved mean reader we observed the potentially counter-intuitive result that, when integrated into the screening pathway, an AI reader that outperforms the mean radiologist in terms of both sensitivity and specificity does not improve the overall pathway performance. While the specificity improves from 96.0% to 97.1%, the sensitivity drops from 79.9% to 77.4%. To understand this observation, we refer to the breakdown of the sensitivity and specificity contribution by the label outcomes (Table 5, 6). With the replacement of a human reader by the AI improved mean reader, the numbers of screen-detected cancers identified by the first reader and the AI reader actually improves from 2,488 to 2,660 compared with the original first two readers. However, many of these screen-detected cancers were identified by the third reader in the original pathway, so overall, there was no improvement. It merely shifted the decisions back in time. In fact, since the sensitivity of the first reader and the AI reader was inferior to that of the third reader, the shift of these episodes incurred additional missed screen-detected cancers, and this deficiency reduced the overall sensitivity.

In terms of interval cancers, the first reader and the AI reader made a marginal improvement

of 42 - 26 = 16 additional detection, and the shift of the screen-detected cancers allowed the third reader to see more interval cancers, also resulting in more detection of interval cancers. However, aggregating the gains and losses, the overall sensitivity did not improve. The reason behind this result is that the AI reader with the improved mean reader threshold had high specificity and recommended recall insufficiently frequently, such that some screen-detected cancers ended up in immediate no-recall. On the flip side, the high specificity results in not recalling many of the benign and NSA episodes, from 33 to 2,726 and 50 to 3,551 respectively, explained the boost in the overall specificity.

For the reader replacement scenario with the AI improved system reader the AI reader operated at a high sensitivity and thus the system sensitivity improved from 79.9% to 82.6% and the specificity improved from 96.0% to 96.1%. The differences from the AI improved mean reader case were that the misses of the screen-detected cancers (by both the first reader and the AI reader as no-recall-and-no-recall) in the AI improved mean reader case were now corrected (as no-recall-and-recall) and sent to the third reader for arbitration. In exchange, the high sensitivity also moved some of the correctly identified screen-detected cancers by the third reader back to the first reader and the AI reader (by turning recall-and-no-recall decisions in the AI improved mean reader case into recall-and-recall decisions in the AI improved system reader case). This subtle exchange of episodes lead to an "improved" detection of screen-detected cancer by the first reader and the AI reader from 2,488 to 2,963 without incurring many misses overall.

With the small loss in screen-detected cancers and the marginal gain of 68-26=42 detection of interval cancers, the real gain of the optimised AI-integrated scenario (using the AI improved system reader) came from the third reader having more access to the interval cancers as the AI reader had been highly sensitive and missed fewer interval cancers. Under the assumption that the simulated third reader performs at the same level as the retrospective performance of the third human reader (at about 30% success rate of identifying interval cancer), there would be a gain of 170-34=136 additional detections of interval cancer. These detected interval cancers translated into the gain in sensitivity.

The higher sensitivity of the optimised AI reader came at a trade-off of lower specificity. Therefore, the numbers of benign and NSA identified reduced, and there was a significant increase of disagreement with the first reader, causing a large increase of third readings. Yet, the number of identified benign and NSA episodes remained sizeable and the savings of reads by the AI reader was so significant that the overall specificity and cost were still better than the original screening pathway. In summary, if an AI reader is as good as a competent human reader but cheaper, then there is scope for human readers to (co)operate productively with the AI reader within the consensus screening model by allowing for higher-recall operating points for the AI reader.

Finally, for the band-pass AI-integrated scenario, we saw from the simulations that it can be an attractive alternative to the replacement scenario, as it also achieved better sensitivity (from 79.9% to 80.6%) and specificity (from 96.0% to 96.2%) than the original screening pathway, with twice the cost savings of the reader replacement with AI improved system reader scenario. Its key drawback, however, may be that in the band-pass scenario the AI reader can decide on episode outcomes without any human intervention (335,371 episodes = 332,122 non-cancer episodes + 3,249 cancer episodes). In contrast, in the reader replacement scenarios, humans are in charge with AI-in-the-loop. Retaining human oversight in the assessment of all images from all episodes may be necessary in clinical implementation of AI readers to allay consumer concerns about giving AI too much decision-making autonomy. Keeping a human in charge throughout the screening pathway also mitigates the operational risk that the AI reader's performance drifts or deteriorates over time or it malfunctions, which may be impossible to discover until a later time.

The design of the consensus reader systems dictates that in the reader replacement scenario, even if an AI reader were to be perfect, it does not necessarily lead to system-wide improvement, because its decisions need to be combined with the other human reader for a collective decision. This effect leads to the important point that for AI-integrated scenarios, instead of optimising the

AI reader to outperform human readers as an individual reader, we should optimise the human-AI collaboration.

As every episode needs to be decided upon, optimising human-AI collaboration is about letting each reader do what they do best. The AI reader threshold can be set to optimise the system performance by either:

- establishing higher sensitivity than the human reader to improve cancer detection, or
- establishing higher specificity than the human reader to reduce false recalls to assessment.

In our reader replacement scenario the optimised AI reader threshold established higher sensitivity accepting some greater level of "third reads" which can be accommodated because of the saving of human reading without a second reader. This approach enabled fewer missed cancers with a lower level of human reading workload. In our AI-integrated band-pass scenario, we were able to establish a high sensitivity threshold to autonomously recall, and a low specificity threshold to autonomously clear, resulting in a model that was able to improve the sensitivity and specificity, with lower false recall and significant reading savings because of a proportion of autonomous reading.

As a curious remark, if the AI reader achieved 100% sensitivity and specificity, there would be no gain from using a multi-reader consensus system. In fact, there would be harm since human readers do make mistakes. Somewhere between the current AI performance and the perfect AI performance, there is a point beyond which it would be better to replace the two-reader consensus system (with AI replacing the second reader) entirely with the AI reader alone. Under some mathematical assumptions, the "super-human" performance point works out to be a true positive rate of 0.972 and true negative rate of 0.967. This is the (TPR, TNR) point beyond which including human readers in the system would be unfavourable to the clinical outcomes. Interested readers may find the full details of the mathematical model of the three-reader system in Supplementary Methods.

In considering the limitations of our study we emphasise that the reader replacement result relies on the assumption that the simulated third reader continues to perform at the level observed in our retrospective dataset when the human reader is replaced by the AI reader. The retrospective performance, that is the rate of correct identification, of the human reader at the third position is 99.93% for normal episodes, 100% for screen-detected cancer, 0% for interval cancer, and 0% for both benign and NSA episodes. These aggregate to 95.9% accuracy for cancers and 55.4% for non-cancers. An appropriately-controlled prospective study would be required to examine that the third reader performance is maintained at the current level in an implemented AI-integrated screening program.

Another limitation of this study is that here, as in any retrospective study, we can study how AI performs on human reads, but not the effects of AI on human performance. Prospective studies are required to examine the impact of integration of an AI reader on human behaviour.

Our desire to have a large, representative two-year testing set while retaining as much training data as possible necessitated that some clients have earlier screening rounds appearing in the training set and subsequent screening rounds in the testing set. Potential bias from this characteristic of the dataset also poses a challenge for human readers. Our strong external validation results and previous internal testing did not indicate any significant issues. Nevertheless, further study is warranted to assess any potential impacts as repeated screening episodes from the same client are a feature of any population screening dataset.

To conclude, we have presented a new AI reader for the detection of breast cancer from mammograms in population screening settings. Our AI reader was trained on a large cohort of screening episodes from Victoria, Australia. The AI reader showed strong performance on our testing set of 509,109 sequential episodes from 2017 and 2018, comparing favourably to the mean and individual radiologist performance. Our AI reader generalised well across different groupings of the testing data and achieved state of the art performance on external benchmarking datasets. We conducted retrospective simulation studies to investigate human-AI collaboration in AI-integrated

screening pathways. Our results showed that human-AI collaboration in a population breast cancer screening program, both with reader-replacement and band-pass AI integration, could achieve improved sensitivity and specificity than current consensus reader programs, while lowering human reader workload and reducing overall costs.

Contributors

Conceptualization: HMLF, DJM, 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 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 Vincent's 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.

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 [16].

Acknowledgments

The authors would like to acknowledge Rita Butera, Luke Neill and Georgina Marr from BreastScreen Victoria and 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 Vincent's 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 Vincent's 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. This preprint was created using the LaPreprint template (https://github.com/roaldarbol/lapreprint) by Mikkel Roald-Arbøl.

Current screening pathway	Decided by reader 1 & 2			Decided by reader 3		
	Screen-detected cancer		Interval cancer	Screen-detected cancer		Interval cancer
Classified correctly	2488		26	1159		34
Total	2492		875	1159		116
	Decided by reader 1 & 2			Decided by reader 3		
	Normal	Benign	NSA	Normal	Benign	NSA
Classified correctly	466722	33	50	17608	0	0
Total	466947	3557	1924	17619	6793	7627

Sensitivity = $79.9\% = ((2488 + 26) + (1159 + 34)) / ((2492 + 875) + (1159 + 116))$

Specificity = $96.0\% = ((466722 + 33 + 50) + (17608 + 0 + 0)) / ((466947 + 3557 + 1924) + (17619 + 6793 + 7627))$

Reader replacement (AI improved mean reader)	Decided by reader 1 & AI reader			Decided by simulated reader 3		
	Screen-detected cancer		Interval cancer	Screen-detected cancer		Interval cancer
Classified correctly	**2660**		42	814		79
Total	**2838**		743	814		249
	Decided by reader 1 & AI reader			Decided by simulated reader 3		
	Normal	Benign	NSA	Normal	Benign	NSA
Classified correctly	458871	**2726**	**3551**	24800	0	0
Total	459749	4548	4169	24817	5801	5382

Sensitivity = $77.4\% = ((2660 + 42) + (814 + 79)) / ((2838 + 743) + (814 + 249))$

Specificity = $97.1\% = ((458871 + 2726 + 3551) + (24800 + 0 + 0)) / ((459749 + 4548 + 4169) + (24817 + 5801 + 5382))$

Reader replacement (AI improved system reader)	Decided by reader 1 & AI reader			Decided by simulated reader 3		
	Screen-detected cancer		Interval cancer	Screen-detected cancer		Interval cancer
Classified correctly	2963		**68**	634		**170**
Total	3017		456	634		535
	Decided by reader 1 & AI reader			Decided by simulated reader 3		
	Normal	Benign	NSA	Normal	Benign	NSA
Classified correctly	371421	**1478**	**2382**	109577	0	0
Total	374914	5822	4815	109652	4528	4736

Sensitivity = $82.6\% = ((2963 + 68) + (634 + 170)) / ((3018 + 456) + (634 + 535))$

Specificity = $96.1\% = ((371421 + 1478 + 2382) + (109577 + 0 + 0)) / ((374914 + 5822 + 4815) + (109652 + 4528 + 4736))$

Table 5. Sensitivity and specificity breakdown of the current screening pathway and the reader-replacement scenario with the AI improved mean reader and the AI improved system reader. Numbers are in episodes.

Band-pass (AI band-pass reader)	Decided by AI reader		Decided by reader 1 & 2		Decided by reader 3	
	Screen-detected cancer	Interval cancer	Screen-detected cancer	Interval cancer	Screen-detected cancer	Interval cancer
Classified correctly	2713	147	380	13	468	20
Total	2800	449	383	473	468	69

	Decided by AI reader			Decided by reader 1 & 2			Decided by reader 3		
	Normal	Benign	NSA	Normal	Benign	NSA	Normal	Benign	NSA
Classified correctly	317116	2868	4335	152612	20	26	8571	0	0
Total	323268	4196	4658	152722	2113	1096	8576	4041	3797

Sensitivity $= 80.6\% = ((2713 + 147) + (380 + 13) + (468 + 20)) \,/\, ((2800 + 449) + (383 + 473) + (468 + 69))$

Specificity $= 96.2\% = \dfrac{((317116 + 2868 + 4335) + (152612 + 20 + 26) + (8571 + 0 + 0))}{((323268 + 4196 + 4658) + (152722 + 2113 + 1096) + (8576 + 4041 + 3797))}$

Table 6. Sensitivity and specificity breakdown of the band-pass scenario with the AI band-pass reader. Numbers are in episodes.

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 radiologists. *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. Frazer, H. M. *et al.* ADMANI - annotated digital mammograms and associated non-image datasets. *In press* (2022).
17. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **115,** 211–252 (2015).
18. Wightman, R. *PyTorch Image Models* https://github.com/rwightman/pytorch-image-models. 2019.
19. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.
20. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. *Densely Connected Convolutional Networks* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

21. Wang, Q. *et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks* 2020. arXiv: 1910.03151 [cs.CV].

22. Tan, M. & Le, Q. *Efficientnet: Rethinking model scaling for convolutional neural networks* in *Proceedings of the 36th International Conference on Machine Learning* (2019), 6105–6114.

23. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. *Inception-v4, inception-resnet and the impact of residual connections on learning* in *Thirty-first AAAI conference on artificial intelligence* (2017).

24. Chollet, F. *Xception: Deep Learning With Depthwise Separable Convolutions* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

25. Chen, Y. *et al. Multi-view Local Co-occurrence and Global Consistency Learning Improve Mammogram Classification Generalisation* in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (Springer Nature Switzerland, Cham, 2022), 3–13. ISBN: 978-3-031-16437-8.

26. Wang, C. *et al. Knowledge Distillation to Ensemble Global and Interpretable Prototype-Based Mammogram Classification Models* in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (Springer Nature Switzerland, Cham, 2022), 14–24. ISBN: 978-3-031-16437-8.

27. Chen, Y. *et al. Fine-Tuning ResNet for Breast Cancer Classification from Mammography* in *Proceedings of the 2nd International Conference on Healthcare Science and Engineering* (Springer Singapore, Singapore, 2019), 83–96.

28. Al-antari, M. A., Han, S.-M. & Kim, T.-S. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. *Computer Methods and Programs in Biomedicine* **196,** 105584. ISSN: 0169-2607. https://www.sciencedirect.com/science/article/pii/S0169260720314176 (2020).

29. Hou, X., Bai, Y., Xie, Y. & Li, Y. Mass segmentation for whole mammograms via attentive multi-task learning framework. *Physics in Medicine and Biology* **66,** 105015 (2021).

30. Lou, Q., Li, Y., Qian, Y., Lu, F. & Ma, J. Mammogram classification based on a novel convolutional neural network with efficient channel attention. *Computers in Biology and Medicine* **150,** 106082. ISSN: 0010-4825. https://www.sciencedirect.com/science/article/pii/S0010482522007909 (2022).

31. Geras, K. J. *et al.* High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. *arXiv e-prints,* arXiv:1703.07047. arXiv: 1703.07047 (2017).

32. Uehara, K., Murakawa, M., Nosato, H. & Sakanashi, H. *Prototype-Based Interpretation of Pathological Image Analysis by Convolutional Neural Networks* in *Pattern Recognition* (Springer International Publishing, Cham, 2020), 640–652. ISBN: 978-3-030-41299-9.

33. Huang, H., Wu, Z., Li, W., Huo, J. & Gao, Y. Local descriptor-based multi-prototype network for few-shot Learning. *Pattern Recognition* **116,** 107935. ISSN: 0031-3203. https://www.sciencedirect.com/science/article/pii/S0031320321001229 (2021).

34. Paszke, A. *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library* in *Advances in Neural Information Processing Systems* (eds Wallach, H. *et al.*) **32** (Curran Associates, Inc., 2019).

35. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

36. Sashank, J. R., Satyen, K. & Sanjiv, K. *On the convergence of adam and beyond* in *International Conference on Learning Representations* **5** (2018), 7.

37. Oyelade, O. N. & Ezugwu, A. E. A deep learning model using data augmentation for detection of architectural distortion in whole and patches of images. *Biomedical Signal Processing and Control* **65,** 102366. ISSN: 1746-8094. https://www.sciencedirect.com/science/article/pii/S1746809420304730 (2021).

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

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

40. 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).

41. 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.

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