BMJ Open

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

M Luke Marinovich ,[1,2] Elizabeth Wylie,[3] William Lotter,[4] Alison Pearce ,[2] Stacy M Carter ,[5] Helen Lund,[3] Andrew Waddell,[3] Jiye G Kim,[4] Gavin F Pereira,[1,6] Christoph I Lee,[7] Sophia Zackrisson,[8] Meagan Brennan,[2] Nehmat Houssami[2,9]

To cite: Marinovich ML, Wylie E, Lotter W, *et al*. Artificial intelligence (AI) to enhance breast cancer screening: protocol for population-based cohort study of cancer detection. *BMJ Open* 2022;**12**:e054005. doi:10.1136/bmjopen-2021-054005

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-054005).

Received 31 May 2021
Accepted 01 December 2021

Check for updates

For numbered affiliations see end of article.

Correspondence to
Dr M Luke Marinovich;
luke.marinovich@curtin.edu.au

ABSTRACT

Introduction Artificial intelligence (AI) algorithms for interpreting mammograms have the potential to improve the effectiveness of population breast cancer screening programmes if they can detect cancers, including interval cancers, without contributing substantially to overdiagnosis. Studies suggesting that AI has comparable or greater accuracy than radiologists commonly employ 'enriched' datasets in which cancer prevalence is higher than in population screening. Routine screening outcome metrics (cancer detection and recall rates) cannot be estimated from these datasets, and accuracy estimates may be subject to spectrum bias which limits generalisabilty to real-world screening. We aim to address these limitations by comparing the accuracy of AI and radiologists in a cohort of consecutive of women attending a real-world population breast cancer screening programme.

Methods and analysis A retrospective, consecutive cohort of digital mammography screens from 109 000 distinct women was assembled from BreastScreen WA (BSWA), Western Australia's biennial population screening programme, from November 2016 to December 2017. The cohort includes 761 screen-detected and 235 interval cancers. Descriptive characteristics and results of radiologist double-reading will be extracted from BSWA outcomes data collection. Mammograms will be reinterpreted by a commercial AI algorithm (DeepHealth). AI accuracy will be compared with that of radiologist single-reading based on the difference in the area under the receiver operating characteristic curve. Cancer detection and recall rates for combined AI–radiologist reading will be estimated by pairing the first radiologist read per screen with the AI algorithm, and compared with estimates for radiologist double-reading.

Ethics and dissemination This study has ethical approval from the Women and Newborn Health Service Ethics Committee (EC00350) and the Curtin University Human Research Ethics Committee (HRE2020-0316). Findings will be published in peer-reviewed journals and presented at national and international conferences. Results will also be disseminated to stakeholders in Australian breast cancer screening programmes and policy makers in population screening.

Strengths and limitations of this study

- With data from over 100 000 distinct, consecutive screening examinations, and including interval cancers, this will be the largest study to date to investigate the accuracy of an artificial intelligence (AI) algorithm for interpreting digital mammograms in a population breast cancer screening programme.
- The consecutive cohort will overcome limitations of previous studies that have used 'cancer enriched' datasets, resulting in accuracy estimates that will be generalisable to screening programmes, thus enabling the estimation of population-based screening outcome metrics.
- The retrospective design requires simulation of the integration of AI into double-reading by analytically pairing AI with a human reader, which may differ from integrated AI–human reading strategies in practice.
- Societal and ethical issues along with the economic implications of AI are beyond the scope of this study protocol, but are being investigated in adjunct projects.

INTRODUCTION

Healthcare systems in developed countries have implemented population breast cancer screening for several decades. This is based on evidence from randomised trials that mammography reduces breast cancer-specific mortality,[1] complemented by observational evidence of benefit from real-world screening.[2] Breast cancer screening involves interpretation of digital mammograms to identify suspicious abnormalities that warrant further investigation ('recall to assessment'), and is a subjective process that can detect cancer, yield false-positive results or miss a cancer because the cancer is not visible to the radiologist. Cancers that are not detected at the screening examination often present

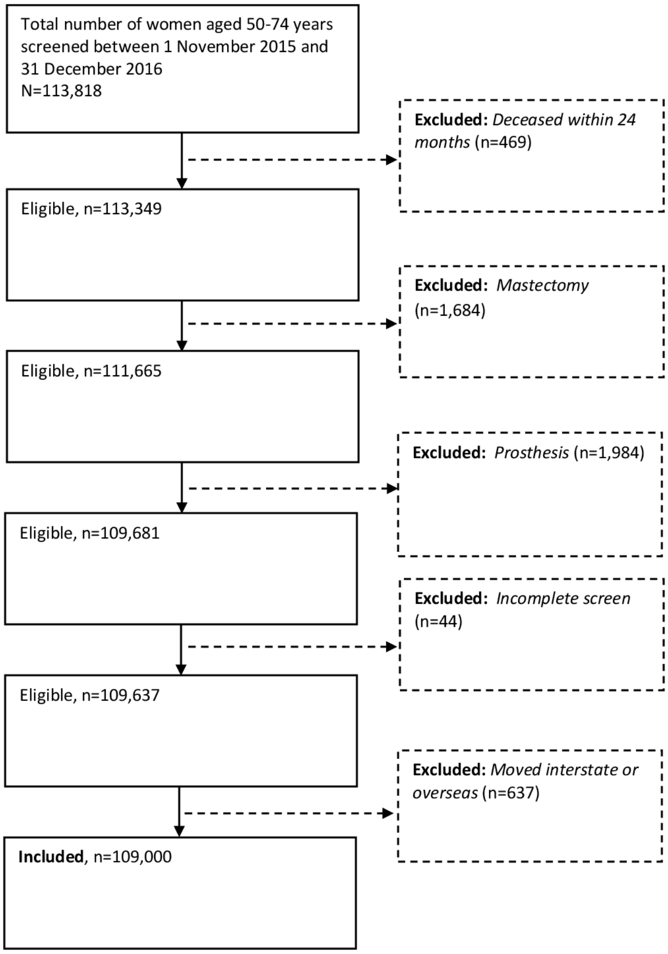
Total number of women aged 50-74 years screened between 1 November 2015 and 31 December 2016
N=113,818
Excluded: *Deceased within 24 months* (n=469)
Eligible, n=113,349
Excluded: *Mastectomy* (n=1,684)
Eligible, n=111,665
Excluded: *Prosthesis* (n=1,984)
Eligible, n=109,681
Excluded: *Incomplete screen* (n=44)
Eligible, n=109,637
Excluded: *Moved interstate or overseas* (n=637)
Included, n=109,000

Figure 1 Flowchart of cohort inclusions and exclusions.

symptomatically in the interval between screening rounds and are known as 'interval cancers'.[3] Interval cancers are more often fast-growing and aggressive compared with screen-detected cancer,[4] and interval cancer rates are routinely monitored by screening programmes as an indicator of screening effectiveness.[5] Population-based breast cancer screening programmes in Australia (BreastScreen), Europe and the UK use 'double-reading', implemented as independent screen-readings by two radiologists (with arbitration for discordance) to reduce screen-reading error. There is, however, variability in the accuracy of screening between radiologists and across screening programmes.[6]

Internationally, there is increasing concern about the ongoing viability of population breast screening programmes due to what has been termed 'a global radiology workforce crisis'.[7] As in the UK and Europe, resourcing screen-reads in Australia is increasingly difficult for publicly funded screening programmes, where reader shortages exist in some locations.[8] The Royal Australian and New Zealand College of Radiologists' Workforce Survey Report identifies screening mammography as an area of practice 'at significant risk of workforce shortage', with this deficit predicted to increase over time.[9] Simultaneously, screening volumes are increasing, corresponding to an ageing population, coupled with recent policy and funding decisions to increase the target age range for breast cancer screening in Australia from 50–69 years to 50–74 years.[5] Artificial intelligence (AI) has the potential to address these resource challenges by making screen-reading more efficient and accurate. AI may particularly improve screening effectiveness if it can detect some interval cancers (cancers missed at screening) without substantially contributing to overdiagnosis (detection of cancers that would not otherwise become clinically apparent).[3]

Deep learning, a rapidly growing field of AI that integrates computer science and statistics, allows computers to learn directly through automatic extraction and analysis of complex data. An AI algorithm can be trained to detect breast cancer given mammography examinations with known outcomes. In doing so, the AI algorithm learns to identify automatically extracted quantitative variables ('features') that are predictive of cancer presence. In this respect, deep learning is a significant advance over earlier computer-aided detection systems that relied on limited sets of human-extracted features, and resulted in unacceptably high false-positive rates.[7]

Studies that have evaluated AI for breast cancer screening suggest the technology can achieve accuracy that is comparable to expert radiologists.[6 10–12] However, such studies commonly employ 'enriched' datasets in which the prevalence of cancer is substantially higher than in population screening (up to 55%, compared with real-world screening populations where breast cancer prevalence is less than 1%).[13] Selected datasets enriched with cancers are likely to be unrepresentative of disease spectrum in screening populations, and may lead to estimates of accuracy for both AI and radiologists that are not generalisable to real-world screening.[13–15] Furthermore, routine screening metrics (cancer detection rate (CDR) and recall rate) cannot be accurately estimated from these datasets. There is therefore a need to generate evidence of AI performance that is generalisable to routine screening practice to inform decisions about adopting the technology.[13 16]

Study aims and hypotheses

This project aims to compare AI reading of digital mammograms with human reading in a real world, population breast cancer screening setting. We hypothesise that the AI algorithm has accuracy that is comparable to human readers, and that integrating the AI into a standard screen-reading strategy will accurately detect cancers including interval cancers. Specifically, we aim to:

1. Compare the accuracy of AI with the average accuracy of single human reading in terms of the area under the receiver operating characteristic curve (AUC-ROC).
2. Compare integrated AI–human screen-reading with human double-reading (standard breast cancer screen-reading practice) in terms of CDR (number of cancers detected per 1000 screens) and case-specific recall rate (percentage of women recalled to further assessment).

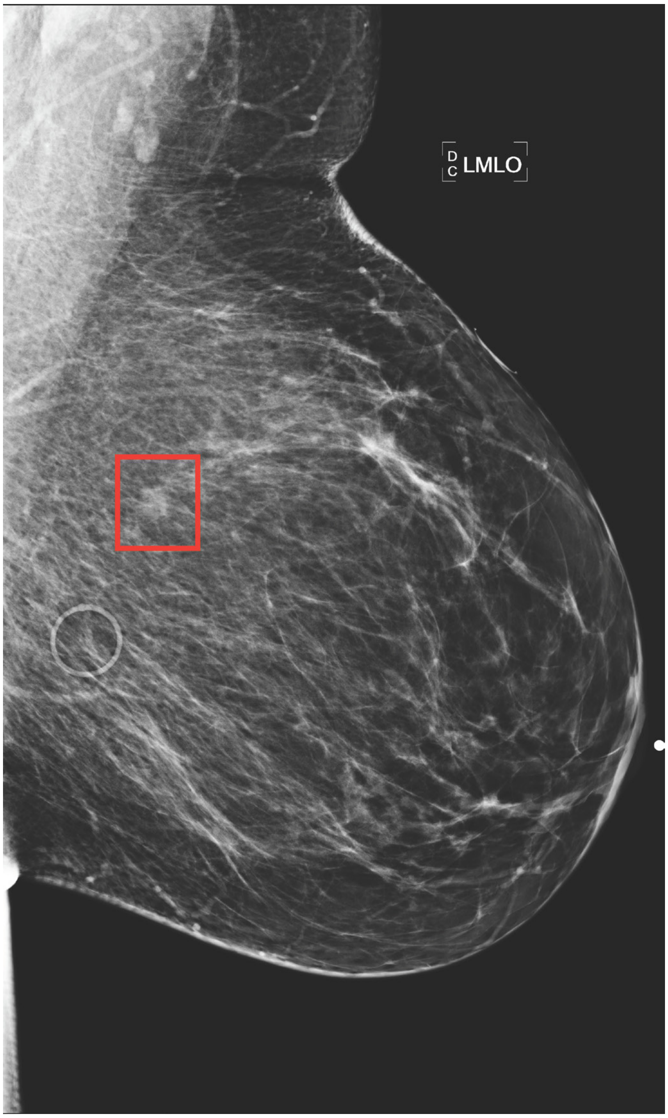

Figure 2 Digital mammogram mediolateral oblique view with region of interest (denoted by bounding box) identified by the AI algorithm as suspicious for malignancy. Cancer was confirmed as invasive ductal carcinoma. AI, artificial intelligence.

METHODS AND ANALYSIS

Study design and inclusion criteria

A retrospective study design was used to assemble a contemporary cohort of unique, consecutive digital mammography screens from BreastScreen WA (BSWA), the population breast cancer screening programme in Western Australia (WA). The study will avoid biases identified in previous research on AI for mammography screening[13] by using consecutive screens (ie, all screening examinations meeting the inclusion criteria in a defined time interval) representative of real-world screening populations, with ascertained outcomes including interval cancers. Consecutive women attending screening at BSWA and fulfilling the following criteria were included in the cohort:

1. Screened between 1 November 2015 and 31 December 2016.
2. Age 50–74 years (the target age range for biennial breast cancer screening in Australia[5]).
3. For women with multiple screening examinations in this time period, only the last will be included.

In order to ensure a minimum follow-up period of 24 months for ascertainment of interval cancers, and adequacy and completeness of screening examinations for reinterpretation by the AI algorithm, the following exclusion criteria were applied:

1. Deaths within 24 months.
2. Out-of-state relocations.
3. Women who have had a previous mastectomy (and therefore cannot contribute bilateral images for reinterpretation by AI).
4. Women with implants (self-reported or radiologist-identified).
5. Incomplete screens (eg, due to physical limitation, fainting or distress, where the screening episode is unable to be completed at a later time).

Study cohort characteristics

A total of 113 818 unique, consecutive screening examinations were identified during the study period. After applying the exclusion criteria, 109 000 screening examinations (95.8%) were eligible for inclusion in the cohort (figure 1). The mean age of the cohort is 61.0 years (SD 6.9 years; range 50–74 years). There were 9076 baseline (first ever) screens (8.3%); the remainder were subsequent screens. A total of 13 954 women (12.8%) were offered annual screening due to a previous history of breast (n=3354) and/or ovarian cancer (n=631); and/or a previous diagnosis of 'benign high risk' disease (n=382) defined as atypical ductal or lobular hyperplasia or lobular carcinoma in situ; and/or a significant family history (n=10 197) defined by BSWA as two or more first-degree relatives with breast cancer, or at least one first-degree relative with breast cancer occurring at <50 years or with bilateral breast cancer.

Measurement

BSWA routinely collects demographic characteristics and risk factors through a self-administered registration form. Details of the screening examination and further assessment are also routinely recorded in the Mammographic Screening Registry. Descriptive variables (age; screening round; time since last screen for repeat screens; mammographic breast density; personal history of breast cancer; first-degree family history of breast cancer; personal history of ovarian cancer; hormone replacement therapy in the past 6 months; a history of removal or biopsy of benign lump and self-reported breast symptoms) will be used to characterise the cohort. Breast density (defined as heterogeneously or extremely dense breasts identified by at least one of two radiologists) is recorded by BSWA only for women with no abnormality identified (ie, women who are not recalled for further testing). A deidentified screen episode ID will be used to link these data to output of the AI algorithm

Table 1 Significant gaps in knowledge needed to develop prospective real-world screening trials or evaluation (adapted from Houssami *et al*[13])

Knowledge gap or limitations of published studies	Addressed by this study?	Description of how addressed in our study
Few studies use commercially available AI systems.	Partly	The AI algorithm used in this study[10] underlies a triage product that is FDA-approved and commercially available in the USA.
Studies have used relatively small datasets, often consisting of mammograms from several hundred women (rarely several thousand). Larger validation datasets are required.	Yes	A large validation dataset including 109 000 women will be used.
The same or selected subsets of the same datasets were used to train and validate models. Validation using independent, external datasets is required.	Yes	The study dataset is external to and independent from the datasets used to train the algorithm.
Datasets were commonly enriched with malignant lesions, with studies often selecting images containing suspicious abnormalities. Studies are required in unselected screening populations.	Yes	The study dataset is a consecutive, unselected population drawn from a real world, biennial population-based breast screening programme (BreastScreen WA). The dataset is not enriched with cancers. The prevalence and disease spectrum of screen-detected and interval cancers are representative of population breast screening.
There is a paucity of studies reporting conventional screening metrics (CDR and recall rate).	Yes	The inclusion of unique, consecutive screening episodes will allow estimation of CDR and recall rate (it is not possible to accurately derive these metrics from case-controlled, cancer-enriched datasets).
There is limited data on AI versus human interpretation. Future studies should compare AI to radiologists' performance or report the incremental improvement for AI algorithms in combination with radiologists.	Yes	The comparative accuracy of AI and radiologists will be estimated in terms of AUC-ROC, sensitivity and specificity. Incremental rates of cancer detection and recall will be estimated for double-reading with and without AI.
There are no studies on women's or societal perspectives on the acceptability of AI.	No	This is beyond the scope of the present study. A parallel stream of social and ethical research by some of the study investigators will explore the acceptability of AI.
Future studies should include images from digital breast tomosynthesis, given the rapid adoption of this technology.	No	This is beyond the scope of the present study. Digital breast tomosynthesis is not currently used in Australian publicly funded population breast screening programmes.

AI, artificial intelligence; AUC-ROC, area under the receiver operating characteristic curve; CDR, cancer detection rate; FDA, Food and Drug Administration.

(see the section 'Reinterpretation of mammograms by AI algorithm').

The final screening outcome (recall or not recall) will be collected, along with findings from each reader and a deidentified radiologist ID. Data on cancer diagnosis (date of diagnosis; screen-detected or interval cancer) and cancer characteristics (histological type; tumour size; grade; nodal status) will also be extracted.

Definitions of screen-detected and interval cancers

Screen-detected breast cancers are defined as either invasive cancer or ductal carcinoma in situ (DCIS) detected at the index screening episode.[17] BSWA collects details on all screening participants recalled for further testing and their subsequent cancer diagnosis. There are 761 screen-detected breast cancers in the study cohort (606 invasive, 155 DCIS; overall CDR 7.0 per 1000 screens). Interval breast cancers are defined as invasive cancers that are diagnosed after a negative index screening episode and before the next scheduled screening episode (ie, within 24 months for biennial screeners, and 12 months for the minority of women scheduled to have an annual screen).[17] Interval cancers are identified through data linkage to the WA Cancer Registry and are reported regularly to BSWA according to national quality and accreditation standards. Interval cancers also include women who present symptomatically to BSWA for early re-screening and a cancer is diagnosed in the same breast. There are 235 interval cancers in the study cohort (2.2 per 1000 screens).

Reinterpretation of mammograms by AI algorithm

The DeepHealth algorithm used in this study underlies a triage product that is Food and Drug Administration (FDA)-cleared and commercially available in the USA. Development of the algorithm has been described previously.[10] In brief, DeepHealth used a progressive, stagewise training strategy motivated by how a radiologist might learn to read an image: by first viewing cropped examples of various lesion types, benign and malignant, before learning to scan an entire screen and make a global decision on whether a suspicious lesion is present. Convolutional neural networks (a deep learning approach to analysing visual data) were trained on five datasets from the USA and UK, making use of both strongly and weakly labelled data. Australian data were not used for algorithm training; therefore, training datasets were independent of the dataset used for the current external validation study. The trained algorithm outputs a 'bounding box' identifying a region of interest (figure 2), along with a malignancy score quantifying the likelihood that the region of interest represents a malignancy. The algorithm evaluates each image in a study independently and aggregates the scores across all potential regions in the study to compute a single study-level malignancy score. The overall accuracy of the algorithm based on this study-level malignancy score has been compared with five individual radiologists, each fellowship-trained in breast imaging, on a cancer-enriched dataset, and was shown to outperform all five readers. At the average radiologist specificity, the algorithm resulted in an absolute increase in sensitivity of 14.2%; at the average radiologist sensitivity, the absolute increase in specificity was 24.0%.[10] The algorithm also outperformed radiologists in detecting malignancy in a set of prior 'normal' mammograms from the same set of cancer cases (increase in sensitivity 17.5%; increase in specificity 16.2%), demonstrating the potential to detect interval cancers 'missed' by radiologists.

All imaging analysis for the study will take place at BSWA to ensure security of images. Images will only be accessed by investigators who are employed by BSWA, and have such access under the usual conditions of their employment; these images will not be used for further refinement of DeepHealth's algorithm. A laptop with the AI algorithm installed and a graphics processing unit supporting its evaluation will be located at BSWA. An external hard drive will be attached containing the cohort of digital mammogram data (DICOM files consisting of four views per breast, two breasts per woman). The algorithm will output data to a csv file including bounding box coordinates, malignancy scores ranging from 0 to 1, and a unique identifier extracted from the DICOM header to enable woman-level matching of results to BSWA routine screening data.

Data de-identification and secure storage

De-identified data on cohort characteristics, screening findings and cancer diagnosis will be transferred by secure online file transfer to the Curtin School of Population Health, Curtin University. No paper-based or portable electronic media storage of these data will take place. Project data will be electronically stored on a secure server, which is backed up daily to prevent any unintentional data loss. The research environment includes a variety of security controls to restrict unauthorised access—these include access controls, role-based delegations, encryption, firewalls and physical access restrictions (authorised access to server rooms and research offices is restricted by key). Automatic screen locking will occur on electronic devices after 5 min of inactivity. Data will not be stored or used in public terminals.

Statistical methods

All statistical analyses will be undertaken at the School of Public Health, Curtin University. To descriptively compare the accuracy of AI with the average accuracy of single human reading, an ROC curve for the AI algorithm will first be plotted from the algorithm's study-level malignancy score and the AUC-ROC derived. The hierarchical summary ROC model proposed by Rutter and Gatsonis[18 19] will be used to model radiologist accuracy and derive an area under the summary ROC curve for radiologists (using numerical integration), along with summary estimates of sensitivity and specificity. The sensitivity and specificity of AI will be descriptively compared with that of radiologists by estimating the AI's sensitivity at the summary radiologist specificity, and the AI's specificity at the summary radiologist sensitivity. The malignancy score derived from the AI algorithm will also be dichotomised using a prospectively defined threshold selected to reflect an expected recall rate of 4% (the overall recall rate of the BSWA programme) based on DeepHealth's (non-Australian) validation data, allowing for a descriptive comparison of sensitivity and specificity at this threshold with summary radiologist estimates.

The CDR and case-specific recall rate of double-reading by radiologists (current population reading practice) will be compared with double-reading strategies integrating AI (McNemar's test), where the first radiologist read per screen will be paired analytically with AI. The following integrated AI–radiologist strategies will be used:

1. Recall to assessment based on an 'either positive' rule (ie, either AI or radiologist is positive for suspicious abnormality). This strategy will maximise CDR.[20]
2. Recall to assessment based on a 'both positive' rule (ie, both AI and radiologist are positive for suspicious abnormality). This strategy will minimise recall rate.[20]
3. Recall to assessment based on results of AI–human reading, where 'both positive' findings for AI and radiologist trigger a decision to recall, and 'either positive' findings (ie, disagreement) are arbitrated by the second radiologist read that occurred in practice. This strategy simulates current screen-reading practice.

The effect on CDR and recall rates of alternative thresholds for dichotomising the AI algorithm score will be explored in sensitivity analyses. CDR results for integrated AI–radiologist reading will be stratified by interval versus non-interval cancers to estimate the incremental CDR for interval (clinically progressive) cancers. Sensitivity analyses will also be conducted to apply a consistent 12-month follow-up period for ascertaining interval cancers.

Sample size and power calculation

Power calculations were derived for the outcome of CDR, based on the sample size and number of screen-detected and interval cancers present in the cohort. The CDR for double-reading by radiologists in the study cohort is 7.0 per 1000 screens. With a sample size of 109 000 unique screening examinations, at an alpha of 0.05 (two-sided) the study has 80% power to detect an increase in CDR to 7.5 per 1000 screens for integrated AI–radiologist reading. This assumes concordance between the reading strategies of 5.5 cancers per 1000 screens, with 1.5 cancers per 1000 detected by radiologist double-reading only (and not by integrated AI–radiologist reading) and 2.0 cancers per 1000 screens detected by integrated AI–radiologist reading only (and not by radiologist double-reading). This 1.5:2 ratio of discordant cases is derived from a UK study comparing AI with radiologist double-reading.[11]

Substudies

In addition to the primary study objectives, substudies will be undertaken to further explore differences in accuracy observed in the main analyses. These will include:

1. Description of cancers for which there are discordant results (ie, cancers detected by the AI algorithm but not by radiologists and vice versa), in terms of radiological and cancer characteristics.
2. Investigation of presumed 'false positive' AI algorithm results in terms of the presence or absence of cancer in the next screening round (when available), to explore the extent to which these may represent true early cancer detection.[11]

Patient and public involvement

The research team includes a consumer advocate who contributed to the development and refinement of the research questions and project plan, and highlighted key ethical implications from a consumer perspective that may arise from the research (eg, data security and privacy). Consumer health representatives external to the research team have been engaged to provide community perspectives on this research (eg, advice on language, including lay summaries; potential utilisation of the research findings and advocacy on behalf of consumers and the community). In addition, several of the study investigators are undertaking a concurrent, parallel stream of research (with separate protocols and ethical approval) to elicit community perspectives about the acceptability of AI and social and ethical issues around its use in breast cancer screening.

ETHICS AND DISSEMINATION

Human research ethics committee approval

This study has ethical approval from the Women and Newborn Health Service Ethics Committee (EC00350) and the Curtin University Human Research Ethics Committee (HRE2020-0316). Both committees provided a waiver of consent for this study. Participants in the BSWA programme provide written consent for their data to be used for research purposes each time they screen.

Intended publications and research dissemination

Datasets generated and/or analysed during the current study are not publicly available due to data confidentiality agreements with data custodians. Results generated by the research will be made publicly available at the summary level. Manuscripts addressing the study aims will be published in peer-reviewed journals. Results will also be presented at relevant national and international conferences. Study outcomes will also be disseminated to stakeholders in Australian breast cancer screening programmes and policy makers in population screening, to inform future evaluation and policy discussions about the potential implementation of AI.

DISCUSSION

Organised population breast screening programmes are facing growing screen-reading resource challenges, so the current global research effort aimed at developing and testing AI algorithms for interpreting screening mammograms can contribute to ensuring future sustainability of screening. Although the field is rapidly evolving, to date there has been a focus on algorithm development with relatively few studies evaluating AI in real-world breast cancer screening settings. A scoping review of the literature on AI for breast screening identified eight key deficiencies of the evidence base (table 1), and concluded that although studies indicate a potential role of AI in this clinical scenario, those evidence gaps should be addressed prior to the initiation of prospective trials and the adoption of the technology in routine practice.[13] The primary concerns raised relate to the quality of datasets used to validate AI models and the paucity of evidence comparing the accuracy of AI and radiologists, potentially affecting the applicability and robustness of AI algorithms and raising the possibility of bias. The study we present in this protocol addresses those evidence gaps by comparing the accuracy of a commercially available algorithm with that of radiologists using a large, external validation dataset representing consecutive, unselected digital mammograms from a real-world screening programme (table 1). This retrospective cohort study is therefore an essential step to build the evidence base to underpin prospective trials and inform their design, and to provide timely evidence to screening stakeholders.

Although this study will overcome most key limitations of the evidence base, there are potential limitations associated with its retrospective design. Data collected for administrative purposes may be more prone to misclassification than data collected specifically for research purposes through a prospective trial. For instance, we have excluded women from the study cohort who relocated outside WA after the index screening examination and therefore were potentially lost to follow-up. Since the date of relocation is not routinely collected, it is possible that some women with complete follow-up were excluded. Given that exclusions for relocation represented <0.6% of women during the study period (figure 1), this is unlikely to represent a significant concern. Data on outcomes (recalls, screen-detected and interval cancers) are meticulously collected according to national quality and accreditation standards and are therefore unlikely to be subject to misclassification. Furthermore, we have defined the end date for study enrolment (31 December 2016) to ensure completeness of notifications for interval cancers (while simultaneously ensuring a contemporary cohort that is representative of the current target population for breast cancer screening in Australia). Errors in the classification of outcome data are therefore considered to be rare.

To estimate CDR and recall rate for integrated AI–human reading, we will take an analytic approach to combining AI and radiologist findings. This pragmatic approach is dictated by the retrospective study design; however, it may not be representative of how AI screening results might be incorporated into practice. Our decision rules for defining recall to further assessment are among several proposed uses of AI information. Some alternative approaches (such as the use of AI to 'triage' women to double-reading if exceeding a threshold probability of malignancy[11]) may potentially be investigated analytically by our study design, but others (such as AI output used by radiologists interactively as a decision support[6]) can only be evaluated in studies using a prospective design. Furthermore, the methods adopted to derive summary ROC curves for radiologists and associated measures of accuracy are dictated by the retrospective design. Although these methods are established and appropriate for our real-world screening data,[18] they allow only for descriptive comparisons with empirical estimates for the AI algorithm.[19]

The lack of studies exploring social and ethical issues, particularly women's perspectives and preferences around AI, has been identified as a critical evidence gap (table 1). Although beyond the scope of this study, a parallel research stream using qualitative methods is being undertaken by some of the study authors to elucidate those perspectives. For instance, women will be provided with information about potential uses of AI in breast screening, and will then discuss potential implementation with a focus on what matters most to them, and how implementation should (or should not) take place. Similarly, economic modelling to estimate incremental costs and benefits from the use of AI is critical to informing policy decisions about adopting the technology. Cost-effectiveness analysis will be undertaken in a future project building on the results of this study.

AI algorithms for interpreting mammograms have the potential to improve the effectiveness of population breast cancer screening programmes if they can detect cancers, including interval cancers, without contributing substantially to overdiagnosis. This will be the largest study to date to investigate the accuracy of an AI algorithm for interpreting consecutive digital mammograms in a population-based breast cancer screening programme. The evidence generated by this study can be used to inform decisions about adopting AI for mammogram interpretation in the future, to improve accuracy, effectiveness and efficiency.

Author affiliations
[1]Curtin School of Population Health, Curtin University, Perth, Western Australia, Australia
[2]Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia
[3]BreastScreen WA, Perth, Western Australia, Australia
[4]DeepHealth Inc, Cambridge, Massachussetts, USA
[5]Australian Centre for Health Engagement, Evidence and Values (ACHEEV), School of Health and Society, University of Wollongong, Wollongong, New South Wales, Australia
[6]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Oslo, Norway
[7]Department of Radiology, University of Washington, Seattle, Washington, USA
[8]Diagnostic Radiology, Department of Translational Medicine, Lund University, Malmö, Sweden
[9]The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, New South Wales, Australia

Twitter Stacy M Carter @stacymcarter

Contributors MLM, NH, EW, HL, AW and WL conceived the idea, planned and designed the study protocol. AP, SC, MB, GFP, JGK, CL and SZ contributed to the development of the protocol, study design and methods. MLM wrote the first draft. NH, EW, HL, AW, WL, AP, SC, MB, GFP, JGK, CL and SZ critically revised the draft for important intellectual content. All authors have approved the final written manuscript.

Funding This work was supported by funding from a National Breast Cancer Foundation Investigator Initiated Research Scheme grant (IIRS-20-011 to MLM, NH, EW, WL, SC and AP). NH receives funding from the National Health & Medical Research Council (NHMRC) (Investigator/Leader Grant #1194410). SC receives funding from the NHMRC (Ideas Grant #1181960). GFP receives funding from the NHMRC (Project Grant #1 099 655 and Investigator Grant #1173991) and the Research Council of Norway through its Centres of Excellence funding scheme (#262700). CL and WL are funded in part by the National Cancer Institute (R37 CA240403).

Competing interests WL and JGK are employees of RadNet, the parent company of DeepHealth. CL reports textbook royalties from McGraw Hill, Wolters Kluwer, and Oxford University Press; and research consulting fees from GRAIL, all outside the submitted work. SZ reports speaker fees from Siemens Healthcare AG. Other authors have no competing interest to declare.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
M Luke Marinovich http://orcid.org/0000-0002-3801-8180
Alison Pearce http://orcid.org/0000-0002-5690-9542
Stacy M Carter http://orcid.org/0000-0003-2617-8694

REFERENCES

1 Marmot MG, Altman DG, Cameron DA, *et al*. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013;108:2205–40.
2 Hanley JA, Hannigan A, O'Brien KM. Mortality reductions due to mammography screening: contemporary population-based data. *PLoS One* 2017;12:e0188947.
3 Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer* 2017;3:12.
4 Baré M, Torà N, Salas D, *et al*. Mammographic and clinical characteristics of different phenotypes of screen-detected and interval breast cancers in a nationwide screening program. *Breast Cancer Res Treat* 2015;154:403–15.
5 Australian Institute of Health and Welfare. *Breastscreen Australia monitoring report 2020*. Canberra, 2020.
6 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, *et al*. Stand-Alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019;111:djy222:916–22.

7 Harvey H, Karpati E, Khara G, *et al*. The role of deep learning in breast screening. *Curr Breast Cancer Rep* 2019;11:17–22.
8 Crouch B. *Shortage of radiologists pushing out breast scan result times for patients*. The Advertiser, 2018 October 24, 2018.
9 The Royal Australian and New Zealand College of Radiologists. *2016 RANZCR clinical radiology workforce census report: Australia*. Sydney, NSW, 2018.
10 Lotter W, Diab AR, Haslam B, *et al*. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med* 2021;27:244–9.
11 McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.
12 Salim M, Wåhlin E, Dembrower K, *et al*. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol* 2020;6:1581–8.
13 Houssami N, Kirkpatrick-Jones G, Noguchi N, *et al*. Artificial intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices* 2019;16:351–62.
14 Leeflang MMG, Rutjes AWS, Reitsma JB, *et al*. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ* 2013;185:E537–44.
15 Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. *Radiology* 2019;290:272–3.
16 Lee CI, Elmore JG. Artificial intelligence for breast cancer imaging: the new frontier? *J Natl Cancer Inst* 2019;111:djy223.
17 Australian Institute of Health and Welfare. *Breastscreen Australia data dictionary version 1.2*, 2019.
18 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865–84.
19 Oakden-Rayner L, Palmer L. Docs are ROCs: a simple off-the-shelf approach for estimating average human performance in diagnostic studies. *arXiv* 2020.
20 Macaskill P, Walter SD, Irwig L, *et al*. Assessing the gain in diagnostic performance when combining two diagnostic tests. *Stat Med* 2002;21:2527–46.