

# BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

# BMJ Open

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

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-054005
Article Type:	Protocol
Date Submitted by the Author:	31-May-2021
Complete List of Authors:	Marinovich, M Luke; Curtin University, Curtin School of Population Health; The University of Sydney, Sydney School of Public Health Wylie, Elizabeth; BreastScreen WA Lotter, William; DeepHealth Inc. Pearce, Alison ; The University of Sydney, Carter, Stacy; University of Wollongong, Australian Centre for Health Engagement, Evidence and Values Lund, Helen; BreastScreen WA Waddell, Andrew; BreastScreen WA Kim, Jiye; DeepHealth Inc. Pereira, G.F; Curtin University, Curtin School of Population Health; Norwegian Institute of Public Health, Centre for Fertility and Health Lee, C; University of Washington, Department of Radiology Zackrisson, Sophia; Lund University, Diagnostic Radiology Brennan, ME; The University of Sydney, Sydney School of Public Health Houssami, Nehmat; The University of Sydney, Sydney School of Public Health; The University of Sydney, The Daffodil Centre
Keywords:	Breast tumours < ONCOLOGY, Breast imaging < RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING

SCHOLARONE™  
Manuscripts

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

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

## Author affiliations:

<sup>1</sup> Curtin School of Population Health, Curtin University, Perth, Western Australia, Australia

<sup>2</sup> Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Camperdown, New South Wales, Australia

<sup>3</sup> BreastScreen WA, Perth, Western Australia, Australia

<sup>4</sup> DeepHealth Inc., RadNet AI Solutions, Cambridge, Massachusetts, USA.

<sup>5</sup> Australian Centre for Health Engagement, Evidence and Values (ACHEEV), School of Health and Society, University of Wollongong, Wollongong, New South Wales, Australia

<sup>6</sup> Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

<sup>7</sup> Department of Radiology, University of Washington School of Medicine, Seattle, Washington, USA

<sup>8</sup> Diagnostic Radiology, Department of Translational Medicine, Lund University, Skåne University Hospital, Malmö, Sweden

<sup>9</sup> The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, New South Wales, Australia.

## Corresponding author:

Dr ML Marinovich, Research Fellow  
School of Public Health, Curtin University  
GPO Box U1987  
Perth Western Australia 6845  
Email: [Luke.Marinovich@curtin.edu.au](mailto:Luke.Marinovich@curtin.edu.au)  
Phone: +61 8 9266 4006

**Protocol version:** 1.0 (May 2021)

**Word count:** 3,528

*Protocol for study of AI to enhance breast cancer screening***ABSTRACT**

**Introduction:** Artificial intelligence (AI) algorithms for interpreting mammograms have the potential to improve the effectiveness of population breast cancer screening programs 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 from these studies may be subject to spectrum bias which limits generalisability to real-world screening. We aim to address these limitations by comparing the accuracy of AI and radiologists in a cohort of consecutive women attending a real-world population breast cancer screening program.

**Methods and Analysis:** A retrospective cohort of 109,000 unique, consecutive digital mammography screens (including 761 screen-detected and 235 interval cancers) was assembled from BreastScreen WA (BSWA), Western Australia’s biennial population screening program. Descriptive characteristics of the cohort 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 to that of radiologist single reading based on the difference in the area under the receiver operating characteristics curve (AUC). Cancer detection and recall rates for combined AI-radiologist reading will be estimated by randomly pairing one 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

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 stakeholders in Australian breast cancer screening programs and policy makers in population  
4  
5 screening.  
6  
7

8 **Keywords:** breast cancer; screening; artificial intelligence mammography; accuracy  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

**ARTICLE SUMMARY****Strengths and limitations of this study**

- With data from over 100,000 unique, consecutive screening examinations, and including interval cancers, this will be the largest study to date to investigate the accuracy of an artificial intelligence algorithm for interpreting digital mammograms in a population breast cancer screening program.
- 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 programs, 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

Health care 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 symptomatically in the interval between screening rounds and are known as “interval cancers”.(3) Interval cancers are more often fast-growing and aggressive compared to screen-detected cancer,(4) and interval cancer rates are routinely monitored by screening programs as an indicator of screening effectiveness.(5) Population-based breast cancer screening programs 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 programs.(6)

Internationally, there is increasing concern about the ongoing viability of population breast screening programs 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 programs, 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 aging population, coupled with recent policy and funding

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 decisions to increase the target age range for breast cancer screening in Australia from 50-69  
4 years to 50-74 years.(5) Artificial intelligence (AI) has the potential to address these resource  
5 challenges by making screen-reading more efficient and accurate. AI may particularly improve  
6 screening effectiveness if it can detect some interval cancers (cancers missed at screening)  
7 without substantially contributing to overdiagnosis (detection of cancers that would not  
8 otherwise become clinically apparent).(3)

9  
10  
11  
12  
13  
14  
15  
16  
17 Deep learning, a rapidly growing field of AI that integrates computer science and statistics,  
18 allows computers to learn directly through automatic extraction and analysis of complex data.  
19 An AI algorithm can be trained to detect breast cancer given mammography examinations with  
20 known outcomes. In doing so, the AI algorithm learns to identify automatically-extracted  
21 quantitative variables (“features”) that are predictive of cancer presence. In this respect, deep  
22 learning is a significant advance over earlier computer-aided detection (CAD) systems that  
23 relied on limited sets of human-extracted features, and resulted in unacceptably high false-  
24 positive rates.(7)

25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36 Studies that have evaluated AI for breast cancer screening suggest the technology can achieve  
37 accuracy that is comparable to expert radiologists.(6, 10-12) However, such studies commonly  
38 employ “enriched” datasets in which the prevalence of cancer is substantially higher than in  
39 population screening (up to 55%, compared with real-world screening populations where breast  
40 cancer prevalence is less than 1%).(13) Selected datasets enriched with cancers are likely to be  
41 unrepresentative of disease spectrum in screening populations, and may lead to estimates of  
42 accuracy for both AI and radiologists that are not generalisable to real-world screening.(13-15)  
43 Furthermore, routine screening metrics (cancer detection rate [CDR] and recall rate) cannot be  
44 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 curve (AUC).
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 1,000 screens) and recall rate (number of women recalled to further assessment per 1,000 screens).

## **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 program in Western Australia (WA). The study will avoid biases identified in previous research on AI for mammography screening(13) by using consecutive screens representative of real-world screening populations, with ascertained outcomes

*Protocol for study of AI to enhance breast cancer screening*

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 to 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 (e.g. 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 (standard deviation 6.9 years; range 50-74 years). There were 9,076 baseline (first ever) screens (8.3%); the remainder were subsequent screens. A total of 13,954 women (12.8%) were offered annual

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 screening due to a previous history of breast (n=3,354) and/or ovarian cancer (n=631); and/or  
4  
5 a previous diagnosis of “benign high risk” disease (n=382) defined as atypical ductal or lobular  
6  
7 hyperplasia or lobular carcinoma in situ; and/or a significant family history (n=10,197) defined  
8  
9 by BSWA as two or more first-degree relatives with breast cancer, or at least one first-degree  
10  
11 relative with breast cancer occurring at <50 years or with bilateral breast cancer.  
12  
13  
14  
15  
16  
17

**Measurement**

18  
19  
20 BSWA routinely collects demographic characteristics and risk factors through a self-  
21  
22 administered registration form. Details of the screening examination and further assessment  
23  
24 are also routinely recorded in the Mammographic Screening Registry. Descriptive variables  
25  
26 (age; screening round; time since last screen for repeat screens; mammographic breast density;  
27  
28 personal history of breast cancer; first-degree family history of breast cancer; personal history  
29  
30 of ovarian cancer; hormone replacement therapy in the past 6 months; a history of removal or  
31  
32 biopsy of benign lump; and self-reported breast symptoms) will be used to characterise the  
33  
34 cohort. Breast density (defined as heterogeneously or extremely dense breasts identified by at  
35  
36 least one of two radiologists) is recorded by BSWA only for women with no abnormality  
37  
38 identified (i.e. women who are not recalled for further testing). A deidentified screen episode  
39  
40 ID will be used to link these data to output of the AI algorithm (see section ‘Reinterpretation  
41  
42 of mammograms by AI algorithm’).  
43  
44  
45  
46  
47

48  
49 The final screening outcome (recall or not recall) will be collected, along with findings from  
50  
51 each reader and a deidentified radiologist ID. Data on cancer diagnosis (date of diagnosis;  
52  
53 screen-detected or interval cancer) and cancer characteristics (histological type; tumour size;  
54  
55 grade; nodal status) will also be extracted.  
56  
57  
58  
59  
60

### **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.<sup>(17)</sup> 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 1,000 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 (i.e. within 24 months for biennial screeners, and 12 months for the minority of women scheduled to have an annual screen).<sup>(17)</sup> 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 1,000 screens).

### **Reinterpretation of mammograms by AI algorithm**

Development of the DeepHealth algorithm that will be implemented in this study has been described previously.<sup>(10)</sup> In brief, DeepHealth used a progressive, stage-wise 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 data sets, making use of both strongly and weakly labelled data. Training data sets were *independent* of the data set 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

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 quantifying the likelihood that the region of interest represents a malignancy. The overall  
4 accuracy of the algorithm has been compared with five individual radiologists, each fellowship-  
5 trained in breast imaging, on a cancer-enriched data set, and was shown to outperform all five  
6 readers. At the average radiologist specificity, the algorithm resulted in an absolute increase in  
7 sensitivity of 14.2%; at the average radiologist sensitivity, the absolute increase in specificity  
8 was 24.0%.<sup>(10)</sup> The algorithm also outperformed radiologists in detecting malignancy in a set  
9 of prior “normal” mammograms from the same set of cases (increase in sensitivity 17.5%;  
10 increase in specificity 16.2%), demonstrating the potential to detect interval cancers “missed”  
11 by radiologists.  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22

23 All imaging analysis for the study will take place at BSWA to ensure security of images.  
24 Images will only be accessed by investigators who are employed by BSWA, and have such  
25 access under the usual conditions of their employment; these images will not be used for further  
26 refinement of DeepHealth’s algorithm. A laptop with the AI algorithm installed and a graphics  
27 processing unit supporting its evaluation will be located at BSWA. An external hard drive will  
28 be attached containing the cohort of digital mammogram data (DICOM files consisting of four  
29 views per breast, two breasts per woman). The algorithm will output data to a csv file including  
30 bounding box coordinates, malignancy scores ranging from 0 to 1, and a unique identifier  
31 extracted from the DICOM header to enable woman-level matching of results to BSWA routine  
32 screening data.  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49

**Data de-identification and secure storage**

50  
51 De-identified data on cohort characteristics, screening findings, and cancer diagnosis will be  
52 transferred by secure online file transfer to the Curtin School of Population Health, Curtin  
53 University. No paper-based or portable electronic media storage of these data will take place.  
54  
55  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 Project data will be electronically stored on a secure server, which is backed up daily to prevent  
4 any unintentional data loss. The research environment includes a variety of security controls to  
5 restrict unauthorised access – these include access controls, role-based delegations, encryption,  
6 firewalls, and physical access restrictions (authorised access to server rooms and research  
7 offices is restricted by key). Automatic screen locking will occur on electronic devices after  
8 five minutes of inactivity. Data will not be stored or used in public terminals.  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19

**Statistical methods**

20  
21  
22 All statistical analyses will be undertaken at the School of Public Health, Curtin University. To  
23 compare the accuracy of AI with the *average* accuracy of single human reading, a receiver  
24 operating characteristics (ROC) curve for the AI algorithm will firstly be plotted from the  
25 malignancy score and the AUC derived. Hierarchical summary ROC modelling(18) will be  
26 used to model radiologist accuracy and derive an area under the summary ROC curve for  
27 radiologists(19), along with summary estimates of sensitivity and specificity. The sensitivity  
28 and specificity of AI will be compared with that of radiologists by estimating AI's sensitivity  
29 of at the summary radiologist specificity, and AI's specificity at the summary radiologist  
30 sensitivity.  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43

44 The CDR and recall rate of double-reading by radiologists (current population reading practice)  
45 will be compared with double-reading strategies integrating AI, where one of the two  
46 radiologist reads per screen will be randomly paired analytically with AI. The following  
47 integrated AI-radiologist strategies will be used:  
48  
49  
50  
51  
52

- 53  
54 1. Recall to assessment based on an “either positive” rule (i.e. either AI or radiologist is  
55 positive for suspicious abnormality). This strategy will maximise CDR.(20)  
56  
57  
58  
59  
60

2. Recall to assessment based on a “both positive” rule (i.e. 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, with arbitration of disagreement by the second radiologist read that occurred in practice (i.e. the radiologist that was not randomly paired with AI). This strategy reflects current screen-reading practice.

The malignancy score derived from the AI algorithm will be dichotomised using a prospectively-defined threshold. The effect of alternative thresholds on CDR and recall rates 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.

### **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 1,000 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 1,000 screens for integrated AI-radiologist reading. This assumes concordance between the reading strategies of 5.5 cancers per 1,000 screens, with 1.5 cancers per 1,000 detected by radiologist double-reading only (and not by integrated AI-radiologist reading), and 2.0 cancers per 1,000 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)

## **Sub-studies**

In addition to the primary study objectives, sub-studies 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 (i.e. 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 (e.g. data security and privacy). Consumer health representatives external to the research team have been engaged to provide community perspectives on this research (e.g. 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 program 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 programs and policy makers in population screening, to inform future evaluation and policy discussions about the potential implementation of AI.

## **DISCUSSION**

Organised population breast screening programs 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

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 deficiencies of the evidence base (Table 1), and concluded that although studies indicate a  
4 potential role of AI in this clinical scenario, those evidence gaps should be addressed prior to  
5 the initiation of prospective trials and the adoption of the technology in routine practice.(13).  
6  
7

8  
9  
10 The primary concerns raised relate to the quality of datasets used to validate AI models and the  
11 paucity of evidence comparing the accuracy of AI and radiologists, potentially affecting the  
12 applicability and robustness of AI algorithms and raising the possibility of bias. The study we  
13 present in this protocol addresses those evidence gaps by comparing the accuracy of a  
14 commercially available algorithm with that of radiologists using a large, external validation  
15 dataset representing consecutive, *unselected* digital mammograms from a real-world screening  
16 program (Table 1). This retrospective cohort study is therefore an essential step to build the  
17 evidence base to underpin prospective trials and inform their design, and to provide timely  
18 evidence to screening stakeholders.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29

30  
31 Although this study will overcome most key limitations of the evidence base, there are potential  
32 limitations associated with its retrospective design. Data collected for administrative purposes  
33 may be more prone to misclassification than data collected specifically for research purposes  
34 through a prospective trial. For instance, we have excluded women from the study cohort who  
35 relocated outside WA after the index screening examination and therefore were potentially lost  
36 to follow-up. Since the date of relocation is not routinely collected, it is possible that some  
37 women with complete follow-up were excluded. Given that exclusions for relocation  
38 represented <0.6% of women during the study period (Figure 1), this is unlikely to represent a  
39 significant concern. Data on outcomes (recalls, screen-detected and interval cancers) are  
40 meticulously collected according to national quality and accreditation standards and are  
41 therefore unlikely to be subject to misclassification. Furthermore, we have defined the end date  
42 for study enrolment (31 December 2016) to ensure completeness of notifications for interval  
43 cancers (while simultaneously ensuring a contemporary cohort that is representative of the  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 current target population for breast cancer screening in Australia). Errors in the classification  
4  
5 of outcome data are therefore considered to be rare.  
6  
7

8 To estimate CDR and recall rate for integrated AI-human reading, we will take an analytic  
9  
10 approach to combining AI and radiologist findings. This pragmatic approach is dictated by the  
11  
12 retrospective study design; however, it may not be representative of how AI screening results  
13  
14 might be incorporated into practice. Our decision rules for defining recall to further assessment  
15  
16 are among several proposed uses of AI information. Some alternative approaches (such as the  
17  
18 use of AI to “triage” women to double-reading if exceeding a threshold probability of  
19  
20 malignancy(11)) may potentially be investigated analytically by our study design, but others  
21  
22 (such as AI output used by radiologists interactively as a decision support(6)) can only be  
23  
24 evaluated in studies using a prospective design.  
25  
26  
27

28  
29 The lack of studies exploring social and ethical issues, particularly women’s perspectives and  
30  
31 preferences around AI, has been identified as a critical evidence gap (Table 1). Although  
32  
33 beyond the scope of this study, a parallel research stream using qualitative methods is being  
34  
35 undertaken by some of the study authors to elucidate those perspectives. For instance, women  
36  
37 will be provided with information about potential uses of AI in breast screening, and will then  
38  
39 discuss this potential implementation, with a focus on what matters most to them, and how  
40  
41 implementation should (or should not) take place. Similarly, economic modelling to estimate  
42  
43 incremental costs and benefits from the use of AI is critical to informing policy decisions about  
44  
45 adopting the technology. Cost-effectiveness analysis will be undertaken in a future project  
46  
47 building on the results of this study.  
48  
49  
50

51  
52 AI algorithms for interpreting mammograms have the potential to improve the effectiveness of  
53  
54 population breast cancer screening programs if they can detect cancers, including interval  
55  
56 cancers, without contributing substantially to overdiagnosis. This will be the largest study to  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 date to investigate the accuracy of an artificial intelligence algorithm for interpreting  
4 consecutive digital mammograms in a population-based breast cancer screening program. The  
5  
6 evidence generated by this study can be used to inform decisions about adopting AI for  
7  
8 mammogram interpretation in the future, to improve accuracy, effectiveness, and efficiency.  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

**List of abbreviations**

AI – artificial intelligence

AUC – area under the receiver operating characteristics curve

BSWA – BreastScreen WA

CAD – computer-aided detection

CDR – cancer detection rate

DCIS – ductal carcinoma *in situ*

ROC – receiver operating characteristics

WA – Western Australia

**DECLARATIONS****Ethics**

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 program provide written consent for their data to be used for research purposes each time they screen.

**Competing interests**

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

**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, BL, SMC and AP). NH

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 receives funding from the National Health & Medical Research Council (NHMRC)  
4  
5 (Investigator/Leader Grant #1194410). SMC receives funding from the NHMRC (Ideas Grant  
6  
7 #1181960). GP receives funding from the NHMRC (Project Grant #1099655 and Investigator  
8  
9 Grant #1173991) and the Research Council of Norway through its Centres of Excellence  
10  
11 funding scheme (#262700). CIL and WL are funded in part by the National Cancer Institute  
12  
13 (R37 CA240403).  
14  
15

**Author contributions**

16  
17  
18 MLM, NH, EW, HL, AW and WL conceived the idea, planned and designed the study protocol.  
19  
20 AP, SMC, MB, GP, JGK, CIL, and SZ contributed to the development of the protocol, study  
21  
22 design and methods. MLM wrote the first draft. NH, EW, HL, AW, WL, AP, SMC, MB, GP,  
23  
24 JGK, CIL, and SZ critically revised the draft for important intellectual content. All authors  
25  
26 have approved the final written manuscript.  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**References**

1. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *British journal of cancer*. 2013;108(11):2205-40.
2. Hanley JA, Hannigan A, O'Brien KM. Mortality reductions due to mammography screening: Contemporary population-based data. *PloS one*. 2017;12(12):e0188947-e.
3. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *npj Breast Cancer*. 2017;3(1):12.
4. Baré M, Torà N, Salas D, Sentís M, Ferrer J, Ibáñez J, et al. Mammographic and clinical characteristics of different phenotypes of screen-detected and interval breast cancers in a nationwide screening program. *Breast Cancer Research and Treatment*. 2015;154(2):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, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute*. 2019;111(9):djj222.
7. Harvey H, Karpati E, Khara G, Korkinof D, Ng A, Austin C, et al. The Role of Deep Learning in Breast Screening. *Current Breast Cancer Reports*. 2019;11(1):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.

*Protocol for study of AI to enhance breast cancer screening*

10. Lotter W, Diab AR, Haslam B, Kim JG, Grisot G, Wu E, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine*. 2021;27(2):244-9.
11. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94.
12. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncology*. 2020;6(10):1581-8.
13. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Review of Medical Devices*. 2019;16(5):351-62.
14. Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*. 2013;185(11):E537-E44.
15. Park SH. Diagnostic Case-Control versus Diagnostic Cohort Studies for Clinical Validation of Artificial Intelligence Algorithm Performance. *Radiology*. 2019;290(1):272-3.
16. Elmore JG, Lee CI. Artificial Intelligence for Breast Cancer Imaging: The New Frontier? *JNCI: Journal of the National Cancer Institute*. 2019;111(9):djj223.
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(19):2865-84.
19. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med*. 2002;21(9):1237-56.



- 1  
2  
3 20. Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic  
4 performance when combining two diagnostic tests. *Statistics in Medicine*. 2002;21(17):2527-  
5  
6  
7 46.  
8  
9

10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

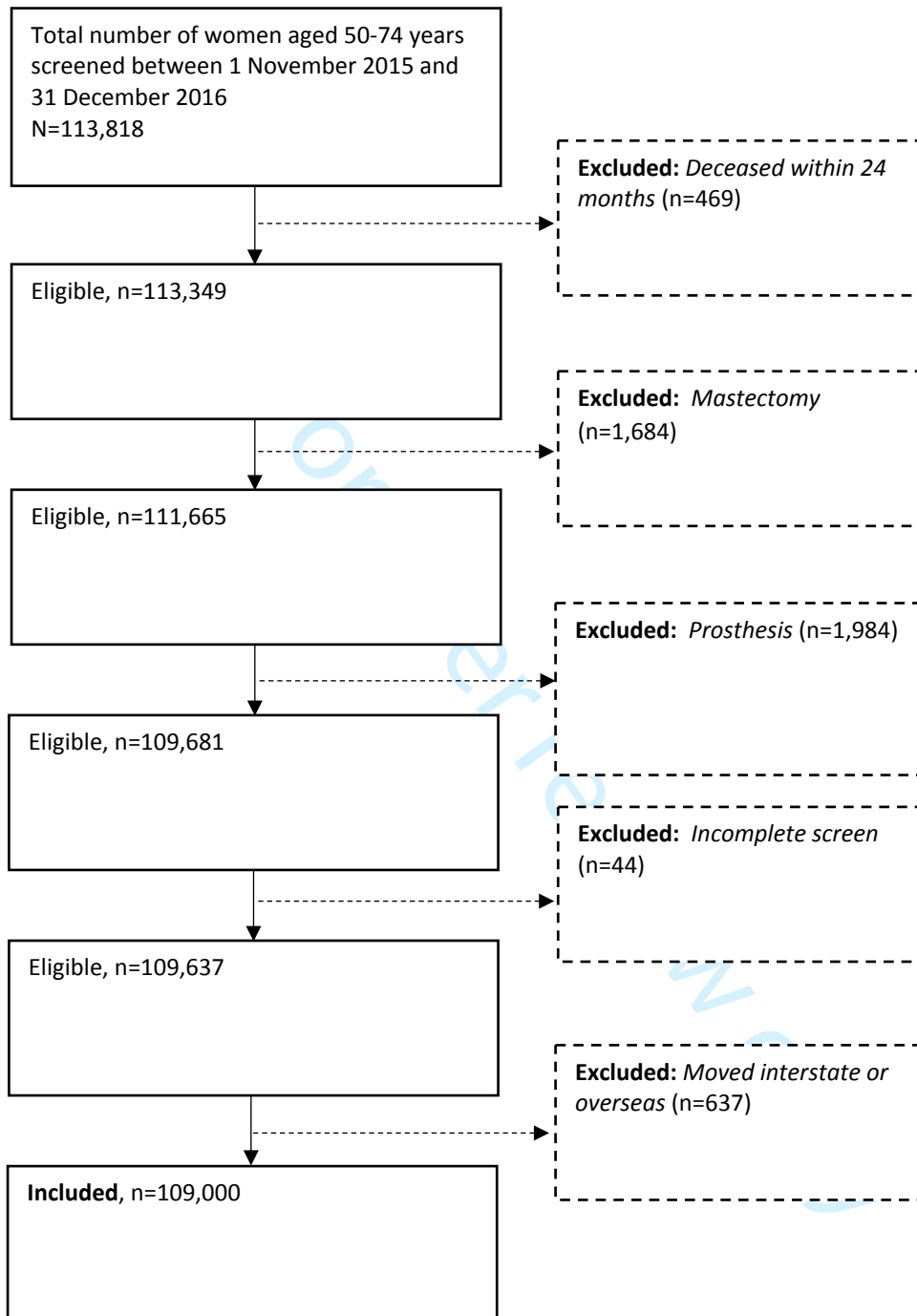
**Table 1.** Significant gaps in knowledge needed to develop prospective real-world screening trials or evaluation (adapted from Houssami et al.(13))

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 commercially-available triage product.
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 data sets were used to train and validate models. Validation using independent, external data sets 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 program (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, 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 programs.

1  
2  
3 **Figure 1:** Flowchart of cohort inclusions and exclusions  
4  
5  
6

7 **Figure 2:** Digital mammogram mediolateral oblique view with region of interest (denoted by  
8 bounding box) identified by the AI algorithm as suspicious for malignancy. Cancer was  
9 confirmed as invasive ductal carcinoma.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



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.

1174x1444mm (72 x 72 DPI)

# BMJ Open

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

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-054005.R1
Article Type:	Protocol
Date Submitted by the Author:	04-Nov-2021
Complete List of Authors:	Marinovich, M Luke; Curtin University, Curtin School of Population Health; The University of Sydney, Sydney School of Public Health Wylie, Elizabeth; BreastScreen WA Lotter, William; DeepHealth Inc. Pearce, Alison ; The University of Sydney, Carter, Stacy; University of Wollongong, Australian Centre for Health Engagement, Evidence and Values Lund, Helen; BreastScreen WA Waddell, Andrew; BreastScreen WA Kim, Jiye; DeepHealth Inc. Pereira, G.F; Curtin University, Curtin School of Population Health; Norwegian Institute of Public Health, Centre for Fertility and Health Lee, C; University of Washington, Department of Radiology Zackrisson, Sophia; Lund University, Diagnostic Radiology Brennan, ME; The University of Sydney, Sydney School of Public Health Houssami, Nehmat; The University of Sydney, Sydney School of Public Health; The University of Sydney, The Daffodil Centre
<b>Primary Subject Heading</b>:	Oncology
Secondary Subject Heading:	Radiology and imaging
Keywords:	Breast tumours < ONCOLOGY, Breast imaging < RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING

SCHOLARONE™  
Manuscripts

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

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

**Author affiliations:**

<sup>1</sup> Curtin School of Population Health, Curtin University, Perth, Western Australia, Australia

<sup>2</sup> Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Camperdown, New South Wales, Australia

<sup>3</sup> BreastScreen WA, Perth, Western Australia, Australia

<sup>4</sup> DeepHealth Inc., RadNet AI Solutions, Cambridge, Massachusetts, USA.

<sup>5</sup> Australian Centre for Health Engagement, Evidence and Values (ACHEEV), School of Health and Society, University of Wollongong, Wollongong, New South Wales, Australia

<sup>6</sup> Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

<sup>7</sup> Department of Radiology, University of Washington School of Medicine, Seattle, Washington, USA

<sup>8</sup> Diagnostic Radiology, Department of Translational Medicine, Lund University, Skåne University Hospital, Malmö, Sweden

<sup>9</sup> The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, New South Wales, Australia.

**Corresponding author:**

Dr ML Marinovich, Research Fellow

School of Public Health, Curtin University

GPO Box U1987

Perth Western Australia 6845

Email: [Luke.Marinovich@curtin.edu.au](mailto:Luke.Marinovich@curtin.edu.au)

Phone: +61 8 9266 4006

**Protocol version:** 1.1 (November 2021)

**Word count:** 3,739

*Protocol for study of AI to enhance breast cancer screening***ABSTRACT**

**Introduction:** Artificial intelligence (AI) algorithms for interpreting mammograms have the potential to improve the effectiveness of population breast cancer screening programs 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 generalisability to real-world screening. We aim to address these limitations by comparing the accuracy of AI and radiologists in a cohort of consecutive women attending a real-world population breast cancer screening program.

**Methods and Analysis:** A retrospective cohort of 109,000 distinct, consecutive digital mammography screens from November 2016 to December 2017 (including 761 screen-detected and 235 interval cancers) was assembled from BreastScreen WA (BSWA), Western Australia’s biennial population screening program. Descriptive characteristics of the cohort 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 to that of radiologist single reading based on the difference in the area under the receiver operating characteristic curve (AUC-ROC). 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



*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 presented at national and international conferences. Results will also be disseminated to  
4  
5 stakeholders in Australian breast cancer screening programs and policy makers in population  
6  
7 screening.  
8  
9

10 **Keywords:** breast cancer; screening; artificial intelligence mammography; accuracy  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

**ARTICLE SUMMARY****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 algorithm for interpreting digital mammograms in a population breast cancer screening program.
- 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 programs, 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

Health care 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 symptomatically in the interval between screening rounds and are known as “interval cancers”.(3) Interval cancers are more often fast-growing and aggressive compared to screen-detected cancer,(4) and interval cancer rates are routinely monitored by screening programs as an indicator of screening effectiveness.(5) Population-based breast cancer screening programs in Australia (BreastScreen), Europe and the United Kingdom (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 programs.(6)

Internationally, there is increasing concern about the ongoing viability of population breast screening programs 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 programs, 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 aging population, coupled with recent policy and funding

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 decisions to increase the target age range for breast cancer screening in Australia from 50-69  
4 years to 50-74 years.(5) Artificial intelligence (AI) has the potential to address these resource  
5 challenges by making screen-reading more efficient and accurate. AI may particularly improve  
6 screening effectiveness if it can detect some interval cancers (cancers missed at screening)  
7 without substantially contributing to overdiagnosis (detection of cancers that would not  
8 otherwise become clinically apparent).(3)

9  
10  
11  
12  
13  
14  
15  
16  
17 Deep learning, a rapidly growing field of AI that integrates computer science and statistics,  
18 allows computers to learn directly through automatic extraction and analysis of complex data.  
19 An AI algorithm can be trained to detect breast cancer given mammography examinations with  
20 known outcomes. In doing so, the AI algorithm learns to identify automatically-extracted  
21 quantitative variables (“features”) that are predictive of cancer presence. In this respect, deep  
22 learning is a significant advance over earlier computer-aided detection (CAD) systems that  
23 relied on limited sets of human-extracted features, and resulted in unacceptably high false-  
24 positive rates.(7)

25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36 Studies that have evaluated AI for breast cancer screening suggest the technology can achieve  
37 accuracy that is comparable to expert radiologists.(6, 10-12) However, such studies commonly  
38 employ “enriched” datasets in which the prevalence of cancer is substantially higher than in  
39 population screening (up to 55%, compared with real-world screening populations where breast  
40 cancer prevalence is less than 1%).(13) Selected datasets enriched with cancers are likely to be  
41 unrepresentative of disease spectrum in screening populations, and may lead to estimates of  
42 accuracy for both AI and radiologists that are not generalisable to real-world screening.(13-15)  
43 Furthermore, routine screening metrics (cancer detection rate [CDR] and recall rate) cannot be  
44 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 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 1,000 screens) and recall rate (number of women recalled to further assessment per 1,000 screens).

## **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 program in Western Australia (WA). The study will avoid biases identified in previous research on AI for mammography screening(13) by using consecutive screens (i.e. all screening examinations meeting the inclusion criteria in a defined time interval)

*Protocol for study of AI to enhance breast cancer screening*

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 to 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 (e.g. 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 (standard deviation 6.9 years; range 50-74 years). There were 9,076 baseline (first ever) screens (8.3%);

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 the remainder were subsequent screens. A total of 13,954 women (12.8%) were offered annual  
4  
5 screening due to a previous history of breast (n=3,354) and/or ovarian cancer (n=631); and/or  
6  
7 a previous diagnosis of “benign high risk” disease (n=382) defined as atypical ductal or lobular  
8  
9 hyperplasia or lobular carcinoma in situ; and/or a significant family history (n=10,197) defined  
10  
11 by BSWA as two or more first-degree relatives with breast cancer, or at least one first-degree  
12  
13 relative with breast cancer occurring at <50 years or with bilateral breast cancer.  
14  
15  
16  
17  
18  
19

**Measurement**

20  
21  
22  
23 BSWA routinely collects demographic characteristics and risk factors through a self-  
24  
25 administered registration form. Details of the screening examination and further assessment  
26  
27 are also routinely recorded in the Mammographic Screening Registry. Descriptive variables  
28  
29 (age; screening round; time since last screen for repeat screens; mammographic breast density;  
30  
31 personal history of breast cancer; first-degree family history of breast cancer; personal history  
32  
33 of ovarian cancer; hormone replacement therapy in the past 6 months; a history of removal or  
34  
35 biopsy of benign lump; and self-reported breast symptoms) will be used to characterise the  
36  
37 cohort. Breast density (defined as heterogeneously or extremely dense breasts identified by at  
38  
39 least one of two radiologists) is recorded by BSWA only for women with no abnormality  
40  
41 identified (i.e. women who are not recalled for further testing). A deidentified screen episode  
42  
43 ID will be used to link these data to output of the AI algorithm (see section ‘Reinterpretation  
44  
45 of mammograms by AI algorithm’).  
46  
47  
48  
49

50  
51 The final screening outcome (recall or not recall) will be collected, along with findings from  
52  
53 each reader and a deidentified radiologist ID. Data on cancer diagnosis (date of diagnosis;  
54  
55 screen-detected or interval cancer) and cancer characteristics (histological type; tumour size;  
56  
57 grade; nodal status) will also be extracted.  
58  
59  
60

## 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.<sup>(17)</sup> 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 1,000 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 (i.e. within 24 months for biennial screeners, and 12 months for the minority of women scheduled to have an annual screen).<sup>(17)</sup> 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 1,000 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 United States (US). Development of the algorithm has been described previously.<sup>(10)</sup> In brief, DeepHealth used a progressive, stage-wise 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 data sets from the US and UK, making use of both strongly and weakly



*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 labelled data. Australian data were not used for algorithm training; therefore, training data sets  
4 were *independent* of the data set used for the current external validation study. The trained  
5  
6 algorithm outputs a “bounding box” identifying a region of interest (Figure 2), along with a  
7  
8 malignancy score quantifying the likelihood that the region of interest represents a malignancy.  
9  
10  
11 The algorithm evaluates each image in a study independently and aggregates the scores across  
12  
13 all potential regions in the study to compute a single study-level malignancy score. The overall  
14  
15 accuracy of the algorithm based on this study-level malignancy score has been compared with  
16  
17 five individual radiologists, each fellowship-trained in breast imaging, on a cancer-enriched  
18  
19 data set, and was shown to outperform all five readers. At the average radiologist specificity,  
20  
21 the algorithm resulted in an absolute increase in sensitivity of 14.2%; at the average radiologist  
22  
23 sensitivity, the absolute increase in specificity was 24.0%.<sup>(10)</sup> The algorithm also  
24  
25 outperformed radiologists in detecting malignancy in a set of prior “normal” mammograms  
26  
27 from the same set of cancer cases (increase in sensitivity 17.5%; increase in specificity 16.2%),  
28  
29 demonstrating the potential to detect interval cancers “missed” by radiologists.  
30  
31  
32  
33  
34  
35

36 All imaging analysis for the study will take place at BSWA to ensure security of images.  
37  
38 Images will only be accessed by investigators who are employed by BSWA, and have such  
39  
40 access under the usual conditions of their employment; these images will not be used for further  
41  
42 refinement of DeepHealth’s algorithm. A laptop with the AI algorithm installed and a graphics  
43  
44 processing unit supporting its evaluation will be located at BSWA. An external hard drive will  
45  
46 be attached containing the cohort of digital mammogram data (DICOM files consisting of four  
47  
48 views per breast, two breasts per woman). The algorithm will output data to a csv file including  
49  
50 bounding box coordinates, malignancy scores ranging from 0 to 1, and a unique identifier  
51  
52 extracted from the DICOM header to enable woman-level matching of results to BSWA routine  
53  
54 screening data.  
55  
56  
57  
58  
59  
60

## 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 five minutes 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, a receiver operating characteristic (ROC) curve for the AI algorithm will firstly 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

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 expected recall rate of 4% (the overall recall rate of the BSWA program) based on  
4  
5 DeepHealth's (non-Australian) validation data, allowing for a descriptive comparison of  
6  
7 sensitivity and specificity at this threshold with summary radiologist estimates.  
8  
9

10 The CDR and recall rate of double-reading by radiologists (current population reading practice)  
11  
12 will be compared with double-reading strategies integrating AI (McNemar's test), where the  
13  
14 first radiologist read per screen will be paired analytically with AI. The following integrated  
15  
16 AI-radiologist strategies will be used:  
17  
18

- 19  
20 1. Recall to assessment based on an "either positive" rule (i.e. either AI or radiologist is  
21  
22 positive for suspicious abnormality). This strategy will maximise CDR.(21)  
23  
24
- 25 2. Recall to assessment based on a "both positive" rule (i.e. both AI and radiologist are  
26  
27 positive for suspicious abnormality). This strategy will minimise recall rate.(21)  
28  
29
- 30 3. Recall to assessment based on results of AI-human reading, where "both positive"  
31  
32 findings for AI and radiologist trigger a decision to recall, and "either positive" findings  
33  
34 (i.e. disagreement) are arbitrated by the second radiologist read that occurred in  
35  
36 practice. This strategy simulates current screen-reading practice.  
37  
38

39 The effect on CDR and recall rates of alternative thresholds for dichotomising the AI algorithm  
40  
41 score will be explored in sensitivity analyses. CDR results for integrated AI-radiologist reading  
42  
43 will be stratified by interval versus non-interval cancers to estimate the incremental CDR for  
44  
45 interval (clinically progressive) cancers. Sensitivity analyses will also be conducted to apply a  
46  
47 consistent 12-month follow-up period for ascertaining interval cancers.  
48  
49  
50  
51  
52  
53

### 54 **Sample size and power calculation**

55

56 Power calculations were derived for the outcome of CDR, based on the sample size and number  
57  
58 of screen-detected and interval cancers present in the cohort. The CDR for double-reading by  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

radiologists in the study cohort is 7.0 per 1,000 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 1,000 screens for integrated AI-radiologist reading. This assumes concordance between the reading strategies of 5.5 cancers per 1,000 screens, with 1.5 cancers per 1,000 detected by radiologist double-reading only (and not by integrated AI-radiologist reading), and 2.0 cancers per 1,000 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)

**Sub-studies**

In addition to the primary study objectives, sub-studies 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 (i.e. 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 (e.g. data security and privacy). Consumer health representatives external to the research team have been engaged to provide community perspectives on this research (e.g. 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 program 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

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 be made publicly available at the summary level. Manuscripts addressing the study aims will  
4  
5 be published in peer-reviewed journals. Results will also be presented at relevant national and  
6  
7 international conferences. Study outcomes will also be disseminated to stakeholders in  
8  
9 Australian breast cancer screening programs and policy makers in population screening, to  
10  
11 inform future evaluation and policy discussions about the potential implementation of AI.  
12  
13  
14  
15  
16  
17

**DISCUSSION**

18  
19  
20 Organised population breast screening programs are facing growing screen-reading resource  
21  
22 challenges, so the current global research effort aimed at developing and testing AI algorithms  
23  
24 for interpreting screening mammograms can contribute to ensuring future sustainability of  
25  
26 screening. Although the field is rapidly-evolving, to date there has been a focus on algorithm  
27  
28 development with relatively few studies evaluating AI in real-world breast cancer screening  
29  
30 settings. A scoping review of the literature on AI for breast screening identified eight key  
31  
32 deficiencies of the evidence base (Table 1), and concluded that although studies indicate a  
33  
34 potential role of AI in this clinical scenario, those evidence gaps should be addressed prior to  
35  
36 the initiation of prospective trials and the adoption of the technology in routine practice.(13).  
37  
38  
39 The primary concerns raised relate to the quality of datasets used to validate AI models and the  
40  
41 paucity of evidence comparing the accuracy of AI and radiologists, potentially affecting the  
42  
43 applicability and robustness of AI algorithms and raising the possibility of bias. The study we  
44  
45 present in this protocol addresses those evidence gaps by comparing the accuracy of a  
46  
47 commercially available algorithm with that of radiologists using a large, external validation  
48  
49 dataset representing consecutive, *unselected* digital mammograms from a real-world screening  
50  
51 program (Table 1). This retrospective cohort study is therefore an essential step to build the  
52  
53  
54  
55  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 evidence base to underpin prospective trials and inform their design, and to provide timely  
4  
5 evidence to screening stakeholders.  
6  
7

8 Although this study will overcome most key limitations of the evidence base, there are potential  
9  
10 limitations associated with its retrospective design. Data collected for administrative purposes  
11  
12 may be more prone to misclassification than data collected specifically for research purposes  
13  
14 through a prospective trial. For instance, we have excluded women from the study cohort who  
15  
16 relocated outside WA after the index screening examination and therefore were potentially lost  
17  
18 to follow-up. Since the date of relocation is not routinely collected, it is possible that some  
19  
20 women with complete follow-up were excluded. Given that exclusions for relocation  
21  
22 represented <0.6% of women during the study period (Figure 1), this is unlikely to represent a  
23  
24 significant concern. Data on outcomes (recalls, screen-detected and interval cancers) are  
25  
26 meticulously collected according to national quality and accreditation standards and are  
27  
28 therefore unlikely to be subject to misclassification. Furthermore, we have defined the end date  
29  
30 for study enrolment (31 December 2016) to ensure completeness of notifications for interval  
31  
32 cancers (while simultaneously ensuring a contemporary cohort that is representative of the  
33  
34 current target population for breast cancer screening in Australia). Errors in the classification  
35  
36 of outcome data are therefore considered to be rare.  
37  
38  
39  
40  
41  
42

43 To estimate CDR and recall rate for integrated AI-human reading, we will take an analytic  
44  
45 approach to combining AI and radiologist findings. This pragmatic approach is dictated by the  
46  
47 retrospective study design; however, it may not be representative of how AI screening results  
48  
49 might be incorporated into practice. Our decision rules for defining recall to further assessment  
50  
51 are among several proposed uses of AI information. Some alternative approaches (such as the  
52  
53 use of AI to “triage” women to double-reading if exceeding a threshold probability of  
54  
55 malignancy(11)) may potentially be investigated analytically by our study design, but others  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 (such as AI output used by radiologists interactively as a decision support(6)) can only be  
4  
5 evaluated in studies using a prospective design. Furthermore, the methods adopted to derive  
6  
7 summary ROC curves for radiologists and associated measures of accuracy are dictated by the  
8  
9 retrospective design. Although these methods are established and appropriate for our real-  
10  
11 world screening data,(18) they allow only for descriptive comparisons with empirical estimates  
12  
13 for the AI algorithm.(19)  
14  
15

16  
17 The lack of studies exploring social and ethical issues, particularly women's perspectives and  
18  
19 preferences around AI, has been identified as a critical evidence gap (Table 1). Although  
20  
21 beyond the scope of this study, a parallel research stream using qualitative methods is being  
22  
23 undertaken by some of the study authors to elucidate those perspectives. For instance, women  
24  
25 will be provided with information about potential uses of AI in breast screening, and will then  
26  
27 discuss this potential implementation, with a focus on what matters most to them, and how  
28  
29 implementation should (or should not) take place. Similarly, economic modelling to estimate  
30  
31 incremental costs and benefits from the use of AI is critical to informing policy decisions about  
32  
33 adopting the technology. Cost-effectiveness analysis will be undertaken in a future project  
34  
35 building on the results of this study.  
36  
37  
38  
39  
40

41 AI algorithms for interpreting mammograms have the potential to improve the effectiveness of  
42  
43 population breast cancer screening programs if they can detect cancers, including interval  
44  
45 cancers, without contributing substantially to overdiagnosis. This will be the largest study to  
46  
47 date to investigate the accuracy of an artificial intelligence algorithm for interpreting  
48  
49 consecutive digital mammograms in a population-based breast cancer screening program. The  
50  
51 evidence generated by this study can be used to inform decisions about adopting AI for  
52  
53 mammogram interpretation in the future, to improve accuracy, effectiveness, and efficiency.  
54  
55  
56  
57  
58  
59  
60



**List of abbreviations**

AI – artificial intelligence

AUC-ROC – area under the receiver operating characteristic curve

BSWA – BreastScreen WA

CAD – computer-aided detection

CDR – cancer detection rate

DCIS – ductal carcinoma *in situ*

ROC – receiver operating characteristic

WA – Western Australia

**DECLARATIONS****Ethics**

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 program provide written consent for their data to be used for research purposes each time they screen.

**Competing interests**

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

**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, BL, SMC and AP). NH

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 receives funding from the National Health & Medical Research Council (NHMRC)  
4  
5 (Investigator/Leader Grant #1194410). SMC receives funding from the NHMRC (Ideas Grant  
6  
7 #1181960). GP receives funding from the NHMRC (Project Grant #1099655 and Investigator  
8  
9 Grant #1173991) and the Research Council of Norway through its Centres of Excellence  
10  
11 funding scheme (#262700). CIL and WL are funded in part by the National Cancer Institute  
12  
13 (R37 CA240403).  
14  
15

**Author contributions**

16  
17  
18 MLM, NH, EW, HL, AW and WL conceived the idea, planned and designed the study protocol.  
19  
20 AP, SMC, MB, GP, JGK, CIL, and SZ contributed to the development of the protocol, study  
21  
22 design and methods. MLM wrote the first draft. NH, EW, HL, AW, WL, AP, SMC, MB, GP,  
23  
24 JGK, CIL, and SZ critically revised the draft for important intellectual content. All authors  
25  
26  
27  
28  
29 have approved the final written manuscript.  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**References**

1. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *British journal of cancer*. 2013;108(11):2205-40.
2. Hanley JA, Hannigan A, O'Brien KM. Mortality reductions due to mammography screening: Contemporary population-based data. *PloS one*. 2017;12(12):e0188947-e.
3. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *npj Breast Cancer*. 2017;3(1):12.
4. Baré M, Torà N, Salas D, Sentís M, Ferrer J, Ibáñez J, et al. Mammographic and clinical characteristics of different phenotypes of screen-detected and interval breast cancers in a nationwide screening program. *Breast Cancer Research and Treatment*. 2015;154(2):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, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute*. 2019;111(9):djj222.
7. Harvey H, Karpati E, Khara G, Korkinof D, Ng A, Austin C, et al. The Role of Deep Learning in Breast Screening. *Current Breast Cancer Reports*. 2019;11(1):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.

*Protocol for study of AI to enhance breast cancer screening*

10. Lotter W, Diab AR, Haslam B, Kim JG, Grisot G, Wu E, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine*. 2021;27(2):244-9.
11. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94.
12. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncology*. 2020;6(10):1581-8.
13. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Review of Medical Devices*. 2019;16(5):351-62.
14. Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*. 2013;185(11):E537-E44.
15. Park SH. Diagnostic Case-Control versus Diagnostic Cohort Studies for Clinical Validation of Artificial Intelligence Algorithm Performance. *Radiology*. 2019;290(1):272-3.
16. Elmore JG, Lee CI. Artificial Intelligence for Breast Cancer Imaging: The New Frontier? *JNCI: Journal of the National Cancer Institute*. 2019;111(9):djj223.
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(19):2865-84.

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 19. Oakden-Rayner L, Palmer L. Docs are ROCs: A simple off-the-shelf approach for  
4 estimating average human performance in diagnostic studies. 2020 Available from:  
5  
6 arXiv:2009.11060v2 [stat.ME]. Accessed 27 Oct 2021.  
7  
8

9  
10 20. Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic  
11 performance when combining two diagnostic tests. *Statistics in Medicine*. 2002;21(17):2527-  
12  
13  
14  
15 46.  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

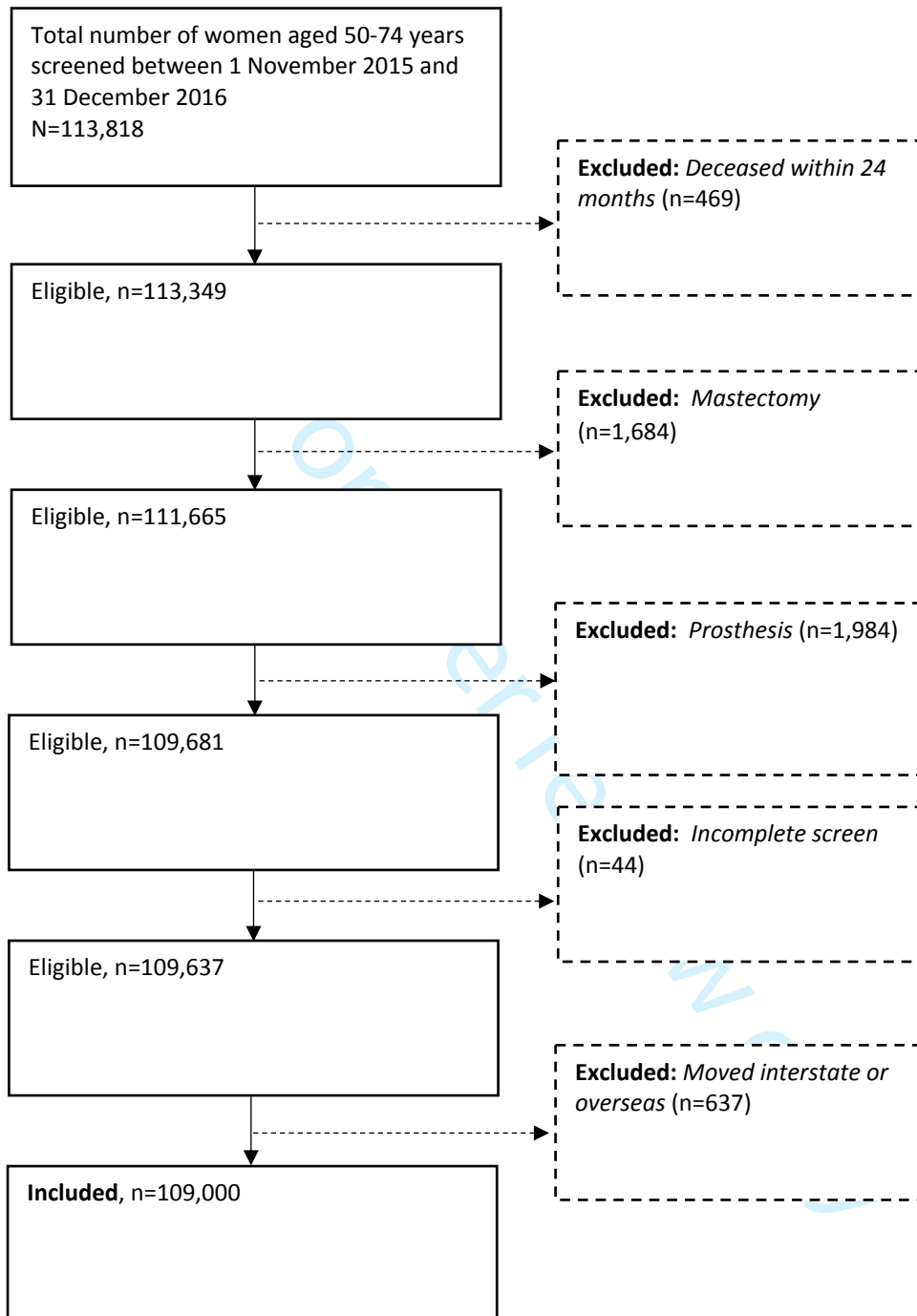
*Protocol for study of AI to enhance breast cancer screening***Table 1.** Significant gaps in knowledge needed to develop prospective real-world screening trials or evaluation (adapted from Houssami et al.(13))

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 US.
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 data sets were used to train and validate models. Validation using independent, external data sets 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 program (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 programs.

1  
2  
3 **Figure 1:** Flowchart of cohort inclusions and exclusions  
4  
5  
6

7 **Figure 2:** Digital mammogram mediolateral oblique view with region of interest (denoted by  
8 bounding box) identified by the AI algorithm as suspicious for malignancy. Cancer was  
9 confirmed as invasive ductal carcinoma.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only





1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



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.

1174x1444mm (72 x 72 DPI)

# BMJ Open

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

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-054005.R2
Article Type:	Protocol
Date Submitted by the Author:	24-Nov-2021
Complete List of Authors:	Marinovich, M Luke; Curtin University, Curtin School of Population Health; The University of Sydney, Sydney School of Public Health Wylie, Elizabeth; BreastScreen WA Lotter, William; DeepHealth Inc. Pearce, Alison ; The University of Sydney, Carter, Stacy; University of Wollongong, Australian Centre for Health Engagement, Evidence and Values Lund, Helen; BreastScreen WA Waddell, Andrew; BreastScreen WA Kim, Jiye; DeepHealth Inc. Pereira, G.F; Curtin University, Curtin School of Population Health; Norwegian Institute of Public Health, Centre for Fertility and Health Lee, C; University of Washington, Department of Radiology Zackrisson, Sophia; Lund University, Diagnostic Radiology Brennan, ME; The University of Sydney, Sydney School of Public Health Houssami, Nehmat; The University of Sydney, Sydney School of Public Health; The University of Sydney, The Daffodil Centre
<b>Primary Subject Heading</b>:	Oncology
Secondary Subject Heading:	Radiology and imaging
Keywords:	Breast tumours < ONCOLOGY, Breast imaging < RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING

SCHOLARONE™  
Manuscripts

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

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

**Author affiliations:**

<sup>1</sup> Curtin School of Population Health, Curtin University, Perth, Western Australia, Australia

<sup>2</sup> Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Camperdown, New South Wales, Australia

<sup>3</sup> BreastScreen WA, Perth, Western Australia, Australia

<sup>4</sup> DeepHealth Inc., RadNet AI Solutions, Cambridge, Massachusetts, USA.

<sup>5</sup> Australian Centre for Health Engagement, Evidence and Values (ACHEEV), School of Health and Society, University of Wollongong, Wollongong, New South Wales, Australia

<sup>6</sup> Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

<sup>7</sup> Department of Radiology, University of Washington School of Medicine, Seattle, Washington, USA

<sup>8</sup> Diagnostic Radiology, Department of Translational Medicine, Lund University, Skåne University Hospital, Malmö, Sweden

<sup>9</sup> The Daffodil Centre, The University of Sydney, a joint venture with Cancer Council NSW, Sydney, New South Wales, Australia.

**Corresponding author:**

Dr ML Marinovich, Research Fellow

School of Public Health, Curtin University

GPO Box U1987

Perth Western Australia 6845

Email: [Luke.Marinovich@curtin.edu.au](mailto:Luke.Marinovich@curtin.edu.au)

Phone: +61 8 9266 4006

**Protocol version:** 1.2 (November 2021)

**Word count:** 3,741

**ABSTRACT**

**Introduction:** Artificial intelligence (AI) algorithms for interpreting mammograms have the potential to improve the effectiveness of population breast cancer screening programs 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 generalisability 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 program.

**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 program, 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 to that of radiologist single-reading based on the difference in the area under the receiver operating characteristic curve (AUC-ROC). 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

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 and presented at national and international conferences. Results will also be disseminated to  
4  
5 stakeholders in Australian breast cancer screening programs and policy makers in population  
6  
7 screening.  
8  
9

10 **Keywords:** breast cancer; screening; artificial intelligence mammography; accuracy  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

**ARTICLE SUMMARY****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 algorithm for interpreting digital mammograms in a population breast cancer screening program.
- 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 programs, 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

Health care 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 symptomatically in the interval between screening rounds and are known as “interval cancers”.(3) Interval cancers are more often fast-growing and aggressive compared to screen-detected cancer,(4) and interval cancer rates are routinely monitored by screening programs as an indicator of screening effectiveness.(5) Population-based breast cancer screening programs in Australia (BreastScreen), Europe and the United Kingdom (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 programs.(6)

Internationally, there is increasing concern about the ongoing viability of population breast screening programs 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 programs, 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 aging population, coupled with recent policy and

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 funding decisions to increase the target age range for breast cancer screening in Australia  
4 from 50-69 years to 50-74 years.(5) Artificial intelligence (AI) has the potential to address  
5 these resource challenges by making screen-reading more efficient and accurate. AI may  
6 particularly improve screening effectiveness if it can detect some interval cancers (cancers  
7 missed at screening) without substantially contributing to overdiagnosis (detection of cancers  
8 that would not otherwise become clinically apparent).(3)

9  
10  
11  
12  
13  
14  
15  
16  
17  
18 Deep learning, a rapidly growing field of AI that integrates computer science and statistics,  
19 allows computers to learn directly through automatic extraction and analysis of complex data.  
20 An AI algorithm can be trained to detect breast cancer given mammography examinations  
21 with known outcomes. In doing so, the AI algorithm learns to identify automatically-  
22 extracted quantitative variables (“features”) that are predictive of cancer presence. In this  
23 respect, deep learning is a significant advance over earlier computer-aided detection (CAD)  
24 systems that relied on limited sets of human-extracted features, and resulted in unacceptably  
25 high false-positive rates.(7)

26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
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



1  
2  
3 a need to generate evidence of AI performance that is generalisable to routine screening  
4  
5 practice to inform decisions about adopting the technology.(13, 16)  
6  
7  
8  
9

## 10 11 **Study aims and hypotheses**

12  
13  
14 This project aims to compare AI reading of digital mammograms with human reading in a  
15  
16 real-world, population breast cancer screening setting. We hypothesise that the AI algorithm  
17  
18 has accuracy that is comparable to human readers, and that integrating the AI into a standard  
19  
20 screen-reading strategy will accurately detect cancers including interval cancers. Specifically,  
21  
22 we aim to:  
23

- 24  
25  
26 1. Compare the accuracy of AI with the average accuracy of single human reading in  
27  
28 terms of the area under the receiver operating curve (AUC-ROC).  
29
- 30  
31 2. Compare integrated AI-human screen-reading with human double-reading  
32  
33 (standard breast cancer screen-reading practice) in terms of CDR (number of  
34  
35 cancers detected per 1,000 screens) and case-specific recall rate (number of  
36  
37 women recalled to further assessment per 1,000 screens).  
38  
39  
40  
41  
42

## 43 **METHODS AND ANALYSIS**

### 44 45 **Study design and inclusion criteria**

46  
47  
48 A retrospective study design was used to assemble a contemporary cohort of unique,  
49  
50 consecutive digital mammography screens from BreastScreen WA (BSWA), the population  
51  
52 breast cancer screening program in Western Australia (WA). The study will avoid biases  
53  
54 identified in previous research on AI for mammography screening(13) by using consecutive  
55  
56 screens (i.e. all screening examinations meeting the inclusion criteria in a defined time  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

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 to 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 (e.g. 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 (standard deviation 6.9 years; range 50-74 years). There were 9,076 baseline (first ever)

1  
2  
3 screens (8.3%); the remainder were subsequent screens. A total of 13,954 women (12.8%)  
4  
5 were offered annual screening due to a previous history of breast (n=3,354) and/or ovarian  
6  
7 cancer (n=631); and/or a previous diagnosis of “benign high risk” disease (n=382) defined as  
8  
9 atypical ductal or lobular hyperplasia or lobular carcinoma in situ; and/or a significant family  
10  
11 history (n=10,197) defined by BSWA as two or more first-degree relatives with breast  
12  
13 cancer, or at least one first-degree relative with breast cancer occurring at <50 years or with  
14  
15 bilateral breast cancer.  
16  
17  
18  
19  
20  
21  
22

### 23 **Measurement**

24  
25 BSWA routinely collects demographic characteristics and risk factors through a self-  
26  
27 administered registration form. Details of the screening examination and further assessment  
28  
29 are also routinely recorded in the Mammographic Screening Registry. Descriptive variables  
30  
31 (age; screening round; time since last screen for repeat screens; mammographic breast  
32  
33 density; personal history of breast cancer; first-degree family history of breast cancer;  
34  
35 personal history of ovarian cancer; hormone replacement therapy in the past 6 months; a  
36  
37 history of removal or biopsy of benign lump; and self-reported breast symptoms) will be used  
38  
39 to characterise the cohort. Breast density (defined as heterogeneously or extremely dense  
40  
41 breasts identified by at least one of two radiologists) is recorded by BSWA only for women  
42  
43 with no abnormality identified (i.e. women who are not recalled for further testing). A  
44  
45 deidentified screen episode ID will be used to link these data to output of the AI algorithm  
46  
47 (see section ‘Reinterpretation of mammograms by AI algorithm’).  
48  
49  
50  
51  
52

53 The final screening outcome (recall or not recall) will be collected, along with findings from  
54  
55 each reader and a deidentified radiologist ID. Data on cancer diagnosis (date of diagnosis;  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

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.<sup>(17)</sup> 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 1,000 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 (i.e. within 24 months for biennial screeners, and 12 months for the minority of women scheduled to have an annual screen).<sup>(17)</sup> 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 1,000 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 United States (US). Development of the algorithm has been described previously.<sup>(10)</sup> In brief, DeepHealth used a progressive, stage-wise 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,

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 before learning to scan an entire screen and make a global decision on whether a suspicious  
4 lesion is present. Convolutional neural networks (a deep learning approach to analysing  
5 visual data) were trained on five data sets from the US and UK, making use of both strongly  
6 and weakly labelled data. Australian data were not used for algorithm training; therefore,  
7 training data sets were *independent* of the data set used for the current external validation  
8 study. The trained algorithm outputs a “bounding box” identifying a region of interest (Figure  
9 2), along with a malignancy score quantifying the likelihood that the region of interest  
10 represents a malignancy. The algorithm evaluates each image in a study independently and  
11 aggregates the scores across all potential regions in the study to compute a single study-level  
12 malignancy score. The overall accuracy of the algorithm based on this study-level  
13 malignancy score has been compared with five individual radiologists, each fellowship-  
14 trained in breast imaging, on a cancer-enriched data set, and was shown to outperform all five  
15 readers. At the average radiologist specificity, the algorithm resulted in an absolute increase  
16 in sensitivity of 14.2%; at the average radiologist sensitivity, the absolute increase in  
17 specificity was 24.0%.<sup>(10)</sup> The algorithm also outperformed radiologists in detecting  
18 malignancy in a set of prior “normal” mammograms from the same set of cancer cases  
19 (increase in sensitivity 17.5%; increase in specificity 16.2%), demonstrating the potential to  
20 detect interval cancers “missed” by radiologists.  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44

45 All imaging analysis for the study will take place at BSWA to ensure security of images.  
46 Images will only be accessed by investigators who are employed by BSWA, and have such  
47 access under the usual conditions of their employment; these images will not be used for  
48 further refinement of DeepHealth’s algorithm. A laptop with the AI algorithm installed and a  
49 graphics processing unit supporting its evaluation will be located at BSWA. An external hard  
50 drive will be attached containing the cohort of digital mammogram data (DICOM files  
51 consisting of four views per breast, two breasts per woman). The algorithm will output data  
52  
53  
54  
55  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 to a csv file including bounding box coordinates, malignancy scores ranging from 0 to 1, and  
4  
5 a unique identifier extracted from the DICOM header to enable woman-level matching of  
6  
7 results to BSWA routine screening data.  
8  
9

**Data de-identification and secure storage**

10  
11  
12  
13  
14  
15  
16 De-identified data on cohort characteristics, screening findings, and cancer diagnosis will be  
17  
18 transferred by secure online file transfer to the Curtin School of Population Health, Curtin  
19  
20 University. No paper-based or portable electronic media storage of these data will take place.  
21  
22 Project data will be electronically stored on a secure server, which is backed up daily to  
23  
24 prevent any unintentional data loss. The research environment includes a variety of security  
25  
26 controls to restrict unauthorised access – these include access controls, role-based  
27  
28 delegations, encryption, firewalls, and physical access restrictions (authorised access to  
29  
30 server rooms and research offices is restricted by key). Automatic screen locking will occur  
31  
32 on electronic devices after five minutes of inactivity. Data will not be stored or used in public  
33  
34 terminals.  
35  
36  
37  
38  
39  
40  
41  
42

**Statistical methods**

43  
44  
45 All statistical analyses will be undertaken at the School of Public Health, Curtin University.  
46  
47 To descriptively compare the accuracy of AI with the *average* accuracy of single human  
48  
49 reading, a receiver operating characteristic (ROC) curve for the AI algorithm will firstly be  
50  
51 plotted from the algorithm's study-level malignancy score and the AUC-ROC derived. The  
52  
53 hierarchical summary ROC model proposed by Rutter and Gatsonis(18, 19) will be used to  
54  
55 model radiologist accuracy and derive an area under the summary ROC curve for radiologists  
56  
57 (using numerical integration), along with summary estimates of sensitivity and specificity.  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 The sensitivity and specificity of AI will be descriptively compared with that of radiologists  
4 by estimating the AI's sensitivity at the summary radiologist specificity, and the AI's  
5 specificity at the summary radiologist sensitivity. The malignancy score derived from the AI  
6 algorithm will also be dichotomised using a prospectively-defined threshold selected to  
7 reflect an expected recall rate of 4% (the overall recall rate of the BSWA program) based on  
8 DeepHealth's (non-Australian) validation data, allowing for a descriptive comparison of  
9 sensitivity and specificity at this threshold with summary radiologist estimates.

10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20 The CDR and case-specific recall rate of double-reading by radiologists (current population  
21 reading practice) will be compared with double-reading strategies integrating AI (McNemar's  
22 test), where the first radiologist read per screen will be paired analytically with AI. The  
23 following integrated AI-radiologist strategies will be used:

- 24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
1. Recall to assessment based on an "either positive" rule (i.e. 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 (i.e. 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 (i.e. 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

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 be conducted to apply a consistent 12-month follow-up period for ascertaining interval  
4  
5 cancers.  
6  
7  
8  
9

**Sample size and power calculation**

10  
11  
12  
13  
14 Power calculations were derived for the outcome of CDR, based on the sample size and  
15  
16 number of screen-detected and interval cancers present in the cohort. The CDR for double-  
17  
18 reading by radiologists in the study cohort is 7.0 per 1,000 screens. With a sample size of  
19  
20 109,000 unique screening examinations, at an alpha of 0.05 (two-sided) the study has 80%  
21  
22 power to detect an increase in CDR to 7.5 per 1,000 screens for integrated AI-radiologist  
23  
24 reading. This assumes concordance between the reading strategies of 5.5 cancers per 1,000  
25  
26 screens, with 1.5 cancers per 1,000 detected by radiologist double-reading only (and not by  
27  
28 integrated AI-radiologist reading), and 2.0 cancers per 1,000 screens detected by integrated  
29  
30 AI-radiologist reading only (and not by radiologist double-reading). This 1.5:2 ratio of  
31  
32 discordant cases is derived from a UK study comparing AI with radiologist double-  
33  
34 reading.<sup>(11)</sup>  
35  
36  
37  
38  
39  
40  
41  
42

**Sub-studies**

43  
44  
45 In addition to the primary study objectives, sub-studies will be undertaken to further explore  
46  
47 differences in accuracy observed in the main analyses. These will include:  
48  
49

- 50 1. Description of cancers for which there are discordant results (i.e. cancers detected by  
51  
52 the AI algorithm but not by radiologists, and vice versa), in terms of radiological and  
53  
54 cancer characteristics.  
55  
56  
57  
58  
59  
60



*Protocol for study of AI to enhance breast cancer screening*

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
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)

For peer review only

*Protocol for study of AI to enhance breast cancer screening***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 (e.g. data security and privacy). Consumer health representatives external to the research team have been engaged to provide community perspectives on this research (e.g. 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 program 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

1  
2  
3 be made publicly available at the summary level. Manuscripts addressing the study aims will  
4  
5 be published in peer-reviewed journals. Results will also be presented at relevant national and  
6  
7 international conferences. Study outcomes will also be disseminated to stakeholders in  
8  
9 Australian breast cancer screening programs and policy makers in population screening, to  
10  
11 inform future evaluation and policy discussions about the potential implementation of AI.  
12  
13  
14  
15  
16  
17

## 18 **DISCUSSION**

19  
20 Organised population breast screening programs are facing growing screen-reading resource  
21  
22 challenges, so the current global research effort aimed at developing and testing AI  
23  
24 algorithms for interpreting screening mammograms can contribute to ensuring future  
25  
26 sustainability of screening. Although the field is rapidly-evolving, to date there has been a  
27  
28 focus on algorithm development with relatively few studies evaluating AI in real-world breast  
29  
30 cancer screening settings. A scoping review of the literature on AI for breast screening  
31  
32 identified eight key deficiencies of the evidence base (Table 1), and concluded that although  
33  
34 studies indicate a potential role of AI in this clinical scenario, those evidence gaps should be  
35  
36 addressed prior to the initiation of prospective trials and the adoption of the technology in  
37  
38 routine practice.(13). The primary concerns raised relate to the quality of datasets used to  
39  
40 validate AI models and the paucity of evidence comparing the accuracy of AI and  
41  
42 radiologists, potentially affecting the applicability and robustness of AI algorithms and  
43  
44 raising the possibility of bias. The study we present in this protocol addresses those evidence  
45  
46 gaps by comparing the accuracy of a commercially available algorithm with that of  
47  
48 radiologists using a large, external validation dataset representing consecutive, *unselected*  
49  
50 digital mammograms from a real-world screening program (Table 1). This retrospective  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1 cohort study is therefore an essential step to build the evidence base to underpin prospective  
2  
3 trials and inform their design, and to provide timely evidence to screening stakeholders.  
4  
5  
6  
7

8 Although this study will overcome most key limitations of the evidence base, there are  
9  
10 potential limitations associated with its retrospective design. Data collected for administrative  
11  
12 purposes may be more prone to misclassification than data collected specifically for research  
13  
14 purposes through a prospective trial. For instance, we have excluded women from the study  
15  
16 cohort who relocated outside WA after the index screening examination and therefore were  
17  
18 potentially lost to follow-up. Since the date of relocation is not routinely collected, it is  
19  
20 possible that some women with complete follow-up were excluded. Given that exclusions for  
21  
22 relocation represented <0.6% of women during the study period (Figure 1), this is unlikely to  
23  
24 represent a significant concern. Data on outcomes (recalls, screen-detected and interval  
25  
26 cancers) are meticulously collected according to national quality and accreditation standards  
27  
28 and are therefore unlikely to be subject to misclassification. Furthermore, we have defined  
29  
30 the end date for study enrolment (31 December 2016) to ensure completeness of notifications  
31  
32 for interval cancers (while simultaneously ensuring a contemporary cohort that is  
33  
34 representative of the current target population for breast cancer screening in Australia). Errors  
35  
36 in the classification of outcome data are therefore considered to be rare.  
37  
38  
39  
40  
41  
42

43 To estimate CDR and recall rate for integrated AI-human reading, we will take an analytic  
44  
45 approach to combining AI and radiologist findings. This pragmatic approach is dictated by  
46  
47 the retrospective study design; however, it may not be representative of how AI screening  
48  
49 results might be incorporated into practice. Our decision rules for defining recall to further  
50  
51 assessment are among several proposed uses of AI information. Some alternative approaches  
52  
53 (such as the use of AI to “triage” women to double-reading if exceeding a threshold  
54  
55 probability of malignancy(11)) may potentially be investigated analytically by our study  
56  
57  
58  
59  
60

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 design, but others (such as AI output used by radiologists interactively as a decision  
4 support(6)) can only be evaluated in studies using a prospective design. Furthermore, the  
5  
6 methods adopted to derive summary ROC curves for radiologists and associated measures of  
7  
8 accuracy are dictated by the retrospective design. Although these methods are established and  
9  
10 appropriate for our real-world screening data,(18) they allow only for descriptive  
11  
12 comparisons with empirical estimates for the AI algorithm.(19)  
13  
14  
15

16  
17 The lack of studies exploring social and ethical issues, particularly women's perspectives and  
18 preferences around AI, has been identified as a critical evidence gap (Table 1). Although  
19  
20 beyond the scope of this study, a parallel research stream using qualitative methods is being  
21  
22 undertaken by some of the study authors to elucidate those perspectives. For instance, women  
23  
24 will be provided with information about potential uses of AI in breast screening, and will then  
25  
26 discuss this potential implementation, with a focus on what matters most to them, and how  
27  
28 implementation should (or should not) take place. Similarly, economic modelling to estimate  
29  
30 incremental costs and benefits from the use of AI is critical to informing policy decisions  
31  
32 about adopting the technology. Cost-effectiveness analysis will be undertaken in a future  
33  
34 project building on the results of this study.  
35  
36  
37  
38  
39

40  
41 AI algorithms for interpreting mammograms have the potential to improve the effectiveness  
42  
43 of population breast cancer screening programs if they can detect cancers, including interval  
44  
45 cancers, without contributing substantially to overdiagnosis. This will be the largest study to  
46  
47 date to investigate the accuracy of an artificial intelligence algorithm for interpreting  
48  
49 consecutive digital mammograms in a population-based breast cancer screening program.  
50  
51 The evidence generated by this study can be used to inform decisions about adopting AI for  
52  
53 mammogram interpretation in the future, to improve accuracy, effectiveness, and efficiency.  
54  
55  
56  
57  
58  
59  
60

**List of abbreviations**

AI – artificial intelligence

AUC-ROC – area under the receiver operating characteristic curve

BSWA – BreastScreen WA

CAD – computer-aided detection

CDR – cancer detection rate

DCIS – ductal carcinoma *in situ*

ROC – receiver operating characteristic

WA – Western Australia

**DECLARATIONS****Ethics**

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 program provide written consent for their data to be used for research purposes each time they screen.

**Competing interests**

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

**Funding**

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 This work was supported by funding from a National Breast Cancer Foundation Investigator  
4 Initiated Research Scheme grant (IIRS-20-011 to MLM, NH, EW, BL, SMC and AP). NH  
5 receives funding from the National Health & Medical Research Council (NHMRC)  
6 (Investigator/Leader Grant #1194410). SMC receives funding from the NHMRC (Ideas Grant  
7 #1181960). GP receives funding from the NHMRC (Project Grant #1099655 and Investigator  
8 Grant #1173991) and the Research Council of Norway through its Centres of Excellence  
9 funding scheme (#262700). CIL and WL are funded in part by the National Cancer Institute  
10 (R37 CA240403).  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20

**Author contributions**

21  
22  
23 MLM, NH, EW, HL, AW and WL conceived the idea, planned and designed the study  
24 protocol. AP, SMC, MB, GP, JGK, CIL, and SZ contributed to the development of the  
25 protocol, study design and methods. MLM wrote the first draft. NH, EW, HL, AW, WL, AP,  
26 SMC, MB, GP, JGK, CIL, and SZ critically revised the draft for important intellectual  
27 content. All authors have approved the final written manuscript.  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**References**

1. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *British journal of cancer*. 2013;108(11):2205-40.
2. Hanley JA, Hannigan A, O'Brien KM. Mortality reductions due to mammography screening: Contemporary population-based data. *PloS one*. 2017;12(12):e0188947-e.
3. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *npj Breast Cancer*. 2017;3(1):12.
4. Baré M, Torà N, Salas D, Sentís M, Ferrer J, Ibáñez J, et al. Mammographic and clinical characteristics of different phenotypes of screen-detected and interval breast cancers in a nationwide screening program. *Breast Cancer Research and Treatment*. 2015;154(2):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, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute*. 2019;111(9):djj222.
7. Harvey H, Karpati E, Khara G, Korkinof D, Ng A, Austin C, et al. The Role of Deep Learning in Breast Screening. *Current Breast Cancer Reports*. 2019;11(1):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, Kim JG, Grisot G, Wu E, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine*. 2021;27(2):244-9.
11. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94.
12. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncology*. 2020;6(10):1581-8.
13. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Review of Medical Devices*. 2019;16(5):351-62.
14. Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*. 2013;185(11):E537-E44.
15. Park SH. Diagnostic Case-Control versus Diagnostic Cohort Studies for Clinical Validation of Artificial Intelligence Algorithm Performance. *Radiology*. 2019;290(1):272-3.
16. Elmore JG, Lee CI. Artificial Intelligence for Breast Cancer Imaging: The New Frontier? *JNCI: Journal of the National Cancer Institute*. 2019;111(9):djj223.
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(19):2865-84.

*Protocol for study of AI to enhance breast cancer screening*

1  
2  
3 19. Oakden-Rayner L, Palmer L. Docs are ROCs: A simple off-the-shelf approach for  
4  
5 estimating average human performance in diagnostic studies. 2020 Available from:  
6  
7 arXiv:2009.11060v2 [stat.ME]. Accessed 27 Oct 2021.  
8  
9

10 20. Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic  
11  
12 performance when combining two diagnostic tests. *Statistics in Medicine*. 2002;21(17):2527-  
13  
14 46.  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

**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 US.
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 data sets were used to train and validate models. Validation using independent, external data sets 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 program (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 programs.

1  
2  
3 **Figure 1:** Flowchart of cohort inclusions and exclusions  
4  
5  
6

7 **Figure 2:** Digital mammogram mediolateral oblique view with region of interest (denoted by  
8 bounding box) identified by the AI algorithm as suspicious for malignancy. Cancer was  
9 confirmed as invasive ductal carcinoma.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

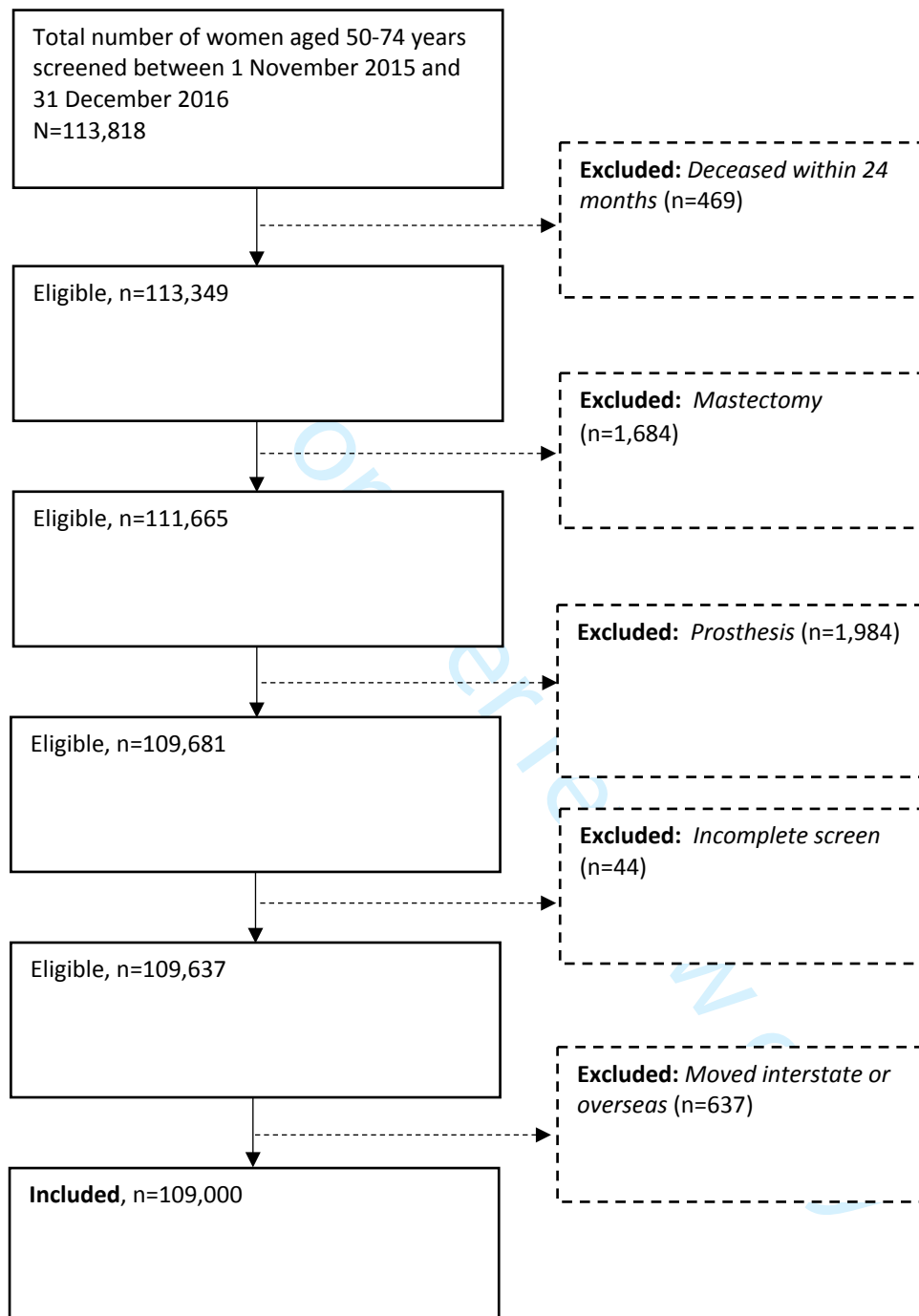




Figure 2: Digital mammogram mediolateral oblique view with region of interest (denoted by bounding box) identified by the AI algorithm as suspicious for malignancy. Cancer was confirmed as invasive ductal carcinoma.

1174x1444mm (72 x 72 DPI)