BMJ Open

Protocol for evaluating the fitness for purpose of an artificial intelligence product for radiology reporting in the BreastScreen New South Wales breast cancer screening programme

Matthew Warner-Smith ,[1] Kan Ren,[1] Chirag Mistry,[1] Richard Walton,[1] David Roder ,[2] Nalini Bhola,[1] Sarah McGill,[1] Tracey A O'Brien[1,3]

To cite: Warner-Smith M, Ren K, Mistry C, *et al*. Protocol for evaluating the fitness for purpose of an artificial intelligence product for radiology reporting in the BreastScreen New South Wales breast cancer screening programme. *BMJ Open* 2024;**14**:e082350. doi:10.1136/bmjopen-2023-082350

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2023-082350).

Received 21 November 2023
Accepted 09 May 2024

[1]Cancer Institute NSW, St Leonards, New South Wales, Australia
[2]Cancer Research Institute, University of South Australia, Adelaide, South Australia, Australia
[3]UNSW Medicine & Health, Sydney, New South Wales, Australia

Correspondence to
Matthew Warner-Smith;
Matthew.WarnerSmith@health.nsw.gov.au

ABSTRACT

Introduction Radiologist shortages threaten the sustainability of breast cancer screening programmes. Artificial intelligence (AI) products that can interpret mammograms could mitigate this risk. While previous studies have suggested this technology has accuracy comparable to radiologists most have been limited by using 'enriched' datasets and/or not considering the interaction between the algorithm and human readers. This study will address these limitations by comparing the accuracy of a workflow using AI alongside radiologists on a large consecutive cohort of examinations from a breast cancer screening programme. The study will combine the strengths of a large retrospective design with the benefit of prospective data collection. It will test this technology without risk to screening programme participants nor the need to wait for follow-up data. With a sample of 2 years of consecutive screening examinations, it is likely the largest test of this technology to date. The study will help determine whether this technology can safely be introduced into the BreastScreen New South Wales (NSW) population-based screening programme to address radiology workforce risks without compromising cancer detection rates or increasing false-positive recalls.

Methods and analysis A retrospective, consecutive cohort of digital mammography screens from 658 207 examinations from BreastScreen NSW will be reinterpreted by the Lunit Insight MMG AI product. The cohort includes 4383 screen-detected and 1171 interval cancers. The results will be compared with radiologist single reading and the AI results will also be used to replace the second reader in a double-reading model. New adjudication reading will be performed where the AI disagrees with the first reader. Recall rates and cancer detection rates of combined AI–radiologist reading will be compared with the rates obtained at the time of screening.

Ethics and dissemination This study has ethical approval from the NSW Health Population Health Services Research Ethics Committee (2022/ETH02397). Findings will be published in peer-reviewed journals and presented at conferences. The findings of this evaluation will be provided to programme managers, governance bodies and other stakeholders in Australian breast cancer screening programmes.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This is likely the most extensive evaluation of this technology on real-world population screening data and will include cancers detected at the time of screening and with 24 months of follow-up in the entire cohort of clients screened by the New South Wales breast cancer screening programme.
⇒ The study will combine the strengths of a large retrospective design with the benefit of prospective data collection in a real-world clinical workflow, providing a real-world test of this technology without risk to screening programme participants nor the need to wait for long-term follow-up data.
⇒ The consecutive cohort will overcome limitations of previous studies that have used 'cancer enriched' datasets, resulting in accuracy estimates that will be generalisable to screening programmes and enabling the estimation of population-based screening outcome metrics.
⇒ The generalisability of the study will be limited to programmes using similar screening workflows and similar mammography technologies.
⇒ This study will not investigate the acceptability of the use of this technology to clients of screening programmes and the clinicians that work in them.

INTRODUCTION

Breast cancer was the second most commonly diagnosed cancer worldwide in 2022 after lung cancer, accounting for 11.6% of diagnoses. Globally, there were over 2.3 million new cases and 665 000 deaths in 2022.[1] This is predicted to increase to over 3 million new cases and 1 million deaths in 2040.[2]

Breast cancer is the second most common cancer recorded by the New South Wales (NSW) Cancer Registry in Australia, with

approximately 6000 new diagnoses (65.4 per 100 000) annually, and the fifth most common cause of cancer mortality.[3] Population-based breast cancer screening programmes have been shown to identify cancers at an earlier stage,[4] enable breast-conserving surgery[5] and reduce breast cancer-specific mortality.[6]

The BreastScreen NSW (BSNSW) breast cancer screening programme provides routine screening to asymptomatic women aged over 40. The standard of care in the programme is radiological reading of mammograms by two independent readers, with adjudication by a third reader in the event of discordant findings. The first two readers are blinded to each other's results. The adjudication reader has access to the results of the first two readers. The majority of readers are radiologists with a small number of breast physicians, both with specialisation in mammography This workflow is heavily reliant on the radiology workforce.

The sustainability of population-based breast cancer screening programmes is at risk due to increasing demand for radiologists which is not matched by the supply of newly qualified radiologists.[7] As in the UK and Europe, ensuring an adequate workforce for interpreting screening results in Australia is becoming challenging for publicly funded screening programmes, particularly in areas with shortages of skilled readers.[8] The Workforce Survey Report from the Royal Australian and New Zealand College of Radiologists identifies screening mammography as a field facing a 'significant risk of workforce shortage', a shortfall expected to worsen over time.[8] The emergence of artificial intelligence (AI) may mitigate this issue by enhancing the efficiency of screen reading. Software products that use image recognition algorithms based on deep learning could replace a portion of the human reading in a multireader operational model such as that used across Australia or to complement radiologist reading.[9]

Studies evaluating such products for breast cancer screening suggest the technology can achieve accuracy comparable to expert radiologists.[10] However, there are limitations in these studies which prevent their generalisability to screening programmes.

The first limitation is that to overcome statistical power limitations resulting from the relative rarity of breast cancer, validation studies often use data enriched with a greater prevalence of cancer than seen in the general population.[9] Marinovich *et al* (2022) addressed this by using a consecutive cohort of 13 months of mammography screens from the BreastScreen Western Australia programme. The study found that when using a recall threshold that achieved a sensitivity equivalent to radiologists, the AI product (Saige-Q V.2.0.0) had substantially lower specificity (81% AI; 98% radiologists).

A second limitation of validation studies is that they have sought to demonstrate the accuracy of these products relative to the accuracy of a skilled radiologist, and not tested the product in a multireader workflow.[9] This does not provide sufficient evidence to support the adoption of this technology in screening as it fails to consider differences in the characteristics of cases that the machine can identify relative to those that radiologists can identify. Two recent studies have attempted to address both these limitations by incorporating an AI product into the reading workflow of a screening programme.[11 12]

Marinovich *et al*[11] used a retrospective design, replacing one of the first two readers in a double-blind workflow with adjudication reading with an AI product ('simulated AI-radiologist reading'). They assessed the performance of AI-radiologist reading relative to the performance of the original reading (two radiologists with a third adjudicating radiologist where necessary). The study found that using AI at a recall threshold that achieved sensitivity comparable to radiologists resulted in a fourfold increase in arbitration reading due to the poorer specificity of the AI product, but still achieved a 41% reduction in total radiologist reads. At all three AI recall thresholds tested the simulated AI-radiologist reader had significantly poorer cancer detection than the standard radiologist workflow.

However, the study used the reading result of the original second reader for adjudication for examinations where the AI product disagreed with the radiologist result (first reader). This is problematic because the first and second reads are not independent when considering the final reading outcome. Using this design, the AI was only able to influence the small minority (5%; 5145 of 108 970) of examinations that originally went to arbitration. The only influence the AI could have had was to overturn the original arbitration result in favour of the result of the second reader. Any cancer that the AI identified and which neither of the first two readers identified could not be recalled. In examinations where the AI read overturned the result of the arbitration read, it is more likely that the arbitration result was correct than the second reader's result, since (a) arbitration reading is unblinded to the first two reads and is effectively the collective result of three radiologists' opinions, and (b) arbitration readers are typically the most experienced readers in a screening programme. This explains why the authors found that AI-radiologist reading had significantly worse cancer detection and that most interval cancers that were detected by AI were 'arbitrated out'.

The need for evidence of the efficacy of AI in a combined AI-radiologist reading workflow and the requirement for independent arbitration between the AI result and the radiologist result has prompted calls for prospective studies. Lång *et al*[12] recently published the first randomised controlled trial of the use of AI in breast cancer screening, in which AI reading was used to triage examinations to single or double radiologist reading. The study concluded that AI-supported reading was non-inferior to exclusively radiologist reading, with comparable cancer detection rates between both groups. The trial found that the AI-supported arm reduced radiologist readings by 44%. However, the final results of the trial

will not be known until follow-up has been completed to determine interval cancer rates.

This paper presents a study protocol addressing the challenges of assessing real-world performance of machine reading by using retrospective unenriched data from a continuous cohort of examinations with interval cancer follow-up from a population-based screening programme with prospective independent adjudication reading by experienced radiologists. The protocol allows for evaluating an AI product's performance and determining the probable effects of introducing machine reading into a population-based breast cancer screening programme.

This project aims to compare AI reading of digital mammograms with the BSNSW standard of care in a real-world, population-based breast cancer screening setting. It will determine whether using AI in place of one of the first two readers results in inferior recall rates, cancer detection rates or interval cancer rates.

Specifically, the study will:

1. Compare the accuracy of AI with the average accuracy of a single human reading as assessed by a range of measures including sensitivity, specificity and positive and negative predictive value (PPV/NPV).
2. Compare the accuracy of standard reading in the BSNSW programme (which consists of double-blind reading with independent adjudication reading in the event of disagreement) with the accuracy of reading using AI instead of a radiologist as one of the first two readers. This comparison will be presented using conventional screening metrics including radiology reading workloads; the number of examinations requiring diagnostic assessment; cancer detection rates; and interval cancer rates.

METHODS AND ANALYSIS

Study design

A retrospective study design will be combined with prospective adjudication reading on a cohort of unique, consecutive, digital mammography screens from BSNSW, the population-based breast cancer screening programme in NSW, Australia. The study will avoid biases identified in previous research on AI[13] for mammography screening by using all screening examinations in a defined time period with verified outcomes including 2-year follow-up of interval cancers. This will provide a sample that is representative of real-world screening populations.

The mammograms for all examinations in the study cohort will be processed (re-read) by the Lunit Insight MMG V.1.1.7.2 product. Insight MMG is certified by the Australian Therapeutic Goods Administration as a 'software device that assists physicians in the interpretation of mammograms. The device has been designed to automatically analyse digital mammograms via deep learning technology' (TGA certificate DV-2020-MC-26241-1, ARTG identifier 345076).

In examinations where Insight MMG and the original first reader disagree (are 'discordant'), and where an adjudication read was done at the time of screening, the original adjudication read will be used, but only if the abnormality that was adjudicated initially is in the same location as the abnormality recalled in the study. For examinations that did not have an eligible adjudicating read at the time of screening, independent prospective adjudication reading will be done by specialised breast radiologists who regularly perform adjudicating readings in BSNSW.

Study cohort characteristics

The cohort comprised all clients who attended screening at BSNSW between 1 January 2016 and 31 December 2017. All screening episodes in this period will be assessed (ie, a person could be included more than once in the cohort), with reading outcomes, episode outcomes and interval cancer follow-up specific to each episode. No exclusion criteria will be applied.

The cohort contains 658 207 screening episodes of 626 851 people, with 92 910 (14%) being a first-time screen and the remainder being second or subsequent screens. The mean age of the cohort at the time of screening was 60.7 years.

33.6% of the cohort was born in a country other than Australia. 1.6% of the cohort identified as indigenous Australians (Aboriginal and Torres Strait Islander).

The cohort includes 5554 cases of cancer. Of these, 4383 were cancers detected during the screening episode, with a further 1171 being interval cancers. Of the screen-detected cancers, 3558 were invasive breast cancers, while 825 were cases of ductal carcinoma in situ.

Insight MMG analysis

All mammograms for the date range of the study will be analysed by Lunit Insight MMG and the results will be recorded and provided in a data file to the researchers (figure 1). The results for each examination will include:

- A measure of breast density on a scale of 1–10 (most dense).
- An abnormality score of 0–100 representing the degree of confidence that the abnormality identified was a malignant lesion for each abnormality identified and the view and laterality of that view.
- The highest abnormality score identified in the examination, the view it was identified in and the laterality of that view.

To compare the Insight MMG result to human readers' results, the highest abnormality score will be converted from a continuous variable representing the degree of confidence into a categorical variable of Recall or Not Recall. This will require selecting an abnormality score threshold above which the examination was deemed to be recalled.

For the purposes of this analysis, if there is at least one lesion in the same breast (left or right) that was recalled, the case will be considered diagnosed. In practice, additional diagnostics would be undertaken before a definitive diagnosis was given.

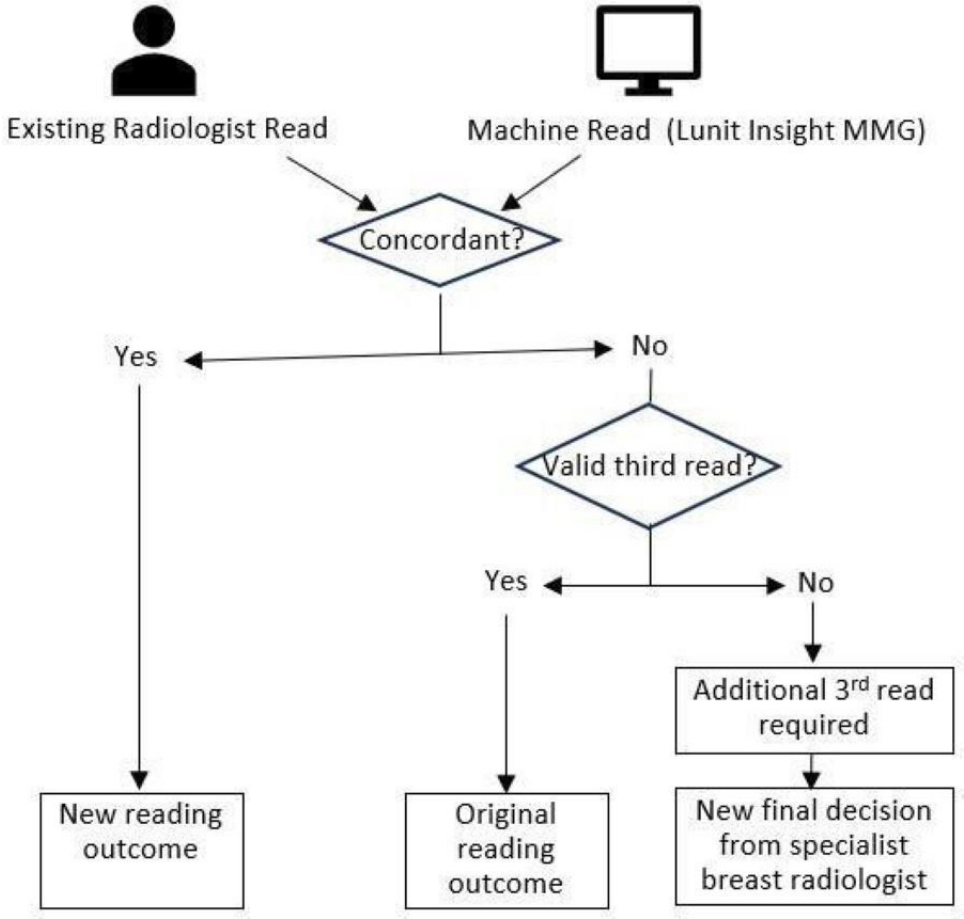

Figure 1 Study workflow.

Threshold determination

Three abnormality thresholds will be tested. These are defined as:

- The score at which approximately 10% of examinations are recalled.
- The score which provides approximately equivalent sensitivity to human readers.
- The score which provides approximately equivalent specificity to human readers.

A recall rate of 10% was selected to maximise cancer detection within the recall limits set by the National Accreditation Standards (NAS) of the BreastScreen Australia programme. The Insight MMG product does not compare the mammogram being analysed to prior mammograms of that person, hence the NAS standard for first-round screening was used.

The abnormality thresholds which provide equivalent sensitivity and specificity to human readers were chosen, as these are the thresholds which will most closely replicate standard screening practice. Ideally, a single threshold would provide equivalent sensitivity and specificity to human readers, However, it is recognised that increasing sensitivity comes at the expense of decreased specificity and visa versa. Choosing the three thresholds will allow for determination of the degree of compromise that would result from prioritising one over the other.

The scores will be determined using the raw results of all examinations for the study period before any cleansing or analysis.

Identification of examinations that require adjudication reading

For those examinations that are discordant between Insight MMG and the first reader, new adjudication reading will be required. Where examinations have multiple abnormalities, the laterality of every abnormality identified by Insight MMG will be compared against the laterality of every abnormality recalled by the first reader. Where tone or more abnormalities on the same side were recalled by both Insight MMG and the first reader, these examinations will be considered a concordant recall for further assessment. If Insight MMG identifies one or more abnormalities in only one breast and the first reader only identifies one or more abnormalities in the opposite breast, the examination will be deemed discordant and adjudication reading will be performed.

To reduce the adjudication workload, any adjudication reading done at the time of the original reading of that screening episode will be reused if it is deemed valid. If an adjudication read was performed at the time of screening, the location of any abnormality recalled by the original second reader (ie, the read that is being replaced by the

AI) will be compared with the location of the abnormality with the highest score in examinations recalled by Insight MMG. If the same abnormality is being recalled, the original adjudicating read will be used to adjudicate between Insight MMG and the first reader. If the abnormalities are different, the original adjudicating read will be deemed to not be specific to the recalled abnormality. For those examinations, a new adjudicating read will be performed.

Adjudication reading

Nine specialist breast radiologists will perform the new adjudicating reading required for the study. The radiologists are experienced at performing adjudicating reading and are current adjudicating readers for the BSNSW programme. The adjudication reading for the study will mimic the adjudication reading that is done by BSNSW. Reading will be done on the same workstations with diagnostic quality monitors that are used for adjudicating reading in BSNSW. Readers will view examinations in the study Picture Archiving and Communications System (PACS), which has the same configuration as the PACS used for BSNSW reading. Alongside the PACS, the adjudicating readers will use a webform developed for the study that mimics the radiology information system used for reading in BSNSW (the BSNSW Information System (BIS)). The webform will provide the same information to the study readers that adjudicating readers are supplied with in the BSNSW programme. This includes the results of the first two readers with the location of the abnormality presented on an annotated diagram; a diagram of surface lesions and scars recorded by the radiographer at the time of screening; and any symptoms or history that were collected from the client at the time of screening.

The webform will include a reading list that the adjudicating readers select from and will be integrated with the PACS using HL7 messaging to ensure that the presentation of each examination in the PACS is synchronised with the examination that is being viewed in the webform (viewed 'in context').

Images for all examinations deemed to require new adjudication reading will be held in a dedicated instance of the BSNSW PACS. For each examination requiring adjudication reading, all mammograms taken prior to the original reading will be held in the same PACS. Prior mammograms for that client will be reviewed during adjudication reading, as they are in standard reading in the screening programme.

Study dataset

A relational database has been created to store the study dataset. The dataset will contain data linked by accession and study identifiers, and will include:

- The deidentified clinical and demographic details of each examination collected at the time of screening taken from the BSNSW BIS.
- The pathology details of each cancer diagnosed either through the programme during that screening episode or outside the programme in the 24 months after that episode (interval cancer) obtained from BIS and the NSW Cancer Registry.
- The results from the analyses by Insight MMG.
- The results of the new adjudicating reading.

Analysis

The analysis for the study will be conducted in two parts. In the first part, the performance of Insight MMG on the entire study cohort and selected subgroups within the study cohort will be compared with the performance of BSNSW readers.

The analysis will consider the results of three readers: the group of radiologists and breast physicians who completed the first read ('Reader A'); the group of radiologists and breast physicians who completed the second read ('Reader B'); and Insight MMG. This analysis will assess performance as an independent 'stand-alone' reader.

The second part of the analysis will compare the overall read outcome at the time of screening (the result of first two readers and the adjudicating read if there was one), with the modelled outcome from the study which combines the original first reader with Insight MMG in place of the second reader and any associated adjudicating read (either an eligible adjudicating read from the time of screening or a new study adjudicating read).

Subgroup analyses will consider variations in the performance of the AI and AI-human ensemble relative to human reading by:

- Age.
- Ethnicity.
- Aboriginality.
- Tumour characteristics.
- Breast cancer risk.
- First screening round versus those who have had at least one prior screen.
- Breast density.
- Screening modality.

The analysis of both the stand-alone performance of Lunit Insight MMG and the adjudicated double reading workflow with Reader A and Lunit Insight MMG will seek to demonstrate non-inferiority relative to standard practice using predetermined non-inferiority margins.

Non-inferiority margins were determined based on levels of variation in sensitivity and specificity currently observed in the programme, along with clinical and operational considerations. CIs to assess non-inferiority will be calculated for sensitivity and specificity using a methodology for binary diagnostic tests with paired data.[14]

With an observed sensitivity for Reader A of 68.9%, the associated non-inferiority margin was derived as 62.4% (OR=0.75). With an observed specificity for Reader A of 95.5%, the associated non-inferiority margin was derived as 94.4% (OR=0.8).

With an observed sensitivity for the human ensemble of 79.9% the associated non-inferiority margin was derived as 74.8% (OR=0.75). With an observed specificity for

human ensemble of 96.5% the associated non-inferiority margin was derived as 95.7% (OR=0.8).

The stand-alone analysis will compute the sensitivity and specificity of the AI and of all pooled readers in the BSNSW programme, weighted by their reading volume. In addition, the following metrics will be computed: PPV, NPV, recall rate, balanced accuracy, Matthew's correlation coefficient and F1 score. Formal comparisons will be made accounting for the paired nature of the comparisons.

The ensemble analysis will similarly compute the sensitivity and specificity of combined AI and radiologist reading and will compare that to the sensitivity and specificity achieved in standard double reading by radiologists.

Programme outcomes relative to the counterfactual will be computed and compared, including: third read volumes, assessments, cancers detected, interval cancers and overall programme costs (reading and assessment).

Patient and public involvement

None.

ETHICS AND DISSEMINATION

This study has ethical approval from the NSW Health Population Health Services Research Ethics Committee (2022/ETH02397).

All identifying information was removed from DICOM headers prior to processing by Lunit Insight MMG. Records were linked using PACS accession numbers. No personal identifying details were extracted from the clinical information system when the analytics dataset was created. The extraction from the clinical information system and the deidentification of DICOM headers was done by individuals outside the research team who are authorised to manage identified data in those systems. The researchers did not have access to any identified data.

Findings will be published in peer-reviewed journals and presented at conferences. The findings of this evaluation will be provided to programme managers, governance bodies and other stakeholders in Australian breast cancer screening programmes.

DISCUSSION

This study will address several of the main limitations of previous studies as articulated by Houssami *et al*.[9] Unlike many previous studies, this study will use a commercially available product that has regulatory approval in Australia and other markets for decision support for radiology reporting of mammograms.

Whereas previous studies have typically used relatively small datasets, this study will use a large dataset comprising over 650 000 screening episodes. By comparison, McKinney *et al*[10] used a dataset with just under 29 000 examinations and Marinovich *et al*[11] used a dataset with 109 000 examinations.

The dataset that will be used is entirely independent of the datasets used to train the algorithm and of the datasets used to validate the algorithm and obtain regulatory approval. The dataset is representative of 2 years of screening in BSNSW. The BSNSW programme uses a 2-year screening interval, and therefore the cohort is effectively the entirety of clients who were actively participating in the programme at the time the study data were collected.

In addition to enabling generalisability, the use of a large continuous cohort of screening examinations avoids the need to enrich the dataset with additional cancers to achieve statistical power. The prevalence of cancer in the study cohort of 0.8% (5554 cancers in 658 207 examinations) is equivalent to the prevalence of cancer in the screening population. By comparison, the dataset used by McKinney *et al*[10] had a cancer prevalence of almost 4% (1100 cancers in 28 953 examinations).

With a small number of exceptions,[11 15 16] previous studies have validated the performance of AI as standalone reader but have not considered the interaction between the AI and human readers in a typical screen reading workflow. This study will use conventional screening metrics alongside performance metrics to assess the probable effect of using Insight MMG in a population-based screening programme. While studies by Marinovich *et al* (2023) and others[16] are limited by the use of an existing read as the adjudicating read, this study will include valid arbitration reading independent of the original read result of the examination. While the study by Lång *et al*[12] used a robust prospective design, it is limited by insufficient follow-up to assess interval cancers. In contrast, this study will have 2 years of interval cancer follow-up. The unique study design of a retrospective cohort with prospective adjudication allows sensitivity and specificity to be calculated precisely in addition to cancer detection rates while also accurately assessing how the AI technology integrates with human readers in a double reading system.

The principal limitation of this study is that it will not assess the performance of the AI on three-dimensional images taken using digital breast tomosynthesis. This is an area that warrants further investigation. Despite the potential for tomosynthesis to increase cancer detection rates, the extended reading time required by radiologists has made using tomosynthesis in population screening programmes cost prohibitive. Using this technology to reduce the number of radiologists required to view each examination may make it cost-effective.

Another limitation is that the study uses examinations taken 6–7 years ago. Imaging technology has improved since. This may limit the generalisability of the study to the imaging currently performed within the programme. Furthermore, the distribution of modality brands and models in the BSNSW fleet has changed. If differences in performance are observed in different brands of modality the results may not be directly applicable to the current programme.

Finally, this study will not investigate the acceptability of the use of this technology to clients of screening programmes and the clinicians that work in them.

Future research should explore the aforementioned limitations. In particular, the results of this study will need to be validated on contemporary breast imaging technology, and the potential for the use of AI on three-dimensional breast imaging should be investigated.

Acknowledgements The authors thank Associate Professor Magnus Dustler and Professor Stephen Duffy for providing independent peer review of the study protocol.

Contributors MW-S, KR, CM, RW and DR planned and designed the study protocol. SM, NB and TAO'B contributed to the development of the protocol, study design and methods. MW-S, CM, KR, RW, DR, SM, NB and TAO'B critically revised the draft. All authors have approved the final written manuscript.

Funding This programme of work is entirely conducted and supported by the Cancer Institute NSW, a pillar of NSW Health dedicated to state-wide cancer control.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Matthew Warner-Smith http://orcid.org/0009-0003-7056-2551
David Roder http://orcid.org/0000-0001-6442-4409

REFERENCES

1 Bray F, Laversanne M, Sung H, *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;74:229–63.
2 Arnold M, Morgan E, Rumgay H, *et al*. Current and future burden of breast cancer: global statistics for 2020 and 2040. *Breast* 2022;66:15–23.
3 Cancer Institute NSW. Cancer statistics. 2023. Available: https://www.cancer.nsw.gov.au/research-and-data/cancer-data-and-statistics/data-available-now/cancer-statistics-nsw/cancer-incidence-and-mortality
4 Tong S, Warner-Smith M, McGill S, *et al*. Effect of mammography screening and sociodemographic factors on stage of female breast cancer at diagnosis in New South Wales. *Aust Health Rev* 2020;44:944–51.
5 Shahabi-Kargar Z, Johnston A, Warner-Smith M, *et al*. Differences in breast cancer treatment pathways for women participating in screening through breastscreen New South Wales (BSNSW). *AMJ* 2020;13:189–200.
6 Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* 2012;380:1778–86.
7 Royal College of Radiologists. Clinical radiology census report 2021. 2022. Available: https://www.rcr.ac.uk/2022-cr-census
8 Royal Australian and New Zealand College of Radiologists. 2020 RANZCR clinical radiology workforce census report: Australia. 2023. Available: https://www.ranzcr.com/doclink/2020-workforce-census-report-australia
9 Houssami N, Kirkpatrick-Jones G, Noguchi N, *et al*. Artificial intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices* 2019;16:351–62.
10 McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.
11 Marinovich ML, Wylie E, Lotter W, *et al*. Artificial intelligence (AI) to enhance breast cancer screening: protocol for population-based cohort study of cancer detection. *BMJ Open* 2022;12:e054005.
12 Lång K, Josefsson V, Larsson A-M, *et al*. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol* 2023;24:936–44.
13 Freeman K, Geppert J, Stinton C, *et al*. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy BMJ. *BMJ* 2021;374:n1872.
14 Roldán-Nofuentes JA. Compbdt: an R program to compare two binary diagnostic tests subject to a paired design. *BMC Med Res Methodol* 2020;20:143.
15 Leibig C, Brehmer M, Bunk S, *et al*. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *Lancet Digit Health* 2022;4:e507–19.
16 Sharma N, Ng AY, James JJ, *et al*. Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. *BMC Cancer* 2023;23:460.