

UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening

Sian Taylor-Phillips, Farah Seedat, Goda Kijauskaite, John Marshall, Steve Halligan, Chris Hyde, Rosalind Given-Wilson, Louise Wilkinson, Alastair K Denniston, Ben Glocker, Peter Garrett, Anne Mackie, Robert J Steele



Artificial intelligence (AI) could have the potential to accurately classify mammograms according to the presence or absence of radiological signs of breast cancer, replacing or supplementing human readers (radiologists). The UK National Screening Committee's assessments of the use of AI systems to examine screening mammograms continues to focus on maximising benefits and minimising harms to women screened, when deciding whether to recommend the implementation of AI into the Breast Screening Programme in the UK. Maintaining or improving programme specificity is important to minimise anxiety from false positive results. When considering cancer detection, AI test sensitivity alone is not sufficiently informative, and additional information on the spectrum of disease detected and interval cancers is crucial to better understand the benefits and harms of screening. Although large retrospective studies might provide useful evidence by directly comparing test accuracy and spectrum of disease detected between different AI systems and by population subgroup, most retrospective studies are biased due to differential verification (ie, the use of different reference standards to verify the target condition among study participants). Enriched, multiple-reader, multiple-case, test set laboratory studies are also biased due to the laboratory effect (ie, radiologists' performance in retrospective, laboratory, observer studies is substantially different to their performance in a clinical environment). Therefore, assessment of the effect of incorporating any AI system into the breast screening pathway in prospective studies is required as it will provide key evidence for the effect of the interaction of medical staff with AI, and the impact on women's outcomes.

Introduction

In the UK, breast cancer screening is offered to women aged 50–70 years every 3 years, with two mammograms taken of each breast. Mammograms are examined separately by two experts to decide whether to recall women for further tests. These experts are breast radiologists, radiography-advanced practitioners, or clinicians with specialist training to examine mammograms—henceforth, for brevity, referred to as radiologists. In case of disagreement, a third radiologist arbitrates. Cancer is diagnosed via histopathology after biopsy. In the USA, single-reader mammography (ie, one radiologist's interpretation) is offered every 1 or 2 years, and, in most of Europe, two-reader mammography is offered every 2 years.¹

Artificial intelligence (AI) is a computer system that can analyse complex data and recognise patterns.² AI-based technologies could potentially have a role in almost every stage of the breast screening pathway. This Health Policy paper considers AI systems that examine screening mammograms for signs of cancer. How AI systems are proposed to be used and how radiologists interact with these systems are very important aspects to consider in the evaluation, because it is the combination of experts and AI in clinical practice that will determine overall accuracy and women's outcomes (figure 1). AI has the potential to improve the quality of care by detecting cancers missed by current practice, reduce shortage of radiologists, and reduce delays in decision making that might have detrimental effects on women's lives.³ However, AI might have the opposite effect, depending on the accuracy of AI systems and how radiologists interact with them.

The use of computers to interpret breast imaging is not novel. In the USA, computer-aided detection (a technology designed to reduce the risk of missing pathologies of interest, which assists radiologists in the interpretation of mammograms by marking specific areas of images that might seem atypical) was implemented into mammographic practice after US Food and Drug Administration approval in 1998, and insurance reimbursement in 2002.⁴ This implementation occurred despite evidence of clinical benefits being derived from retrospective studies, which were not replicated in real-life situations.^{4–6} Renewed interest in automated interpretation of mammograms follows the development and promotion of algorithms that are based on deep learning, where the AI system learns complex associations in the data, which are mathematically modelled by use of artificial neural networks.⁷ A computer processes hundreds of thousands of images and continuously learns to classify images by analysing artificial neural networks. These networks are layers of mathematical operations, which the computer learns on its own, in a structured way, so that multiple layers can be analysed and combined, similar to how the human brain operates.^{8,9} The key difference between computer-aided detection and deep learning is that deep learning does not need to be explicitly programmed by a person, but trains using its own network until it can correctly classify images unaided.

A 2021 review examining the accuracy of AI systems for the detection of breast cancer in mammography screening concluded that the current evidence is a long way from having the quality and quantity required for the implementation of AI systems into clinical practice.¹⁰

Lancet Digit Health 2022; 4: e558–65

Warwick Medical School, University of Warwick, Coventry, UK (Prof S Taylor-Phillips PhD); UK National Screening Committee, Office for Health Improvement and Disparities, Department of Health and Social Care, London, UK (F Seedat PhD, G Kijauskaite MSc, J Marshall MA, Prof A Mackie PhD); Centre for Medical Imaging, Division of Medicine, University College London, London, UK (Prof S Halligan FMedSci); Exeter Test Group, College of Medicine and Health, University of Exeter, Exeter, UK (Prof C Hyde MD); St George's University Hospitals NHS Foundation Trust, London, UK (R Given-Wilson FRCR); Oxford Breast Imaging Centre, Churchill Hospital, Oxford, UK (L Wilkinson FRCR); Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, UK (Prof A K Denniston PhD); Department of Computing, Imperial College London, London, UK (B Glocker PhD); Department of Chemical Engineering and Analytical Science, University of Manchester, Manchester, UK (P Garrett PhD); Ninewells Hospital and Medical School, University of Dundee, Dundee, UK (Prof R J Steele MD)

Correspondence to: Prof Sian Taylor-Phillips, Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK s.taylor-Phillips@warwick.ac.uk

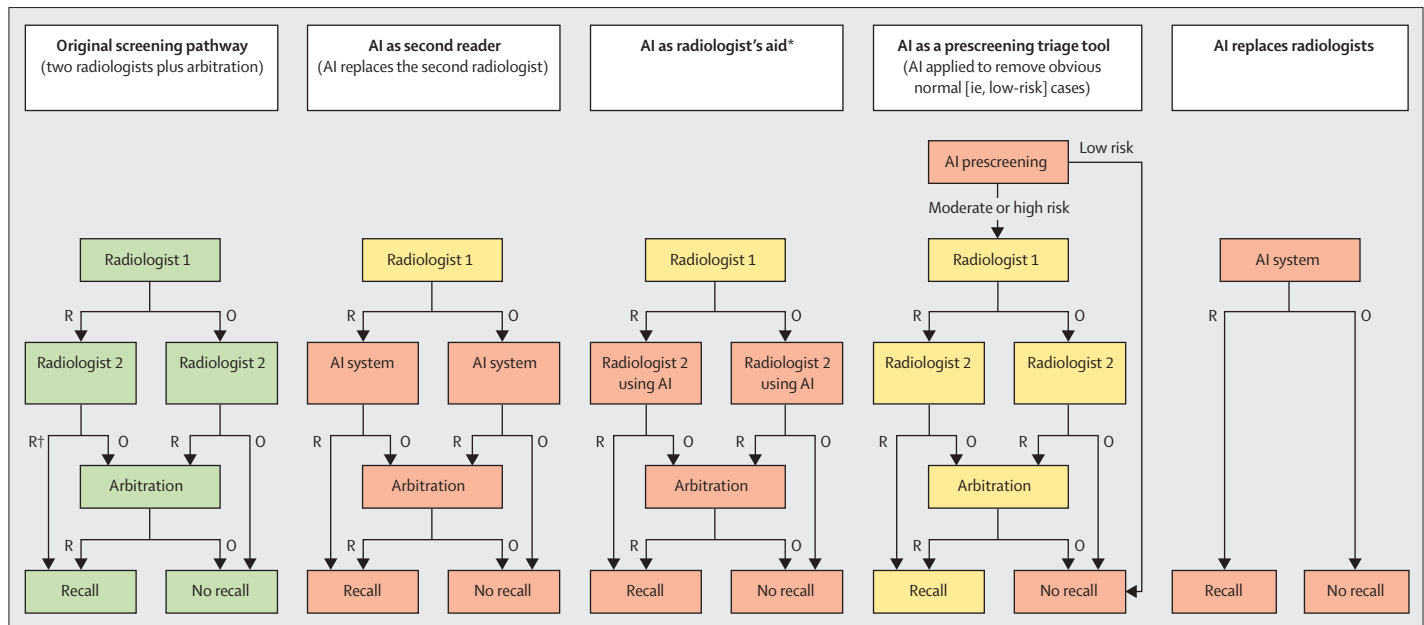


Figure 1: Selection of potential roles of AI in screening pathway

Red boxes indicate instances where AI replaces radiologists completely or could have a strong direct influence on radiologists' behaviour, such as providing prompts on the mammograms with the aim of assisting the radiologist. Yellow boxes denote instances where AI could have an indirect effect on radiologists' behaviour, for example by removing straightforward, normal mammograms and hence increasing cancer prevalence and average case difficulty in the mammograms examined by the radiologist. Green boxes indicate no influence of AI. In the AI prescreening role, in addition to the scenario shown, an additional scenario (not shown) exists where AI can be used to triage low-risk examinations to a single radiologist for review, and moderate-risk or high-risk examinations to two radiologists for independent review. AI=artificial intelligence. O=recommendation not to recall for further tests. R=recommendation to recall for further tests because there are indications of cancer. *AI is used as radiologist's aid to support the decision of the second radiologist, but it could equally be used to support the decision of the first radiologist or arbitration, or both. †Some centres opt for arbitration decision if patients are recalled by both radiologists, to reduce recall rate.

Previous research has sought to compare national approaches to screening evidence synthesis and policy making.^{11–13} In this Health Policy paper, we describe the considerations and evidence required to allow judgement on whether to recommend the implementation of AI into the UK National Health Service (NHS) Breast Screening Programme. In particular, we focus on the potential key outcomes that require investigation and the study designs needed to investigate them.

Goals of AI in breast screening

The UK National Screening Committee (NSC) is responsible for advising the four UK Governments on changes to screening programmes. There is consensus among the UK NSC,¹⁴ the National Institute of Health and Care Excellence,¹⁵ the US and Canadian Preventive Services Task Force,¹⁶ and the Australian Medical Services Advisory Committee¹⁷ that medical policy decisions should be driven by patient and public health outcomes, and that test accuracy alone is insufficient. These views are supported by existing methodological literature on test evaluation.^{18–20} The UK NSC has received much interest from manufacturers, researchers, and other stakeholders to assess the potential for AI to examine mammograms in the NHS Breast Screening Programme because of developments in the technology. Any proposed change to breast screening should be acceptable to women invited for screening and should

improve their health outcomes. The goals of AI in the NHS Breast Screening Programme should therefore be considered in this context. An independent review estimated that breast screening in the UK prevents 1300 deaths from breast cancer annually but also risks harming around 4000 women with overdiagnosis—the unnecessary diagnosis and treatment for cancer that would never have become symptomatic in a woman's lifetime.²¹ Additionally, 70 000 women could also experience anxiety because of false positive results (ie, when a screening mammogram shows suspicious findings but further tests show no cancer).^{21,22}

Breast cancer screening could be improved by reducing mortality, treatment-related morbidity, false positive recalls, overdiagnosis, or by maintaining current clinical outcomes with clear, additional advantages over the current screening programme, such as logistic efficiencies, increases in workforce capacity, or cost reduction. The effect of AI on some of these health outcomes, such as mortality, is unlikely to be measurable directly in a study, because of the sample size requirement and cost of such a study. We discuss approaches to overcome these issues, while maintaining focus on the benefits and harms to women screened for breast cancer.

Key outcomes to measure

The outcomes of interest to the UK NSC relate to the benefits and harms of breast cancer screening.

Consideration is given to the overall (net) effect of any proposed change across different benefits and harms, taking into account how many people are affected by each benefit or harm, and the nature and magnitude of the effect. Assessment methods and measured outcomes depend on the magnitude and type of proposed change.¹⁴ Changes to tests can cause a wide range of downstream consequences, such as an increase in the number of women with false positive results being referred for further testing, which in turn causes unnecessary anxiety in women, delays in workflow, and reduced capacity by staffing additional assessment clinics.¹⁹ Here, we suggest some potential outcomes of clinical significance in breast screening that can be used in studies, but exact outcomes chosen in any evidence review will depend on the nature of the proposed change.

Evidence of test accuracy is an important step in evaluation. Test accuracy of AI systems used in any new proposed pathway (figure 1) informs about the ability of these systems to correctly identify women with cancer (sensitivity) and without cancer (specificity). Specificity is a key outcome because small changes in specificity might have a large effect on the number of false positive recalls in a population screening programme, with each recall causing harm to the woman screened and a financial and workforce cost to the NHS.²³ The specificity of the NHS Breast Screening Programme is higher than 96%.²⁴

Sensitivity is a more complex outcome to interpret. Women who have breast cancer detected at screening might receive mortality or morbidity benefit, or overdiagnosis harm, or they might neither benefit, nor be harmed by detection of cancer at screening because they would have had the same outcomes after symptomatic detection. Therefore, the types of cancer detected (ie, spectrum of disease) are as important as the number of cancers detected, as they are both linked to the benefits and harms of screening.^{20,25} For example, early detection of grade 3 cancer has been linked to mortality benefit,²⁶ whereas detection of low-grade ductal carcinoma in situ might be linked to harm from overdiagnosis and overtreatment.²¹ An AI system able to detect the same types or more aggressive types of cancer with similar sensitivity to a radiologist is preferable to an AI system with higher sensitivity that detects extra patients who are, however, predominantly diagnosed with low-grade ductal carcinoma in situ. Avoiding drift towards detection of less clinically significant cancer will result in AI systems that are more likely to improve the balance between benefits and harms.

Reduction in numbers of cancers detected symptomatically between screening rounds (ie, interval cancers) is another key outcome, because these cancers tend to be more aggressive, higher grade, and with poorer prognosis than cancers detected at screening, and might benefit from more sensitive screening tests leading to earlier detection.²⁷ By definition, interval cancers are not associated with harm from overdiagnosis

at screening. Similarly, information about the number of cancers and spectrum of disease detected symptomatically in the years after screening or detected at subsequent screening rounds could be important. For example, the mechanism by which breast screening might reduce mortality and morbidity is through a stage shift, in which cancer is detected at an earlier stage than the stage of cancer found if there had been no screening.²⁸ Therefore, if an AI system was substantially more sensitive than current practice, evidence about whether that extra detection of patients with cancer at screening led to fewer symptomatic cancers or fewer late-stage cancers would be important.

Study designs to assess test accuracy

Studies to assess the accuracy of AI systems can be retrospective or prospective, with a range of study designs (figure 2, table). Studies of most interest to inform test accuracy for breast screening are comparative, directly comparing AI with other AI and radiologists within the same study, because these direct comparisons are not affected by between-study differences that introduce bias. Analyses at clinically relevant thresholds are more useful than area under the receiver operating characteristic (ROC) curve, because the shape of the ROC curve and the area under it do not affect clinical outcomes; only the test accuracy at the threshold used in practice is relevant to women's outcomes at screening.²⁹

Large retrospective, test accuracy studies are an important step of the assessment process.³⁰ These studies use test sets of mammograms with known outcomes (ie, reference standard), against which AI systems can be assessed. The reference standard is established by use of retrospective clinical data (eg, biopsy results) from screening databases. Such databases enable direct comparison of several different AI systems on the same mammograms, with either the decision of the original radiologist made as part of clinical practice (retrospective, comparative, test accuracy studies) or the decisions of multiple radiologists outside of clinical practice, under laboratory conditions (enriched, multiple-reader, multiple-case, test set laboratory studies). In this paper, we focus on retrospective, comparative, test accuracy studies because the accuracy of radiologists to read test sets in the laboratory is not generalisable to clinical practice (ie, the laboratory effect).³¹

Large test sets can be established at relatively low cost and should use external validation in either consecutive or randomly selected mammograms; these test sets should also be generalisable to the UK screening population. Very large retrospective studies of consecutively enrolled women attending their breast screening can be undertaken: electronic health records can provide data from further testing of women with positive screening results and allow long-term symptomatic follow-up of women with negative screening results during the screening interval (ie, 3 years). The

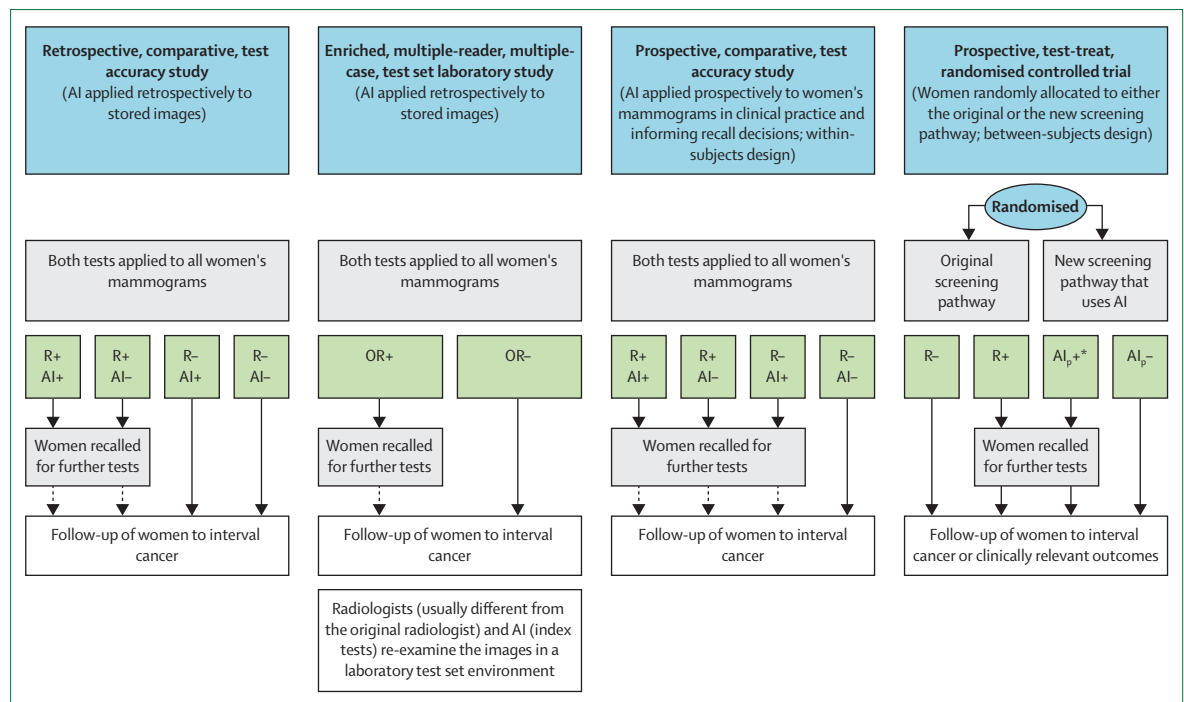


Figure 2: Four comparative study designs used to assess accuracy of AI and radiologists in breast screening

Solid arrows indicate uninterrupted follow-up to determine whether women develop interval cancer. Dashed arrows denote that, although follow-up of women to interval cancer is possible, the absolute number of interval cancers in women with negative AI screening results will be underestimated, because follow-up to interval cancers is truncated when cancer is detected at screening by the radiologist comparator test. Decisions are either to recall women for further tests (+) or not (-). AI=artificial intelligence. OR=original radiologist. R=radiologist. *AI_p denotes the decision of the new screening pathway that uses AI, rather than the decision of the AI system alone. The two retrospective study designs can only measure accuracy of the AI system in isolation. Prospective, comparative, test accuracy studies can measure accuracy of AI alone or the new screening pathway that uses AI, whereas prospective, test-treat, randomised controlled trials evaluate the new screening pathway that uses AI.

large sample size of these studies allows for tight confidence intervals around specificity estimates required for population screening, determination of the spectrum of disease detected (when concordance with radiologists is high), and estimation of accuracy in population subgroups. Accuracy of AI by population subgroup, such as age or ethnicity, is important for assessing the effect of AI on inequality, which could be an issue because, if AI systems have been trained in one ethnic group (eg, White people), they might be less accurate in other ethnic groups.³²

Retrospective studies cannot quantify the overall effect on accuracy or interval cancers when AI is integrated into the screening pathway, because they cannot measure the effect of AI when it is incorporated into the workflow with radiologists. Furthermore, retrospective studies might not correctly establish test accuracy or the spectrum of disease detected if concordance between AI and radiologists is low or if AI has substantially higher sensitivity. These issues arise because test sets are subject to verification bias, which occurs when two different reference standards for the outcome are used to verify the disease of interest in different groups of patients. The first reference standard is diagnosis of cancer at assessment, and the second reference standard is

follow-up and diagnosis of cancer at symptomatic (interval) presentation or next screen. Women recalled for further tests by the original radiologists have both standards applied as they could be diagnosed with cancer either at assessment or at follow-up. Women not recalled for further tests by the original radiologist receive either no reference standard (partial verification) or only the second reference standard of follow-up to subsequent cancer (differential verification), which is less likely to detect cancer when present.³³

Verification bias is introduced when women are recalled for further tests on the basis of the radiologist's decision, because cancer, if present, is more likely to be found in patients receiving follow-up tests after recall from screening, than in patients who are not recalled for further tests and, instead, simply receive follow-up when cancer presents symptomatically (second reference standard). This is because some cancers detected at screening follow-up tests would never result in symptoms or might present symptomatically only after the follow-up period of the study. Therefore, women with positive AI screening results who are not recalled by the original radiologists (ie, defined as negative in the test sets) cannot be characterised; the AI results might be true positive with unknown disease spectrum, which

Value relative to other design options		Limitations and biases relative to other design options
Enriched, multiple-reader, multiple-case, test set laboratory study		
AI systems applied to test set of mammograms with known outcomes from screening (ie, reference standard). The reference standard is established by use of retrospective data (eg, biopsy and follow-up results) from screening databases. The test set is usually enriched with extra patients with cancer to enable estimation of sensitivity in smaller test sets. The comparator is independent radiologists, who are invited to read mammograms outside clinical practice. The decisions of AI and independent radiologists are compared with the reference standard.	Quick to complete. Can directly compare the performance of multiple AI systems on the same images, or AI as radiologist's aid. Introduces less incorporation bias than retrospective, comparative, test accuracy study because both AI and independent radiologists do not form part of the reference standard, as they have neither been used in the original decision of whether to recall for further tests, nor used in deciding which further tests to do or for identifying the location of anomalies for biopsy. Relatively rapid to perform compared with other study designs because enrichment provides adequate numbers of patients with cancer.	Biased by the laboratory effect: accuracy of independent radiologists in these studies cannot be generalised to clinical practice because of differences existing between laboratory and clinical practice in reading conditions and prevalence of cancer. Low generalisability to clinical practice. Pronounced selection and spectrum biases possible, if selection of women and enrichment are not on a consecutive or random basis. Temporal separation usually needed between AI-aided and unaided reads, to diminish recall bias. Cannot assess and compare the effect of decisions of AI with those of radiologists on clinically significant outcomes, because of the difficulty in extrapolating the complete pathway change (ie, using AI) in clinical practice, where prevalence of cancer is lower, from the single decision taken in a laboratory environment.
Retrospective, comparative, test accuracy study		
AI systems applied to test set of mammograms with known outcomes from screening (ie, reference standard). The reference standard is established by use of retrospective data (eg, biopsy and follow-up results) from screening databases. The AI is compared with the reference standard and with the original decisions of radiologists made in clinical practice.	Quick to complete. Good starting point for estimating accuracy of AI. Can directly compare the performance of multiple AI systems on the same images. Not subject to the laboratory effect.	Prone to partial verification bias when women who are not recalled by the original radiologist do not receive any follow-up to check if they have cancer (they do not receive any further tests such as a biopsy [the reference standard], and they are not followed up to check for symptomatic presentation of cancer [interval cancer]). Prone to differential verification bias when women receive different reference standards depending on whether they were recalled for further tests by the original radiologist. If they were recalled, they receive follow-up tests such as biopsy, whereas, if they were not recalled, they are simply followed up to record subsequent cancers (future symptomatically detected cancers [interval cancers] or next-round screen detected cancers). Women who were recalled for further tests are more likely to have cancer detected, if present. Prone to incorporation bias, which occurs when results of the radiologist index test form part of the reference standard. Presence of differential verification bias and incorporation bias results in uncertainty on true status of women with positive results from AI but negative results from original radiologist. Cannot assess complex interaction between AI and radiologists; for example, although it is often assumed that radiologists will recall women who receive AI positive results, in clinical practice they might override AI decisions depending on how AI is implemented. Cannot assess and compare the effect of decisions of AI with those of radiologists on clinically significant outcomes.
Prospective, comparative, test accuracy study		
Mammograms of women examined via both the original and new screening pathways (without and with AI). If women receive positive screening results from either pathway, they are recalled for further tests (eg, taking a biopsy where clinically indicated). The reference standard corresponds to further tests (including biopsy) in women recalled by either AI or radiologists, and it might also include longer-term follow-up to symptomatically detected cancer (interval cancer) or to cancer detected at next screening.	Least biased method for measuring accuracy. Reduces or removes partial and differential verification biases because women with positive results from both AI and radiologists receive the same reference standard. Reduces or removes incorporation bias only if follow-up tests are blinded to the type of index tests (eg, by changing AI's and radiologist's annotations of images to the same appearance). Can assess the interaction between AI and radiologists.	Low prevalence of anomalies in screening means many women are required to achieve adequate study power, increasing study time. Comparing the performance of multiple AI systems in this study design is likely to be more logistically challenging than in retrospective study designs. Cannot accurately measure the effect of implementing the AI test on interval cancers, because uninterrupted follow-up is not possible since women are recalled for further tests by other index tests (eg, either by the radiologist or other AI systems that are used as comparator in the study).
Prospective, test-treat, randomised controlled trial		
Women randomly allocated to either the original (standard) pathway or the new, proposed screening pathway that incorporates AI. Follow-up of women to symptomatic (interval) cancer or to cancer detected at subsequent screening rounds.	Least biased method for measuring impact on women's outcomes. Women's clinical trajectory reflects exactly what would happen in clinical practice with and without AI (depending on which group they are randomly allocated to). Can measure the effect of the new pathway on the number of interval cancers and other clinically relevant outcomes. Can assess the interaction between AI and radiologists.	Low prevalence of anomalies in screening means many women are required to achieve adequate study power, increasing study time. Mammograms of different women are examined with different tests, so no direct comparison of test accuracy on the same women is possible, which greatly increases the numbers needed to achieve adequate study power. Inefficient method for measuring accuracy.
Designs are listed in order of complexity, from least to greatest. AI=artificial intelligence.		
Table: Summary of test accuracy study designs considered by the UK National Screening Committee and their contribution to overall evidence		

therefore might be clinically significant or overdiagnosed cancer, or false positive. This verification bias is reduced by including interval cancers and cancers detected at next screening, but it cannot be removed entirely because not

all cancers diagnosed by radiologists at screening would be detected in the interval or at the next screen. For example, overdiagnosed cancer might never appear symptomatically, and cancers detected by the AI system

but missed by the original radiologists might also be missed by the radiologist at the next screen.

Although retrospective studies do not provide sufficient evidence to implement AI systems into screening pathways, they can be useful as a starting point to establish whether a prospective study integrating AI into the screening pathway is worth investing in, and to optimise the tests and the screening pathway required for the study. Additionally, despite being subject to the previously described biases (partial verification and differential verification), large, consecutive, retrospective studies can provide other useful information, such as accuracy in population subgroups (eg, ethnicity), for which prospective studies might be underpowered.

Prospective, comparative, test accuracy studies can measure test accuracy when concordance between AI and radiologists is low, or sensitivity differs between them. In these instances, women are referred for further tests if either radiologists or AI systems suggest recall. Prospective, comparative, test accuracy studies cannot measure the effect that AI would have on interval cancers or clinical outcomes, because women with negative AI screening results are not always followed up uninterrupted to clinical outcomes—they might receive further tests based on the radiologist's decision, preventing interval cancers from occurring.

Prospective, comparative, test accuracy studies can be used to measure accuracy of the whole testing pathway with AI integrated in clinical practice, and compare it with the accuracy of the pathway without AI. By contrast, retrospective studies can only assess the accuracy of single AI reads (ie, interpretations) of mammograms and not the whole testing pathway. Accuracy of the AI testing pathway in breast screening practice is substantially different to that of a single read and depends on how the radiologists interact with AI systems. Figure 1 indicates some mechanisms through which the AI system could affect the behaviour and accuracy of radiologists. For example, an AI system that supports the radiologist's decision by highlighting suspicious areas has a direct effect on accuracy. Replacing the second reader with an AI system can affect decisions made by arbitration, and change the behaviour of the radiologist examining mammograms. Direct comparisons, although more challenging to implement, remain important for prospective studies, because they enable comparison of accuracy between the original screening pathway, in which mammograms are examined by radiologists only, and the new screening pathway, in which AI is incorporated, to assess the differences in their performances.

Study designs to assess test impact

The key question for the UK NSC is the effect of the AI system on women's outcomes, as part of the balance between benefits and harms of screening. This question is not answered by test accuracy alone; rather it

necessitates information on clinically meaningful outcomes and their proxies, such as stage shift of cancer, characteristics of cancer detected, and subsequent development of interval cancers. The effects of the change to the whole testing pathway when AI is incorporated should be measured, rather than the AI test alone. These wider effects could include effects on the radiologist's behaviour and the workflow. Downstream outcomes and broader implications require prospective assessment; this requirement is increasingly recognised internationally in the field of AI in health care or medicine.^{34,35}

The least biased study designs for measuring impact on women's outcomes are prospective, test-treat, randomised controlled trials that randomly allocate women to the original pathway (two radiologists plus arbitration) or the new proposed pathway, with follow-up of women to clinically significant outcomes. These studies avoid the need to link together evidence from different studies, and thus minimise the risk of bias associated with differences among studies in a linked chain of evidence.²⁵ In the UK, very large pragmatic, randomised controlled trials of AI systems can be undertaken, in which the trial randomisation and outcome collection are integrated into the existing reporting practices and software of the NHS Breast Screening Programme. The homogeneity of the screening pathway, quality assurance, reporting standards, and screening software across the UK makes this possible. Such trials have been used previously in the UK to investigate other variations to breast screening, such as the extension of the breast screening age range (NCT01081288) or an intervention to reduce radiologists' fatigue.³⁶ These larger trials, which included 1·2–6 million participants, are examples of studies that can provide power to detect more clinically meaningful outcomes.

Quasi-experimental studies, such as prospective cohort studies, and implementation evaluations can provide additional information about the test's fit with, and its effect on, the testing pathway; however, considerations of these study designs are beyond the scope of this Health Policy paper.

Any study should also be considered for likelihood of generalisability to the UK, including characteristics of women screened, radiologists (eg, accuracy and recall threshold of radiologists differ substantially between the UK and USA³⁷), and digital mammography and broader software and hardware systems. Although the principles of test assessment, which underpin the approach we have underlined in this Health Policy paper, apply to all countries, variations in current practice, decision-making structures, and broader health-care and societal differences will probably lead to international differences in approach to assessment.

Conclusions

The UK NSC's assessments of the use of AI systems to examine screening mammograms will continue to focus

Search strategy and selection criteria

This Health Policy paper represents the UK National Screening Committee's (NSC's) approach to reviewing evidence on artificial intelligence (AI) in mammography and builds on the 2019 NSC paper *Interim Guidance for Those Wishing to Incorporate Artificial Intelligence into the National Breast Screening Programme*. Rather than a complete description of the literature, key concepts and approaches were summarised after performing a literature search. Searches were designed to identify papers assessing changes to screening tests, because the literature specific to AI screening tests is less well developed. Searches were originally done between May 1 and Dec 31, 2019, and updated on May 14, 2021. We searched MEDLINE from inception to May 14, 2021 for articles describing methods that assess new screening tests, by combining medical subject headings terms for "mass screening" (methods, standards, and use), with terms for humans (thus excluding animal studies), English language, and systematic review. Two people (including ST-P) screened papers for inclusion. To identify methodological guidance of policy makers in the grey literature, we searched websites of 28 national screening organisations or advisory bodies from 16 countries, using search terms for "methods" and for specific screening tests, with no date or language limits. These searches were done between May 1 and Dec 31, 2019, with targeted updates in May, 2021. We identified the websites using two previous systematic reviews, and ST-P screened guidance for inclusion. We included papers from both searches if they presented a concept, principle, criterion, or approach that was considered applicable either directly or indirectly to evaluating the use of AI to screening mammography. We compiled 41 methodological papers from both searches (35 from the first literature search, 11 from the grey literature search (of which 7 were already identified by the first search), and an additional 2 from the updates to these searches). 17 additional papers related to AI applied to medical imaging and breast screening were selected by experts, including members of the UK NSC AI Task Group. We used saturation-type approaches to data extraction to identify relevant concepts, without the requirement to locate and cite more than one source for each concept.

on any changes to the balance between benefits and harms of breast cancer screening. Although retrospective studies have an important role in early-stage assessment, they also have biases and limitations; therefore, prospective studies are required to assess the effect of incorporating any AI system into the breast screening pathway.

Contributors

ST-P and JM conceptualised and developed this Health Policy paper. ST-P made the final decision on which papers to include, with advice from other authors. All authors contributed to the method, participated in the analysis, interpretation, and drafting of the manuscript, and checked its accuracy. All authors approved the final version.

Declaration of interests

This Health Policy paper was sent to all members of the UK NSC and the AI Task Group of the UK NSC for comment, before submission. BG is a part-time employee of HeartFlow and Kheiron Medical Technologies and holds stock options as part of the standard compensation package for both companies. BG had an advisory role with stock options for Kheiron Medical Technologies (from January, 2018, to September, 2021). BG was a Visiting Researcher and part-time employee of Microsoft Research until May, 2021. BG has received grants from Innovate UK, the EU Commission, and jointly from the National Institute for Health and Care Research (NIHR) and the Medical Research Council. Kheiron Medical Technologies currently has a breast screening AI product available. As such, BG contributed to the text as an expert in AI, focusing on the accuracy of the technical descriptions of the AI method. BG did not comment on or suggest changes to the described requirements of evidence. LW declares that she is a member of the Optimam Steering Group, a collaboration between research scientists at the University of Surrey and Cancer Research UK, which, among other activities, is collating a database of mammography images to be used for research and assessment of AI. AM, JM, GK, and FS declare that they are employed by the UK NSC. SH is Chair of the AI Task Group of the UK NSC. RG-W is Chair of the Adult Reference Group of the UK NSC. RJS was the Chair of the UK NSC at the time of writing this paper. ST-P has previously received funds from the UK NSC to undertake evidence reviews for the UK NSC, including for Breast AI, and is a member of the Adult Reference Group and AI Task Group of the UK NSC. The opinions are those of the authors and not the NIHR, the NHS, or the Department of Health and Social Care. All other authors declare no competing interests.

Acknowledgments

Members of the UK NSC, the AI Task Group of the UK NSC, and other stakeholders made comments. This Health Policy paper was funded by an NIHR Career Development Fellowship (CDF-2016-09-018) awarded to ST-P. SH's institution is supported by an NIHR Biomedical Research Centre funding scheme.

References

- 1 Ebell MH, Thai TN, Royalty KJ. Cancer screening recommendations: an international comparison of high income countries. *Public Health Rev* 2018; **39**: 7.
- 2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25**: 44–56.
- 3 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; **18**: 500–10.
- 4 Kohli A, Jha S. Why CAD failed in mammography. *J Am Coll Radiol* 2018; **15**: 535–37.
- 5 Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? *J Am Coll Radiol* 2010; **7**: 802–05.
- 6 Sechopoulos I, Mann RM. Stand-alone artificial intelligence—the future of breast cancer screening? *Breast* 2020; **49**: 254–60.
- 7 Georgevici AI, Terblanche M. Neural networks and deep learning: a brief introduction. *Intensive Care Med* 2019; **45**: 712–14.
- 8 Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019; **103**: 167–75.
- 9 Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *Br J Cancer* 2021; **125**: 15–22.
- 10 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* 2021; **374**: n1872.
- 11 Taylor-Phillips S, Stinton C, Ferrante di Ruffano L, Seedat F, Clarke A, Deeks JJ. Association between use of systematic reviews and national policy recommendations on screening newborn babies for rare diseases: systematic review and meta-analysis. *BMJ* 2018; **361**: k1612.
- 12 Seedat F, Cooper J, Cameron L, et al. International comparisons of screening policy-making: a systematic review. October, 2014. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/444227/FINAL_REPORT_International_Screening.pdf (accessed July 27, 2021).

- 13 Dobrow MJ, Hagens V, Chafe R, Sullivan T, Rabeneck L. Consolidated principles for screening based on a systematic review and consensus process. *CMAJ* 2018; **190**: E422–29.
- 14 UK National Screening Committee. UK NSC evidence review process. 2015. <https://www.gov.uk/government/publications/uk-nsc-evidence-review-process> (accessed Feb 20, 2020).
- 15 National Institute for Health and Care Excellence. The guidelines manual: process and methods. Nov 30, 2012. <https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-pdf-2007970804933> (accessed July 27, 2021).
- 16 US Preventative Services Task Force. Procedure manual. 2021. <https://uspreventiveservicestaskforce.org/uspstf/about-uspstf/methods-and-processes/procedure-manual> (accessed July 27, 2021).
- 17 Medical Services Advisory Committee. Guidelines for preparing assessments for the Medical Services Advisory Committee. May, 2021. [http://www.msac.gov.au/internet/msac/publishing.nsf/Content/E0D4E4EDDE91EAC8CA2586E0007AFC75/\\$File/MSAC%20Guidelines-complete-16-FINAL\(18May21\).pdf](http://www.msac.gov.au/internet/msac/publishing.nsf/Content/E0D4E4EDDE91EAC8CA2586E0007AFC75/$File/MSAC%20Guidelines-complete-16-FINAL(18May21).pdf) (accessed July 27, 2021).
- 18 Schünemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies—from test accuracy to patient-important outcomes and recommendations. *J Clin Epidemiol* 2019; **111**: 69–82.
- 19 Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012; **344**: e686.
- 20 Lord SJ, Irwig L, Bossuyt PMM. Using the principles of randomized controlled trial design to guide test evaluation. *Med Decis Making* 2009; **29**: E1–12.
- 21 Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* 2012; **380**: 1778–86.
- 22 NHS Digital. NHS Breast Screening Programme, England 2020–21. 2022. <https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme/england---2020-21> (accessed March 4, 2022).
- 23 Cole P, Morrison AS. Basic issues in population screening for cancer. *J Natl Cancer Inst* 1980; **64**: 1263–72.
- 24 Burnside ES, Vulkan D, Blanks RG, Duffy SW. Association between screening mammography recall rate and interval cancers in the UK Breast Cancer Service Screening Program: a cohort study. *Radiology* 2018; **288**: 47–54.
- 25 Merlin T, Lehman S, Hiller JE, Ryan P. The “linked evidence approach” to assess medical tests: a critical analysis. *Int J Technol Assess Health Care* 2013; **29**: 343–50.
- 26 Tabar L, Chen TH-H, Yen AM-F, et al. Effect of mammography screening on mortality by histological grade. *Cancer Epidemiol Biomarkers Prev* 2018; **27**: 154–57.
- 27 Kirsh VA, Chiarelli AM, Edwards SA, et al. Tumor characteristics associated with mammographic detection of breast cancer in the Ontario breast screening program. *J Natl Cancer Inst* 2011; **103**: 942–50.
- 28 Connor RJ, Chu KC, Smart CR. Stage-shift cancer screening model. *J Clin Epidemiol* 1989; **42**: 1083–95.
- 29 Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *Eur Radiol* 2015; **25**: 932–39.
- 30 Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open* 2020; **3**: e200265.
- 31 Gur D, Bandos AI, Cohen CS, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008; **249**: 47–53.
- 32 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; **1**: e271–97.
- 33 O’Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. *BMJ Evid Based Med* 2018; **23**: 54–55.
- 34 Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018; **286**: 800–09.
- 35 Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; **17**: 195.
- 36 Taylor-Phillips S, Wallis MG, Jenkinson D, et al. Effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection: a randomized clinical trial. *JAMA* 2016; **315**: 1956–65.
- 37 Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United Kingdom. *JAMA* 2003; **290**: 2129–37.

Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.