UK
National
Screening
Committee

Use of artificial intelligence for mammographic image analysis in breast cancer screening

Rapid review and evidence map

Version: Final

Author: Karoline Freeman
Julia Geppert
Chris Stinton
Dan Todkill
Sam Johnson
Aileen Clarke
Sian Taylor-Phillips

Date: February 2022

The UK National Screening Committee secretariat is hosted by the Department of Health and Social Care

About the UK National Screening Committee (UK NSC)

The UK NSC advises ministers and the NHS in the 4 UK countries about all aspects of population screening and supports implementation of screening programmes. Conditions are reviewed against evidence review criteria according to the UK NSC's evidence review process.

Read a complete list of UK NSC recommendations.

UK National Screening Committee, Southside, 39 Victoria Street, London, SW1H 0EU

www.gov.uk/uknsc

Blog: https://nationalscreening.blog.gov.uk/

For queries relating to this document, please contact: https://view-health-screening-recommendations.service.gov.uk/helpdesk/

Published: February 2022

Contents

Plain English summary

Breast cancer is the most common cancer in women in the UK. The National Health Service breast cancer screening programme invites women aged 50 – 70 years for screening every 3 years. The screening involves taking x-rays (mammograms) of each breast. Two readers look at each x-ray image to see if there are signs of cancer. The readers decide whether the woman is offered extra tests to confirm if cancer is present. In 2018-2019, the NHS breast cancer screening programme screened 1.82 million women in England and found breast cancer in 15,285 women. The aim of the programme is to reduce deaths from breast cancer by detecting cancer earlier when it is more treatable. Breast cancer screening programmes also miss between 15%–35% of cancers present in screened women. This is either due to error or because the cancer is not visible to the reader.

Computer image recognition programmes or artificial intelligence (AI) that can learn to spot changes in breast mammograms have been developed to assist humans in breast cancer screening programmes. The interest in using AI for clinical practice is growing because it can offer many advantages. In breast screening for instance, fewer cancers may be missed because an AI programme does not lose concentration or become tired. AI could also reduce the workload of breast screening by reducing the effort needed to read thousands of mammograms for instance by replacing one of the mammogram readers. But there is also concern that AI may detect changes which would never cause the woman any harm. The current UK breast cancer screening programme does not use AI. If AI is to be considered in the UK breast screening programme, we need to understand the benefits and harms of adding AI into the current screening programme.

The current review looked at the evidence on:

- how good AI is at finding cancers in breast cancer screening
- what benefits and harms AI has for the women who are screened or for the screening programme and the health professionals involved

Based on the current evidence, the UK NSC does not recommend using AI in the NHS breast cancer screening programme. This is because:

- the use of AI systems would change the current screening programme therefore it is important to assess how accurate AI is in breast screening clinical practice before changing it
- the performance of AI systems varies in different settings but there are no good quality studies in the UK
- it is unclear how good AI is at finding different types of breast cancer or at finding breast cancers in different groups of women (for example different ethnic groups)

- AI might reduce the workload of staff, the number of cancers missed at screening, and the number of women called back for further tests when they do not have cancer, however, the quality of evidence is very low.

Executive summary

Purpose of the review

This document presents the findings of a rapid review and an evidence map on the impact of using artificial intelligence (AI) to examine women's breast cancer screening mammograms for signs of cancer. The aim was to highlight the current evidence and gaps in the evidence in terms of test accuracy and clinical utility outcomes. The literature searches for this review retrieved 10 relevant studies for the key areas of interest for this topic. Based on the findings of the review and evidence map, a review in 1-3 years' time may be necessary, when the evidence base has developed further. The methods in this report may be used as a baseline to build upon in that future review.

Background

In the UK NHS Breast cancer screening programme (NHSBSP), women aged 50 – 70 years are invited to breast cancer screening every 3 years. This involves taking digital mammography images of each breast from two views. The interpretation of the images taken is serially carried out by two readers. Each reader makes a decision about whether the image appears normal or if a woman should be recalled for further assessment. In case of disagreement, arbitration is employed. Women who are recalled are offered additional testing to determine whether they have cancer. The aim is to detect cancer earlier at screening when treatment is more effective. However, some cancers detected at screening never would have given the woman symptoms or caused harm within her lifetime, called overdiagnosis. This results in unnecessary treatment, called overtreatment. Some cancers are missed during screening, so the women are falsely reassured.

It has been suggested that mammographic image recognition using AI within breast screening programmes can have potential benefits. For instance, fewer cancers may be missed because an AI algorithm is unaffected by fatigue or subjective diagnosis and AI could reduce the workload involved by replacing the second reader role in the NHSBSP leading to greater efficacy and efficiency in screening.

AI is a computer system that can perform complex data analysis and tasks of image recognition; made possible by both immense computational power and the use of deep learning algorithms which are now being applied to the healthcare sector. Several potential places of AI in the breast screening pathway have been envisaged:
1) AI could replace one or all human readers;
2) AI could be used to pre-screen images with only high-risk images being subsequently read by human readers;

3) AI could be used as a reader aid, where the human reader uses the AI system as a decision support during a read; and
4) AI could be used to post-screen negative images following double reading to recall women at highest risk of undetected cancer.

In addition to potential benefits AI may also exacerbate harm from screening. AI could alter the spectrum of disease detected at breast screening and lead to an increase in overdiagnosis. AI could decrease the specificity of screening, which would increase the number of women incorrectly recalled for further tests. This would cause those women anxiety and increase the overall workload in breast screening through increased assessment appointments. This could alter the balance of benefits and harms of breast screening.

Focus of the review

The aim of this review was to synthesise the evidence on the use of deep learning AI algorithms to read mammograms (as reader aid or stand-alone) of women attending routine breast screening for digital (full field digital mammography, FFDM) mammograms. The evidence is presented in the form of a rapid review (question 1) and an evidence map (question 2). The review included studies published between January 2010 and September 2020 and aimed to address the following two questions answering the UK NSC criteria as outlined.

Question 1 – What is the accuracy of AI algorithms to detect breast cancer in women attending screening mammography?

> *Criterion 4 - Accuracy of the tests*
> *There should be a simple, safe, precise and validated screening test.*
>
> *Criterion 5 - Distribution of test values in the target population*
> *The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.*

Question 2 – What is the clinical impact of the use of AI algorithms to detect breast cancer in mammograms compared to current practice in breast screening programmes?

> *Criterion 11 — Effectiveness of the screening programme*
> *There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an "informed choice" (such as Down's syndrome or cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.*

> *Criterion 12 — Benefits and harms of the screening programme*
> *The benefit gained by individuals from the screening programme should outweigh any harms, for example from overdiagnosis, overtreatment, false positives, false reassurance, uncertain findings and complications.*

The focus of the review was on retrospective database studies estimating the accuracy of AI, and prospective studies measuring the accuracy in clinical practice and impact of incorporating AI. Within the retrospective database studies, the review focused on the validation rather than the development of AI systems. Internal validation (where the same dataset is used for training and validation (e.g. cross validation or split sample validation)) overestimates accuracy and has limited generalisability and is excluded from this review. Training and validation test sets should not overlap. With that in mind, geographical validation (using a validation test set from different screening centres to the training set) is thought to be the least biased method even though in urban areas women invited to screening may move into different screening catchment areas between screening rounds. Geographical validation has the benefits of understanding the effectiveness across potentially different technical parameters (such as different machines) and operating personnel and is included in this review. Temporal validation has some features of both but is excluded here because it often uses the same machines, personnel and women as the training set so may overestimate accuracy. To investigate the impact of this decision the excluded temporal validation studies were summarised and discussed at the end of the review. Often prospective studies are also required to understand the accuracy and impact of AI in clinical practice (interacting with human readers), and to measure the true disease status of AI positive / human reader negative test results in order to characterise additional cancers detected by AI in terms of benefits and harms.

Recommendation under review

The UK NSC recommends screening for breast cancer. National screening programmes are in place in each of the four countries of the UK. No prior review has been conducted on the use of artificial intelligence in breast cancer screening by the UK NSC.

Findings and gaps in the evidence of this review

Database searches yielded 4,969 results, of which 7 studies were judged to be relevant to question 1 and 8 studies were included in question 2 (5 studies contributed to both questions). Studies at full text stage were mainly excluded for the following reasons:

- Image type was not screening mammograms or FFDM
- Internal validation test sets
- Detecting only subtypes of breast cancer
- Lack of detection / classification
- Intervention not AI

- Outcomes not relevant

Question 1 - What is the accuracy of AI algorithms to detect breast cancer in women attending screening mammography?

The evidence base on the accuracy of AI to detect breast cancer was of low quality and applicability. There were no studies that described the accuracy of AI integrated into any breast screening pathway; all 7 studies reported accuracy of AI to detect cancer in mammograms as a single read. Therefore, there is no direct evidence on how AI may affect accuracy if integrated into UK breast screening practice. There were no prospective test accuracy studies in clinical practice, only retrospective comparative test accuracy studies and enriched test set multiple reader multiple case (MRMC) laboratory accuracy studies. Enrichment led to breast cancer prevalence which is atypical of a screening population. All studies included a human comparator which was either the original decision as part of clinical practice or prospective single or average reads of independent radiologists without AI under laboratory conditions. Risk of bias, assessed using the modified QUADAS-2 tool, was considered high in all 7 studies. The index test (stand-alone AI or AI as reader aid) was the area with the greatest risk of bias because studies were either biased by the laboratory effect or by the lack of a pre-specified test threshold (the threshold for classifying images was derived from the dataset which was then used to evaluate the AI system). There were significant concerns regarding the applicability of the research identified to the UK screening population in 6 out of the 7 included studies mainly because the cancer prevalence did not match the screening context and the studies did not represent a complete testing pathway applicable to the UK. None showed the impact or accuracy of AI in UK clinical practice.

Three enriched test set MRMC laboratory studies reported test accuracy for a single read of AI as a reader aid. Another 3 retrospective comparative test accuracy studies (only one using a non-enriched dataset) and one enriched test set MRMC study reported the test accuracy for a single read of AI as a stand-alone system.

Stand-alone systems

None of the 4 studies reporting the test accuracy of AI as a stand-alone system recruited women prospectively. Instead, they included mammograms from available databases to be read by the AI system which included commercially available as well as in-house AI systems.

The comparator in 3 studies was the original decision on recall / no recall without AI recorded in the database based on either a single reader or 2 readers with consensus. In one study AI performance was compared to the decision from human readers who read the mammograms prospectively under laboratory conditions.

The studies undertook non-inferiority analyses for test accuracy without pre-specified thresholds for the interpretation of AI scores. Study point estimates of sensitivity or specificity were higher for some AI than single human readers but did not perform as well as consensus.

AI as reader aid

None of the 3 studies reporting the test accuracy of AI as a reader aid recruited women prospectively. Instead, they included mammograms from available databases to be read by the AI system. All 3 studies used an enriched test set to manage the number of mammograms to be read by the human readers. The 3 AI systems evaluated were commercially available systems. Sensitivity and specificity were reported as an average across the readers and compared to the same readers reading the same mammograms without AI reader aid. Point estimates of the means of sensitivity and specificity were slightly higher for readers with AI support in 2 of the 3 studies. However, confidence intervals overlapped. These studies will be affected by the laboratory effect so absolute accuracy is not generalisable to clinical practice, but relative accuracy may be somewhat informative.

Secondary analyses

Evidence on the detection of interval cancers separately was scarce but one study suggested that AI algorithms may be able to promote earlier cancer detection.

Combination of different AIs may increase overall AI performance in classifying mammograms. Simulated integration of AI systems into the screening pathway resulted in similar specificities. Subgroup analyses by lesion type, patient characteristics and breast density were poorly reported without any clear messages on the test accuracy in these groups.

Evidence from temporal validation studies

At present, excluding temporal validation studies did not exclude useful evidence for this review. The two studies identified used an enriched test set MRMC laboratory study design in which retrospectively collected images were read prospectively by human readers and by a stand-alone AI system outside clinical practice.

Recommendations on screening

There is insufficient evidence in quality and quantity to recommend implementation of AI into clinical practice of the NHS breast screening programme. Overall, the evidence on the test accuracy of AI algorithms to detect breast cancer in women attending screening mammography using geographical validation test sets was sparse and lacked applicability to the UK context (no study used a UK dataset). Except for one study, study populations were small with a cancer prevalence atypical of the screening context.

Question 2 - What is the clinical impact of the use of AI algorithms to detect breast cancer in mammograms compared to current practice in breast screening programmes?

The evidence on the clinical utility of AI algorithms when used to read mammograms in a breast screening programme is limited. None of the 8 identified studies evaluated the AI algorithm as a change of a screening pathway in a randomised controlled trial or prospective cohort study.

There is some limited evidence from 8 simulation studies that:

- AI may potentially reduce the recall rate if used as a pre-screen (2 studies) or as a second reader (1 study).
- AI has the potential to promote earlier detection of interval cancers compared to double reading with consensus (1 study) and interval cancers as well as next-round screen detected cancers assessed as negative by human double reading when used as post-screen (1 study)
- AI as reader aid may reduce reading time for low risk examinations but increase reading time for high suspicion examinations (2 studies). AI may not prolong the workflow of the radiologists (1 study) or even decrease the average reading time per case (1 study).
- AI may decrease the screen reading workload for the single human reader (2 studies) or the second reader (1 study).
- AI may detect less DCIS but more invasive (stage 2 or higher) cancers than the first or second reader (1 study). If used as a pre-screen, AI may miss ~10% of screen-detected, invasive cancers, but no DCIS (1 study).

All of these simulation studies rely on assumptions which are not yet reliably measured in clinical practice and refer to single AI systems in hypothetical situations.

Recommendations for a future review

At present there is an insufficient volume and quality of evidence on clinical utility related to the use of AI in the NHSBSP or analogous populations to justify commissioning an evidence review. In light of the recent increase in interest in and funding of AI research a rapid review may be important in 1-3 years' time.

Gaps in the evidence

Overall, there is some evidence from early stage evaluation studies that AI has the potential to be an accurate tool to detect cancer in breast screening mammograms. The simulation studies show potential for AI to reduce radiologist's workload without compromising performance. However, there is no direct evidence on how AI may affect accuracy if integrated into UK breast screening practice. There were no studies that described accuracy of AI integrated into any breast screening pathway, and no prospective studies of test accuracy in clinical practice. No breast screening dataset from the UK was used in any of the included studies. Therefore, applicability of the current evidence to the UK screening context is limited. Furthermore, the available evidence is highly biased because 1) all but one study used enriched test sets, 2) about half the studies were biased by the laboratory effect, 3) the choice of the reference standard (biopsy/follow-up) was based on the outcome of the original read alone, 4) the threshold for interpreting the AI read was not pre-specified in any of the studies using the AI system as a stand-alone tool and 5) the reference standard was not equally accurate for all

tests (index and comparators). Seven out of 10 studies were not undertaken independently from the AI manufacturer.
There is no evidence from high quality randomised controlled trials or prospective cohort studies that compared the benefit of a breast cancer screening programme using AI to a screening programme without AI on clinical outcomes, patient management and practical implication outcomes. There is insufficient evidence how AI works for different subpopulations of women considering age, breast density, prior cancer and breast implants. Furthermore, evidence is missing on the types of cancers detected by AI to allow an assessment of potential changes to the balance of benefits and harms including potential overdiagnosis. Finally, there is no evidence on the impact of different mammogram machines or other sources of variability in current practice on the accuracy of AI systems, or on how the AI system may work within the breast screening IT system in the UK.

Algorithms are short lived and consistently improve. Assessments of AI systems may be out of date by the time of study publications and their assessments may not be applicable to AI systems available at the time.

Limitations

There were some limitations to the approach to the review. The review is not a full systematic review which means that only 20% of the search results, data extractions and quality appraisal were double checked. This may have increased the risk of error. For question 2 only citation searches were undertaken and included studies did not undergo quality appraisal.

The study inclusion criteria may have led to the exclusion of some relevant or more applicable studies. The exclusion of traditional CAD studies may mean that relevant studies were excluded because the distinction between AI CAD and traditional CAD is not clear cut. Internal validation studies and studies using temporal validation were excluded from the review because they are known to overestimate test accuracy. However, this has led to the exclusion of the more generalisable studies using UK datasets. Exclusion of internal and temporal validation studies in future reviews may potentially exclude large UK based screening studies because there are only 94 UK screening centres and many of them are involved in study development.

Finally, studies scored poorly in the QUADAS-2 assessment throughout. The QUADAS-2 adaptation was a first iteration and needs further refinement taking into consideration the QUADAS-2 AI version and AI reporting guides such as STARD-AI and CONSORT-AI which are expected to be published in due course.

Evidence uncertainties

Based on the rapid review and evidence map, the following evidence uncertainties remain to be addressed:

- how accurately a commercial AI system may classify women into recall / no recall when imbedded into the UK breast screening programme
- how accurately AI systems read mammograms at a pre-set threshold to classify women into recall / no recall
- if and how UK radiologists will change their recall behaviour when AI is incorporated into the screening pathway
- the impact on interval cancers of incorporating AI into the pathway
- what types of cancers AI detects preferentially
- comparative accuracy and impact of different AI systems
- whether images from different mammogram machines have an impact on the accuracy of AI systems
- whether variability in current practice has an impact on cancer detection with AI
- how the AI system would work with the breast screening IT systems in the UK

Introduction and approach

Background

Breast cancer is the most common cancer in women in the UK.[1] Approximately 55,000 women are diagnosed with breast cancer annually, which accounts for 15% of all new cases of cancer.[2] The age standardised incidence rate of breast cancer was an estimated 204.9 per 100,000 women in 2014, and is projected to increase to 209.5 per 100,000 women by 2035.[1]

Breast cancer is a heterogeneous disease that varies considerably between tumours, in terms of appearance, biology, and behaviour.[3] A range of risk factors for breast cancer have been identified, including sex, age, breast density, family history of breast cancer, genetic mutations, reproductive history, BMI, inactivity, and the use of hormone replacement therapy.[4 5]

Breast cancers are classified in relation to histopathology, grade, stage, receptor status, and genetic information. These classifications are informative for predicting outcomes and guiding treatment choices.[6 7] There are many histopathological types of breast cancer, the most common of which are in situ and invasive carcinomas. In situ carcinoma refers to the presence of cancer cells within ducts or lobules that have not spread to the surrounding tissue. Ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS) increase the risk that a women will go on to develop an invasive cancer,[8 9]. Low grade DCIS and low grade invasive cancer are associated with overdiagnosis (the detection of disease that would not have become symptomatic during a person's lifetime).[10-12] Approximately 1% of women die with occult invasive cancer in their breast, 9% with occult DCIS.[13] In contrast, invasive cancers are those where cancer cells have broken thorough the boundary of the duct or lobule into the surrounding breast tissue...

Both in situ and invasive carcinomas are graded according to the extent that tumour cells resemble normal cells. Grading provides an indication of how quickly tumours might grow and spread. In general, grading is conducted on a 3-point scale from 1 (low grade, slow growing) to 3 (high grade, quick growing). The process of staging provides information how advanced a tumour is, and far it has spread. One of the most frequently used cancer staging systems is the TNM system. In this, cancer stage is determined based on the size and extent of the primary tumour (T), the number of affected nearby lymph nodes (N), and whether the tumour has metastasised (M). In the TNM systems, cancers are categorised into five stages, from zero (carcinomas in situ) to four (cancer that has metastasised). More recently, classification of breast cancer also includes assessment of the expression of proteins and genes. For example, most breast cancer cells have receptors that attach to the hormones oestrogen and progreserone.[14] By attaching to these receptors, the hormones can cause the cancerous cells to grow. Cell growth is also caused by an over-expression of the protein HER2.[15] Target therapies are available to women who have oestrogen-, progesterone-, HER2, and BRCA1/2 positive breast cancers.[16]

Breast Screening Programme in the UK

The UK NHS Breast cancer screening programme (NHSBSP) began in 1988, with the aim of reducing breast cancer-related morbidity and mortality through the earlier detection and treatment of disease. In the current programme, women who are registered with a general practitioner are invited to breast cancer screening every 3 years from age 50 until their 71st birthday. Screening is conducted at 94 breast screening centres and their associated mobile screening units across the UK. The screening process involves digital mammography, in which x-rays are taken of each breast from two views: craniocaudal, and mediolateral oblique. The images from digital mammograms are stored and accessed via the Picture Archiving and Communications System. The interpretation of the images is carried out by two readers serially, with each reader making a decision about whether the image appears normal or if a woman should be recalled for further assessment. Arbitration (by either a 3rd reader or a panel of arbitrators) is employed where the readers do not agree on whether a women should be recalled.[17] In some centres arbitration is also employed in cases identified for recall by both readers, with the aim of reducing the number of false positive recalls to assessment. The number and identity of readers involved in arbitration, the pairing of readers, and whether arbitration is used when both readers are recalled are all decided by each breast screening centre individually, with the aim of maintaining sensitivity to detect cancer whilst reducing the number of false positive recalls to assessment and meeting national targets.

All readers (radiologists, radiography advanced practitioners, breast clinicians) undergo formal training, read a minimum of 5,000 mammograms per year, participate in assessment clinics and continuing professional development, and audit their performance.[18] Rarely, repeat examinations are required due to technical failures, e.g. inadequate images.[19] Women who have normal screening results are invited to return for routine screening after 3 years. Women who are recalled are offered additional test to determine whether they have cancer. These tests include clinical assessment, imaging (e.g. x-rays, ultrasound, digital breast tomosynthesis), and needle biopsy (to confirm or exclude malignancy).[20] Women who test negative at this stage return to routine screening, while those whose cancer is confirmed are referred for treatment, such as surgery, endocrine therapy, chemotherapy, and radiotherapy.[16]

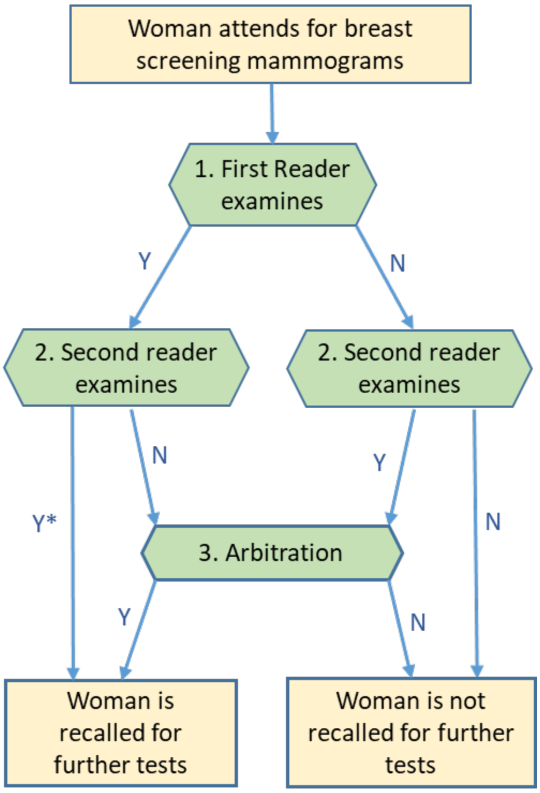

Woman attends for breast screening mammograms
1. First Reader examines
Y
N
2. Second reader examines
2. Second reader examines
Y*
N
Y
N
3. Arbitration
Y
N
Woman is recalled for further tests
Woman is not recalled for further tests

Figure 1. The current breast screening pathway in the UK.

Y=Recommendation to recall for further tests as there are indications of cancer.

N=Recommendation not to recall for further tests.

*Some centres send cases recalled by both radiologists for arbitration decision to reduce recall rate. AI is most commonly proposed in the following roles: to replace step 2 (the second reader); to assist the decision-making of the first and/or second reader for example by providing localisation prompts; to entirely replace steps 1 to 3, or as a pre-screening tool to select women whose mammograms have no signs of cancer and do not require human reading.

The breast cancer screening programme is supported by key information systems, which are used to manage women through the process (e.g. manage call/recall of women), to record screening data including the outcomes of screening appointments, to produce annual return data (e.g. screening coverage), and to enable the evaluation of the programme. In England the information systems are the Breast Screening Select system, the Breast Screening Information System, and the National Breast Screening System. Wales and Ireland use versions of the National Breast Screening System linked to the National Health Application and Infrastructure Services. In Scotland the system is called the Scottish Breast Screening Programme.

Data from the NHSBSP indicated that in the period from 2018 to 2019, 2.56 million women aged 50 – 70 were invited to screening in England, of whom 1.82 million women attended. Recent data from the UK indicated that national coverage was 77.1% in England (2018 – 2019), 72.8% in Wales (2018 - 2019), and 73.6% in Northern Ireland (2018-2019) compared to an acceptable threshold of 70%.[21 22] These data are not readily available for Scotland. Recent data (2017) have suggested that there are few differences between the 4 nations in terms of age-standardised (1) incidence of invasive breast cancer (Scotland: 164.6, Wales: 165.7, England 166.7, Northern Ireland: 166.8; all per 100,000 population), or (2) mortality (Scotland: 32.5, England: 33.3, Northern Ireland: 34.3, Wales: 35.2; all per 100,000 population) for women.[23 24] In contrast, the age-standardised incidence rates of situ cancers in women (all per 100,000 population) was lower in Scotland (16.6) than England (25.9), Wales (26.2), and Northern Ireland (27.6).[25]

Further differences between the UK nations are uptake and achievement of the 3-year round length. Key performance indicators show that the national average performance for screening round length (proportion of eligible women with 1st appointment being offered within 36 months of previous round) was below the acceptable threshold of 90% in England (81.8%; though 52 out of 78 screening centres exceeded this threshold), and above the threshold in Northern Ireland (98.1%).[26 27] These data are not available for Scotland and Wales. National uptake exceeded the acceptable uptake threshold (70%) in England (71.1%), Northern Ireland (75.0%), Scotland (72.3%), and was below the threshold in Wales (69.1%).[21 26-29] In England, uptake was lower amongst women invited to screening for the first time (61.1%) compared to those previously invited (72%).[21]

Nationally, of those attending screening in England, 3.8% (84,559) of women were referred for further assessment, with 15,285 cases of breast cancer diagnosed (8.2 per 1,000 women screened); the majority of cancers were invasive (78.8%).[21] The remaining cancers were non-invasive of which 80% are generally classed as DCIS.[30] DCIS is the earliest form of breast cancer and is sometimes a precursor of invasive breast cancer. Approximately 6,000 additional interval cancers (cancer detected symptomatically between screening rounds) are diagnosed annually in England, of which an estimated 20% were present but not detected during screening (false negative test results).[31]

Breast cancer screening has been estimated to result in a 20% reduction in breast cancer mortality in women invited for screening. For the UK screening programmes, this corresponds to prevention of around 1300 deaths from breast cancer each year.[32]

In addition to the anxiety associated with false positive test results and the false reassurance associated with false negative test results, breast cancer screening is associated with a number of other harms, most notably overdiagnosis and overtreatment. Approximately 19% of screen-detected cancers are estimated to be overdiagnosed.[32] For example, DCIS represents approximately 21% of cancers diagnosed at breast screening.[33] Although DCIS may progress to invasive disease and hence may result in death from breast cancer, in many cases, especially with lower grade DCIS and in older women, the diagnosis and treatment for DCIS has no impact on breast cancer specific mortality. This is difficult to quantify, because there are few women with untreated DCIS in whom to measure the natural history. A recent review of 89 women with unresected DCIS followed up for 59 (range 12-180) months showed that 29 women were diagnosed with invasive cancer.[34] Within this cohort high grade DCIS was associated with significantly higher rates of progression to invasive cancer, over shorter timescales, and higher grade of subsequent invasive cancer. A review of 35,000 women diagnosed with DCIS within the NHSBSP showed increased death rates from breast cancer at 5 years from diagnosis and beyond compared to the national mortality rates, with larger size of DCIS associated with higher breast cancer mortality.[9]

Ninety percent of all cases of DCIS are detected only on imaging, particularly digital mammography, because of its association with microcalcifications.[30] Digital mammography also increasingly detects a heterogeneous group of lesions of "uncertain malignant potential"(or B3 lesions), which often present as clustered microcalcifications.[35] These are sometimes considered a type of 'pre-DCIS', although they often do not progress to DCIS or invasive cancer. There are several types of B3 lesions including atypical ductal hyperplasia (ADH), flat epithelial atypia (FEA), lobular neoplasia (LN), papillary lesions and radial scars.[36] These B3 lesions are diagnosed in 5-10% of core needle biopsies performed as part of the NHSBSP.[35] They have a 17% risk of upgrade to malignancy on core needle biopsy [36] and a suggested 4-times increased risk of subsequent breast cancer over time.[37] [38] However, upgrade ranges from <2% to 39.5% by type of lesion raising concerns over overdiagnosis.[35] Advancements in imaging technology has led to an increase in the detection of DCIS and B3 lesions and there is potential that this trend could be promoted by artificial intelligence (AI) if AI differentially detects more microcalcifications. If AI preferentially identified lower grade DCIS and B3 lesions, it could increase rates of over-diagnosis and over-treatment, whereas if it preferentially identified higher grades of DCIS and higher grade invasive cancers, it may reduce over-diagnosis and over-treatment.

Similarities and differences between breast screening programmes across countries

Breast cancer screening programmes are in place in many countries, with variation between them in terms of organisation, screening processes, age at which screening commences and stops, and round length. For example, screening is organised at a national level or local/regional levels, offered to all eligible women or via health insurance plans, intervals between rounds range from 1 – 3 years, eligible age at which screening is first offered varies from 40 – 50, mammograms are interpreted by 2 readers or a single reader with/without computer-aided detection, and mammography interpretation is classified as either a recall/no recall decision or on the basis of risk assessment tools (e.g. Breast Imaging-Reporting and Data System).[39-42]

Artificial intelligence

AI is a capacious umbrella term, but broadly refers to the branch of computer science dedicated to the creation of systems that perform tasks that usually require the input of human intelligence.[43] The field of research into AI is largely credited as having begun subsequent to the Dartmouth Conference in 1956,[44] and subsequently the popular discourse has become one of the promises of 'Strong AI' with the ability to perform understanding, cognitive and intellectual tasks at (at least) the ability of a human (and/or experience consciousness).

Whilst ‘Strong’ or General AI which exceeds human performance, intellect and reasoning remains theoretical, there has been substantial progress in using AI to interpret complex data and provide a quantitative assessment of particular problems. This is evidenced by the rapid progress in natural language processing and image recognition tasks which are now pervasive in daily life, such as voice or face recognition algorithms.

This progress has been comparatively recent; made possible by both increased computational power and the use of deep learning algorithms which are now being applied to the healthcare sector and are the focus of this review.

Development process of AI
AI development is a complex process; involving data collection, preparation, model construction (which includes training of an algorithm and fine-tuning) and subsequent evaluation.[45] The process of developing a model requires data; and different types of machine learning approaches are used depending on the type of data available. These groups of machine learning techniques are generally considered to be either unsupervised or supervised learning. In unsupervised learning, the goal is to discover patterns in the data such as the existence of sub-groups (clustering) or to separate informative from non-informative information (dimensionality reduction, noise removal). The goal is typically not to make any predictions from the data, but to learn about the hidden characteristics of the data. In practice, achieving good results with unsupervised learning remains difficult, [43] and this also applies to image recognition. In contrast, in supervised learning the data is pre-labelled, and the goal is to infer rules that map the input to the target output labels. These rules can then be applied to new data without labels in order to make predictions. This is the method most commonly used for image recognition.

The dataset(s) used to develop an algorithm are generally divided into training set, tuning set and validation test set.[46] The training set allows optimisation of the parameters used in the model, for example, the weights in a deep neural network. The tuning set is used for tuning hyperparameters and for model selection, and finally when all model parameters are set the validation test set is used to report on final model performance on unseen data.[43] Park et al. describe many of the issues related to these steps.[47] This includes the issue of internal validation whereby the validation dataset used to assess a model uses data which were used to develop that model. During the validation step, bootstrapping and/or cross-validation (resampling techniques which use the original data) are used to assist with preliminary assessment and fine-tuning of the model. The issue of using data on which an algorithm was trained with is that models can be prone to overfitting; whereby the model fits the trained data extremely well, but to the detriment of the model’s ability to perform when presented with new data, which is known as poor generalization. The split-sample approach is generally an inefficient form of internal validation because it does not accurately reflect a model’s generalisability.[48]

Due to these limitations of internal validation, external validation as part of the testing (assessment of the model using unseen data that was not available during method development) is critical to understanding a model's effectiveness. Separate and unused data need to be used for this stage, and can come from a range of sources; prospectively recruited women from the same site (temporal validation) or collected from different sites (geographical validation).Temporal validation has been reported by some to be sufficient in meeting the expectations of an external validation set[47] and regarded as an approach that lies midway between internal and external validation by others.[48] However in a screening context often there are additional issues for temporal validation, in particular the same women attending repeat screens, in addition to use of the same machines by the same personnel. Geographical validation has the benefits of understanding the effectiveness across potentially different technical parameters (such as different machines) and operating personnel.[47] While geographic validation will generally ensure separate and unseen data is used for training and validation, there may be instances where women have moved between screening catchment areas between screening rounds. Determining the test accuracy of AI models may also require prospective test accuracy studies.

The use of artificial intelligence in breast screening programme

The use of computers to assist in healthcare, and breast screening is not new. During the late 1980s, computer assisted detection (CAD); was introduced to the healthcare setting, and mammography,[49] these systems rely on rule-based algorithms designed by domain experts or less complex machine learning models that use hand-crafted features which might be suggestive of a diagnosis of interest. For example, on a mammogram, the raw data comprising an image (such as pixel values) would need to be transformed such that the learning system could detect patterns (such as shape, volume or texture) which could then be classified into benign or malignant. Transforming that initial data requires substantial human expertise in design and is limited in terms of the feature complexity. Examples of this type of machine learning include support vector machines or random forests; and are broadly considered 'traditional' or conventional forms of AI.[50] Whilst the use of traditional CAD in mammography broadly acted as either a second opinion or as an assist to radiologists, there was a lack of good quality evidence that traditional CAD had a significant effect on cancer detection rates.[51] They were not used widely in UK breast screening programmes.

The rapid developments in recent years of AI have largely been a result of the adoption of 'modern' forms of AI employing more complex machine learning models, made possible by both increased computational power, and in image recognition; the transfer to digital capture and storage. Almost all modern AI approaches make use of 'deep learning'. Deep learning refers to a part of machine learning, that employs 'representation-learning' methods typically using deep artificial neural networks (dANN). In representation learning the goal is to automatically learn the task-relevant features from the raw input data (e.g., the raw pixel values of an image), making the hand-crafted feature engineering of traditional machine learning obsolete.

Certain types of deep learning have excelled at image recognition, and were piloted as early as the mid-1990's for breast screening; Chan et al. and Sahiner et al. conducted some of the first studies to demonstrate using a convolutional neural network (CNN) with mammography images, with some success.[52] [53] CNN's are a class of Deep Learning, based loosely on the structure of neurons in the animal visual cortex. These start with the raw input (for example, an image) and repeatedly transform into a representation at a slightly higher, more abstract level; with enough subsequent transformations complex, highly nonlinear functions can be learned that map input images to output target labels. LeCun et al. describe this process clearly; [54] in the case of an image (or mammogram); the array of pixel values and the first layer of representation may be distinct edges at orientations or locations in the image. The second layer might detect motifs through spotting the arrangement of edges, and the third layer might be the assembly of motifs in large combinations which correspond to familiar objects and subsequent layers detect combinations of the parts. The key difference to traditional CAD is that the layers of features are not designed by human expertise; instead, they are learnt from the underlying data. The 'deep' refers to these 'hidden' layers (which in recent neural network architectures for visual recognition can comprise tens or even hundreds of layers); the non-linearities facilitate the learning of complex mathematical functions represented by the neural network.[55]

For classification the different features are combined to determine regions with suspicious findings. A value is assigned to each region, representing the level of suspicion that cancer is present. The scores from all mammography views are combined into the examination-based score, and calibrated such that the number of mammograms in each category is roughly equal in a screening setting with a higher score (Transpara score) representing an elevated risk of cancer.[56] Other outputs from different systems include giving a binary recall/no recall decision on the whole examination, specifying the images which gave rise to a recall outcome and marking the features giving rise to increased suspicion which can be accompanied by an indication of the percentage probability of malignancy.

In transfer learning the knowledge from one image domain can be transferred to another image domain and depends on the extent of similarity between the databases on which a CNN is pre-trained and the database to which the image features are being transferred. Transfer learning may be used to fine-tune the CNN model pre-trained on a large mammography dataset to detect masses in small mammography datasets.[57]

There are numerous potential benefits of AI within breast screening programmes, and roles for AI have been suggested along the radiology pathway,[58] alongside clinical decision support; processing,[59] quality control,[60] and understanding narrative radiology reports.[61] The primary drivers for AI in medical imaging have been cited as the desire for greater efficacy and efficiency in clinical care.[50] In this review we only consider the role in examining breast screening mammograms for signs of cancer.

Potential benefits of the implementation of AI in breast screening
For women, a false positive result can lead to psychological distress,[62] anxiety and impact upon the rates of return for mammography.[63] The impact of false negatives on the woman may be delay in detection of cancer and potentially worse outcomes. The impact of false negatives for the screening programme are potential legal action and erosion of public confidence.[64] Radiologists' errors that lead to false negative results fall into three categories: Search, perception and interpretation[65] In search and perception errors, the lesion is evident but not identified by the radiologist; this may be in part due to subtle or architectural features of malignancy which make the lesion difficult to perceive. Interpretation errors are where the radiologist identifies the lesion but decides it is not sufficiently suspicious for recall for further tests. The causes of errors are not fully known, but case difficulty, inattention, fatigue or lack of experience can be contributory.

There is the potential for improvements in accuracy through computer aid; AI can be trained on many thousands of images, many more than could be seen through a radiologist's career. Secondly, an algorithm is unaffected by fatigue or subjective diagnosis; there are increasing workloads for radiologists with imaging volumes which have grown at disproportionate rates to imaging utilisation.[66] The increasing workload is accompanied by a shortage of radiologists in the UK.[67] It has been suggested that in the UK setting, AI could reduce the workload involved by replacing the second reader role, AI could also potentially reduce workload by increasing test specificity, thereby reducing the significant extra work in assessment in women recalled. Specificity is a key driver of workload because examining a set of mammograms as second reader may take around 30 seconds, whereas an assessment appointment may take much longer and includes invasive and unpleasant procedures.

The impact of AI on waiting times for mammography has not been demonstrated or modelled, the logical impact of reduced workload would be reduced waiting lists and time for results; both of which are considered barriers to uptake and a cause of anxiety.[68]

Barriers to implementation of AI in breast screening
There are, however, a number of social, ethical and legal questions which AI's role in mammography raise. These are comprehensively described in the review by Carter et.al. [69] The first of these issues is that of the intrinsic values which an algorithm may in itself adopt; the deep learning algorithm may perform differently depending on the characteristics of what is being examined and arise from how the algorithm is trained. Biases may develop through features of the mammogram, or different demographics of the women screened. This speaks to the importance of understanding the validity of studies involving AI and algorithms' transferability to other settings, but also the crucial problem of interpretability. Unlike human interpretation, it can be difficult to understand how or why an algorithm has made a decision (known as the 'black box' problem[70]). Carter et al. argue that AI systems will inevitably encode values, and that those values may be in turn difficult to discern.[69]

The ‘black box’ nature of some deep learning algorithms present legal and governance challenges. The traditional clinician-led decision making model has clear lines of accountability; if an AI was to cause harm, it is currently unclear where the responsibility would lie;[71] and although strides are being made towards reforming the legislation both in Europe and the USA, this is a gradual process.[72]

Development of AI also can raise many ethical questions. As individual patient data is required for the training of the AI, how this data is collected becomes increasingly important. Have individuals attending screening been able to provide consent that their data may be used for this purpose? Attempts have been made to address this, particularly in Europe where the General Data Protection Regulation is explicit in rules for storing, using data and adopts an ‘opt-in‘ as the default approach, making consumer consent for data use clear.[72] The sharing of data has significant monetary implications, and governmental release of data to private providers without consent raises significant ethical questions.[69]

The impact of AI on the radiology workforce, and the perception of the speciality is likely to be significant, and also have both positive and negative effects. The possibility of AI replacing radiologists is already leading to a significant proportion of medical students discounting the speciality as a career choice; Sit et al. recently surveyed 484 medical students from 19 UK medical schools, and found 49% were less likely to consider a career in radiology due to AI.[73] Related to this is a clear need for education for health professionals to enable them to work with digital tools (including AI), highlighted in the recent governmental Topol review.[74] Amongst radiologists the perception of AI is, however, optimistic, following a large survey of radiologists in France by Waymel et al. the majority thought that AI will have a positive impact on future practice with the expectation that imaging related medical errors and interpretation time would be reduced, allowing for an increase in the time spent with patients.[58]

Kotter and Ranschaert call for a proactive attitude amongst radiologists towards AI to identify existing needs allowing AI developers to train algorithms with a clearly clinical purpose, to recognise the need of AI to be included in medical training and to evaluate these systems in terms of how the use of AI enhances the performance of radiologists rather than as stand-alone systems.[75]

Current policy context and previous reviews

The UK NSC recommends screening for breast cancer. National screening programmes are in place in each of the four countries of the UK. No prior review has been conducted on the use of artificial intelligence in breast cancer screening.

Objectives

The aim of this review is to synthesise the evidence on the use of deep learning AI algorithms (assistive or stand-alone) in breast cancer screening.

This review consists of 2 types of evidence products: 1 rapid review and 1 evidence map. The rapid review gauges significant evidence on the diagnostic accuracy of AI algorithms in detecting breast cancer and evidence gaps. The evidence map gauges the volume and type of evidence on the clinical utility of AI algorithms when used to read mammograms in a breast screening programme.

These evidence products provide the basis for discussion on whether a systematic review on the topic is justified at this time.

The key questions for this review are shown in **Table 1.**

Table 1. Key questions for the evidence summary, and relationship to UK NSC screening criteria

	Criterion	Key questions	Studies Included
	THE TEST		
4	There should be a simple, safe, precise and validated screening test.	1) What is the accuracy of AI algorithms to detect breast cancer in women attending screening mammography?	N=7 (8 records)[56 76-82]
5	The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.		
	THE SCREENING PROGRAMME		
11	There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an “informed choice” (eg. Down’s syndrome, cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.	2) What is the clinical impact of the use of AI algorithms to detect breast cancer in mammograms compared to current practice in breast screening programmes?	N=8 (9 records)[56 77 79-85]
12	There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public.		

Methods

The current review was conducted by Warwick Screening (University of Warwick), in keeping with the UK National Screening Committee evidence review process. Database searches were conducted on 9 September 2020 to identify studies from 1 January 2010 relevant to the review questions 1 and 2, detailed in **Table 1**. To identify further studies relevant to question 2, the reviewers performed additional citation searches and trial name searches for all included studies for question 1 on 20 November 2020.

Databases/sources searched

The search strategy comprised the following elements:
1) Searching of electronic bibliographic databases,
2) Contacting experts in the field,
3) Scrutiny of references of included studies and relevant systematic reviews,
4) Identification of additional evidence for question 2, by performing citation searches and trial name searches for all studies included for question 1.

One systematic literature search was undertaken on 9 September 2020 to cover both review questions. The search strategy was developed in MEDLINE (Ovid) using terms relating to breast AND AI AND (screening OR mammography) limited to test accuracy studies or randomised controlled trials. The search was adapted as appropriate for the other bibliographic databases: MEDLINE In-Process & Other Non-Indexed Citations (Ovid); EMBASE (Ovid); Web of Science (Ovid), and the Cochrane Library: Cochrane Database of Systematic Reviews (Wiley). A copy of the search strategies that were used in the major databases is provided in Appendix 1. The searches were limited to studies published since 2010 (i.e. the date of the first identified study on the use of AI for mammography reading).

Citation searches were conducted in Web of Science and Google Scholar on 13 identified relevant papers (including UK internal validation studies).[56 76-78 80-82 85-90] Two papers mentioned specific trials - Malmö Breast Tomosynthesis Screening Trial, (NCT01091545)[85] and the CSAW (Swedish Cohort of Screen-Age Women) data set.[80] These trials were searched for in Google/Google Scholar and cited references followed.

Eligibility for inclusion in the review

The following review process was followed for questions 1 and 2:

1. Each abstract was reviewed against the inclusion/exclusion criteria by one reviewer. Where the applicability of the inclusion criteria was unclear, the article was included at this stage in order to ensure that all potentially relevant studies were captured. A second independent reviewer provided input in cases of uncertainty, and independently validated 20% of the first reviewer's screening decisions. Any disagreements were resolved by discussion until a consensus was met.
2. Full-text articles required for the full-text review stage were acquired.
3. Each full-text article was reviewed against the inclusion/exclusion criteria by one reviewer, who determined whether the article was relevant to one or more of the review questions. A second independent reviewer provided input in cases of uncertainty and validated 20% of the first reviewer's screening decisions. Any disagreements were resolved by discussion until a consensus was met.

Eligibility criteria for each question are presented in **Table 2** below.

Table 2. Inclusion and exclusion criteria for the key questions

Key question	Inclusion criteria						
	Population	**Target condition**	**Intervention**	**Reference standard**	**Comparator**	**Outcome**	**Study type**
Question 1	Women's breast cancer screening mammograms obtained by digital mammography.	Breast cancer (overall and stratified by spectrum of disease)	Deep learning AI algorithms incorporated into the breast screening pathway or its approximations (AI applied to women's mammograms, but not evaluated as part of pathway change) (see **Table 3** below).	Cancer confirmed by histological analysis of biopsy samples (or follow-up with no cancer diagnosis). Definition of cancer may include DCIS or invasive cancer only.	No comparator or head-to-head comparisons (human reader, pathway without AI or different AI system).	• Test accuracy (decision is made for a whole woman or a whole image) for breast cancer detection overall and by spectrum of disease. • Spectrum of disease detected (for example grade, stage, size, nodal involvement) • Subsequent symptomatic cancer detection including interval cancers and their characteristics (for example grade, stage, size, nodal involvement) • Subsequent screening (next round) cancer detection and their characteristics (for example grade, stage, size, nodal involvement)	• Randomised test accuracy studies • Prospective test accuracy studies • Retrospective test accuracy studies using geographical validation only • Comparative cohort studies (and single arm cohort studies in the UK only) • Enriched test set multiple reader multiple case laboratory studies (deprioritised and included only if other evidence was lacking)
Question 2	Women's breast cancer screening mammograms obtained by digital mammography.	Breast cancer (overall and stratified by spectrum of disease)	A breast screening programme that uses deep learning AI algorithms (whole pathway).	Cancer confirmed by histological analysis of biopsy samples (or follow-up with no cancer diagnosis). Definition of cancer may include DCIS or invasive cancer only.	A breast screening programme that does not use deep learning AI-based algorithms (whole pathway).	Any clinical utility outcomes, including: • Morbidity, mortality, quality of life • Interval cancers (or proxy if interval cancer is not possible) • Spectrum of disease Patient management and practical implication outcomes such as: • Workforce (e.g. workload, training) • Costs	• Randomised controlled trials • Cohort studies (retrospective / prospective) • Systematic reviews

Key question	**Exclusion criteria**						
	Population	**Target condition**	**Intervention**	**Reference standard**	**Comparator**	**Outcome**	**Study type**
Question 1	• Diagnostic mammograms or <90% screening mammograms or type of mammogram not specified • Images of cancer for grading or staging • Subpopulations only (e.g. women with dense breasts, women with interval cancers only, women with single density, images of masses) • Images not from mammography • Mammography types using intravenous injection of a contrast agent • Non-digital mammography • Images not of whole mammogram (e.g. region of interest only) • Simulated cancers / mammograms	Not breast cancer	• AI for personalised (future) cancer risk • Improvement of AI system (e.g. comparison of different training modalities) • Image processing algorithms (e.g. for visual enhancement, de-noising, pixel resolution) • AI not for cancer detection (e.g. for segmentation of pectoral muscle or mammary glands, parenchymal patterns, breast density) • Detection of subtypes only (e.g. spiculated masses, architectural distortion, asymmetries, microcalcifications) • AI for cancer segmentation without classification by AI or human reader • “Old” CAD	NA	NA	• Area under curve (AUC), diagnostic odds ratio (DOR), or other measures which are not expressed at the clinically relevant threshold. • Accuracy without outcomes characterising the trade-off between false positive and false negative results. • Classification not relating to whole woman / whole image (e.g. regions of interest, false positive marks per image). • Study only reports sensitivity or only reports specificity.	Studies using an internal validation test set (e.g. x-fold cross validation, leave-one-out method), split sample validation or temporal validation.
Question 2	See Question 1	Not breast cancer	See Question 1	NA	No comparator	NA	See Question 1

The rationale for the review's eligibility criteria are as follows:

Population

This review focused on the UK screening context, therefore, studies or sub-studies that did not use the same imaging technique or images that represented a different use case to the UK were excluded. In the UK, screening mammography is undertaken using full field digital mammography (FFDM) in women attending breast screening. Other imaging techniques not classed as mammography or not available in the UK or not digital are not relevant. Results from imaging other than FFDM may not be applicable to FFDM in the UK. Images of part of a mammogram or of diagnostic mammograms do not represent the use case in the UK screening context, which requires recall or not decisions to be made on women's (craniocaudal and mediolateral oblique) screening mammograms for both breasts.

In addition, studies or sub-studies that only included images with cancer were excluded as this is not sufficient to estimate test accuracy of AI for screening mammograms, as it excludes specificity, and the trade-off between sensitivity and specificity.

Finally, studies or sub-studies on images of subpopulations by screening risk or screening outcome were excluded as they do not represent the screening population and no inference on the performance of the AI system in a screening population can be drawn (however, subpopulations by ethnicity or socioeconomic status are included as the impact of any change on equity is important). If the population represents a group of women at any stage within the screening pathway (e.g. recalled women without selection on final diagnosis) the study is included on the assumption that AI could be incorporated for this subgroup only.

Intervention / Index test

Studies were included if the index test was AI technology to assist or replace human readers in deciding whether to recall women for further tests from screening mammography. The review aimed to provide evidence sufficiently robust to enable the formulation of decisions on the future integration of AI into the UK breast cancer screening programme. The most applicable evidence was considered to be from studies where the index test is the AI system integrated into the screening pathway, as it would be used in screening practice (**Table 3** column 1 and

Figure 2). For example, if the AI is designed to replace the second reader, then the index test is the combination of reader 1, AI as reader 2 and arbitration to decide whether to recall the woman. Nevertheless, studies which reported the test accuracy of the index test as AI for a 'full single read' (i.e. test accuracy reported is applicable to the whole process of detection and classification of mammograms into cancer / no cancer on which a recall decision can be based) were also included (**Table 3** column 2). Whilst this version of the index test does not translate directly into screening practice (unless AI fully replaces all humans in the decision) these studies are important in providing direct (head to head) comparisons between different AI systems.

Table 3. Interventions (index tests) included in the review

Intervention (index test) of interest as a pathway change	Approximations (Index test applied to women's mammograms, but not evaluated as part of pathway change)
AI replaces second reader (Study has to report test accuracy of the whole process; for example first reader + AI system + arbitration.)	AI classifying women's mammograms in any manner that can be interpreted simply into a decision to recall or not (preferably with localisation to guide follow-up tests, but without localisation also included). (Study has to report test accuracy of AI system in classifying women's mammograms.)
AI as a reader aid (Study has to report test accuracy of whole process integrated into pathway; for example first reader + second reader using AI + arbitration; or AI could equally be used by the first reader and/or arbitration.)	AI in combination with radiologists (or equivalent) classifying women's mammograms in any manner that can be interpreted simply into a decision to recall or not. (Study has to report test accuracy of AI system and radiologist combined in classifying women's mammograms.)
AI pre-screening AI to classify low risk mammograms which require single or no human readers. (Study has to report accuracy of whole process integrated into pathway.)	AI system classifying women's mammograms into different risk categories, in a manner that can be interpreted simply into dichotomous low risk and high risk categories. (Study has to report test accuracy of AI system by risk category.)
AI completely replaces human reader(s) AI classifies women / images into cancer / no cancer (with or without localisation). (Study has to report test accuracy of AI system.)	No approximations required, AI would replace existing pathway.
Other AI to classify mammograms in a role not previously covered. (Study has to report test accuracy of whole process integrated into pathway)	Any relevant approach where the AI classifies women's mammograms. (Study has to report test accuracy of AI system in classifying women's mammograms)

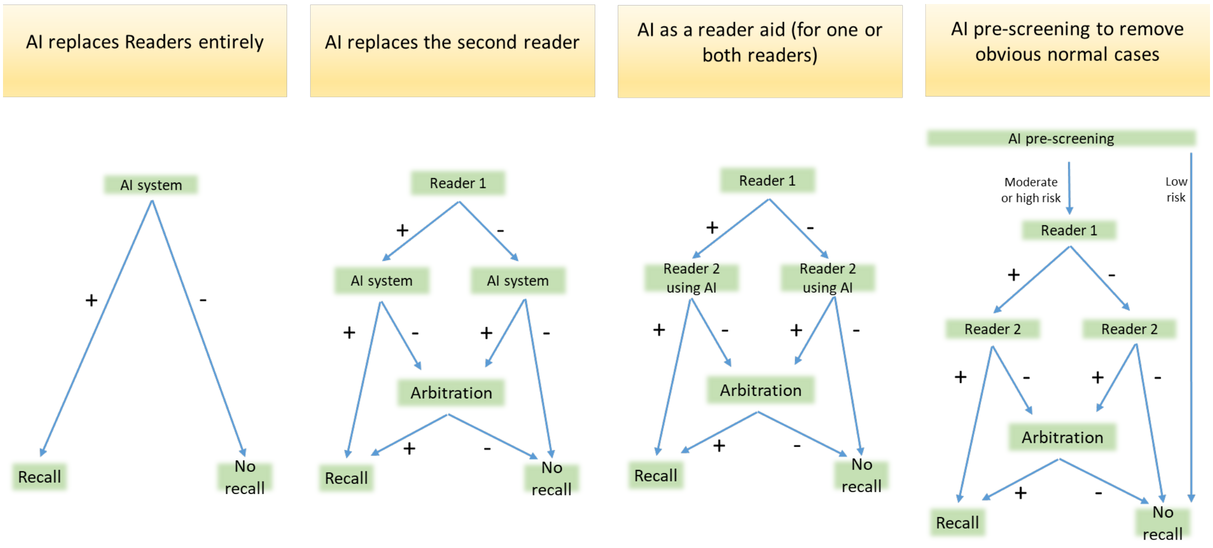

AI replaces Readers entirely
AI replaces the second reader
AI as a reader aid (for one or both readers)
AI pre-screening to remove obvious normal cases
AI system
+
-
Recall
No recall
Reader 1
AI system
AI system
Arbitration
Reader 2 using AI
AI pre-screening
Moderate or high risk
Low risk
Reader 2

Figure 2. Some potential roles of AI in screening pathway

+ denotes recommendation to recall for further tests,

- denotes recommendation not to recall for further tests

On the other hand, studies on the prediction of future cancer risk including the detection of breast density and parenchymal patterns as risk factors were excluded as the review only considered AI for the detection of cancer on screening mammograms. Similarly, the detection of cancer subtypes does not present the complete picture of cancer detection (e.g. microcalcifications are associated mainly with DCIS and not with cancer; detecting microcalcifications only will miss some types of cancers). On their own, these AI systems do not provide the information on cancer present/ not present to inform a decision whether to recall or not recall. Systems reporting single features could be combined to provide a more complete picture, however, studies would need to report an overall outcome of test accuracy of the combination of systems.

Studies reporting AI systems that read regions of interest (ROIs) taken from databases where the accuracy of the identification of the ROI is not reported were classed as not sufficient. Studies reporting the segmentation of cancers compared with ROIs from a database without classification were classed as not sufficient. Studies using AI for segmentation and a second AI system or a human reader for classification were included.

Studies reporting image processing algorithms without segmentation (or segmentation of pectoral muscle or mammary glands) and without AI or human reader for classification of cancers do not provide test accuracy of AI systems in the detection of cancers and were excluded.

Studies reporting the test accuracy and effectiveness of traditional CAD systems were considered significantly different from studies assessing deep learning AI systems and were excluded.

Comparator

Studies were included if the comparator was the current breast screening pathway where decisions of whether to recall women for further tests after mammography are made by two human readers plus arbitration, or an approximation to that pathway (for example a single human reader).

Outcomes

Studies were included if the primary outcome for question 1 was test accuracy using biopsy proven breast cancer diagnosis as the reference standard. The trade-off between false positive and false negative results is critical to test accuracy, so studies reporting only statistics related to one or the other or a global measure of test accuracy were excluded.

Studies were included in question 2 if they reported any outcomes relevant to the impact of adopting AI to examine mammograms into the breast screening pathway (e.g. clinical utility outcomes, patient management and practical implication outcomes).

Interval cancers, cancer spectrum, and number of false positives are important outcomes for questions 1 and 2 and were covered in different study designs for both questions. There is, therefore, some overlap in reporting these outcomes in discussion of findings for question 1 and question 2.

Study design

For question 1 to evaluate accuracy, studies with any of the following five study designs were included: prospective test accuracy studies; retrospective test accuracy studies using geographical validation only; randomised test accuracy studies; comparative cohort studies and single arm cohort studies in the UK only (**Table 4** and **Figure 3**).[91] Direct (head to head) comparisons between several AI systems and current practice is important to evaluate accuracy, so that differences between tests and testing pathways can be established without the confounding effects of differences between studies in participating women, participating centres, and study design.

For question 2 randomised controlled trials, cohort studies (retrospective and prospective), and systematic reviews were included.

Table 4. Definitions of included study designs

Study design	Definition
Prospective test accuracy studies	A (consecutive) cohort of women attending routine breast screening has their mammograms read by human readers and AI. Screen positive women on either test are recalled for further testing. Women may be followed-up for missed cancers on screening.
Retrospective test accuracy studies	Available mammograms from routine breast screening databases are retrospectively read by AI. Reference standard from biopsy from routine records and follow up for missed cancers on screening. Comparison to the original reader decision in clinical practice. (Without comparison also eligible but significantly less informative)
Randomised test accuracy study	Women of screening age are randomised to having their mammograms read with AI (AI pathway) or without AI (human comparator pathway). Screen positive women are recalled for further testing. All women are followed-up for missed cancers on screening.
Comparative cohort studies	Similar to randomised test accuracy study but allocation to AI or human pathway not at random. Includes a range of quasi-experimental designs. Before after study is one of the simpler designs.
Single arm cohort studies (Research pilots)	Implementation of AI into screening practice and measurement of outcomes without a comparator.
Enriched test set multiple reader multiple case laboratory study	Retrospective test set examined prospectively in a laboratory setting by human readers and AI.

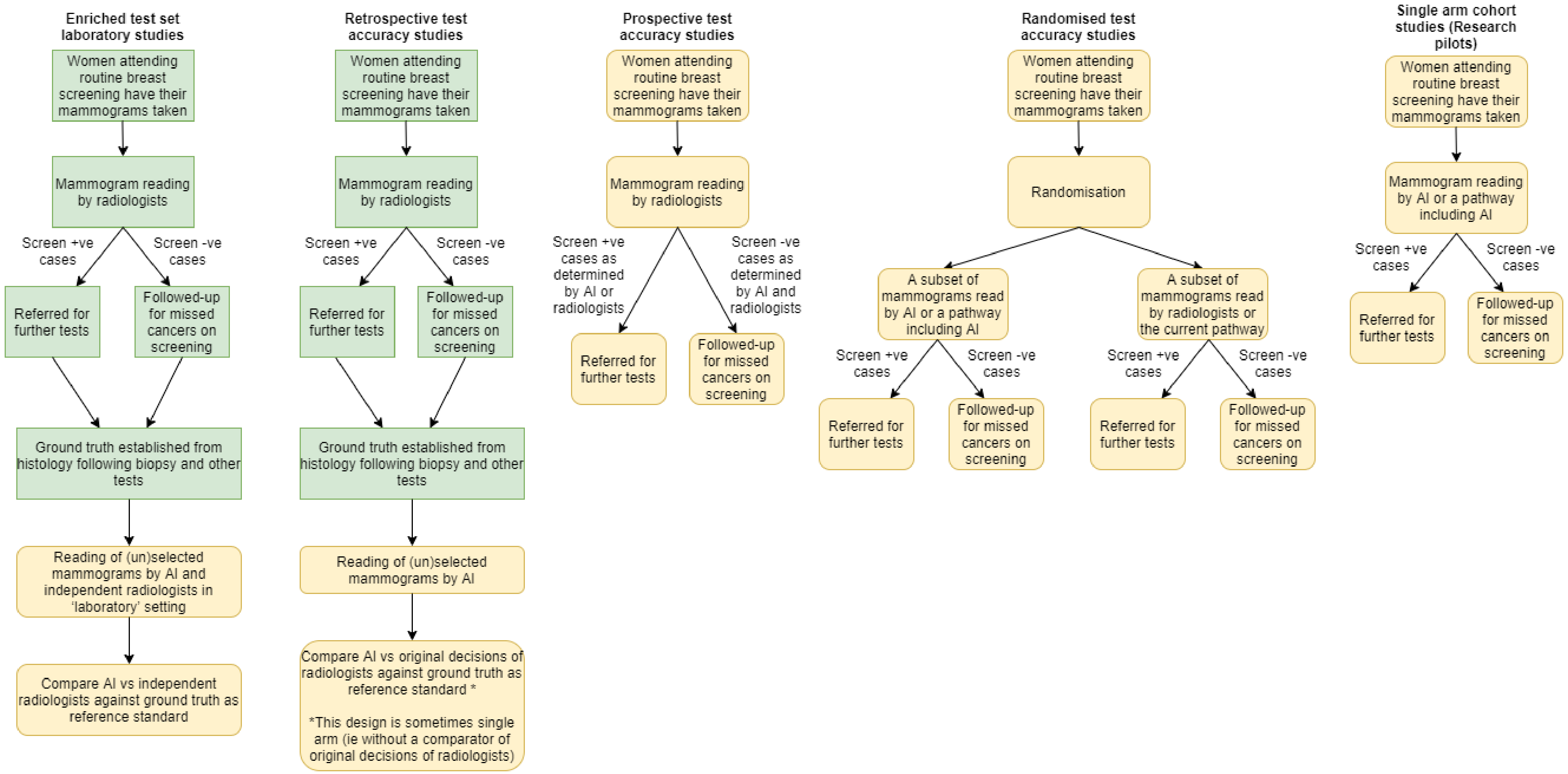

Enriched test set laboratory studies
Women attending routine breast screening have their mammograms taken
Mammogram reading by radiologists
Screen +ve cases
Screen -ve cases
Referred for further tests
Followed-up for missed cancers on screening
Ground truth established from histology following biopsy and other tests
Reading of (un)selected mammograms by AI and independent radiologists in 'laboratory' setting
Compare AI vs independent radiologists against ground truth as reference standard
Retrospective test accuracy studies
Women attending routine breast screening have their mammograms taken
Mammogram reading by radiologists
Screen +ve cases
Screen -ve cases
Referred for further tests
Followed-up for missed cancers on screening
Ground truth established from histology following biopsy and other tests
Reading of (un)selected mammograms by AI
Compare AI vs original decisions of radiologists against ground truth as reference standard *
*This design is sometimes single arm (ie without a comparator of original decisions of radiologists)
Prospective test accuracy studies
Women attending routine breast screening have their mammograms taken
Mammogram reading by radiologists
Screen +ve cases as determined by AI or radiologists
Screen -ve cases as determined by AI and radiologists
Referred for further tests
Followed-up for missed cancers on screening
Randomised test accuracy studies
Women attending routine breast screening have their mammograms taken
Randomisation
A subset of mammograms read by AI or a pathway including AI
A subset of mammograms read by radiologists or the current pathway
Screen +ve cases
Screen -ve cases
Referred for further tests
Followed-up for missed cancers on screening
Screen +ve cases
Screen -ve cases
Referred for further tests
Followed-up for missed cancers on screening
Single arm cohort studies (Research pilots)
Women attending routine breast screening have their mammograms taken
Mammogram reading by AI or a pathway including AI
Screen +ve cases
Screen -ve cases
Referred for further tests
Followed-up for missed cancers on screening

Figure 3. Illustration of study designs for the evaluation of test accuracy of AI systems in mammography

Prospective test accuracy studies may provide the least biased reference standard, because the same women receive the existing pathway and one or more AI pathways and receive further tests including biopsy if they test positive in any of these pathways. Further, these studies can be designed to test the accuracy of the pathway incorporating AI, rather than simply the standalone accuracy of AI. For example, for the proposed use of AI to replace the second reader, prospective test accuracy studies can measure the accuracy of the combination of human first reader, AI second reader, and human arbitration, as would be implemented in practice. Such results are more clinically relevant.

Retrospective studies using validation test sets can enable large studies at low cost, and comparison of many AI systems on the same women's mammograms. However, which cases are recalled for further tests is decided by the human reader not by any of the AI systems. Therefore, the true status of AI positive / human reader negative cases is not known. Further, these studies do not provide information on the accuracy of the pathway incorporating AI, or human interaction with AI. These retrospective studies should use geographical validation test sets. This is where the test set is made up of images from different centres than the training set, where different women attend, different readers operate, and different imaging equipment is used to fully assess the generalisability of the AI system. The reviewers recognise that temporal validation has been reported by some to be sufficient in meeting the expectations of a validation set.[47] However, in the breast screening context this might not be sufficient because women attend breast screening repeatedly so the temporal validation set may be made up of the same women attending subsequent screening rounds, using the same mammography machines operated by the same radiographers. In the UK, women are invited for screening for 20 years. Consequently, temporal validation in the breast screening context may have similar issues to split sample validation which can partially address the internal validity of a model but not its generalisability.[47] The reviewers surmise that temporal validation, like split sample validation and internal validation, may lead to overfitting. Therefore, only geographical validation of test sets was formally accepted, and studies of temporal validation were formally excluded.

However, there is the possibility that an AI system has been developed using images from large parts of the UK, and a UK validation test set may also include those centres. Large UK studies with some overlap between training and test centres may potentially be useful as the risk of overfitting bias may be lower than in smaller temporal validation studies, and outcomes would be generalisable to the UK screening context. Such studies would be of importance for decision making, thus it may not be useful to have an *a priori* rule excluding or including all studies with temporal validation. On this basis, the reviewers have separately reported the identified temporal validation studies to contrast the evidence with geographically validated studies. This will provide a reference point on the issue for further discussion and help to inform future eligibility decisions.

Randomised controlled trials provide the least biased evidence for the impact of the new AI pathway, and in particular evidence for the outcomes of interval cancers, so are most applicable to question 2. Randomised controlled trials and randomised test accuracy studies do not usually give the two tests (AI and human readers) to the same women, so are of limited use in measurement of accuracy, unless the design is adapted to include test accuracy sub-studies.

Comparative cohort studies may provide information about the impact of AI on clinical practice, and so provide information for question 2. Comparative cohort studies may provide some information for question 1, but with significant additional bias in comparison to prospective test accuracy studies. Single arm cohort studies are more difficult to interpret but may provide some useful information for question 2 if they are in a context very applicable to the NHSBSP. Single arm cohort studies are difficult to interpret for question 1, because there is no comparator so no information on whether accuracy is better or worse than current practice.

Enriched test set multiple reader multiple case (MRMC) test accuracy studies where the human read is in a laboratory setting were deprioritised due to the biases associated with the 'laboratory effect'[92], so were included only if other evidence was lacking.

Data extraction

For questions 1 and 2, data were extracted by one reviewer, with a random 20% checked by a second reviewer for studies on question 1 only. All data extractions were entered into a piloted electronic data collection form. Any disagreements were resolved by consensus or discussion with a third reviewer.

Appraisal for quality/risk of bias tool

For question 1, quality appraisal of diagnostic accuracy studies was conducted using a modified QUADAS-2 tool[93] with signalling questions tailored for the critical appraisal of machine learning studies.

The following signalling questions were removed from the tailored tool:
Flow and timing domain: Was there an appropriate interval between index test(s) and reference standard?
Flow and timing domain: Did all women receive the same reference standard?

The following questions were added:
Patient selection domain: Were the women and mammograms included in the study independent of those used to train the AI algorithm?
Index test domain: Were the index test results interpreted without knowledge of the results of any other index tests?
Index test domain: Where human readers are part of the test, were their decisions made in a clinical practice context? (i.e. avoidance of the laboratory effect)
Flow and timing domain: Did the study avoid choosing which reference standard women receive based on results of just one of the index tests?

The justification of the tailoring is explained in detail in **Appendix 5.**

Quality appraisal was conducted by one reviewer, with a random 20% checked by a second reviewer. The adaptations to QUADAS-2 used here are intended as the first version, for finalising in advance of a subsequent review. It will be updated when the QUADAS-AI becomes available, and converted to QUADAS-C[94] once it becomes available to account for comparative test accuracy studies.

For question 2, no formal quality assessment was conducted.

Question level synthesis

Question 1

Criterion 4 - Accuracy of the tests

There should be a simple, safe, precise and validated screening test.

Criterion 5 - Distribution of test values in the target population

The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.

Question 1 – What is the accuracy of AI algorithms to detect breast cancer in women attending screening mammography?

The interest in AI for clinical practice is growing. This is promoted by a huge volume of published studies on the development of AI systems, for instance, in the area of breast screening which is expected to further increase due to the large UK investment in evaluating AI systems.[95] This may inappropriately convey the message that AI systems are well developed and fit for purpose in breast screening programmes. There is therefore an urgent need to evaluate the evidence in terms of the effect of the pathway change by integrating AI into the UK screening pathway and to assess the applicability of reported test accuracy estimates to the UK screening context.

Eligibility for inclusion in the review

Two types of AI use cases were included in the review: stand-alone AI and AI as a reader aid. The population considered were women attending routine breast screening for full field digital (FFDM) mammograms. The index test was a combination of AI with readers (radiologist or equivalent) in any configuration to make a decision based on the mammograms whether to recall the women for further tests. This is explained in detail in the methods section. As comparators the reviewers considered a screening pathway or a read without AI (single or double human readers with or without “traditional” CAD). The target condition of consideration was breast cancer confirmed by histology following biopsy. The reviewers included test accuracy outcomes which characterise the trade-off between false positive and false negative test results.

Reasons for exclusion of full text articles are detailed in **Appendix 2 Table 19**. The main reasons for exclusion of full text articles included:

1) Image type was not screening mammograms or FFDM
2) Internal validation test sets
3) Detecting only subtypes of breast cancer
4) Lack of detection / classification
5) Intervention not AI
6) Outcomes not relevant

Furthermore, sub-studies within included articles were also assessed against the inclusion and exclusion criteria. Seven sub-studies were excluded from 5 articles because the sub-studies used a) internal validation, b) included >10% or unclear proportion of diagnostic mammograms or c) lacked outcomes relevant to test accuracy. The sub-studies and reasons for exclusion are reported in **Appendix 2 Table 20**.

Description of the evidence

Database searches yielded 4,969 results, of which 6 were judged to be relevant to this key question (**Figure 1**). An additional 2 relevant articles were identified through hand-searching the reference lists of relevant systematic reviews and contact of experts, so 8 articles reporting 7 studies were ultimately included in this review.

Rodriguez-Ruiz et al. (2018)[79] reported preliminary results of the complete study reported in Rodriguez-Ruiz et al. (2019).[56] In-text citations from hereon only refer to Rodriguez-Ruiz et al. (2019).[56]

Appendix 2 contains a full PRISMA flow diagram (**Figure 7**), along with a table of the included publications and details of which questions these publications were relevant to (**Table 18**).

The review identified 4 studies (1 non-enriched[81] and 3 enriched[76 78 80]) reporting the test accuracy for a single read of AI as a stand-alone system, of which 3 used a retrospective cohort design (retrospective test accuracy study)[76 80 81] and one an enriched test set MRMC laboratory study design.[78] Furthermore, the review identified 3 enriched test set MRMC laboratory studies reporting test accuracy for a single read of AI as a reader aid.[56 77 82] In studies using AI as stand-alone systems, images were collected retrospectively and AI readings compared to the recorded diagnosis in the original dataset to confirm the final disease status. AI test accuracy was compared to the original human decision (single reader or consensus) whether to recall for further testing or to the human decision when images were read prospectively under “laboratory” conditions. In studies using AI as a reader aid, human readers used AI to read mammograms prospectively under “laboratory” conditions, whereas images and final disease status were collected retrospectively.

Two studies evaluated datasets from Sweden,[80 81] one from the Netherlands,[78] 3 from the U.S.[76 77 82] and one from Germany and the U.S.[56] The study populations included in the studies ranged from 199[78] to 68,008[81] women in the studies using AI as a stand-alone system and from 122[82] to 240[77 56] women in the studies using AI as a reader aid.

Five studies evaluated commercially available AI systems[56 77 78 80 82] and 2 evaluated in-house systems.[76 81]

Methodological quality of the evidence

The methodological quality of the 7 included studies comprising in total 80,711 women was assessed by tailored QUADAS-2.[93] Assessment results are summarised in **Figure 4**, **Figure 5** and **Appendix 3**.

Risk of bias

Risk of bias was considered high in 3 or more domains in all 7 studies. **Figure 4** shows that risk of bias was high in all domains.

The risk of bias was classed as high in the *patient selection domain* in 6 out of 7 studies as the studies did not enrol a consecutive or random sample of women and did not avoid a case-control design.[56 76-78 80 82] One big study that comprised 84% of all 80,711 included women in the rapid review was classed a low risk of bias for patient selection as it consecutively enrolled women (**Figure 4B**).[81]

The major problem in the *index test domain* was that the threshold used for stand-alone AI systems was not pre-specified in any of the 4 studies.[76 78 80 81] For AI as reader aid, the main issue was that in these 3 enriched test set MRMC laboratory studies,[56 77 82] the mammograms were not read in clinical practice, so bias might have been introduced due to the ‘laboratory effect’ for the index test and comparator.[92]

For the *comparator domain* (human readers without AI), risk of bias was classed as high in 4 out of 7 studies.[56 77 78 82] These 4 studies were all small enriched test set MRMC laboratory studies (comprising <1% of all included women, **Figure 4B**). The reason for the high risk rating was that in these 4 studies,[56 77 78 82] the mammograms were not read in clinical practice, so human reader’s decisions might have been biased due to the ‘laboratory effect’.[92]

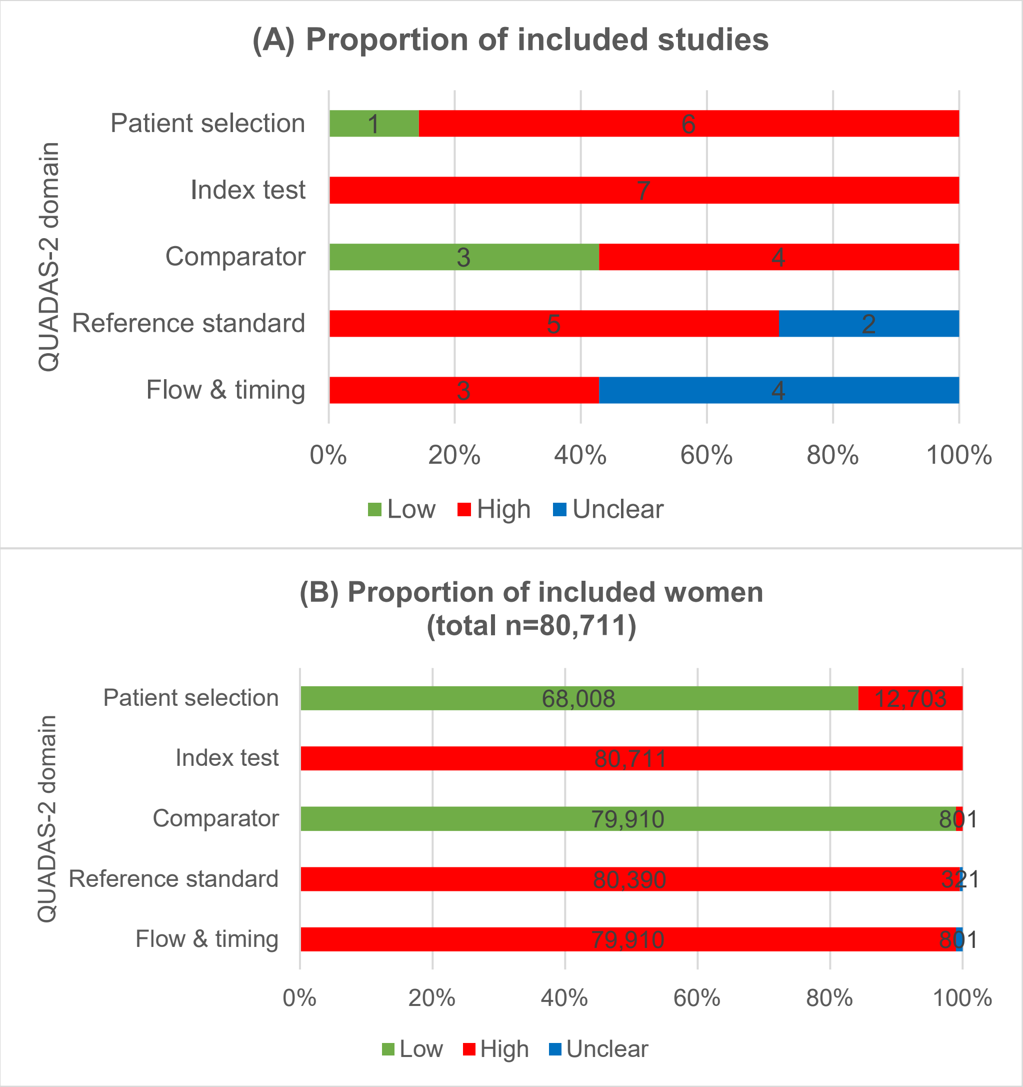


Figure 4. Risk of bias assessment results based on
(A) Proportion of included studies.
(B) Proportion of included women in the rapid review.

In the *reference standard domain*, 5 studies (comprising 99.6% of all included women) were classified as high risk of bias as follow-up of screen-negatives was less than 2 years (at least one year,[56 80 81] at least 18 months[77] and at least 21 months,[76] respectively) and/or studies included retrospective datasets where the reference standard results were interpreted with knowledge of the index test (original human reader comparator).[76 80 81] The remaining 2 enriched test set MRMC laboratory studies were rated as being at unclear risk of bias.[78 82] Despite having a follow-up time for screen-negatives of at least 2 years, it was unclear if the original human readers were the same as the readers taking part in the laboratory reader study.

The *flow & timing domain* was classed as high risk of bias in 3 retrospective test accuracy studies (comprising together 99% of all included women) as the choice of the reference standard was based on the comparator test (original human readers) only.[76 80 81] This represents incorporation bias and differential verification bias. The human reader comparator test is incorporated into the reference standard because it is used to choose which reference standard is used: when the human readers' decision is positive, the woman is referred for further tests whereas if they are negative they are simply followed up to symptomatic detection. This is problematic as when cancer is present it is much more likely to be detected if follow-up tests are undertaken, thus artificially increasing the chance for the human comparator to correctly detect a cancer than AI. The risk of bias was unclear in 4 small (covering together only 1% of all included women), enriched test set MRMC laboratory studies.[56 77 78 82] These studies avoided choosing a reference standard based on results of just one of the index/comparator tests, but the original human readers' recall decisions might agree more with the prospective reader's recall decisions than with the AI system's calls.

Applicability concerns

There were significant concerns regarding the applicability of the research identified to the UK screening *population* in 6 out of the 7 (86%) included studies[56 76-78 81 82] (see **Figure 5,** A and B). The main reason for this rating was that the study populations were enriched with >3% cancer cases (range 8.4%[80] to 73.8%[82]). One big study (comprising 84% of all included women in the review) was based on a consecutive screening cohort with 1.1% cancer and was classed as unclear applicability concerns as the women were recruited from one centre in Stockholm (Sweden) and no information was available on ethnicity and the mammography system used (i.e. manufacturer).[81]

Concerns regarding the applicability of the *index test* to the situation in the UK were classified as high in all 7 studies as the used AI algorithms were not commercially available and/or did not have pre-specified thresholds, and the AI systems were not used in a complete testing pathway applicable to UK (for example AI accuracy for single read, but not integrated into screening centre decisions, e.g. arbitration).

There were high concerns regarding the applicability of the *comparator* to the screening pathway in the UK (human double reading with arbitration at UK threshold) in 5 out of the 7 (72%) included studies as the human comparator was not a complete testing pathway applicable to the UK but a single reader only.[56 76-78 82] In the 2 remaining studies comprising 95% of all included women, 2 different human comparators were used: original single reader (first and second reader, respectively, classed as high applicability concerns) and original consensus reading (classed as low applicability concerns).[80 81]

Concerns regarding the applicability of the *reference standard* were rated as high in all 7 studies due to a shorter screening interval (between one and 2 years) than in the UK screening programme (3 years) for follow-up affecting the detection of interval cancers.

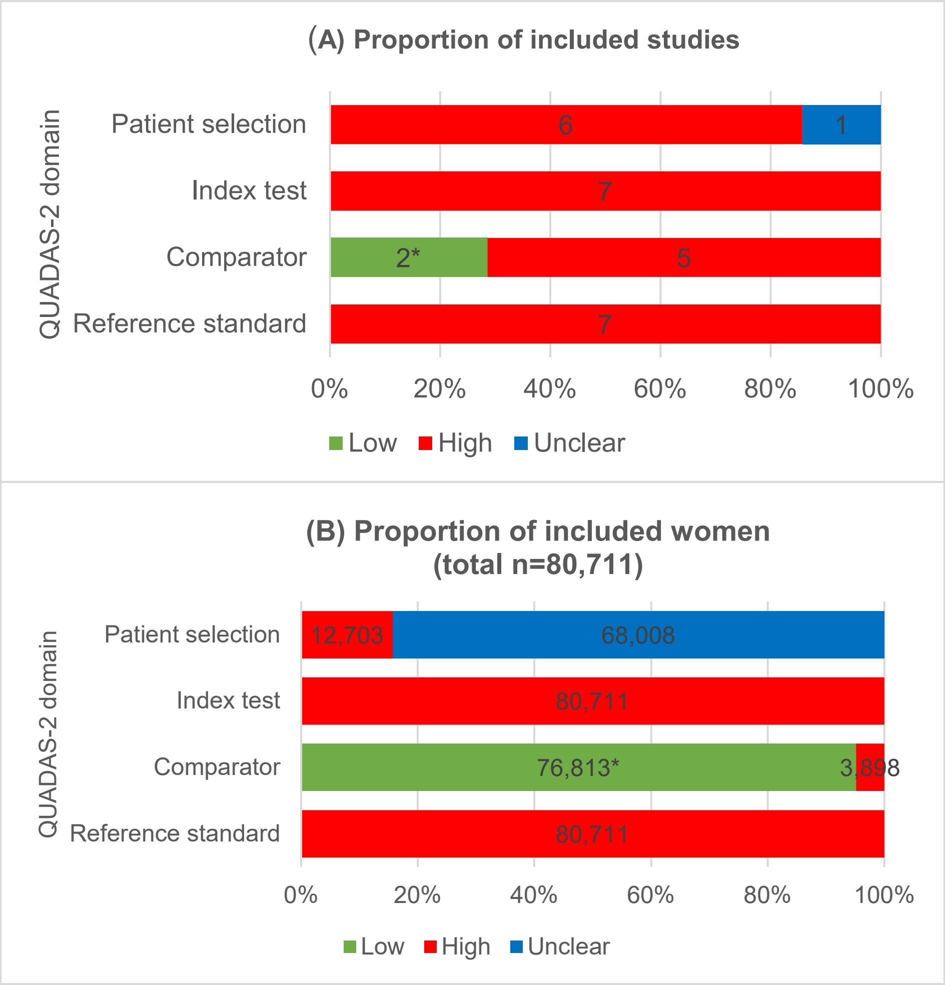


Figure 5. Applicability concerns assessment results based on
(A) Proportion of included studies.
(B) Proportion of included women in the rapid review.

* Low applicability concerns for consensus reading; high applicability concerns for single reader.

Discussion of findings

There were no studies that described accuracy of AI integrated into any breast screening pathway. Therefore, there is no direct evidence on how AI may affect accuracy if integrated into UK breast screening practice (**Table 5**).

Seven studies reported accuracy of AI to detect cancer in mammograms, but as a single read not incorporated into the breast screening pathway.[56 76-78 80-82] There were no prospective test accuracy studies in clinical practice, only retrospective test accuracy studies[76 80 81] and enriched test set MRMC laboratory studies.[56 77 78 82] Of these, 3 enriched test set MRMC laboratory studies reported test accuracy for a single read of AI as a reader aid.[56 77 82] Another 4 studies reported the test accuracy for a single read of AI as a stand-alone system in a retrospective test accuracy study [76 80 81] or an enriched test set MRMC laboratory study.[78] The latter studies reported AI algorithms that provide a cancer risk score which can be turned into a binary operating point to classify women into high risk (recall) or low risk (no recall).

Table 5. Number of eligible studies identified by intervention of interest

Intervention (index test) of interest as a pathway change	**Number of eligible studies**	**Approximations (Index test applied to women's mammograms, but not evaluated as part of pathway change)**	**Number of eligible studies**
1. AI replaces second reader (Study has to report test accuracy of the whole process; for example first reader + AI system + arbitration.)	N=0	AI classifying women's mammograms in any manner that can be interpreted simply into a decision to recall or not (preferably with localisation to guide follow-up tests, but without localisation also included). (Study has to report test accuracy of AI system in classifying women's mammograms.)	N=4[76 78 80 81]
2. AI pre-screening AI to classify low risk mammograms which require single or no human readers. (Study has to report accuracy of whole process integrated into pathway.)	N=0	AI system classifying women's mammograms into different risk categories, in a manner that can be interpreted simply into dichotomous low risk category. (Study has to report test accuracy of AI system by risk category.)	
3. AI completely replaces human reader(s) AI classifies women / images into cancer / no cancer (with or without localisation). (Study has to report test accuracy of AI system.)	N=0	No approximations required, AI would replace existing pathway.	
4. Other AI to classify mammograms in a role not previously covered. (Study has to report test accuracy of whole process integrated into pathway)	N=0	Any relevant approach where the AI classifies women's mammograms. (Study has to report test accuracy of AI system in classifying women's mammograms)	
5. AI as a reader aid (Study has to report test accuracy of whole process integrated into pathway; for example first reader + second reader using AI + arbitration; or AI could equally be used by the first reader and/or arbitration.)	N=0	AI in combination with radiologists (or equivalent) classifying women's mammograms in any manner that can be interpreted simply into a decision to recall or not. (Study has to report test accuracy of AI system and radiologist combined in classifying women's mammograms.)	N=3[56 77 82]

These AI algorithms may be used for pre-screening (point 2 in **Table 5**), replacing individual or all human readers (points 1 and 3) or post-screening of negatives after radiological assessment (point 4). The evaluation of a single read rather than AI incorporated into a screening pathway, and lack of prospective studies, indicates how little we know in terms of the potential impact of AI algorithms in breast screening pathways depending on the role of AI.

The evidence on the test accuracy of AI algorithms to detect breast cancer in women attending screening mammography using geographical validation test sets is sparse and lacked applicability to the UK context. Two studies evaluated datasets from Sweden,[80 81] one from the Netherlands,[78] 3 from the U.S.[76 77 82] and one from Germany and the U.S.[56] No UK data was available in geographical test sets.

The evidence is presented separately for stand-alone AI systems and AI systems as reader aids in the following sections. Temporal validation studies were excluded (see methods), but at the end of the discussion the reviewers added a brief exploration of 2 studies with temporal validation test sets. This is for the purposes of exploring what was excluded on that basis. A study-level summary of data extracted from each included publication is presented in **Appendix 3**, **Table 21**. In **Appendix 3** publications are stratified by question.

Stand-alone AI algorithms

None of the 4 studies reporting the test accuracy of AI as a stand-alone system[76 78 80 81] recruited women prospectively but included mammograms from available databases to be read by the AI system. Only one study recruited all women where the study population resembled a true screening population (n=68,008).[81] The remaining studies used an enriched sample where the number of included women was low for the screening context (n=8,805,[80] 3,097[76] and 199[78]) and the cancer prevalence was uncharacteristic for a screening population (8.4%,[80] 22.2%[76] and 39.7%[78]). Two studies applied an inverse probability weighting to adjust for this and to simulate a study population with a cancer prevalence matching a screening cohort.[76 80] This should be considered in the interpretation of the reported test accuracy measures because their applicability to the real screening population is uncertain.

The used AI systems assessed were various. Rodriguez-Ruiz et al. evaluated the commercial AI algorithm Transpara (v 1.4.0, Screenpoint Medical BV, Nijmegen, the Netherlands).[78] Schaffter et al. reported on the performance of the top performing model submitted to a crowdsourced AI for breast screening challenge (Therapixel, Paris, France) as well as on an ensemble method of the 8 best performing models.[81] Salim et al. explored 3 anonymised commercially available AI algorithms[80] while McKinney et al. used an in-house system in their study.[76] Three studies compared the performance of the AI system to the original decision on recall / no recall recorded in the database based on either a single reader[76] or 2 readers with consensus.[80 81] One study compared the performance of the AI system to the decision of 9 human readers who read the images prospectively under laboratory conditions.[78]

Test accuracy of the included studies is summarised in **Table 6**. The studies evaluating AI algorithms as stand-alone systems undertook non-inferiority analyses for test accuracy without pre-specified thresholds for the interpretation of AI scores. McKinney et al. set the binary operating point using the validation set where the AI system achieved superiority for both sensitivity and specificity compared to the decision of a single reader.[76] Salim et al.[80] and Rodriguez-Ruiz et al.[78] reported sensitivity at the same specificity as the comparator (single reader or consensus, respectively) and Schaffter et al.[81] reported the specificity at the same sensitivity as the comparator. Depicting the reported study estimates of sensitivity and specificity in ROC space may suggest that some AI systems are better than a single human reader but do not perform as well as consensus, but these estimates are not generalisable to clinical practice (**Figure 6**). Between 8.5%[78] and 31%[81] of cancers included in these studies were interval cancers. Salim et al. was the only study to separately report 121/739 (16%) clinically detected (interval) cancer cases within 12 months of the screening examination in comparison with screen detected cancers only. However, outcomes were only reported as area under the curve (AUC) for 3 anonymised commercially available AI systems (AI-1: AUC 0.81 (95% CI, 0.77-0.85, AI-2: AUC 0.73 (95% CI, 0.68-0.78), AI-3: AUC 0.74 (95% CI, 0.70-0.79) compared to AUCs for 618 screen detected cancers: AI-1: 0.98 (95%CI, 0.98-0.99), AI-2: 0.96 (95%CI, 0.95-0.97), AI-3: 0.95 (95% CI, 0.94-0.96).[80] They also reported an AUC of 0.92 (95% CI 0.91-0.93) when the study population was extended with 174 women who received a diagnosis of cancer between 12 and 23 months after screening. They concluded that AI algorithms may be able to promote earlier cancer detection. McKinney et al. did not report the proportion of interval cancers within included cancer cases. However, they considered follow-up intervals spanning a subsequent round of screening (27 months) to include cancers that may have been initially missed by readers.[76] They reported point estimates of sensitivity and specificity that were superior to the decision of a single reader at the BI-RADS cut-off of 3.

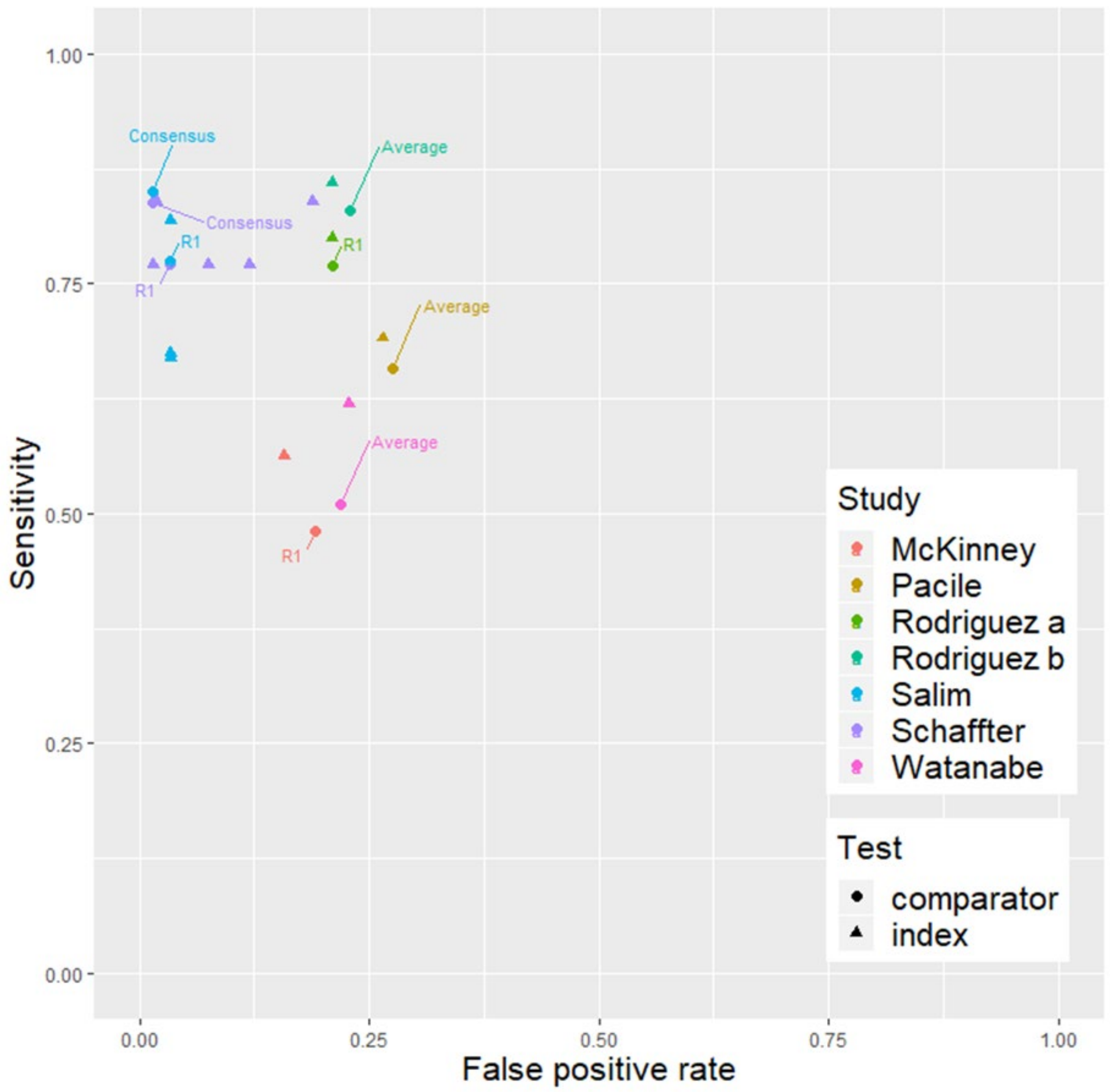


Figure 6. Study estimates of sensitivity and false positive rate (1-specificity) in ROC space by index test (AI) and comparator (human reader) for 7 included studies.

Comparators defined as consensus of 2 readers + arbitration (consensus), single reader decision (R1) or average of several readers (average).

False positive rate in the NHS Breast Screening Programme is less than 5%, so applicability of some of these studies is limited.

Retrospective test accuracy studies: Salim et al.,[80] Schaffter et al.,[81] and McKinney et al.[76]

Enriched test set MRMC laboratory studies: Pacilè et.,[77] Watanabe et al.,[82]

Rodriguez-Ruiz et al.[78] – Rodriguez a in diagram and

Rodriguez-Ruiz et al.[56] – Rodriguez b in diagram.

Comparison of AI systems and combination of AI systems

A comparison of different algorithms in a single study suggested that AI systems have different sensitivities in the detection of cancers and tend to have lower sensitivity than screening programmes with double reading followed by consensus.[80] Combination of 3 AI systems resulted in a sensitivity of 86.7% (95% CI 84.2-89.2) and specificity of 92.5% (95% CI 92.3-92.7) where the joint assessment was considered abnormal if at least one of the AI systems made an abnormal assessment.[80] This resulted in a higher sensitivity than the best performing single AI system. However, the AI systems were anonymised, which means which AI systems and version the results apply to was not reported.

Simulated integration of AI systems with reader decisions

In an analysis which simulated the combination of 3 AI systems with 2 reader decisions (at least 2 had to make a positive assessment), the estimated sensitivity was 87.4% (95%CI 85.0-89.8), and the estimated specificity was 95.9% (95%CI 95.7-96.0). The sensitivity but not the specificity of the combined algorithms and radiologists was higher than the best performing single AI system.[80]

Schaffter et al. integrated the original decision of the first reader or the consensus decision with the AI score (overall score changed to 1 [recall] if the original decision by the first reader or consensus decision was to recall). This simulated integration into the screening pathway resulted in similar specificities to the consensus interpretation alone (**Table 5**).[81]

Subgroup analyses

Schaffter et al. considered subgroup analyses of women with invasive cancer versus DCIS, by age group and time since examination using an AI system integrated with the original readers' decision.[81] The simulated integration of the radiologist's decision with the AI system resulted in significantly higher specificities (at the readers' sensitivity) compared with the single radiologist's assessment alone in all subgroup analyses except for women in the oldest age group (≥ 70 years). Considering the consensus decision, the simulated integration only achieved higher specificity in the subgroup of DCIS (n=92) and all cancer negatives (n=67,128). However, specificity estimates were not reported for these analyses.

Salim et al. performed subgroup analyses by age, mammographic density and cancer detection mode but only reported AUCs. They concluded that AI performance is decreased for younger (<55 years) versus older women (≥55 years) and for higher versus lower breast density on mammography.[80] Subgroup analyses were not compared to human readers.

AI as a reader aid

None of the 3 studies reporting the test accuracy of AI as a reader aid recruited women prospectively but included mammograms from available databases to be read by the AI system. All 3 studies used an enriched sample to manage the number of mammograms to be read by the human readers.[56 77 82] The resulting study populations were small with a cancer prevalence of 41.7%,[56] 50%[77] and 73.8%,[82] respectively.

The 3 AI systems evaluated were commercially available versions of Transpara (version 1.3.0, ScreenPoint Medical, Nijmegen, the Netherlands),[56] MammoScreen V1 (Therapixel, Nice, France)[77] and cmAssist^{TM} (CureMetrix, Inc., La Jolla, CA, USA).[82] cmAssist^{TM} provides markings and their corresponding quantitative scores of areas with high suspicion of cancer. MammoScreen reports the image positions with a related suspicion score for suspicion of breast cancer. Transpara provides a level of suspicion (on a scale of 1 to 100) for the area clicked. All 3 studies compared the test accuracy of the AI-aided read with an unaided read by the same radiologists. In Watanabe et al. readers were asked to view each mammogram without cmAssist markings and make a clinical decision about recall and were subsequently given the opportunity to change their clinical decision when provided with the cmAssist markings and quantitative scores.[82] In Rodriguez-Ruiz and Pacilè et al. readers read half the mammograms with AI and half without AI for the first reading session and vice versa for the second reading session.[56 77] Reading session were separated 4 weeks apart. Reading of mammograms by radiologists was assessed under laboratory conditions.

The experience of the radiologists ranged from 1-24 years (average 14 years of 9 radiologists) in Rodriguez-Ruiz et al.,[56] from 0-25 years (median 8.5 years of 14 American Board of Radiology and Mammography Quality Standards Act (MQSA) certified radiologists) in Pacilè et al.[77] and from <5-42 years (7 MQSA certified radiologists, <5 years n=2, 42 years n=1, 17 years n=1, mammography fellowship trained n =3) in Watanabe et al.[82]

Sensitivity and specificity were reported as an average of the 14[56 77] or 7[82] readers in the studies with and without AI reader aid. Point estimates of the means of sensitivity and specificity were slightly higher for readers with AI in 2 of the 3 studies (**Figure 6**).[56 77] However, confidence intervals overlapped. Watanabe et al. reported mean sensitivity with ranges.[82] The mean sensitivity was higher for readers with AI but the ranges overlapped. The mean specificity of readers with AI was slightly lower than without AI, however, no confidence intervals or ranges were reported.

Subgroup analyses

Pacilè et al. reported AUCs for subgroups according to lesion type (soft tissue lesion or calcifications), breast density (lower density [BI-RADS categories a and b] or higher density [BI-RADS categories c and d]), radiologists' years of experience (less than 10 years or more than 10 years), and reading time which is considered in the evidence map for question 2.[77]

Watanabe et al. analysed data based on lesion type (mass versus micro-calcifications) and tissue density which is considered in the evidence map for question 2.[82]

Table 6. Summary of reported test accuracy measures* of included studies assessing AI systems to detect breast cancer in women attending screening mammography

Reference and country	Study design and N, AI system, Comparator	Differential	Reference standard	Sensitivity, % (95% CI)	Specificity, % (95% CI)	TP**	FP**	FN**	TN**
Stand-alone AI systems									
Schaffter 2020,[81] Multinational	Retrospective test accuracy study, not enriched, n=68,008, DREAM challenge: 2 non-commercial stand-alone AI systems (TOP AI and CEM); 2 non-commercial AIs combined with reader decisions (CEM+R1; CEM+C); compared to original reader decision of (1) first reader (R1), (2) double reading + consensus	Cancer / no cancer	Cancer: tissue diagnosis of screen detected (69%) and interval cancers (31%) within 12 months of screening Non-cancer: no cancer diagnosis ≥12 months after screening	Index test Threshold set to match sensitivity of R1 (~77.1) Comparator R1: 77.1 Index test Threshold set to match sensitivity of consensus (~83.9) Comparator Consensus: 83.9	Index test TOP AI: 88 CEM: 92.5 CEM+R: 98.5 (98.4-98.6) Comparator R1: 96.7 (96.6-96.8) Index test TOP AI: 81.2 CEM+C: 98.1 Comparator Consensus: 98.5	NR	NR	NR	NR
Salim 2020,[80] Sweden	Retrospective test accuracy study (case control), enriched, n=8,805, 3 commercial stand-alone AI systems (anonymised: AI-1, AI-2 and AI-3) compared to original reader decision of 1) single reader (R1; R2), (2) double reading + consensus	Cancer / no cancer	Cancer: pathology-confirmed screen detected (84%) and clinically detected (16%) cancer within 12 months of screening Non-cancer: ≥2 years cancer free follow-up	Index test AI-1: 81.9† (78.9-84.6) AI-2: 67.0† (63.5-70.4) AI-3: 67.4† (63.9-70.8) Comparator R1: 77.4 (74.2-80.4) R2: 80.1 (77.0-82.9) Consensus: 85.0 (82.2-87.5)	Index test Threshold set to match specificity of R1 (~96.6) Comparator R1: 96.6 (96.5-96.7) R2: 97.2 (97.1-97.3) Consensus: 98.5 (98.4-98.6)	Index test AI-1: 605 AI-2: 495 AI-3: 498 Comparator R1: 572 R2: 592 Consensus: *628‡*	Index test AI-1*: 3,836‡* AI-2: *3,836‡* AI-3*: 3,738‡* Comparator R1: *3,836‡* R2*: 3,136‡* Consensus: *1,681‡*	NR	NR

Reference and country	Study design and N, AI system, Comparator	Differential	Reference standard	Sensitivity, % (95% CI)	Specificity, % (95% CI)	TP**	FP**	FN**	TN**
McKinney 2020,[76] USA, UK	Retrospective test accuracy study, enriched, n=2,738, in-house stand-alone AI system compared to original single reader decision in form of BI-RADS score (scores 0, 4, 5 were treated as positive)	Cancer / no cancer	Cancer: Biopsy-confirmed cancer within 27 months of imaging Non-cancer: One follow-up non-cancer screen or biopsied negative (benign pathologies) after ≥21 months	Index test 56.24¶ Comparator 48.10¶	Index test 84.29¶ Comparator 80.83¶	NR	NR	NR	NR
Rodriguez-Ruiz 2019,[78] Multinational	Enriched test set MRMC laboratory study, n=199 from dataset C[96], stand-alone AI system Transpara 1.4.0 compared to 9 single readers as part of a previously completed MRMC laboratory study[96]	Cancer / no cancer	Cancer: Histopathology-proven screen detected (79%) and interval (21%) cancer Non-cancer: at least one normal follow-up screening examination (2-year screening interval)	Index test 80 (70-90) Comparator 77 (70-83)	Index test Threshold set to match specificity of radiologist (~79) Comparator 79 (73-86)	NR	NR	NR	NR
AI as a reader aid									
Pacilè 2020,[77] France, USA	Enriched test set MRMC laboratory study, n=240, reader aid MammoScreen V1 compared to the same 14 radiologists reading without AI support	Cancer / no cancer	Cancer: histopathology Non-cancer: negative biopsy or negative result at follow-up for at least 18 months	Index test Average 69.1 (60.0-78.2) Comparator Average 65.8 (57.4-74.3)	Index test Average 73.5 (65.6-81.5) Comparator Average 72.5 (65.6-79.4)	NA	NA	NA	NA

Reference and country	Study design and N, AI system, Comparator	Differential	Reference standard	Sensitivity, % (95% CI)	Specificity, % (95% CI)	TP**	FP**	FN**	TN**
Rodriguez-Ruiz 2018[79] and 2019,[56] Netherlands, USA, Germany	Enriched test set MRMC laboratory study, n=240, reader aid Transpara v 1.3.0 compared to the same 14 radiologists reading without AI support	Cancer / no cancer	Cancer: Histopathology-confirmed cancer False positives: histopathologic evaluation (n = 11) or negative follow-up for ≥1 year (n = 29). Non-cancer: ≥1 year of negative follow-up findings	Index test Average 86 (84-88) Comparator Average 83 (81-85)	Index test Average 79 (77-81) Comparator Average 77 (75-79)	NA	NA	NA	NA
Watanabe 2019,[82] USA	Enriched test set MRMC laboratory study, n=122, reader aid cmAssist^{TM} compared to the same 7 radiologists reading without AI support	Cancer / no cancer	Cancer: Biopsy-proven cancer 0.76 to 5.8 years (mean, 2.1 years) after earliest actionable prior screening mammogram Non-cancer: BI-RADS 1 and 2 women with a 2-year follow-up of negative diagnosis	Index test Average 62 (Range 41-75) Comparator Mean 51 (Range 25-71)	Index test Average *77.2* Comparator Average *78.1*‡	NA	NA	NA	NA

*None of the included studies reported predictive values or complete data to calculate the positive and negative predictive value.
**In enriched test set MRMC laboratory studies where multiple readers asses the same images, there are significant issues in summing 2x2 test data across readers.
†Applied an inverse probability weighted bootstrapping (1,000 samples) with a 14:1 ratio of healthy women to women receiving a diagnosis of cancer to simulate a study population with a cancer prevalence matching a screening cohort.
‡Numbers in italics have been calculated by the authors.
¶Applied an inverse probability weighting to adjust for enrichment using Monte Carlo Simulation.
CEM Challenge ensemble method of 8 top-performing AIs from DREAM challenge, CEM+R1 Challenge ensemble method combined with first reader, CEM+C Challenge ensemble method combined with consensus reader, DREAM Dialogue on Reverse Engineering Assessment and Methods, FN False negative, FP False positive, TN True negative, TOP AI Top-performing individual AI from DREAM challenge, TP True positive, R1 reader 1, R2 reader 2.

Evidence from temporal validation studies

Two studies were excluded because they used a temporal validation test set. Becker et al. (2017)[97] reviewed all women undergoing mammography in the year 2012 in their hospital. After applying exclusion criteria, they included all eligible cases from January to September 2012 (125 cancer cases, 770 controls) into the training/tuning dataset and cases from the months October to December 2012 (18 cancer cases, 233 controls) into the validation test set.

Kim et al.[87] validated the AI system using 5 datasets from 3 different countries, 3 from South Korea (data collection from 2004-2016), one from the US (2000-2018) and one from the UK (2010-2018). The validation test set was derived from 2 datasets from South Korea. One dataset overlapped with the training dataset with a slightly different time window of data collection (2014-2018). Authors confirmed that there was no overlap in time. This part of the dataset was, therefore, judged as temporal validation which is why this study is discussed in this section.

The studies were enriched test set MRMC laboratory studies reporting test accuracy for a single read of AI as a reader aid[87] or of AI as a stand-alone system.[87 97] The datasets were from Switzerland (enriched, n=251, 7.2% cancer)[97] and Korea (enriched, n=320, 50% cancer).[87] Becker et al. used a commercially available general-purpose dANN (ViDi Suite Version 2.0; ViDi Systems Inc, Villaz-Saint-Pierre, Switzerland) used for quality inspection purposes in the fabrication of solar panels, textiles, and various high precision mechanical parts with complex shapes which was not approved for diagnostic use in the clinical routine at the time of the study. The system provides a score from 0 to 1 for the whole image and a heat map overlay with suspicious anomalies highlighted. Kim et al. used an in-house system (Lunit, Seoul, South Korea) which provides a per breast abnormality score between 0 and 1 as well as an abnormality score as a heat map.[87]

Compared to 3 single readers under laboratory conditions, the point estimate for sensitivity of ViDi Suite as a stand-alone AI system was within the range of sensitivities of the single readers but specificity was lower than any of the three reader decisions.[97] The study did not report confidence intervals. Cancer diagnoses were considered if proven by histology within 3 weeks of mammography which precludes any conclusions on the detection of interval cancers.

Using the in-house system as a stand-alone system, Kim et al. compared the sensitivity and specificity of the AI system at the threshold of 0.1 which achieved 90% sensitivity in the tuning dataset with a reader representative score (a cancer-positive case was deemed correctly detected by readers if more than half of the 14 readers identified it correctly).[87] They reported greater sensitivity and specificity of the AI system compared to 14 readers. The follow-up time for the definition of a positive cancer diagnosis was not reported. It is therefore unknown whether interval cancers were or were not included. When used as a reader aid, sensitivity for the aided read with the in-house system was higher than for the unaided read but specificity was similar considering the confidence intervals.[87]

In addition to limitations due to using a temporal validation test set, the 2 studies were of poorer quality than the geographical validation studies because the stand-alone systems were evaluated as enriched test set MRMC laboratory studies rather than retrospective test accuracy studies and the reporting / inclusion of interval cancers was insufficient.

Table 7. Summary of test accuracy outcomes (temporal validation studies)

Reference and country	Study design and N, AI system	Differential	Reference standard	Sensitivity % (95% CI)	Specificity, % (95% CI)	TP*	FP*	FN*	TN*
Becker 2017,[97] Switzerland	Enriched test set MRMC laboratory study, n=251, commercial stand-alone system (ViDi Suite Version 2.0) compared to 3 single readers	Cancer / no cancer	Cancer: Histology-proven cancer Non-cancer: ≥2 years of follow-up examinations	Index test 73.7 Comparator R1: 60.0 R2: 60.0 R3: 80.0	Index test 72.0 Comparator R1: 94.4 R2: 93.6 R3: 90.2	Index test *13* Comparator *R1: 11* *R2: 11* *R3: 14*	Index test *65* Comparator *R1: 13* *R2: 15* *R3: 23*	Index test *5* Comparator *R1: 7* *R2: 7* *R3: 4*	Index test *168* Comparator *R1: 220* *R2: 218* *R3: 210*
Kim 2020a,[87] South Korea	Enriched test set MRMC laboratory study, n=320, in-house AI (Lunit) as reader aid compared to 14 single radiologists	Cancer / no cancer	Cancer: confirmed by biopsy Non-cancer: confirmed by ≥1 year of follow-up imaging Benign: confirmed by biopsy or ≥1 year follow-up imaging	Index test† 84.78 (83.22-86.24) Comparator† 75.27 (73.43-77.04)	Index test† 74.64 (72.79-76.43) Comparator† 71.96 (70.05-73.82)	NA	NA	NA	NA
Kim 2020b,[87] South Korea	Enriched test set MRMC laboratory study, n=320, in-house stand-alone AI system (Lunit) compared to the same 14 radiologists reading without AI support	Cancer / no cancer	Cancer: confirmed by biopsy Non-cancer: confirmed by ≥1 year of follow-up imaging Benign: confirmed by biopsy or ≥1 year follow-up imaging	Index test 88.75 (82.80-93.19) Comparator 75.27 (73.43-77.04)	Index test 81.87 (75.02-87.51) Comparator 71.96 (70.05-73.82)	Index test 142 Comparator 122	NR	Index test *18‡* Comparator *38‡*	NR

*Enriched test set MRMC laboratory studies where different readers asses the same images, there are significant issues in summing 2x2 test data across readers.

†Sensitivity and specificity for reader representative score: more than half of the readers identified cancer/no cancer correctly.

‡Numbers in italics were calculated by the reviewers.

Summary of Findings Relevant to Criteria 4 and 5: Not Met

There were no studies that described accuracy of AI integrated into any breast screening pathway, and no prospective studies of test accuracy in clinical practice. Therefore, there is no direct evidence on how AI may affect accuracy if integrated into UK breast screening practice. There were three enriched test set MRMC laboratory studies reporting test accuracy for a single read of AI as a reader aid, but these will be subject to the laboratory effect bias where radiologists act differently in test sets than clinical practice. There were four studies examining AI accuracy in tests sets, of which only one was a consecutive or random sample of women attending breast cancer screening, and this study did not use an AI algorithm with a pre-set threshold. There is some evidence from early stage evaluation studies that AI has the potential to be an accurate tool to detect cancer in breast screening mammograms. However, the current evidence is a long way from the quality and quantity required for implementation into clinical practice.

Question 2

Criterion 11 — Effectiveness of the screening programme

There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an "informed choice" (such as Down's syndrome or cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.

Criterion 12 — Benefits and harms of the screening programme

The benefit gained by individuals from the screening programme should outweigh any harms, for example from overdiagnosis, overtreatment, false positives, false reassurance, uncertain findings and complications.

Question 2 – What is the clinical impact of the use of AI algorithms to detect breast cancer in mammograms compared to current practice in breast screening programmes?

Description of the evidence

The electronic database searches returned 4,969 results. After automatic and manual de-duplication, 3,634 unique references were sifted for relevance to the questions, 423 full texts were assessed and 5 references[56 80-82 85] were included in the final evidence map. Another 4 references were identified via reference list screening (n=1[79]), expert suggestions (n=1[77] and the question 1 citation searches (n=2[83 84]), respectively, resulting in a total of 9 included references (reporting on 8 studies) for the evidence map.

Rodriguez-Ruiz et al. (2018)[79] reported preliminary results of the complete study reported in Rodriguez-Ruiz et al. (2019).[56] Only results from the completed study will be presented and in-text citations from hereon only refer to Rodriguez-Ruiz et al. (2019).[56]

A flow diagram summarising the number of studies included and excluded is presented in **Figure 7**.

Summary of findings

The evidence on the clinical utility of AI algorithms when used to read mammograms in a breast screening programme is limited. None of the 8 identified studies evaluated the impact of an AI algorithm as a change of a screening pathway in a randomised controlled trial or prospective cohort study on clinically significant outcomes (such as spectrum of disease detected, interval cancers, quality of life, morbidity) or patient management and practical implication outcomes (such as workforce or costs).

Specifically, among those 8 studies there were:

- 1 enriched, retrospective case-control study using a screening FFDM data set from Sweden that simulated the use of 3 commercial **stand-alone AI** systems in place of single reading or double reading with consensus (original reader decisions).[80]
- 1 non-enriched, retrospective cohort study using a screening FFDM data set from Sweden that simulated a **combination of AI and radiologist** in place of single reading or double reading with consensus (original reader decisions).[81]
- 2 non-enriched, retrospective cohort studies and 1 retrospective, simulated screening cohort study simulating the impact of using **AI algorithms for triage** (pre-screening[83-85] before radiological assessment or post-screening of negatives after radiological assessment[84]) in place of human double reading with consensus for all mammograms. They used screening FFDM datasets from Germany[83] and Sweden,[84 85] respectively, and the original reader decisions (double reading with consensus).
- 3 enriched test set MRMC laboratory studies comparing screening FFDM reading with and without **AI as reader aid** under laboratory conditions.[56 77 82]

Table 8 presents the number of studies per relevant clinical utility outcome. The 8 identified studies reported on human reader workload and cancer detection by screening (3 studies[83-85]), spectrum of cancer detected (2 studies[80 85]), interval cancers (3 studies[80 82 84]), next-round screen-detected cancers (1 study[84]), recall rate / false positive recall (3 studies[81 83 85]) and reading time (2 studies[56 77]).

Abstract reporting tables are available for all 8 included studies in **Appendix 4**. Reported results for relevant clinical utility outcomes are also summarised in **Table 9**.

Table 8. Summary of studies included in the evidence map with relevant outcomes

Reference	Study type	Relevant outcomes for question 2					
		Human reader workload	**Screen-detected cancers**	**Interval cancers**	**Next round screen-detected cancers**	**Recall rate / false positive recalls**	**Reading time**
Stand-alone AI							
Salim 2020[80]	Enriched, retrospective case-control study		(↓) In-situ (↑) Invasive	(↓)			
Integration of AI systems with reader decisions							
Schaffter 2020[81]	Non-enriched, retrospective cohort study					(↓) Recall rate	
AI for triage (pre-screen unless otherwise noted)							
Balta 2020[83]	Non-enriched, retrospective cohort study	↓	(=)			↓ Recall rate; ↓ False positive recalls	
Dembrower 2020[84]	Retrospective simulated screening cohort	(↓)	(=)	(↓) AI as post-screen	(↓) AI as post-screen		
Lang 2020[85]	Non-enriched, retrospective cohort study	↓	↓ All cancers: (=) In-situ (↓) Invasive			↓ False positive recalls	
AI as reader aid							
Pacilè 2020[77]	Enriched test set MRMC laboratory study						↑ Enriched test set MRMC laboratory study; (=) Prediction for screening programme
Rodriguez-Ruiz 2018[79] and 2019[56]	Enriched test set MRMC laboratory study						= Enriched test set MRMC laboratory study; (↓) Prediction for screening programme
Watanabe 2019[82]	Enriched test set MRMC laboratory study		(↑) Cancerous micro-calcification (↑) Other cancers				

MRMC study, Multireader multicase study.

↑ Significant increase.
= No significant change.
↓ Significant decrease.

(↑) Increase but no p-value or 95% CI reported.
(=) No change but no p-value or 95% CI reported.
(↓) Decrease but no p-value reported.

Human reader workload and screen-detected cancers

AI as pre-screen

A non-enriched, retrospective screening cohort of 18,015 women from Germany was used to simulate the use of an AI system to preselect likely normal mammograms where double reading could be replaced with single reading.[83] Balta et al. estimated that the screen reading workload for the second reader would decrease by 32.6% compared to double reading of all mammograms, with no screen-detected cancers missed. The number of cases going through consensus would decrease by 20.8% ($p<0.0001$).

When an AI system is used as first and only reader to dismiss likely normal mammograms to no radiologist reading in a simulated (11-times up-sampling of healthy women) Swedish screening cohort of 75,536 women (with double reading of the remainder), 60% with the lowest AI scores could be excluded from radiologist reading with no screen-detected cancers missed.[84]

A second study estimated the impact of using an AI algorithm as pre-screening to identify normal screening mammograms that do not need human reading. In a non-enriched, retrospective screening cohort of 9,581 women from Sweden, Lang et al. estimate that 53.0% (95% CI 52.0-54.0%) exams could be removed from human reading with 10.3% (95% CI 3.1-17.5%) screen-detected cancers missed.[85]

Spectrum of cancer detected

Stand-alone AI

Of the 3 anonymised, commercial AI systems, AI-1 was the best performing AI system in an enriched (8.4% cancer), retrospective case-control study including 8,805 Swedish women. AI-1 detected 83.5% of 85 in-situ cancer cases , while the first and second reader both detected 89.4%.[80] For invasive cancers, AI-1 detected 82.8% (first reader: 76.7%; second reader: 79.7%) and for Stage 2 or higher invasive cancers, AI-1 detected 78.4% (first reader: 68.1%; second reader: 68.1%). The apparent higher sensitivity of AI-1 to invasive than in-situ cancers implies that using AI-1 in screening might reduce overdiagnosis and might result in improvements in clinical outcomes such as morbidity, mortality and quality of life.

AI as pre-screen

The used AI system assigned screening exams a risk score of 1–10, with 10 indicating the highest probability of malignancy. In a non-enriched, retrospective screening cohort of 9,581 Swedish women, Lang et al. found that 7 of 68 screen-detected cancers would have been missed if mammograms with a low AI risk score (1-5) were excluded from screen-reading.[85] All 7 missed cancers were invasive: 3 missed cancers were small, low-grade invasive tubular cancers with excellent prognosis, whereas 3 other missed cancers were large (20 mm), one of which was histologic grade 3, with less-favourable prognosis. This implies that using this AI system might lead to impaired clinical outcomes in some women due to delayed cancer detection.

AI as reader aid

In an enriched test set MRMC laboratory study, 7 American radiologists read 122 mammograms that included 90 false negative cases and 32 normal cases with and without AI support.[82] Out of 17 cancers with microcalcifications as leading lesion type and 73 cancers without microcalcifications as leading lesion type (such as focal asymmetries or masses with microcalcifications), the readers recalled an average of 3.4 additional cancerous microcalcifications (+20%) and 6.4 additional other cancers (+9%) with AI support compared to unaided reading. This implies that the use of this AI system as reader aid could in particular increase the sensitivity for microcalcifications, which is a common, and often the single, radiographic feature of ductal carcinoma in situ, and might thereby add to overdiagnosis in some cases.

Interval cancers and next-round screen-detected cancers

Stand-alone AI

Salim et al. evaluated 3 commercial AI systems in an enriched, retrospective case-control study including 8,805 Swedish women (618 screen-detected cancers and 121 interval cancers within 12 months of screening).[80] The AI vendors asked to remain anonymous. They found that AI-1 was the best performing AI algorithm that achieved an AUC of 0.810 for the detection of interval cancer within 12 months after negative radiologist assessment suggesting that there is potential for AI algorithms to promote earlier cancer detection.

AI as post-screen

Dembrower et al. evaluated an AI system as final reader after negative double reading to triage women at highest risk of undetected cancer into an enhanced assessment stream (e.g. MRI) in a retrospective, simulated Swedish screening cohort (347 screen-detected cancers, 200 interval cancers) attending 2 consecutive screening rounds within 2.5 years.[84] They found that if 1% (5%) with the highest AI scores were triaged to enhanced assessment, potentially 12% (27%) of subsequent interval cancers and 14% (35%) of next-round screen-detected cancers could be detected suggesting that there is potential for AI algorithms to promote earlier cancer detection.

Recall rate / false positive recall

AI combined with radiologist

In a non-enriched, retrospective screening cohort from Sweden, Schaffter et al. found that an AI algorithm combined with the single radiologist assessment was associated with higher specificity compared to single reading alone. The authors predict that this could lead to a reduction in recall rate from 9.5% to 8% in a single radiologist environment as the USA.[81]

AI as pre-screen

In the 2 identified studies, the used AI system assigned screening exams a risk score of 1–10, with 10 indicating the highest likelihood of malignancy.

In a non-enriched, retrospective cohort from Germany, Balta et al. simulated the impact on the recall rate if single-reading, instead of double reading, would have been performed for the mammograms with the lowest AI scores.[83] They predicted that the recall rate would decrease from 5.35% to 4.79% (p<0.0001) at the optimal threshold of the AI score (≥8). Balta et al. also found a reduction of false-positive assessments resulting in an increase in the positive predictive value of screening from 11.90% to 13.3% (p<0.0001) at the optimal threshold of the AI score for double reading, with single human reading for low risk mammograms.[83]

In a non-enriched, retrospective cohort from Sweden, the use of AI as pre-screen to identify normal screening mammograms that do not need human reading was estimated to avoid 27.8% (95% CI 21.4-34.2%) of false positive recalls at an AI threshold of ≥6 compared to independent double reading of all mammograms.[85]

These findings imply that the use of AI in screening might reduce the number of follow-up assessments needed as well as reducing anxiety and improving quality of life of women by avoiding false positive mammograms.

Reading time

AI as reader aid

Reading time was evaluated in 2 enriched test set MRMC laboratory studies where, in each study, 14 American radiologists read 240 mammograms with and without AI support.[56 77] Pacilè et al. found that reading time was increased in both reading sessions with AI (p<0.01),[77] whereas Rodriguez-Ruiz et al. report a similar reading time (unaided: 146 seconds; supported by AI: 149 seconds; p=0.15).[56]

Reading time changed in both studies dependently to the AI score, with reading time the same or slightly decreased with AI support for low-suspicion examinations and increased average reading times with AI support for the high-suspicion examinations. Both studies conclude that these findings could imply that the introduction of the AI tool into screening programmes (with the vast majority being low-suspicion mammograms) may not prolong the workflow of the radiologists[77] or even decrease the average reading time per case by approximately 4.5%.[56]

Table 9. Summary of clinical utility outcomes by use of AI (9 articles reporting on 8 studies)

Reference, Country	Study characteristics	AI in pathway and comparator pathway	Reported results
Stand-alone AI (1 study)			
Salim 2020,[80] Sweden	Enriched, retrospective case-control study. 8,805 women (739 cancer, 8.4%) between 40 and 74 years from the Swedish Cohort of Screen-Age Women, screened with FFDM at 1 academic hospital in Stockholm (Sweden) from 2008 to 2015. 618 (84%) screen-detected cancer cases and 121 (16%) interval cancers within 12 months of the screening examination.	Intervention: 3 commercial, anonymised AI systems (AI-1, AI-2 and AI-3) yielding a prediction score for each breast ranging between 0 and 1. Comparator: Retrospective comparison to original reader decision (double reading with consensus). 25 different first-reader radiologists and 20 different second-reader radiologists from 1 academic hospital in Stockholm (Sweden).	**Spectrum of disease detected** In situ cancers (n=85; 78 screen-detected and 7 interval cancers): AI-1: 71 (83.5%) First reader: 76 (89.4%) AI-2: 65 (76.5%) Second reader: 76 (89.4%) AI-3: 65 (76.5%) First and second reader: 80 (94.1%) Invasive cancers (all stages) (n=640; 534 screen-detected and 106 interval cancers) AI-1: 530 (82.8%) First reader: 491 (76.7%) AI-2: 426 (66.6%) Second reader: 510 (79.7%) AI-3: 431 (67.3%) First and second reader: 553 (86.4%) Invasive cancers (Stage 2 or higher) (n=204; 142 screen-detected and 62 interval cancers) AI-1: 160 (78.4%) First reader: 139 (68.1%) AI-2: 119 (58.3%) Second reader: 139 (68.1%) AI-3: 124 (60.8%) First and second reader: 153 (75.0%) **Detection of interval cancer within 12 months after negative radiologist assessment:** AI-1: AUC 0.810 (95% CI, 0.767-0.852) AI-2: AUC 0.728 (95% CI, 0.677-0.779) AI-3: AUC 0.744 (95% CI, 0.696-0.792) AI-1 achieved an AUC of 0.810, suggesting that there is potential for the AI algorithms to promote earlier cancer detection and that there are suspicious findings present in many of those mammograms.
AI combined with radiologist's recall assessment (1 study)			
Schaffter 2020,[81] Multinational	Non-enriched, retrospective cohort study. 68,008 women screened with FFDM at the Karolinska Institute (Stockholm, Sweden) between April 2008 and December 2012. 780 (1.1%) cancer cases within 12 months of mammogram.	Intervention: DREAM challenge; An ensemble method aggregating top-performing AI algorithms and consensus radiologists' recalls (CEM+C). Comparison: Retrospective comparison to original reader decision (double reading with consensus, Sweden).	**Recall rate** Our study suggests that a collaboration between radiologists and an ensemble algorithm may reduce the recall rate from 0.095 to 0.08, an absolute 1.5% reduction. Considering that approximately 40 million women are screened for breast cancer in the United States each year, this would result in more than half a million women annually who would not have to undergo unnecessary diagnostic work-up.

Reference, Country	Study characteristics	AI in pathway and comparator pathway	Reported results
AI for triage (3 studies)			
Balta 2020,[83] Netherlands, Germany	Non-enriched, retrospective cohort study. 18,015 consecutively acquired screening exams (FFDM) from 1 centre in Germany, 114 (0.64%) screen-detected cancer.	Intervention: AI (Transpara^{TM} 1.6.0, Screenpoint Medical BV, Nijmegen, Netherlands); pre-selection of likely normal screening mammograms where double-reading could be safely replaced with single-reading. Transpara risk score of 1–10, different cutoffs evaluated. Comparator: Independent double reading with consensus of all mammograms (Germany).	At optimal threshold of Transpara score (≥8) for double reading, with single reading for low risk mammograms (Transpara scores 1-7): **Cancer detection** No screen-detected cancers missed. **Recall rate** Decreased from 5.35% to 4.79% (p<0.0001) **False positive recall** Reduced; associated with increase in positive predictive value from 11.90% to 13.30% (p<0.0001) **Screen reading workload** Decreased by 32.6% (11,656 not read by reader 2) **Number of cases going through consensus** Decreased by 20.79% (from 2,400 to 1,987; p<0.0001)
Dembrower 2020,[84] Sweden	Retrospective simulation study. 7,364 women with screening exams obtained during 2 consecutive screening rounds in 1 centre in Sweden, 547 (7.4%) cancers: 347 screen-detected, 200 interval cancers. Simulated screening cohort by 11-times up-sampling of healthy women: 75,534 (0.74% incident cancer per screening interval).	Intervention: AI (Lunit, Seoul, South Korea, version 5.5.0.16) as 1) first and only reader to dismiss the majority of normal mammograms into a no radiologist work stream. 2) final reader after negative double reading of the remainder to triage women at highest risk of undetected cancer into an enhanced assessment work stream (e.g. MRI). Comparator: Independent double reading with consensus of all mammograms (Sweden).	**Reduction in workload and screen-detected cancers missed** 1) No radiologist stream 60% with lowest AI scores excluded: 0 cancer missed. 70% with lowest AI scores excluded: 0.3% (95%CI 0.0-4.3) cancer missed. 80% with lowest AI scores excluded: 2.6% (95% CI 1.1–5.4) cancer missed. **Potential additional cancer detection (Interval cancers and next-round screen-detected cancers)** 2) Enhanced assessment work stream 1% with highest AI scores included: potential additional detection of 24 (12%) subsequent interval cancers, 48 (14%) next-round screen-detected cancers 5% with the highest AI scores included: potential additional detection of 53 (27%) subsequent interval cancers, 121 (35%) next-round screen-detected cancers. Triage mammograms into no radiologist assessment and enhanced assessment could potentially reduce radiologist workload by more than half and pre-emptively detect a substantial proportion of cancers otherwise diagnosed later.

Reference, Country	Study characteristics	AI in pathway and comparator pathway	Reported results
Lang 2020,[85] Sweden	Non-enriched, retrospective cohort study. Subcohort of the Malmö Breast Tomosynthesis Screening Trial; 9,581 consecutive non-pregnant women between 40–74 years attending national breast cancer screening (FFDM) at 1 centre in Sweden; 68 screen-detected cancers (0.71%); 187 false-positive recalls.	Intervention: AI (Transpara v.1.4.0, Screenpoint Medical BV, Nijmegen, Netherlands) as pre-screen to identify normal screening mammograms that do not need human reading. Transpara risk score of 1–10, different cutoffs evaluated. Comparator: Independent double reading with consensus of all mammograms (Sweden).	**Reduction in workload, false positives avoided and cancers missed:** Transpara scores 1-2 excluded from human reading: 1,829 (19.1%; 95% CI 18.3–19.9) exams removed, 10 (5.3%; 95% CI 2.1–8.6) false positives avoided, 0 cancers missed. Transpara scores 1-5 excluded from human reading: 5,082 (53.0%; 95% CI 52.0–54.0) exams removed, 52 (27.8%; 95% CI 21.4–34.2) false positives avoided, 7 (10.3%; 95% CI 3.1–17.5) cancers missed. **Spectrum of cancers detected** (Transpara scores 6-10): 30/33 invasive ductal carcinomas, 10/11 invasive lobular cancers, 7/10 Invasive tubular cancer, 11/11 DCIS, 3/3 Other (e.g. papillary carcinoma, apocrine tumour). Of 56 invasive cancers: 20/24 Grade 1 23/25 Grade 2 6/7 Grade 3.

Reference, Country	Study characteristics	AI in pathway and comparator pathway	Reported results
AI as reader aid (3 studies)			
Pacilè 2020,[77] France, USA	Enriched test set MRMC laboratory study. 240 women from 1 centre in the USA (120 cancer, 120 non-cancer), FFDMs acquired between 2013 and 2016. 14 reader participants read cases over 2 reading sessions separated by a washout period of 4 weeks (counterbalance design).	AI (MammoScreen V1, Therapixel, Nice, France) as reader-aid. 14 American Board of Radiology and MQSA certified radiologists. Comparator: 14 American Board of Radiology and MQSA certified radiologists without AI reader aid.	**Reading time** Reading time increased in both reading sessions when using AI. First reading session: Without AI: mean 62.79 sec (95% CI: 60.77, 64.80), With AI: mean 71.93 sec (95% CI: 69.52, 74.33) (p<0.001). Second reading session: Without AI: mean 57.22 sec (95% CI: 55.10, 59.33) With AI: mean 62.16 sec (95% CI: 60.04, 64.29) (p<0.001). Reading time changed dependently to the AI-tool score. For low likelihood of malignancy (<2.5%), the time was about the same in the first reading session and slightly decreased in the second reading session. For higher likelihood of malignancy, the reading time was on average increased with the use of AI. The learning curve observed between the first and the second session, together with the fact that the maximum increment of time did not exceed 15 seconds, suggested that the introduction of this tool into screening programs may not prolong the workflow of the radiologists and possibly even lead to a shorter average reading time.
Rodriguez-Ruiz 2018,[79] Rodriguez-Ruiz 2019,[56] Netherlands, USA, Germany	Enriched test set MRMC laboratory study. Screening FFDM examinations from 240 women performed between 2013 and 2017 at 2 centres (Centre A: USA, Centre B: Germany) were included (100 showing cancers, 40 leading to false-positive recalls, 100 normal).	Intervention: Transpara (version 1.3.0, ScreenPoint Medical, Nijmegen, the Netherlands) as reader aid. 14 MQSA–qualified radiologists, with AI support. Comparison: 14 MQSA–qualified radiologists without AI support.	**Reading time** Reading time per case was similar: Unaided: 146 seconds; Supported by AI: 149 seconds (p=0.15). Reading unaided and with AI support differed as a function of the AI Transpara score (p<0.001). For AI scores 1-5: average reading time per case decreased by 11%. For AI scores 6-10: average reading time per case increased by 2%. Given the high workload of screening programs, from a cost-effectiveness point of view the performance benefit of using AI support is further enhanced by the fact that radiologists do not lengthen their reading time when using this system. In fact, in a real screening scenario, the average reading time per case would actually decrease by approximately 4.5%.

Reference, Country	Study characteristics	AI in pathway and comparator pathway	Reported results
Watanabe 2019,[82] USA	Enriched test set MRMC laboratory study. 122 women with FFDMs performed at 1 community healthcare facility in Southern California between February 7, 2008 (earliest) and January 8, 2016 (latest) that were all originally interpreted as negative in conjunction with R2 ImageChecker CAD, version 10.0. All 90 false-negative cases that were missed by their original interpreting radiologists for up to 5.8 years), 32 normal cases. 7 radiologists read all mammograms first without then with AI aid.	Intervention: Commercially available cmAssist™ (CureMetrix, Inc., La Jolla, CA) as reader aid. 7 American Board of Radiology and MQSA certified radiologists were provided with the cmAssist markings and their corresponding quantitative scores (neuScore™, scale of 0–100). Comparator: 7 American Board of Radiology and MQSA certified radiologists without AI support.	**Additional cancers detected (by type)** **Calcifications** (=Microcalcifications as the leading lesion type, n=17) With AI-CAD assistance, the 7 readers recalled an average of 3.4 additional cancerous calcifications but ignored on average 6.1 flagged malignant calcification cases. **Masses** (all remaining cases without microcalcifications as the leading lesion type, such as focal asymmetry or mass with micro-calcifications; n=73) With AI-CAD assistance, readers recalled an average of 6.4 additional cases of malignant masses but ignored on average 11.4 flagged malignant mass cases. It is noted that all readers in this study appeared to ignore relatively significant number of flagged actionable lesions that would have improved their sensitivity even further. This suggests that even further improvement in reader accuracy and cancer detection rate could occur as radiologists gain experience in using cmAssist and develop more confidence in its markings and use of the neuScore (quantitative probability of malignancy calculated by cmAssist).

AI artificial intelligence; AUC Area under the curve; CAD Computer aided detection; CEM+R Challenge ensemble method plus radiologist assessment; CI Confidence interval; DCIS Ductal carcinoma in situ; DREAM Dialogue on Reverse Engineering Assessment and Methods; FFDM Full field digital mammography; MQSA Mammography Quality Standards Act; MRMC Multireader multicase; MRI Magnetic resonance imaging; NHSBSP National Health Service Breast Screening Programme.

In summary, at present there is an insufficient volume of evidence on clinical utility related to the use of AI in the NHSBSP or analogous populations to justify commissioning an evidence review.

No evidence from high quality randomised controlled trials or prospective cohort studies was identified that compared the benefit of a breast cancer screening programme using AI to a screening programme without AI on clinical outcomes and patient management and practical implication outcomes.

The limited evidence currently available from retrospective simulation studies, retrospective cohort / case-control or enriched test set MRMC laboratory reader studies show potential for AI to reduce radiologist workload without compromising performance. However, these studies do not allow evaluation of the influence that the knowledge of AI scores has on radiologists in a prospective clinical setting, making the quality of the evidence unsuitable for drawing conclusions on the effectiveness of AI use in screening practice.

Summary of limitations

First the limitations of the review methodology are considered. This is followed by a discussion of the limitations of the evidence identified and included in the review.

Limitations of the review
The rapid review and evidence map were conducted in line with the UK NSC requirements for evidence summaries. The inclusion and exclusion criteria were detailed which required a considerable number of records to be assessed at full text. However, 7/423 records were not available as full text for assessment. The review is not a full systematic review which has the following implications. For question 1 only 20% of the search results, data extractions and quality appraisal were double checked. This may have increased the risk of error. For question 2, no separate search was undertaken to identify studies in addition to test accuracy studies and RCTs. Instead, citation searches were undertaken. This might have missed some relevant studies. Additional references identified from other sources were mostly conference papers (e.g. IEEE conferences, Springer Link) that our searches did not pick up. Furthermore, studies included for the evidence map for question 2 did not undergo quality appraisal and data extractions were not double checked. However, the majority of evidence for question 2 came from studies from question 1 which followed the rapid review methodology.

Internal validation studies (e.g. with cross validation) and studies using split sample validation or temporal validation were excluded from the review because they are known to overestimate test accuracy and limit the generalisability of an AI model.[47] Algorithms tested and trained with data from the same centre have the capacity to learn centre-specific biases, often indistinguishable for humans, and performances of such models tend to be overestimated when evaluated on data originating from the training centres. However, this has led to the exclusion of the more generalisable studies using UK datasets. This included the substudy by McKinney et al. using the OPTIMAM dataset consisting of data from 3 UK breast screening sites.[76] The study missed the opportunity to split the data by screening site to create geographical validation sets which would have reduced bias and increased applicability to the review question. Furthermore, this excluded studies using the TOMMY dataset.[88 89] In addition to the issue of internal validation the TOMMY dataset has questionable use for the UK screening context as it includes a mix of recalled, high risk and symptomatic women.[98] Future studies considering the TOMMY dataset should carefully evaluate which patient data to include and which sub population / part of the screening pathway the study results may be applicable to.

A theoretical issue with the exclusion of internal and temporal validation studies is that a large future study based on UK screening data may have some internal or temporal validation because there are only 94 UK screening centres and many of them are involved in study development. Relaxation of the strict exclusion criterion of internal and temporal validation studies, if based on large UK datasets, may be necessary.

This would mean that studies similar to those by McKinney et al. and Kyono et al. may be considered in the evidence synthesis. McKinney et al. used screening images of 123,964 women from the OPTIMAM dataset sampled from 3 UK breast screening sites that were split into the test set (25,856 women sampled randomly from 2 sites), training set and tuning set (women from all 3 sites).[76] Compared to the first reader, the AI system had a 1.2% higher absolute specificity and a 2.7% higher absolute sensitivity. Compared to the second reader and consensus reading, respectively, the AI system showed non-inferiority (at a 5% margin) for both specificity and sensitivity. However, in UK breast screening 5% is too large a margin for non-inferiority of specificity. As an example, a 5% reduction in specificity would increase the number of women recalled for further tests in England from 69,000 per year to 159,000 per year. The additional cost and resources for these appointments, and the additional anxiety for women recalled unnecessarily would be unacceptable. In considering future similar studies savings in reading time will be considered in combination with considerations of specificity and false positive recall to assessment, because there is much greater time and resources invested in each recall for assessment than in examination of each woman's mammograms.

In a simulation study, McKinney et al. estimated that a combination of human as first reader and AI as second reader would result in performance equivalent to that of double reading while saving 88% of the second reader's workload.[76] It may be also feasible to use an AI algorithm at a very-low decision threshold to dismiss 41% of normal cases at an NPV of 99.9%. They also estimate that with an AI at a very-high decision threshold, around 40% of cancer cases could rapidly be prioritised for human reading while maintaining a PPV of 85.6%.

The 2 studies by Kyono et al. used 1,000 and 2,000 randomly selected mammograms from the UK Tommy dataset that contains over 8,000 women, respectively.[88 89] The Tommy dataset was collected through 6 NHSBSP sites throughout the UK and included women (aged 47–73 years) recalled for further assessment after routine breast screening, and women (aged 40–49 years) with moderate/high of risk of developing breast cancer due to family history who were attending annual mammography screening. It was designed to challenge the radiologist with overlapping breast tissue cases.[98]

In the 2018 study, an AI system called Man and Machine Mammography Oracle (MAMMO) was used as a clinical decision support system that aimed to reduce the number of mammograms (both positive and negative) the radiologists read by excluding mammograms that it was confident in and deferring the uncertain decisions to a radiologist.[89] Results showed that MAMMO reduced the number of radiologist readings by 42.8% while improving the overall diagnostic accuracy in comparison to readings done by radiologists alone. In the 2020 study, MAMMO was then redesigned into a new system called Autonomous Radiologist Assistant (AURA) that aimed to exclude normal mammograms from human reading.[88] Using 10-fold cross validation, the proposed AURA model was able to identify 34% and 91% of the negative mammograms for test sets with a cancer prevalence of 15% and 1%, respectively, while maintaining an NPV of 0.99.

At present, excluding temporal validation studies did not exclude useful evidence for this review. Both studies identified[87 97] used an enriched test set MRMC laboratory study for a stand-alone AI system.

The review also excluded computer aided detection (CAD) for breast screening using systems that were categorised as traditional CAD. The definition was based on expert opinion as well as the literature.[50] However, the distinction is not clear cut, and this approach may have excluded relevant studies which poorly reported the AI methods or used a combination of methods.

Due to the limited evidence subgroups according to age, breast density, prior breast cancer, breast implants, larger breast size (more than the standard 4 mammographic images) could not be considered. One study suggested that AI may perform poorer in younger women and women with denser breasts.[80] This would limit the usefulness of AI for two reasons. Firstly, because older women with fatty breasts are much easier for human readers to examine, and so there is less potential for time saving or accuracy increases from AI. Second, younger women with denser breasts are a group with more potential benefit from screening. Women with dense breasts have an increased risk of breast cancer. Younger women have longer life expectancy, so there is more potential benefit and less potential harm from detecting the relatively slower growing cancers at screening.

The review did not specifically consider differences in test accuracy of AI systems because the only direct comparison of 3 AI systems did not disclose the names of the AI systems evaluated and because the evidence was generally so poor that it appears too early to consider differences by AI type.

The adaptation of the QUADAS-2 tool for this review was a first iteration and requires refinement taking into consideration the QUADAS-2 AI version and AI reporting guides such as STARD-AI and CONSORT-AI which are expected to come out in due time. For instance, the adaptation aimed to capture a variation of differential verification bias caused when the choice reference standard is based on results of just one of the index tests or comparators. If women were recalled for further tests on the basis of one of the index tests and not other(s) then this will cause bias because cancer, when present, is more likely to be found if the person receives follow-up tests after recall from screening. In retrospective studies, the decision whether to recall for follow-up tests/biopsy was made on the basis of the human readers' decision. We do not know whether AI positive, human reader negative women are false positive or true positive. Follow-up to development of interval cancers will detect some, but not all of these cancers, so reduces, but does not eliminate this bias. In retrospective reader studies (enriched test set studies) in which readers prospectively read retrospective data, the reference standard is not based on any index test, but the reference standard is based on the original human reader decision. The reviewers are unclear about the risk of bias in these studies and have opted to score these studies as unclear which is unsatisfactory and needs further discussion and exploration.

Using the adapted QUADAS-2 tool, the included studies scored poorly throughout, and the temporal validation studies did not receive a lower assessment in comparison. AI studies need to be improved in many areas before a down grading for temporal validation may be noticeable in the quality assessment of AI test accuracy studies.

Limitations of the evidence

This review focused on the identification of evidence which would allow the evaluation of the future integration of AI into the UK breast cancer screening programme. The most applicable evidence to address this question comes from studies where the index test is the AI system integrated into the screening pathway, as it would be used in screening practice. These studies need to report the change of the whole screening pathway when AI is added either as a second reader, the only reader, as a pre-screen or as a reader aid. However, the review did not identify any studies of this type and there is subsequently no direct evidence on AI as a change in the screening pathway. Furthermore, the evidence from studies reporting the test accuracy of the index test as AI for a ‘full single read’ was scarce and heterogeneous. There were no prospective test accuracy studies of a consecutive screening cohort in clinical practice. The majority of studies were small and used enriched datasets. No breast screening dataset from the UK was used in any of the included studies. And the current evidence on the influence that the knowledge of AI risk scores has on radiologists is limited and comes from retrospective simulation studies, retrospective cohort / case-control or laboratory reader studies. Simulations of the effect of AI on screening pathways are insufficient as they do not measure the impact of AI on readers and their decisions / interactions of readers with AI.

Studies evaluating AI algorithms as stand-alone systems used a retrospective cohort design. They used images collected during routine screening to assess the test accuracy of the AI system to classify women into high risk of cancer / no cancer. Studies used the recorded diagnosis in the original dataset to confirm the final disease status and compared AI test accuracy to the original human decision (single reader or consensus) whether to recall for further testing. The reading of the mammograms by humans was under routine clinical conditions. Therefore, studies avoided the laboratory effect of reading mammogram under study conditions. However, the readers were ‘gatekeepers’ for biopsy. This means that only women recalled for further tests by the original reader(s) receive those further tests and/or biopsy. Women who would have been recalled for further tests by AI, but not by the human reader(s) would not have received those tests, so would not have had cancer detected at screening even if it were present. These women may not have had breast cancer (so represented a false positive result from AI) or they may have had cancer (and represent a true positive result from AI). Within those true positives we do not know the spectrum of disease that would have been detected. For example, an AI system which detects additional small grade 3 cancers will offer more benefit and less harm than one detecting additional low-grade DCIS. Follow-up to symptomatic (interval) cancer detection reduces this bias, but the bias remains considerable even with such follow-up. This means that screened cases will only receive a cancer diagnosis if a mammogram raises the suspicions of a reader. Missed cases may be picked up by including a period of follow-up to detect interval cancers because breast cancer often has a lead time to symptomatic development longer than the follow-up time of the studies.

Studies evaluating AI algorithms as reader aids used enriched test set MRMC laboratory study designs. These studies used images collected during routine screening and, under study conditions, requested readers to prospectively read the mammograms unaided and AI aided. This results in the well-known laboratory effect, where readers under study conditions behave differently to how they would under routine clinical conditions.[92] Furthermore, readers were generally made aware of the higher cancer prevalence in the enriched test set leading to a bias caused by reader expectation (the threshold for declaring a finding as cancer in response to the perceived cancer prevalence). These studies were mainly performed with US radiologists. Consequently, study results have limited applicability to the UK context.

Further methodological issues of the included studies include the repeated use of the same test set to select AI systems for further evaluation and combination into a new AI system,[81] the focus on reporting AUCs (which do not characterise the trade-off between false positive and false negative results) rather than sensitivity and specificity at clinically meaningful thresholds, the focus on single centres studies, the varying length of follow-up which affects the detection of interval cancers, and the inclusion of subpopulations of the screening pathway only which reduces the applicability of the study findings to the screening context.[82] Studies evaluating anonymised commercially available AI systems are futile for policy-makers. Finally, the very low specificity of the human reader (81%) reported in McKinney et al.[76] questions the meaning of the study results.

None of the studies used a pre-specified threshold at which the AI score was interpreted as positive. This is not only an issue of bias but also of applicability. From the included studies we cannot tell how well the AI systems work at a pre-specified internal threshold. While the 4 studies using AI as a stand-alone system did not use the data established with AI to set the threshold, 3 studies based the threshold on the performance of the human readers reading the mammograms from the same dataset.[78 80 81] The studies further did not justify why they used either the sensitivity or the specificity of the reader as the benchmark. One study used the validation set rather than the test set to set the threshold.[76] However, without specifying the temporal relationship for instance by publishing the threshold in a protocol, it is unclear whether several thresholds were explored prior to the one reported. Finally, only one study reported the actual threshold used in the study.[78] This prevents replication of the other three studies.

The applicability of the current evidence to the UK screening context is limited (**Table 10** and **Table 11**). In addition to the aspects already discussed this was mainly because 1) the cancer prevalence was higher than typically found in the UK screening context, 2) the studies did not resemble the complete screening pathway of the UK including a three-year screening round and 3) the AI system was not a commercially available system. In order to adjust for the enrichment two studies used an inverse probability bootstrapping method to approach a screening population for analysis. However, it is unknown to what extent this method is successful and whether the results are more applicable to the enriched or to the screening population. Furthermore, 7/10 studies were not undertaken independently from the AI manufacturer.

Further heterogeneity within the studies hampers the comparison and questions the applicability of study results. Firstly, in studies using AI as a reader aid the reading workstations were set up differently, with AI integrated into the readers' workstations or with markings displayed on a separate screen and only some systems allowed for interactions between the reader and the AI system. Secondly, there was variation of how much information readers and AI systems had available for decision making. This is in contrast with clinical practice where readers use prior mammograms as well as clinical information about women for recall decisions.

Finally, algorithms are short lived, and they consistently change. Assessments of AI systems are most likely out of date by the time of study publications and their assessments not applicable to the version available at the time. This will be an issue if and when commercially available AI systems are considered for implementation into clinical practice.

One unpublished study is in line with our findings from the included studies.[99] This large retrospective study (n= 275,900 women) reported point estimates of 1.8% higher sensitivity of AI compared to the original first reader decision but 4.8% lower specificity, and the AI system was less accurate than consensus reading.[99] The study had the same retrospective design as the included studies and inclusion of the results would not have changed the conclusions of the review concerning the accuracy of AI algorithms to detect breast cancer in women attending screening mammography. The simulation of using AI as a second reader in a UK population did not measure the impact of AI on readers and their decisions and would therefore not add new evidence for the clinical impact of the use of AI algorithms to detect breast cancer in mammograms compared to current practice in breast screening programmes.

Table 10. Study limitations (10 included studies for questions 1 and 2 with geographical validation test set)

Reference and country	Applicability of population to UK screening population	Cancer prevalence*	Proportion originally recalled in screening population	AI assessed in pathway	Screening programme	Manufacturer funded / AI development by authors	Commercial AI	Choice of reference standard based on results of just one of the index tests / follow-up of screen negatives	Laboratory effect
Balta 2020,[83] Netherlands, Germany	Unselected screening cohort (unenriched), access to additional information NR	0.64% (DCIS NR)	5.35%	No, only by simulation	Germany, 4 mammography views, screening interval NR, 2 readers and consensus	NR/yes	Yes	NA	No
Dembrower 2020,[84] Sweden	Selected screening cohort (enriched), access to additional information NR	7.4% (DCIS NR)	2.0% to 2.6%	No, only by simulation	Sweden, 4 mammography views, 18-24 months screening interval, 2 readers and consensus	No/no	Yes	NA	No
Lang 2020,[85] Sweden	Unselected screening cohort (unenriched), access to additional information NR	0.71% (DCIS 16.2%)	2.7%	No, only by simulation	Sweden, 4 mammography views, 18-24 months screening interval, 2 readers and consensus	No/no	Yes	NA	No

Reference and country	Applicability of population to UK screening population	Cancer prevalence*	Proportion originally recalled in screening population	AI assessed in pathway	Screening programme	Manufacturer funded / AI development by authors	Commercial AI	Choice of reference standard based on results of just one of the index tests / follow-up of screen negatives	Laboratory effect
McKinney 2020,[76] USA, UK	Selected screening cohort (enriched), AI had access to women's age but not to previous mammograms	12 months: 11.6% (DCIS 27.9%) 27 months: 22.2% (DCIS 29.5%)	NR	No	US, 4 mammography views, 12 or 24 months screening interval	Yes/yes	No	Choice of biopsy based on original reader assessment / yes	No
Pacilè 2020,[77] France, USA	Selected screening cohort (enriched), no access to prior mammograms	50% (22.5 DCIS)	NR	No	US, 4 mammography views, 18 months screening interval	Yes/yes	Yes	Choice of biopsy based on original reader assessment which was not part of the index test / 18 months	Yes
Salim 2020,[80] Sweden	Selected screening cohort (enriched), access to (current) mammograms but no other data	8.4% (12% of cancers DCIS)	NR	No	Sweden, 4 mammography views, 18-24 months screening interval, 2 readers and consensus	No/no	Yes (not approved by the US Food and Drug Administration for use as independent readers)	Choice of biopsy based on original reader assessment / 2 years follow-up	No
Schaffter 2020,[81] Multinational	Unselected screening population, access to prior mammograms in challenge 2	1.1% (12.7% DCIS)	3.3%	No	Sweden, 4 mammography views, 18-24 months screening interval, 2 readers and consensus	No/yes	No	Choice of biopsy based on original reader assessment / 1 year follow-up	No
Rodriguez-Ruiz 2019,[78] Multinational	Selected screening cohort (enriched), no use of information from prior mammograms	39.7% (DCIS NR)	NR	No	Netherlands, 4 mammography views, 24 months screening interval	Yes/yes	Yes	Choice of biopsy based on original reader assessment / 2 years follow-up	Yes

Reference and country	Applicability of population to UK screening population	Cancer prevalence*	Proportion originally recalled in screening population	AI assessed in pathway	Screening programme	Manufacturer funded / AI development by authors	Commercial AI	Choice of reference standard based on results of just one of the index tests / follow-up of screen negatives	Laboratory effect
Rodriguez-Ruiz 2018[79] and 2019,[56] Netherlands, Germany, USA	Selected screening cohort (enriched), access to additional information NR	41.7% (DCIS NR)	NR	No	German and US screening centres, 4 mammography views, screening interval NR	Yes/no	Yes	Choice of biopsy based on original reader assessment which was not part of the index test / 1 year follow-up	Yes
Watanabe 2019,[82] USA	Selected screening cohort (enriched), mammograms originally interpreted as FN and normal	73.8% (DCIS NR)	NR	No	US, 4 mammography views, 24 months screening interval	Yes/yes	Yes	Choice of biopsy based on original reader assessment which was not part of the index test / 2 years follow-up	Yes

MC microcalcifications, AD architectural distortions, NR not reported, DCIS Ductal carcinoma in situ.

*Expected prevalence in an unenriched screening dataset should be around 0.8%.

Table 11. Study limitations (2 excluded studies using temporal validation)

Reference and country	Applicability of population to UK screening population	Cancer prevalence	Proportion originally recalled in screening population	AI assessed in pathway	Screening programme	Manufacturer funded / AI development by authors	Commercial AI	Choice of reference standard based on results of just one of the index tests / follow-up of screen negatives	Laboratory effect
Becker 2017,[97] Switzerland	Unselected screening cohort (enriched), analysis was on a per-image basis	7.2% (12.5% and DCIS 17.5% in eligible women)	NR	No	Unclear, indication for screening included routine screening and symptomatic screening	No/no	Yes (currently not approved for diagnostic use in clinical practice)	Choice of biopsy based on original reader assessment which was not part of the index test / 2 years follow-up)	Yes
Kim 2020,[87] South Korea	Selected screening cohort (enriched), high prevalence (68%) of dense breast, no consideration of clinical factors, prior mammograms NR	50% (DCIS NR)	NR	No	South Korea, 4 views mammography, mammography and ultrasound at the same time for breast cancer screening is common	Yes/yes	No	Choice of biopsy based on original reader assessment / 1 year follow-up	Yes

NR not reported, DCIS Ductal carcinoma in situ.

Review summary

Conclusions and implications for policy

The current evidence is a long way from the quality and quantity required for implementation of AI into clinical practice of breast screening programmes. This is because there are significant gaps in the evidence.

The gaps are as follows:

- There is no direct evidence on how AI may affect accuracy if integrated into UK breast screening practice.
- There were no studies that described accuracy of AI integrated into any breast screening pathway.
- There were no prospective studies of test accuracy in clinical practice.
- There is no evidence from high quality randomised controlled trials or prospective cohort studies that compared the benefit of a breast cancer screening programme using AI to a screening programme without AI on clinical outcomes and patient management and practical implication outcomes.
- There is insufficient evidence how AI works for different subpopulations of women considering age, breast density, prior cancer and breast implants.
- There is no evidence on the types of cancers detected by AI to allow an assessment of potential changes to the balance of benefits and harms including potential overdiagnosis.
- There is no evidence on the impact of different mammogram machines or other sources of variability in current practice on the accuracy of AI systems.
- There is no evidence on how the AI system may work within the breast screening IT systems in the UK.

Therefore, a rapid review on the clinical effectiveness of AI in breast screening is not recommended. (See limitations of the evidence section above for more detail.)

A rapid review may be warranted soon for three reasons. Firstly, the review has highlighted some potential advantages of AI in reading breast screening mammograms. There is some evidence from early stage evaluation studies that AI has the potential to be an accurate tool to detect cancer in breast screening mammograms. The simulation studies show potential for AI to reduce radiologist workload without compromising performance. The only study reporting the type of cancer detected by AI in comparison with the first or second human reader found spectrum of disease moving towards higher stage cancer and less DCIS, so fears of overdiagnosis are not substantiated at present.[80] Secondly, there is a growing interest in AI for breast screening. And thirdly, there is an expected increase in publications on AI due to the large UK investment in evaluating AI systems and the quality of published studies is slowly improving following a number of critical evaluations and AI specific reporting standards.

Appendix 1 — Search strategy

Electronic databases

The search strategy included searches of the databases shown in Table 12.

Table 12. Summary of electronic database searches and dates

Database	Platform	Searched on date	Date range of search
MEDLINE, MEDLINE In-Process, MEDLINE Daily, Epub Ahead of Print	Ovid SP	09 September 2020	2010 to Present
Embase	Ovid SP	09 September 2020	2010 to 2020 Week 36
The Cochrane Library, including: - Cochrane Database of Systematic Reviews (CDSR) - Cochrane Central Register of Controlled Trials (CENTRAL) - Database of Abstracts of Reviews of Effects (DARE)	Wiley Online	09 September 2020	January 2010 to September 2020
Web of Science	Ovid SP	09 September 2020	2010 to 2020

Search Terms

Search terms included combinations of free text and subject headings (Medical Subject Headings [MeSH] for MEDLINE, and Emtree terms for Embase), grouped into the following categories:

- disease area: **breast**
- intervention: **artificial intelligence**
- study design: **test accuracy studies OR randomised controlled trials**
- other term group: **screening OR mammography**

Search terms for MEDLINE, MEDLINE In-Process, MEDLINE Daily, Epub Ahead of Print, Embase, Web of Science and Cochrane Library databases are shown in **Table 13** to **Table 16**.

Table 13. Search strategy for MEDLINE, MEDLINE In-Process, MEDLINE Daily, Epub Ahead of Print (Ovid SP)

Database: Ovid MEDLINE(R) ALL <1946 to September 08, 2020>
Search Strategy:
----
1 exp Breast Neoplasms/ (293428)
2 (breast adj5 (cancer* or neoplasm* or tumor* or tumour*or carcino* or malignan*or disease*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (394921)
3 1 or 2 (395368)
4 exp artificial intelligence/ or exp machine learning/ or exp deep learning/ or exp supervised machine learning/ or exp support vector machine/ or exp unsupervised machine learning/ (99304)
5 ai.mp. (28888)
6 ((artificial or machine or deep) adj5 (intelligence or learning or reasoning)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (70647)
7 exp Neural Networks, Computer/ or exp Algorithms/ or neural network*.mp. (354996)
8 exp Diagnosis, Computer-Assisted/ (83632)
9 4 or 5 or 6 or 7 or 8 (452038)
10 3 and 9 (10492)
11 exp Mammography/ (30025)
12 mammogra*.mp. (41086)
13 screen*.mp. or exp Mass Screening/ (844672)
14 exp "Early Detection of Cancer"/ or early detect*.mp. (86182)
15 11 or 12 or 13 or 14 (921015)
16 10 and 15 (3324)
17 exp "Sensitivity and Specificity"/ or sensitivity.mp. or specificity.mp. (1898406)
18 exp "Predictive Value of Tests"/ (203774)
19 exp roc curve/ or roc.mp. or receiver operating characteristic*.mp. (119948)
20 exp Area Under Curve/ or auc.mp. (96772)
21 exp False Positive Reactions/ (27763)
22 exp False Negative Reactions/ (17783)
23 exp Observer Variation/ (42540)
24 exp Diagnostic Errors/ (116740)
25 (false adj4 (negativ* or positiv*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (100096)
26 (true adj4 (positiv* or negativ*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (10701)
27 likelihood ratio*.mp. (15918)
28 ((predict* or test*) adj1 (value* or accura* or error*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] (342222)
29 exp Reproducibility of results/ (403133)
30 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 (2397952)
31 Randomized controlled trials as Topic/ (135939)
32 Randomized controlled trial/ (512638)
33 Random allocation/ (103549)
34 Double blind method/ (159672)
35 Single blind method/ (28987)
36 Clinical trial/ (524613)
37 exp Clinical Trials as Topic/ (345470)
38 (clinic$ adj trial$1).tw. (373312)
39 ((singl$ or doubl$ or treb$ or tripl$) adj (blind$3 or mask$3)).tw. (174345)

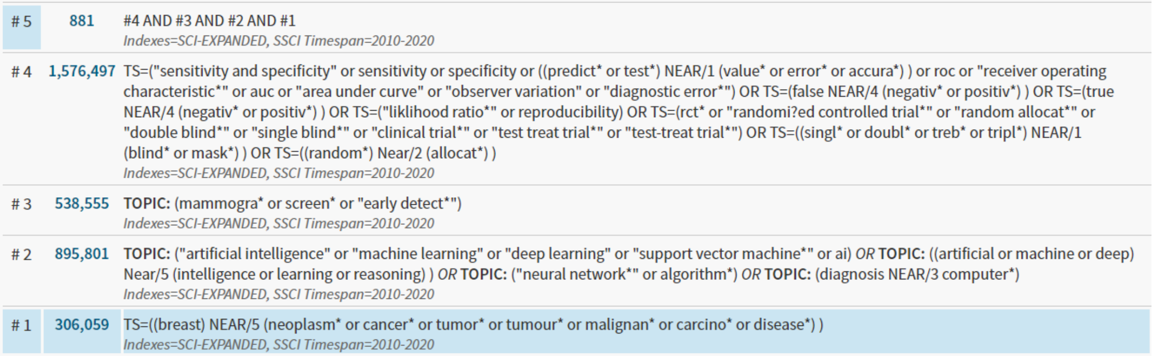
40 Randomly allocated.tw. (29185)
41 (allocated adj2 random).tw. (802)
42 (test-treat trial* or test treat trial*).mp. (1)
43 or/31-42 (1423959)
44 30 or 43 (3676443)
45 3 and 9 and 15 and 44 (2179)
46 Case report.tw. (316143)
47 Letter/ (1098543)
48 Historical article/ (359991)
49 Review of reported cases.pt. (0)
50 Review, multicase.pt. (0)
51 or/46-50 (1758549)
52 45 not 51 (2164)
53 30 or 52 (2398016)
54 3 and 9 and 15 and 44 (2179)
55 54 not 51 (2164)
56 limit 55 to (english language and yr="2010 -Current") (1228)

Table 14. Search strategy for Embase (Ovid SP)

Database: Embase <1980 to 2020 Week 36>
Search Strategy:
----
1 exp breast tumor/ (522987)
2 exp breast cancer/ (459643)
3 (breast adj5 (neoplasm* or cancer* or tumor* or tumour* or malignanc* or carcino* or disease*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (609633)
4 or/1-3 (617527)
5 exp artificial intelligence/ (41063)
6 exp machine learning/ (215028)
7 exp deep learning/ (9250)
8 exp supervised machine learning/ (1511)
9 exp support vector machine/ (22089)
10 exp unsupervised machine learning/ (745)
11 ai.mp. (37967)
12 ((artificial or machine or deep) adj5 (intelliegence or learning or reasoning)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (64514)
13 exp artificial neural network/ or neural network*.mp. (76598)
14 exp algorithm/ (381894)
15 exp computer assisted diagnosis/ (1123074)
16 or/5-15 (1645988)
17 exp mammography/ or mammogra*.mp. (62455)
18 screen*.mp. (1308438)
19 exp mass screening/ or exp screening/ (661930)
20 exp early cancer diagnosis/ or early detect*.mp. (97077)
21 or/17-20 (1419229)
22 exp "sensitivity and specificity"/ or sensitivity.mp. or specificity.mp. (1803827)
23 exp reproducibility/ (217747)
24 exp receiver operating characteristic/ or exp roc curve/ or roc.mp. (163201)
25 exp predictive value/ or ((predict* or test*) adj1 (value* or error* or accura*)).mp. (420596)
26 auc.mp. or exp area under the curve/ (211311)
27 exp false positive result/ (30242)
28 exp false negative result/ (18705)
29 exp observer variation/ (19992)
30 exp diagnostic error/ (97269)
31 (false adj4 (negativ* or positiv*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (120409)
32 (true adj4 (positiv* or negativ*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] (15519)
33 likelihood ratio.mp. (16617)
34 or/22-33 (2487428)
35 clinical trial/ (971925)
36 Randomized controlled trial/ (614311)
37 Randomization/ (87593)
38 Single blind procedure/ (40000)
39 Double blind procedure/ (172538)
40 Crossover procedure/ (64123)
41 Randomi?ed controlled trial$.tw. (236002)
42 Rct.tw. (38207)
43 Random allocation.tw. (2050)
44 Randomly allocated.tw. (35853)
45 Allocated randomly.tw. (2566)
46 (allocated adj2 random).tw. (822)
47 Single blind$.tw. (25171)

48 Double blind$.tw. (204835)
49 ((treble or triple) adj blind$).tw. (1182)
50 Prospective study/ (622435)
51 (test-treat trial* or test treat trial*).mp. (2)
52 or/35-51 (2076152)
53 34 or 52 (4332240)
54 Case study/ (71546)
55 Case report.tw. (410369)
56 Abstract report/ or letter/ (1114503)
57 or/54-56 (1585513)
58 53 not 57 (4229367)
59 4 and 16 and 21 and 58 (5034)
60 limit 59 to (english language and yr="2010 -Current") (3562)
61 limit 60 to (article or article in press or "review") (2808)

Table 15. Search strategy for Web of Science (Ovid SP)



# 5	881	#4 AND #3 AND #2 AND #1 *Indexes=SCI-EXPANDED, SSCI Timespan=2010-2020*
# 4	1,576,497	TS=("sensitivity and specificity" or sensitivity or specificity or ((predict* or test*) NEAR/1 (value* or error* or accura*)) or roc or "receiver operating characteristic*" or auc or "area under curve" or "observer variation" or "diagnostic error*") OR TS=(false NEAR/4 (negativ* or positiv*)) OR TS=(true NEAR/4 (negativ* or positiv*)) OR TS=("liklihood ratio*" or reproducibility) OR TS=(rct* or "randomi?ed controlled trial*" or "random allocat*" or "double blind*" or "single blind*" or "clinical trial*" or "test treat trial*" or "test-treat trial*") OR TS=((singl* or doubl* or treb* or tripl*) NEAR/1 (blind* or mask*)) OR TS=((random*) Near/2 (allocat*)) *Indexes=SCI-EXPANDED, SSCI Timespan=2010-2020*
# 3	538,555	**TOPIC:** (mammogra* or screen* or "early detect*") *Indexes=SCI-EXPANDED, SSCI Timespan=2010-2020*
# 2	895,801	**TOPIC:** ("artificial intelligence" or "machine learning" or "deep learning" or "support vector machine*" or ai) *OR* **TOPIC:** ((artificial or machine or deep) Near/5 (intelligence or learning or reasoning)) *OR* **TOPIC:** ("neural network*" or algorithm*) *OR* **TOPIC:** (diagnosis NEAR/3 computer*) *Indexes=SCI-EXPANDED, SSCI Timespan=2010-2020*
# 1	306,059	TS=((breast) NEAR/5 (neoplasm* or cancer* or tumor* or tumour* or malignan* or carcino* or disease*)) *Indexes=SCI-EXPANDED, SSCI Timespan=2010-2020*

Table 16. Search strategy for Cochrane Library (CENTRAL) (Wiley online)

Search Name: AI and Breast Cancer
Last Saved: 09/09/2020 14:10:48
Comment: Numbers for individual search lines are not captured by the saved search strategy.

ID Search
#1 MeSH descriptor: [Breast Neoplasms] explode all trees
#2 ((breast NEAR/5 (cancer* or neoplasm* or carcino* or malignan* or tumor* or tumour* or disease*))):ti,ab,kw
#3 #1 or #2
#4 MeSH descriptor: [Artificial Intelligence] explode all trees
#5 MeSH descriptor: [Machine Learning] explode all trees
#6 MeSH descriptor: [Deep Learning] explode all trees
#7 MeSH descriptor: [Supervised Machine Learning] explode all trees
#8 MeSH descriptor: [Support Vector Machine] explode all trees
#9 MeSH descriptor: [Unsupervised Machine Learning] explode all trees
#10 (ai):ti,ab,kw
#11 ((artificial or machine or deep) NEAR/5 (intelligence or learning or reasoning)):ti,ab,kw
#12 MeSH descriptor: [Neural Networks, Computer] explode all trees
#13 MeSH descriptor: [Algorithms] explode all trees
#14 (neural network*):ti,ab,kw
#15 MeSH descriptor: [Diagnosis, Computer-Assisted] explode all trees
#16 #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15
#17 MeSH descriptor: [Mammography] explode all trees
#18 (mammogra*):ti,ab,kw
#19 MeSH descriptor: [Mass Screening] explode all trees
#20 (screen*):ti,ab,kw
#21 MeSH descriptor: [Early Detection of Cancer] explode all trees
#22 (early detect*):ti,ab,kw
#23 #17 or #18 or #19 or #20 or #21 or #22
#24 #3 and #16 and #23
#25 MeSH descriptor: [Sensitivity and Specificity] explode all trees
#26 (sensitivity or specificity):ti,ab,kw
#27 MeSH descriptor: [Predictive Value of Tests] explode all trees
#28 MeSH descriptor: [ROC Curve] explode all trees

#29 (roc or "receiver operating characteristic"):ti,ab,kw
#30 MeSH descriptor: [Area Under Curve] explode all trees
#31 (auc):ti,ab,kw
#32 MeSH descriptor: [False Positive Reactions] explode all trees
#33 MeSH descriptor: [False Negative Reactions] explode all trees
#34 MeSH descriptor: [Observer Variation] explode all trees
#35 MeSH descriptor: [Diagnostic Errors] explode all trees
#36 (false NEAR/4 (negativ* or positiv*)):ti,ab,kw
#37 (true NEAR/4 (positiv* or negativ*)):ti,ab,kw
#38 (liklihood ratio*):ti,ab,kw
#39 ((predict* or test*) NEAR/1 (value* or accura* or error*)):ti,ab,kw
#40 MeSH descriptor: [Reproducibility of Results] explode all trees
#41 #25 or #26 or #27 or #28 or #29 or #30 or #31 or #32 or #33 or #34 or #35 or #36 or #37 or #38 or #39 or #40
#42 #24 and #41

Results were imported into EndNote and de-duplicated (see **Table 17**).

Table 17. Number retrieved and de-duplicated numbers per database (search date: 9th September 2020)

	Number retrieved	**De-duplicated numbers**
Medline ALL	1228	1220
Embase	2808	1992
Web of Science	881	380
Cochrane Library (CENTRAL)	52	42
TOTAL	**4969**	**3634**

Appendix 2 — Included and excluded studies

PRISMA flowchart

Figure 7 summarises the volume of publications included and excluded at each stage of the review. Eleven publications were ultimately judged to be relevant to one or both review questions and were considered for extraction. Publications that were included or excluded after the review of full-text articles are detailed below.

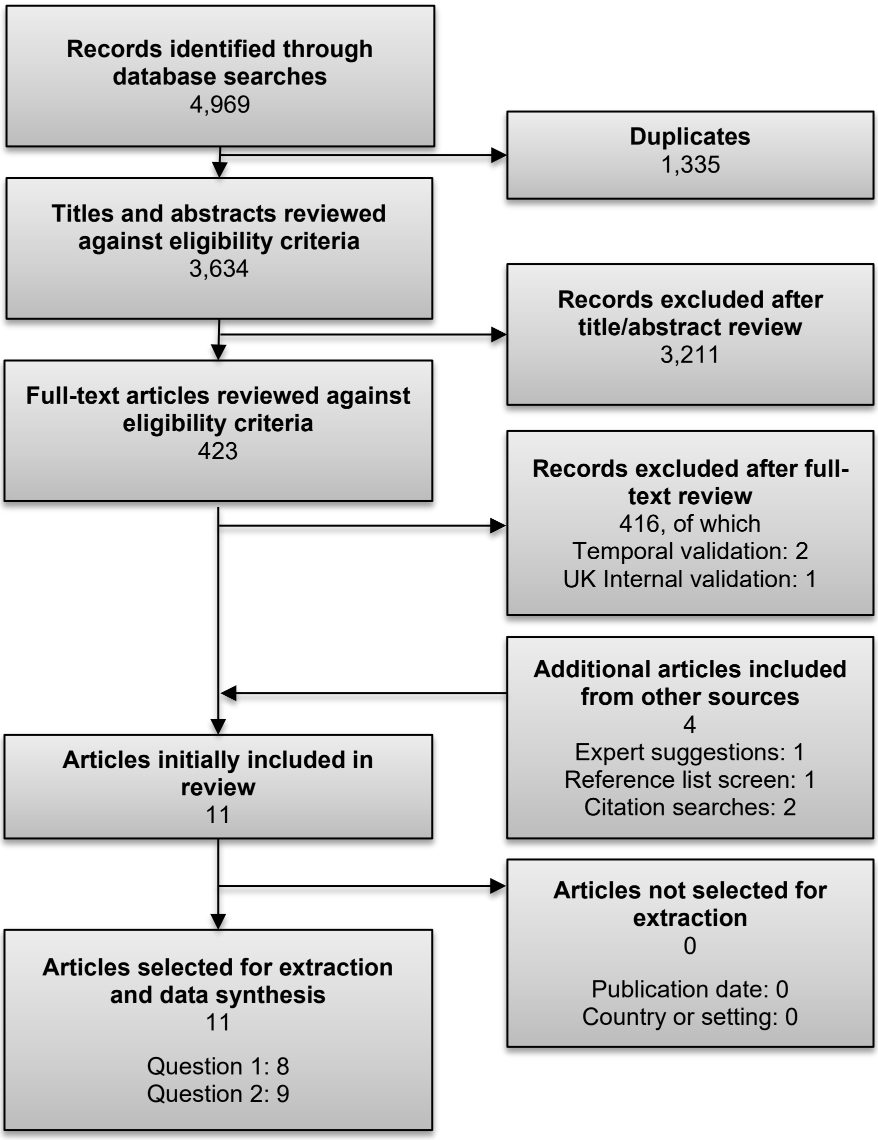

Records identified through database searches
4,969
Duplicates
1,335
Titles and abstracts reviewed against eligibility criteria
3,634
Records excluded after title/abstract review
3,211
Full-text articles reviewed against eligibility criteria
423
Records excluded after full-text review
416, of which
Temporal validation: 2
UK Internal validation: 1
Additional articles included from other sources
4
Expert suggestions: 1
Reference list screen: 1
Citation searches: 2
Articles initially included in review
11
Articles not selected for extraction
0
Publication date: 0
Country or setting: 0
Articles selected for extraction and data synthesis
11
Question 1: 8
Question 2: 9

Figure 7. Summary of publications included and excluded at each stage of the review

Publications included after review of full-text articles

The 11 publications included after review of full-texts are summarised in **Table 18** below.

Table 18. Summary of publications included after review of full-text articles, and the question(s) each publication was identified as being relevant to

Study	The condition	The test	The intervention	The screening programme	Implementation criteria	Comments
Balta 2020[83]	-	-	-	Q2	-	-
Dembrower 2020[84]	-	-	-	Q2	-	-
Lang 2020[85]	-	-	-	Q2	-	-
McKinney 2020[76]	-	Q1	-	-	-	-
Pacilè 2020[77]	-	Q1	-	Q2	-	-
Rodriguez-Ruiz 2018[79]	-	Q1	-	Q2	-	-
Rodriguez-Ruiz 2019a[78]	-	Q1	-	-	-	-
Rodriguez-Ruiz 2019b[56]	-	Q1	-	Q2	-	-
Salim 2020[80]	-	Q1	-	Q2	-	-
Schaffter 2020[81]	-	Q1	-	Q2	-	-
Watanabe 2019[82]	-	Q1	-	Q2	-	-

Q1 Question 1; Q2 Question 2.

Publications excluded after review of full-text articles

Of the 423 publications included after the review of titles and abstracts, 416 were ultimately judged not to be relevant to this review. These publications, along with the main reason for exclusion, are listed in **Table 19**.

Table 19. Publications excluded after review of full-text articles

Reference		Main reason for exclusion
Population – Image type (e.g. digitised film images; not FFDM images) (n=150)		
1.	Abbas Q, Fondo'n I, Celebi E. A Computerized System for Detection of Spiculated Margins based on Mammography. International Arab Journal of Information Technology. 2015;12(6):582-8.	Population – Image type
2.	Agnes SA, Anitha J, Pandian SIA, Peter JD. Classification of Mammogram Images Using Multiscale all Convolutional Neural Network (MA-CNN). J Med Syst. 2019;44(1):30.	Population – Image type
3.	Anitha J, Dinesh Peter J, Immanuel Alex Pandian S. A dual stage adaptive thresholding (DuSAT) for automatic mass detection in mammograms. Computer Methods and Programs in Biomedicine. 2017;138:93-104.	Population – Image type
4.	Bartolotta TV, Orlando A, Cantisani V, Matranga D, Ienzi R, Cirino A, et al. Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support. La Radiologia medica. 2018;123(7):498-506.	Population – Image type
5.	Beheshti SM, AhmadiNoubari H, Fatemizadeh E, Khalili M. An efficient fractal method for detection and diagnosis of breast masses in mammograms. J Digit Imaging. 2014;27(5):661-9.	Population – Image type
6.	Chakraborty J, Midya A, Mukhopadhyay S, Rangayyan RM, Sadhu A, Singla V, et al. Computer-Aided Detection of Mammographic Masses Using Hybrid Region Growing Controlled by Multilevel Thresholding. Journal of Medical and Biological Engineering. 2019;39(3):352-66.	Population – Image type
7.	Chithra Devi M, Audithan S. Analysis of different types of entropy measures for breast cancer diagnosis using ensemble classification. Biomedical Research (India). 2017;28(7):3182-6.	Population – Image type

Reference		Main reason for exclusion
8.	Choi JY, Ro YM. Multiresolution local binary pattern texture analysis combined with variable selection for application to false-positive reduction in computer-aided detection of breast masses on mammograms. Phys Med Biol. 2012;57(21):7029-52.	Population – Image type
9.	Chougrad H, Zouaki H, Alheyane O. Deep Convolutional Neural Networks for breast cancer screening. Comput Methods Programs Biomed. 2018;157:19-30.	Population – Image type
10.	Chowdhary CL, Mittal M, P K, Pattanaik PA, Marszalek Z. An Efficient Segmentation and Classification System in Medical Images Using Intuitionist Possibilistic Fuzzy C-Mean Clustering and Fuzzy SVM Algorithm. Sensors (Basel). 2020;20(14):13.	Population – Image type
11.	Cunningham CA, Drew T, Wolfe JM. Analog Computer-Aided Detection (CAD) information can be more effective than binary marks. Atten Percept Psychophys. 2017;79(2):679-90.	Population – Image type
12.	de Oliveira Silva LC, Barros AK, Lopes MV. Detecting masses in dense breast using independent component analysis. Artif Intell Med. 2017;80:29-38.	Population – Image type
13.	Dheeba J, Albert Singh N, Tamil Selvi S. Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach. J Biomed Inform. 2014;49:45-52.	Population – Image type
14.	Dheeba J, Jaya T, Singh NA. Breast cancer risk assessment and diagnosis model using fuzzy support vector machine based expert system. Journal of Experimental & Theoretical Artificial Intelligence. 2017;29(5):1011-21.	Population – Image type
15.	Dheeba J, Tamil Selvi S. An improved decision support system for detection of lesions in mammograms using Differential Evolution Optimized Wavelet Neural Network. J Med Syst. 2012;36(5):3223-32.	Population – Image type
16.	Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. Int. 2016;11(11):2033-47.	Population – Image type
17.	El Fahssi K, Elmoufidi A, Abenaou A, Jai-Andaloussi S, Sekkaki A. Novel approach to classification of Abnormalities in the mammogram image. International Journal of Biology and Biomedical Engineering. 2016;10:72-9.	Population – Image type

Reference		Main reason for exclusion
18.	El-Shazli AMA, Youssef SM, Elshennawy M. Computer-aided model for breast cancer detection in mammograms. International Journal of Pharmacy and Pharmaceutical Sciences. 2016;8(Supplement 2):31-4.	Population – Image type
19.	Elmoufidi A, El Fahssi K, Jai-andaloussi S, Sekkaki A, Gwenole Q, Lamard M. Anomaly classification in digital mammography based on multiple-instance learning. Iet Image Processing. 2018;12(3):320-8.	Population – Image type
20.	Eltoukhy MM, Faye I, Samir BB. Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. Comput Med Imaging Graph. 2010;34(4):269-76.	Population – Image type
21.	Ericeira DR, Silva AC, de Paiva AC, Gattass M. Detection of masses based on asymmetric regions of digital bilateral mammograms using spatial description with variogram and cross-variogram functions. Comput Biol Med. 2013;43(8):987-99.	Population – Image type
22.	Ganesan K, Acharya RU, Chua CK, Min LC, Mathew B, Thomas AK. Decision support system for breast cancer detection using mammograms. Proc Inst Mech Eng [H]. 2013;227(7):721-32.	Population – Image type
23.	Garma FB, Hassan MA. Classification of breast tissue as normal or abnormal based on texture analysis of digital mammogram. Journal of Medical Imaging and Health Informatics. 2014;4(5):647-53.	Population – Image type
24.	Gedik N. Breast cancer diagnosis system via contourlet transform with sharp frequency localization and least squares support vector machines. Journal of Medical Imaging and Health Informatics. 2015;5(3):497-505.	Population – Image type
25.	Gedik N, Atasoy A. Performance evaluation of the wave atom algorithm to classify mammographic images. Turkish Journal of Electrical Engineering and Computer Sciences. 2014;22(4):957-69.	Population – Image type
26.	Gedik N, Atasoy A, Sevim Y. Investigation of wave atom transform by using the classification of mammograms. Applied Soft Computing. 2016;43:546-52.	Population – Image type
27.	Gorgel P, Sertbas A, Ucan ON. Mammographical mass detection and classification using Local Seed Region Growing-Spherical Wavelet Transform (LSRG-SWT) hybrid scheme. Computers in Biology and Medicine. 2013;43(6):765-74.	Population – Image type
28.	Guan JS, Lin LY, Ji GL, Lin CM, Le TL, Rudas IJ. Breast Tumor Computer-aided Diagnosis using Self-Validating Cerebellar Model Neural Networks. Acta Polytechnica Hungarica. 2016;13(4):39-52.	Population – Image type

Reference		Main reason for exclusion
29.	Hu K, Gao XP, Li F. Detection of Suspicious Lesions by Adaptive Thresholding Based on Multiresolution Analysis in Mammograms. Ieee Transactions on Instrumentation and Measurement. 2011;60(2):462-72.	Population – Image type
30.	James JJ, Gilbert FJ, Wallis MG, Gillan MG, Astley SM, Boggis CR, et al. Mammographic features of breast cancers at single reading with computer-aided detection and at double reading in a large multicenter prospective trial of computer-aided detection: CADET II. Radiology. 2010;256(2):379-86.	Population – Image type
31.	Jebamony J, Jacob D. Classification of Benign and Malignant Breast Masses on Mammograms for Large Datasets using Core Vector Machines. Curr Med Imaging. 2020;16(6):703-10.	Population – Image type
32.	Kadam VJ, Jadhav SM, Vijayakumar K. Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression. J Med Syst. 2019;43(8):263.	Population – Image type
33.	Kanadam KP, Chereddy SR. Mammogram classification using sparse-ROI: A novel representation to arbitrary shaped masses. Expert Systems with Applications. 2016;57:204-13.	Population – Image type
34.	Kanchana M, Varalakshmi P. Computer aided system for breast cancer in digitized mammogram using shearlet band features with LS-SVM classifier. International Journal of Wavelets Multiresolution and Information Processing. 2016;14(3).	Population – Image type
35.	Kanchanamani M, Perumal V. Performance evaluation and comparative analysis of various machine learning techniques for diagnosis of breast cancer. Biomedical Research (India). 2016;27(3):623-31.	Population – Image type
36.	Kashyap KL, Bajpai MK, Khanna P. Globally supported radial basis function based collocation method for evolution of level set in mass segmentation using mammograms. Comput Biol Med. 2017;87:22-37.	Population – Image type
37.	Kashyap KL, Bajpai MK, Khanna P. An efficient algorithm for mass detection and shape analysis of different masses present in digital mammograms. Multimedia Tools and Applications. 2018;77(8):9249-69.	Population – Image type
38.	Kashyap KL, Bajpai MK, Khanna P, Giakos G. Mesh-free based variational level set evolution for breast region segmentation and abnormality detection using mammograms. Int j numer method biomed eng. 2018;34(1):01.	Population – Image type

Reference		Main reason for exclusion
39.	Kayode AA, Akande NO, Adegun AA, Adebiyi MO. An automated mammogram classification system using modified support vector machine. Medical Devices: Evidence and Research. 2019;12:275-84.	Population – Image type
40.	Kelder A, Zigel Y, Lederman D, Zheng B. A new computer-aided detection scheme based on assessment of local bilateral mammographic feature asymmetry - a preliminary evaluation. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:6394-7.	Population – Image type
41.	Khan AA, Arora AS. Computer aided diagnosis of breast cancer based on level set segmentation of masses and classification using ensemble classifiers. Biomedical Research (India). 2018;29(19):3610-5.	Population – Image type
42.	Khan HN, Shahid AR, Raza B, Dar AH, Alquhayz H. Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network. Ieee Access. 2019;7.	Population – Image type
43.	Krishnan MMR, Banerjee S, Chakraborty C, Chakraborty C, Ray AK. Statistical analysis of mammographic features and its classification using support vector machine. Expert Systems with Applications. 2010;37(1):470-8.	Population – Image type
44.	Li JB, Wang YH, Tang LL. Mammogram-based discriminant fusion analysis for breast cancer diagnosis. Clin Imaging. 2012;36(6):710-6.	Population – Image type
45.	Li P, Bi T, Huang J, Li S. Breast cancer early diagnosis based on hybrid strategy. Biomed Mater Eng. 2014;24(6):3397-404.	Population – Image type
46.	Li Z, Sun J, Zhang J, Hu D, Wang Q, Peng K. Quantification of acoustic radiation force impulse in differentiating between malignant and benign breast lesions. Ultrasound Med Biol. 2014;40(2):287-92.	Population – Image type
47.	Lo CM, Moon WK, Huang CS, Chen JH, Yang MC, Chang RF. INTENSITY-INVARIANT TEXTURE ANALYSIS FOR CLASSIFICATION OF BI-RADS CATEGORY 3 BREAST MASSES. Ultrasound in Medicine and Biology. 2015;41(7):2039-48.	Population – Image type
48.	Onan A. A stochastic gradient descent based SVM with fuzzy-rough feature selection and instance selection for breast cancer diagnosis. Journal of Medical Imaging and Health Informatics. 2015;5(6):1233-9.	Population – Image type
49.	Singh WJ, Nagarajan B. Automatic diagnosis of mammographic abnormalities based on hybrid features with learning classifier. Comput Methods Biomech Biomed Engin. 2013;16(7):758-67.	Population – Image type

Reference		Main reason for exclusion
50.	Valarmathi P, Robinson S. An Improved Neural Network for Mammogram Classification Using Genetic Optimization. Journal of Medical Imaging and Health Informatics. 2016;6(7):1631-5.	Population – Image type
51.	Liu X, Zeng Z. A new automatic mass detection method for breast cancer with false positive reduction. Neurocomputing. 2015;152:388-402.	Population – Image type
52.	Liu XM, Tang JS. Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method. Ieee Systems Journal. 2014;8(3):910-20.	Population – Image type
53.	Liu XM, Zhai LL, Zhu T, Liu J, Zhang K, Hu W. Multiple TBSVM-RFE for the detection of architectural distortion in mammographic images. Multimedia Tools and Applications. 2018;77(12):15773-802.	Population – Image type
54.	Mahersia H, Boulehmi H, Hamrouni K. Development of intelligent systems based on Bayesian regularization network and neuro-fuzzy models for mass detection in mammograms: A comparative analysis. Comput Methods Programs Biomed. 2016;126:46-62.	Population – Image type
55.	Mencattini A, Salmeri M. Breast masses detection using phase portrait analysis and fuzzy inference systems. Int. 2012;7(4):573-83.	Population – Image type
56.	Meselhy Eltoukhy M, Faye I, Belhaouari Samir B. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. Comput Biol Med. 2012;42(1):123-8.	Population – Image type
57.	Midya A, Rabidas R, Sadhu A, Chakraborty J. Edge Weighted Local Texture Features for the Categorization of Mammographic Masses. Journal of Medical and Biological Engineering. 2018;38(3):457-68.	Population – Image type
58.	Milosevic M, Jovanovic Z, Jankovic D. A comparison of methods for three-class mammograms classification. Technol Health Care. 2017;25(4):657-70.	Population – Image type
59.	Mohammadi-Sardo S, Labibi F, Shafiei SA. A new approach for detecting abnormalities in mammograms using a computer-aided windowing system based on Otsu's method. Radiol Phys Technol. 2019;12(2):178-84.	Population – Image type
60.	Mohammed SHA, Yousuf SEK. A computer-aided diagnosis system for the detection and classification of breast cancer. Journal of Clinical Engineering. 2016;41(2):97-100.	Population – Image type

Reference		Main reason for exclusion
61.	Mohammed SHAA, Mustafa ZA. Breast Tumors Classification Using Adaptive Neuro-Fuzzy Inference System. Journal of Clinical Engineering. 2017;42(2):68-72.	Population – Image type
62.	Mohanty F, Rup S, Dash B. Automated diagnosis of breast cancer using parameter optimized kernel extreme learning machine. Biomedical Signal Processing and Control. 2020;62 (no pagination).	Population – Image type
63.	Mohanty F, Rup S, Dash B, Majhi B, Swamy MNS. Mammogram classification using contourlet features with forest optimization-based feature selection approach. Multimedia Tools and Applications. 2019;78(10):12805-34.	Population – Image type
64.	Mohanty F, Rup S, Dash B, Majhi B, Swamy MNS. A computer-aided diagnosis system using Tchebichef features and improved grey wolf optimized extreme learning machine. Applied Intelligence. 2019;49(3):983-1001.	Population – Image type
65.	Mohanty F, Rup S, Dash B, Majhi B, Swamy MNS. An improved scheme for digital mammogram classification using weighted chaotic salp swarm algorithm-based kernel extreme learning machine. Applied Soft Computing. 2020;91.	Population – Image type
66.	Mohanty F, Rup S, Dash B, Majhi B, Swamy MNS. Digital mammogram classification using 2D-BDWT and GLCM features with FOA-based feature selection approach. Neural Computing & Applications. 2020;32(11):7029-43.	Population – Image type
67.	Moslemi H, Kazerouni IA, Hourali F. Breast cancer diagnosis in mammogram images using coordinate logic filters. Biomedical Research (India). 2017;28(22):10108-11.	Population – Image type
68.	Muduli D, Dash R, Majhi B. Automated breast cancer detection in digital mammograms: A moth flame optimization based ELM approach. Biomedical Signal Processing and Control. 2020;59 (no pagination).	Population – Image type
69.	Mughal B, Muhammad N, Sharif M. Adaptive hysteresis thresholding segmentation technique for localizing the breast masses in the curve stitching domain. Int J Med Inf. 2019;126:26-34.	Population – Image type
70.	Nagarajan V, Britto EC, Veeraputhiran SM. Feature extraction based on empirical mode decomposition for automatic mass classification of mammogram images. Medicine in Novel Technology and Devices. 2019;1 (no pagination).	Population – Image type
71.	Nagthane DK, Rajurkar AM. An improved diagnosis technique for breast cancer using LCFS and TreeHiCARe classifier model. Sensor Review. 2019;39(1):107-20.	Population – Image type

Reference		Main reason for exclusion
72.	Nanni L, Brahnam S, Lumini A. A very high performing system to discriminate tissues in mammograms as benign and malignant. Expert Systems with Applications. 2012;39(2):1968-71.	Population – Image type
73.	Narvaez F, Alvarez J, Garcia-Arteaga JD, Tarquino J, Romero E. Characterizing Architectural Distortion in Mammograms by Linear Saliency. J Med Syst. 2017;41(2):26.	Population – Image type
74.	Naseem MT, Sulong GZB, Jaffar MA. MRT letter: Quantum noise removal and classification of breast mammogram images. Microscopy Research and Technique. 2012;75(12):1609-12.	Population – Image type
75.	Naveed N, Jaffar MA, Choi TS. MRT Letter: Segmentation and Texture-Based Classification of Breast Mammogram Images. Microscopy Research and Technique. 2011;74(11):985-7.	Population – Image type
76.	Neto OPS, Silva AC, Paiva AC, Gattass M. Automatic mass detection in mammography images using particle swarm optimization and functional diversity indexes. Multimedia Tools and Applications. 2017;76(18):19263-89.	Population – Image type
77.	Nishikawa RM, Schmidt RA, Linver MN, Edwards AV, Papaioannou J, Stull MA. Clinically missed cancer: how effectively can radiologists use computer-aided detection? AJR Am J Roentgenol. 2012;198(3):708-16.	Population – Image type
78.	Nugroho HA, Fajrin HR, Soesanti I, Budiani RL. Analysis of texture for classification of breast cancer on mammogram images. International Journal of Medical Engineering and Informatics. 2018;10(4):382-91.	Population – Image type
79.	P S, R T. Aiding the Digital Mammogram for Detecting the Breast Cancer Using Shearlet Transform and Neural Network. Asian Pac J Cancer Prev. 2018;19(9):2665-71.	Population – Image type
80.	Pak F, Kanan HR, Alikhassi A. Breast cancer detection and classification in digital mammography based on Non-Subsampled Contourlet Transform (NSCT) and Super Resolution. Comput Methods Programs Biomed. 2015;122(2):89-107.	Population – Image type
81.	Panetta K, Zhou Y, Agaian S, Jia H. Nonlinear unsharp masking for mammogram enhancement. IEEE Trans Inf Technol Biomed. 2011;15(6):918-28.	Population – Image type
82.	Paquerault S, Hardy PT, Wersto N, Chen J, Smith RC. Investigation of optimal use of computer-aided detection systems: the role of the "machine" in decision making process. Acad Radiol. 2010;17(9):1112-21.	Population – Image type

Reference		Main reason for exclusion
83.	Paramkusham S, Rao KMM, Prabhakar Rao BVVSN, Sharma S. Application of TAR signature for breast mass analysis. Biomedical Research (India). 2018;29(10):2030-4.	Population – Image type
84.	Parmeggiani D, Avenia N, Sanguinetti A, Ruggiero R, Docimo G, Siciliano M, et al. Artificial intelligence against breast cancer (A.N.N.E.S-B.C.-Project). Ann Ital Chir. 2012;83(1):1-5.	Population – Image type
85.	Patel BC, Sinha GR, Soni D. Detection of masses in mammographic breast cancer images using modified histogram based adaptive thresholding (MHAT) method. International Journal of Biomedical Engineering and Technology. 2019;29(2):134-54.	Population – Image type
86.	Pawar MM, Talbar SN, Dudhane A. Local Binary Patterns Descriptor Based on Sparse Curvelet Coefficients for False-Positive Reduction in Mammograms. J. 2018;2018:5940436.	Population – Image type
87.	Perre AC, Alexandre LA, Freire LC. Lesion classification in mammograms using convolutional neural networks and transfer learning. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization. 2019;7(5-6):550-6.	Population – Image type
88.	Pezeshki H, Rastgarpour M, Sharifi A, Yazdani S. Extraction of spiculated parts of mammogram tumors to improve accuracy of classification. Multimedia Tools and Applications. 2019;78(14):19979-20003.	Population – Image type
89.	Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. Med Decis Making. 2013;33(1):98-107.	Population – Image type
90.	Qasim KR, Ouda AJ. An accurate breast cancer detection system based on deep learning cnn. Medico Legal Update. 2020;20(1).	Population – Image type
91.	Quellec G, Lamard M, Cozic M, Coatrieux G, Cazuguel G. Multiple-Instance Learning for Anomaly Detection in Digital Mammography. IEEE Trans Med Imaging. 2016;35(7):1604-14.	Population – Image type
92.	Rabidas R, Arif W. Characterization of mammographic masses based on local photometric attributes. Multimedia Tools and Applications. 2020;79(29-30):21967-85.	Population – Image type
93.	Rabidas R, Chakraborty J, Midya A. Analysis of 2D singularities for mammographic mass classification. Iet Computer Vision. 2017;11(1):22-32.	Population – Image type
94.	Rabidas R, Midya A, Chakraborty J. Neighborhood Structural Similarity Mapping for the Classification of Masses in Mammograms. IEEE j. 2018;22(3):826-34.	Population – Image type

Reference		Main reason for exclusion
95.	Rabottino G, Mencattini A, Salmeri M, Caselli F, Lojacono R. Performance evaluation of a region growing procedure for mammographic breast lesion identification. Computer Standards & Interfaces. 2011;33(2):128-35.	Population – Image type
96.	Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. Peerj. 2019;7:e6201.	Population – Image type
97.	Raghavendra U, Acharya UR, Fujita H, Gudigar A, Tan JH, Chokkadi S. Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identification of breast cancer using digitized mammogram images. Applied Soft Computing. 2016;46:151-61.	Population – Image type
98.	Rangayyan RM, Banik S, Chakraborty J, Mukhopadhyay S, Desautels JE. Measures of divergence of oriented patterns for the detection of architectural distortion in prior mammograms. Int. 2013;8(4):527-45.	Population – Image type
99.	Rangayyan RM, Banik S, Desautels JE. Computer-aided detection of architectural distortion in prior mammograms of interval cancer. J Digit Imaging. 2010;23(5):611-31.	Population – Image type
100.	Rangayyan RM, Banik S, Desautels JE. Detection of architectural distortion in prior mammograms via analysis of oriented patterns. J. 2013;78:30.	Population – Image type
101.	Rangayyan RM, Oloumi F. Fractal analysis and classification of breast masses using the power spectra of signatures of contours. Journal of Electronic Imaging. 2012;21(2).	Population – Image type
102.	Rangayyan RM, Oloumi F, Nguyen TM. Fractal analysis of contours of breast masses in mammograms via the power spectra of their signatures. Conf Proc IEEE Eng Med Biol Soc. 2010;2010:6737-40.	Population – Image type
103.	Rauch T, Rieger J, Pelzer G, Horn F, Erber R, Wunderle M, et al. Discrimination analysis of breast calcifications using x-ray dark-field radiography. Med Phys. 2020;47(4):1813-26.	Population – Image type
104.	Reyad YA, Berbar MA, Hussain M. Comparison of statistical, LBP, and multi-resolution analysis features for breast mass classification. J Med Syst. 2014;38(9):100.	Population – Image type
105.	Roberts T, Newell M, Auffermann W, Vidakovic B. Wavelet-based scaling indices for breast cancer diagnostics. Stat Med. 2017;36(12):1989-2000.	Population – Image type
106.	Roseline R, Manikandan S. Determination of Breast Cancer Using KNN Cluster Technique. Indian Journal of Public Health Research and Development. 2018;9(2):418-23.	Population – Image type

Reference		Main reason for exclusion
107.	Rouhi R, Jafari M. Classification of benign and malignant breast tumors based on hybrid level set segmentation. Expert Systems with Applications. 2016;46:45-59.	Population – Image type
108.	Rouhi R, Jafari M, Kasaei S, Keshavarzian P. Benign and malignant breast tumors classification based on region growing and CNN segmentation. Expert Systems with Applications. 2015;42(3):990-1002.	Population – Image type
109.	S RS, Rajaguru H. Comparison Analysis of Linear Discriminant Analysis and Cuckoo-Search Algorithm in the Classification of Breast Cancer from Digital Mammograms. Asian Pac J Cancer Prev. 2019;20(8):2333-7.	Population – Image type
110.	Safdar Gardezi SJ, Faye I. Mammogram classification based on morphological component analysis (MCA) and Curvelet decomposition. Neuroscience and Biomedical Engineering. 2015;3(1):27-33.	Population – Image type
111.	5Saki F, Tahmasbi A, Soltanian-Zadeh H, Shokouhi SB. Fast opposite weight learning rules with application in breast cancer diagnosis. Comput Biol Med. 2013;43(1):32-41.	Population – Image type
112.	Salazar-Licea LA, Pedraza-Ortega JC, Pastrana-Palma A, Aceves-Fernandez MA. Location of mammograms ROI's and reduction of false-positive. Comput Methods Programs Biomed. 2017;143:97-111.	Population – Image type
113.	Sampaio WB, Diniz EM, Silva AC, de Paiva AC, Gattass M. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. Comput Biol Med. 2011;41(8):653-64.	Population – Image type
114.	Samulski M, Karssemeijer N. Optimizing Case-based detection performance in a multiview CAD system for mammography. IEEE Trans Med Imaging. 2011;30(4):1001-9.	Population – Image type
115.	Saranyaraj D, Manikandan M, Maheswari S. A deep convolutional neural network for the early detection of breast carcinoma with respect to hyper- parameter tuning. Multimedia Tools and Applications. 2020;79(15-16):11013-38.	Population – Image type
116.	Saraswathi D, Srinivasan E. A CAD system to analyse mammogram images using fully complex-valued relaxation neural network ensembled classifier. J Med Eng Technol. 2014;38(7):359-66.	Population – Image type
117.	Saybani MR, Teh YW, Aghabozorgi SR, Shamshirband S, Kiah MLM, Balas VE. Diagnosing breast cancer with an improved artificial immune recognition system. Soft Computing. 2016;20(10):4069-84.	Population – Image type

Reference		Main reason for exclusion
118.	Schonmeyer R, Athelogou M, Sittek H, Ellenberg P, Feehan O, Schmidt G, et al. Cognition Network Technology prototype of a CAD system for mammography to assist radiologists by finding similar cases in a reference database. Int. 2011;6(1):127-34.	Population – Image type
119.	Selvamani I, Arasu GT. Computer aided system for detection and classification of breast cancer. Current Medical Imaging Reviews. 2015;11(2):77-84.	Population – Image type
120.	Selvi C, Suganthi M. A Novel Enhanced Gray Scale Adaptive Method for Prediction of Breast Cancer. J Med Syst. 2018;42(11).	Population – Image type
121.	Senthilkumar B, Umamaheswari G. Combination of novel enhancement technique and fuzzy C means clustering technique in breast cancer detection. Biomedical Research (India). 2013;24(2):252-6.	Population – Image type
122.	Sha ZJ, Hu L, Rouyendegh BD. Deep learning and optimization algorithms for automatic breast cancer detection. International Journal of Imaging Systems and Technology. 2020;30(2):495-506.	Population – Image type
123.	Shahin OR, Alruily M, Alsmarah M, Alruwaill M. Breast cancer detection using modified hough transform. Biomedical Research (India). 2018;29(16):3188-91.	Population – Image type
124.	Sharma S, Khanna P. Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM. J Digit Imaging. 2015;28(1):77-90.	Population – Image type
125.	Shen LZ, He MF, Shen N, Yousefi N, Wang C, Liu GQ. Optimal breast tumor diagnosis using discrete wavelet transform and deep belief network based on improved sunflower optimization method. Biomedical Signal Processing and Control. 2020;60.	Population – Image type
126.	Shirazinodeh A, Noubari HA, Rabbani H, Dehnavi AM. Detection and classification of breast cancer in wavelet sub-bands of fractal segmented cancerous zones. Journal of Medical Signals and Sensors. 2015;5(3):162-70.	Population – Image type
127.	Shobha Rani N, Rao CS. Exploration and evaluation of efficient pre-processing and segmentation technique for breast cancer diagnosis based on mammograms. International Journal of Research in Pharmaceutical Sciences. 2019;10(3):2071-81.	Population – Image type
128.	Singh B, Kaur M. An approach for classification of malignant and benign microcalcification clusters. Sadhana-Academy Proceedings in Engineering Sciences. 2018;43(3).	Population – Image type
129.	Singh L, Jaffery ZA. Computer-aided diagnosis of breast cancer in digital mammograms. International Journal of Biomedical Engineering and Technology. 2018;27(3):233-46.	Population – Image type

Reference		Main reason for exclusion
130.	Singh SP, Urooj S. An Improved CAD System for Breast Cancer Diagnosis Based on Generalized Pseudo-Zernike Moment and Ada-DEWNN Classifier. J Med Syst. 2016;40(4):105.	Population – Image type
131.	Singh SP, Urooj S, Lay-Ekuakille A. Breast Cancer Detection Using PCPCET and ADEWNN: A Geometric Invariant Approach to Medical X-Ray Image Sensors. Ieee Sensors Journal. 2016;16(12):4847-55.	Population – Image type
132.	Singh VP, Srivastava S, Srivastava R. Effective mammogram classification based on center symmetric-LBP features in wavelet domain using random forests. Technology and Health Care. 2017;25(4):709-27.	Population – Image type
133.	Singh WJ, Nagarajan B. Automatic diagnosis of mammographic abnormalities based on hybrid features with learning classifier. Comput Methods Biomech Biomed Engin. 2013;16(7):758-67.	Population – Image type
134.	Soulami KB, Saidi MN, Honnit B, Anibou C, Tamtaoui A. Detection of breast abnormalities in digital mammograms using the electromagnetism-like algorithm. Multimedia Tools and Applications. 2019;78(10):12835-63.	Population – Image type
135.	Sriramkumar D, Malmathanraj R, Mohan R, Umamaheswari S. Mammogram tumour classification using modified segmentation techniques. International Journal of Biomedical Engineering and Technology. 2013;13(3):218-39.	Population – Image type
136.	Suganthi M, Madheswaran M. An improved medical decision support system to identify the breast cancer using mammogram. J Med Syst. 2012;36(1):79-91.	Population – Image type
137.	Suvetha K, Sultana M. Analysis of breast cancer using tetrolet transform. Indian Journal of Public Health Research and Development. 2017;8(3 Supplement):19-21.	Population – Image type
138.	Tahmasbi A, Saki F, Shokouhi SB. Classification of benign and malignant masses based on Zernike moments. Comput Biol Med. 2011;41(8):726-35.	Population – Image type
139.	Tai SC, Chen ZS, Tsai WT. An automatic mass detection system in mammograms based on complex texture features. IEEE j. 2014;18(2):618-27.	Population – Image type
140.	Tan M, Pu J, Zheng B. Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model. Int. 2014;9(6):1005-20.	Population – Image type
141.	Thivya KS, Sakthivel P, Sai PMV. Analysis of framelets for breast cancer diagnosis. Technology and Health Care. 2016;24(1):21-9.	Population – Image type

Reference		Main reason for exclusion
142.	Timp S, Varela C, Karssemeijer N. Computer-aided diagnosis with temporal analysis to improve radiologists' interpretation of mammographic mass lesions. IEEE Trans Inf Technol Biomed. 2010;14(3):803-8.	Population – Image type
143.	Ting FF, Tan YJ, Sim KS. Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications. 2019;120:103-15.	Population – Image type
144.	Uppal MTN. Classification of mammograms for breast cancer detection using fusion of discrete cosine transform and discrete wavelet transform features. Biomedical Research (India). 2016;27(2):322-7.	Population – Image type
145.	Vaijayanthi N, Caroline BE, Murugan VS. Automatic detection of masses in mammograms using bi-dimensional empirical mode decomposition. Journal of Medical Imaging and Health Informatics. 2018;8(7):1326-41.	Population – Image type
146.	Velikova M, Lucas PJ, Karssemeijerb N. Using local context information to improve automatic mammographic mass detection. Stud Health Technol Inform. 2010;160(Pt 2):1291-5.	Population – Image type
147.	Vikhe PS, Thool VR. Mass Detection in Mammographic Images Using Wavelet Processing and Adaptive Threshold Technique. J Med Syst. 2016;40(4):82.	Population – Image type
148.	Wang H, Feng J, Bu Q, Liu F, Zhang M, Ren Y, et al. Breast Mass Detection in Digital Mammogram Based on Gestalt Psychology. J. 2018;2018:4015613.	Population – Image type
149.	Wang X, Li L, Liu W, Xu W, Lederman D, Zheng B. An interactive system for computer-aided diagnosis of breast masses. J Digit Imaging. 2012;25(5):570-9.	Population – Image type
150.	Wei J, Chan HP, Zhou C, Wu YT, Sahiner B, Hadjiiski LM, et al. Computer-aided detection of breast masses: four-view strategy for screening mammography. Med Phys. 2011;38(4):1867-76.	Population – Image type
Population – Mammography type not reported (n=8)		
151.	Li Y, Chen H, Yang Y, Cheng L, Cao L. A bilateral analysis scheme for false positive reduction in mammogram mass detection. Comput Biol Med. 2015;57:84-95.	Population – Mammography type not reported
152.	Moin P, Deshpande R, Sayre J, Messer E, Gupte S, Romsdahl H, et al. An observer study for a computer-aided reading protocol (CARP) in the screening environment for digital mammography. Acad Radiol. 2011;18(11):1420-9.	Population – Mammography type not reported

Reference		Main reason for exclusion
153.	Mutasa S, Chang P, Nemer J, Van Sant EP, Sun M, McIlvride A, et al. Prospective Analysis Using a Novel CNN Algorithm to Distinguish Atypical Ductal Hyperplasia From Ductal Carcinoma in Situ in Breast. Clinical Breast Cancer. 2020.	Population – Mammography type not reported
154.	Padmavathy TV, Vimalkumar MN, Bhargava DS. Adaptive clustering based breast cancer detection with ANFIS classifier using mammographic images. Cluster Computing-the Journal of Networks Software Tools and Applications. 2019;22:13975-84.	Population – Mammography type not reported
155.	Ribli D, Horvath A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Sci. 2018;8(1):4165.	Population – Mammography type not reported
156.	Sasikala S, Bharathi M, Ezhilarasi M, Reddy MR, Arunkumar S. Fusion of MLO and CC view binary patterns to improve the performance of breast cancer diagnosis. Current Medical Imaging Reviews. 2018;14(4):651-8.	Population – Mammography type not reported
157.	Sasikala S, Ezhilarasi M. Comparative analysis of serial and parallel fusion on texture features for improved breast cancer diagnosis. Current Medical Imaging Reviews. 2018;14(6):957-68.	Population – Mammography type not reported
158.	Vimalkumar MN, Helenprabha K. Adaptive neuro-fuzzy inference system for classification of mammographic image using electromagnetism-like optimisation. International Journal of Biomedical Engineering and Technology. 2018;26(3-4):376-84.	Population – Mammography type not reported
Population – Incomplete images (e.g. regions of interest) (n=8)		
159.	Sun W, Tseng TL, Zhang J, Qian W. Computerized breast cancer analysis system using three stage semi-supervised learning method. Comput Methods Programs Biomed. 2016;135:77-88.	Population – Incomplete images
160.	Tan M, Pu J, Zheng B. A new and fast image feature selection method for developing an optimal mammographic mass detection scheme. Med Phys. 2014;41(8):081906.	Population – Incomplete images
161.	Wang XH, Park SC, Zheng B. Assessment of performance and reliability of computer-aided detection scheme using content-based image retrieval approach and limited reference database. J Digit Imaging. 2011;24(2):352-9.	Population – Incomplete images
162.	Yu X, Kang C, Guttery DS, Kadry S, Chen Y, Zhang YD. ResNet-SCDA-50 for breast abnormality classification. IEEE/ACM transactions on computational biology and bioinformatics. 2020;13.	Population – Incomplete images
163.	Zhang Y, Tomuro N, Furst J, Raicu DS. Building an ensemble system for diagnosing masses in mammograms. Int. 2012;7(2):323-9.	Population – Incomplete images

Reference		Main reason for exclusion
164.	Zyout I, Togneri R. Empirical mode decomposition of digital mammograms for the statistical based characterization of architectural distortion. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:109-12.	Population – Incomplete images
165.	Zyout I, Togneri R. A new approach for the detection of architectural distortions using textural analysis of surrounding tissue. Conf Proc IEEE Eng Med Biol Soc. 2016;2016:3965-8.	Population – Incomplete images
166.	Zyout I, Togneri R. A computer-aided detection of the architectural distortion in digital mammograms using the fractal dimension measurements of BEMD. Comput Med Imaging Graph. 2018;70:173-84.	Population – Incomplete images
Population – Subpolulation (e.g. only cancer cases) (n=3)		
167.	Bolivar AV, Gomez SS, Merino P, Alonso-Bartolome P, Garcia EO, Cacho PM, et al. Computer-aided detection system applied to full-field digital mammograms. Acta Radiol. 2010;51(10):1086-92.	Population - Subpopulation
168.	Cho KR, Seo BK, Woo OH, Song SE, Choi J, Whang SY, et al. Breast Cancer Detection in a Screening Population: Comparison of Digital Mammography, Computer-Aided Detection Applied to Digital Mammography and Breast Ultrasound. Journal of Breast Cancer. 2016;19(3):316-23.	Population - Subpopulation
169.	Hamza AO, El-Sanosi MD, Habbani AK, Mustafa NA, Khider MO. Computer-aided detection of benign tumors of the female breast. Journal of Clinical Engineering. 2013;38(1):32-7.	Population - Subpopulation
Population – <90% screening mammograms or unclear proportion (n=10)		
170.	Al-Najdawi N, Biltawi M, Tedmori S. Mammogram image visual enhancement, mass segmentation and classification. Applied Soft Computing. 2015;35:175-85.	Population – <90% screening mammograms or unclear proportion
171.	Angayarkanni N, Kumar D, Arunachalam G. The application of image processing techniques for detection and classification of cancerous tissue in digital mammograms. Journal of Pharmaceutical Sciences and Research. 2016;8(10):1179-83.	Population – <90% screening mammograms or unclear proportion
172.	Cascio D, Fauci F, Iacomi M, Raso G, Magro R, Castrogiovanni D, et al. Computer-aided diagnosis in digital mammography: Comparison of two commercial systems. Imaging in Medicine. 2014;6(1):13-20.	Population – <90% screening mammograms or unclear proportion
173.	Diz J, Marreiros G, Freitas A. Applying Data Mining Techniques to Improve Breast Cancer Diagnosis. J Med Syst. 2016;40(9).	Population – <90% screening mammograms or unclear proportion

Reference		Main reason for exclusion
174.	Langarizadeh M, Mahmud R, Bagherzadeh R. Detection of masses and microcalcifications in digitalmammogram images using fuzzy logic. Asian Biomedicine. 2016;10(4):345-50.	Population – <90% screening mammograms or unclear proportion
175.	Mutasa S, Chang P, Van Sant EP, Nemer J, Liu M, Karcich J, et al. Potential Role of Convolutional Neural Network Based Algorithm in Patient Selection for DCIS Observation Trials Using a Mammogram Dataset. Acad Radiol. 2020;27(6):774-9.	Population – <90% screening mammograms or unclear proportion
176.	Sasaki M, Tozaki M, Rodriguez-Ruiz A, Yotsumoto D, Ichiki Y, Terawaki A, et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer. 2020;27(4):642-51.	Population – <90% screening mammograms or unclear proportion
177.	Rodriguez-Ruiz A, Lang K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol. 2019;29(9):4825-32.	Population – <90% screening mammograms or unclear proportion
178.	Soulami KB, Kaabouch N, Saidi MN, Tamtaoui A. An evaluation and ranking of evolutionary algorithms in segmenting abnormal masses in digital mammograms. Multimedia Tools and Applications. 2020;79(27-28):18941-79.	Population – <90% screening mammograms or unclear proportion
179.	Zheng J, Lin DA, Gao ZJ, Wang S, He MJ, Fan JP. Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis. Ieee Access. 2020;8:96946-54.	Population – <90% screening mammograms or unclear proportion
Internal validation – Cross validation (n=91)		
180.	Abdar M, Zomorodi-Moghadam M, Zhou XJ, Gururajan R, Tao XH, Barua PD, et al. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters. 2020;132:123-31.	Internal validation – Cross validation
181.	Agarwal R, Diaz O, Yap MH, Llado X, Marti R. Deep learning for mass detection in Full Field Digital Mammograms. Comput Biol Med. 2020;121:103774.	Internal validation – Cross validation
182.	Ahmadi A, Afshar P. Intelligent breast cancer recognition using particle swarm optimization and support vector machines. Journal of Experimental & Theoretical Artificial Intelligence. 2016;28(6):1021-34.	Internal validation – Cross validation
183.	Al-Antari MA, Al-Masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. Int J Med Inf. 2018;117:44-54.	Internal validation – Cross validation

Reference	Main reason for exclusion
184. Al-Masni MA, Al-Antari MA, Park JM, Gi G, Kim TY, Rivera P, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Comput Methods Programs Biomed. 2018;157:85-94.	Internal validation – Cross validation
185. Aminikhanghahi S, Shin S, Wang W, Jeon SI, Son SH. A new fuzzy Gaussian mixture model (FGMM) based algorithm for mammography tumor image classification. Multimedia Tools and Applications. 2017;76(7):10191-205.	Internal validation – Cross validation
186. Arzehgar A, Khalilzadeh MM, Varshoei F. Assessment and classification of mass lesions based on expert knowledge using mammographic analysis. Current Medical Imaging Reviews. 2019;15(2):199-208.	Internal validation – Cross validation
187. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Jr., Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer. 2010;116(14):3310-21.	Internal validation – Cross validation
188. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Computing & Applications. 2013;23(7-8):2387-403.	Internal validation – Cross validation
189. Azar AT, El-Said SA. Probabilistic neural network for breast cancer classification. Neural Computing & Applications. 2013;23(6):1737-51.	Internal validation – Cross validation
190. Azar AT, El-Said SA. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Computing & Applications. 2014;24(5):1163-77.	Internal validation – Cross validation
191. Beura S, Majhi B, Dash R, Roy S. Classification of mammogram using two-dimensional discrete orthonormal S-transform for breast cancer detection. Healthc. 2015;2(2):46-51.	Internal validation – Cross validation
192. Bouyer A. Breast cancer diagnosis using data mining methods, cumulative histogram features, and gary level co-occurrence matrix. Current Medical Imaging Reviews. 2017;13(4):460-70.	Internal validation – Cross validation
193. Cao P, Liu X, Bao H, Yang J, Zhao D. Restricted Boltzmann machines based oversampling and semi-supervised learning for false positive reduction in breast CAD. Bio-Medical Materials and Engineering. 2015;26(Supplement 1):S1541-S7.	Internal validation – Cross validation
194. Carneiro G, Nascimento J, Bradley AP. Automated Analysis of Unregistered Multi-View Mammograms With Deep Learning. IEEE Trans Med Imaging. 2017;36(11):2355-65.	Internal validation – Cross validation

Reference		Main reason for exclusion
195.	Casti P, Mencattini A, Salmeri M, Ancona A, Lorusso M, Pepe ML, et al. Towards localization of malignant sites of asymmetry across bilateral mammograms. Computer Methods and Programs in Biomedicine. 2017;140:11-8.	Internal validation – Cross validation
196.	Casti P, Mencattini A, Salmeri M, Ancona A, Mangeri F, Pepe ML, et al. Contour-independent detection and classification of mammographic lesions. Biomedical Signal Processing and Control. 2016;25:165-77.	Internal validation – Cross validation
197.	Casti P, Mencattini A, Salmeri M, Ancona A, Mangieri F, Rangayyan RM. Development and validation of a fully automated system for detection and diagnosis of mammographic lesions. Conf Proc IEEE Eng Med Biol Soc. 2014;2014:4667-70.	Internal validation – Cross validation
198.	Celaya-Padilla J, Martinez-Torteya A, Rodriguez-Rojas J, Galvan-Tejada J, Trevino V, Tamez-Pena J. Bilateral Image Subtraction and Multivariate Models for the Automated Triaging of Screening Mammograms. Biomed Res Int. 2015;2015:231656.	Internal validation – Cross validation
199.	Celaya-Padilla JM, Guzman-Valdivia CH, Galvan-Tejada CE, Galvan-Tejada JI, Gamboa-Rosales H, Garza-Veloz I, et al. Contralateral asymmetry for breast cancer detection: A CADx approach. Biocybernetics and Biomedical Engineering. 2018;38(1):115-25.	Internal validation – Cross validation
200.	Chakraborty J, Midya A, Rabidas R. Computer-aided detection and diagnosis of mammographic masses using multi-resolution analysis of oriented tissue patterns. Expert Systems with Applications. 2018;99:168-79.	Internal validation – Cross validation
201.	Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications. 2011;38(7):9014-22.23.	Internal validation – Cross validation
202.	Chen HL, Yang B, Wang G, Wang SJ, Liu J, Liu DY. Support vector machine based diagnostic system for breast cancer using swarm intelligence. J Med Syst. 2012;36(4):2505-19.	Internal validation – Cross validation
203.	Chen X, Zargari A, Hollingsworth AB, Liu H, Zheng B, Qiu Y. Applying a new quantitative image analysis scheme based on global mammographic features to assist diagnosis of breast cancer. Comput Methods Programs Biomed. 2019;179:104995.	Internal validation – Cross validation
204.	Choi JY. A generalized multiple classifier system for improving computer-aided classification of breast masses in mammography. Biomedical Engineering Letters. 2015;5(4):251-62.	Internal validation – Cross validation

Reference		Main reason for exclusion
205.	Choi JY, Kim DH, Plataniotis KN, Ro YM. Combining multiple feature representations and AdaBoost ensemble learning for reducing false-positive detections in computer-aided detection of masses on mammograms. Conf Proc IEEE Eng Med Biol Soc. 2012;2012:4394-7.	Internal validation – Cross validation
206.	Choi JY, Kim DH, Plataniotis KN, Ro YM. Computer-aided detection (CAD) of breast masses in mammography: combined detection and ensemble classification. Phys Med Biol. 2014;59(14):3697-719.	Internal validation – Cross validation
207.	Costa DD, Campos LF, Barros AK. Classification of breast tissue in mammograms using efficient coding. Biomed. 2011;10:55.	Internal validation – Cross validation
208.	Dhungel N, Carneiro G, Bradley AP. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med Image Anal. 2017;37:114-28.	Internal validation – Cross validation
209.	do Nascimento MZ, Martins AS, Neves LA, Ramos RP, Flores EL, Carrijo GA. Classification of masses in mammographic image using wavelet domain features and polynomial classifier. Expert Systems with Applications. 2013;40(15):6213-21.	Internal validation – Cross validation
210.	Dong M, Lu X, Ma Y, Guo Y, Ma Y, Wang K. An Efficient Approach for Automated Mass Segmentation and Classification in Mammograms. J Digit Imaging. 2015;28(5):613-25.	Internal validation – Cross validation
211.	Drukker K, Giger ML, Joe BN, Kerlikowske K, Greenwood H, Drukteinis JS, et al. Combined Benefit of Quantitative Three-Compartment Breast Image Analysis and Mammography Radiomics in the Classification of Breast Masses in a Clinical Data Set. Radiology. 2019;290(3):621-8.	Internal validation – Cross validation
212.	Eltrass AS, Salama MS. Fully automated scheme for computer-aided detection and breast cancer diagnosis using digitised mammograms. Iet Image Processing. 2020;14(3):495-505.	Internal validation – Cross validation
213.	Esmaeili M, Ayyoubzadeh SM, Ahmadinejad N, Ghazisaeedi M, Nahvijou A, Maghooli K. A decision support system for mammography reports interpretation. Health Inf Sci Syst. 2020;8(1):17.	Internal validation – Cross validation
214.	Fanizzi A, Basile TMA, Losurdo L, Bellotti R, Bottigli U, Dentamaro R, et al. A machine learning approach on multiscale texture analysis for breast microcalcification diagnosis. BMC Bioinformatics. 2020;21(Suppl 2):91.	Internal validation – Cross validation

Reference		Main reason for exclusion
215.	Ganesan K, Acharya UR, Chua CK, Lim CM, Abraham KT. One-Class Classification of Mammograms Using Trace Transform Functionals. Ieee Transactions on Instrumentation and Measurement. 2014;63(2):304-11.	Internal validation – Cross validation
216.	Ganesan K, Acharya UR, Chua CK, Min LC, Abraham TK. Automated diagnosis of mammogram images of breast cancer using discrete wavelet transform and spherical wavelet transform features: a comparative study. Technol Cancer Res Treat. 2014;13(6):605-15.	Internal validation – Cross validation
217.	Gao F, Wu T, Li J, Zheng B, Ruan L, Shang D, et al. SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis. Comput Med Imaging Graph. 2018;70:53-62.	Internal validation – Cross validation
218.	Garcia-Manso A, Garcia-Orellana CJ, Gonzalez-Velasco H, Gallardo-Caballero R, Macias MM. Consistent performance measurement of a system to detect masses in mammograms based on blind feature extraction. Biomed. 2013;12:2.	Internal validation – Cross validation
219.	Ghasemzadeh A, Azad SS, Esmaeili E. Breast cancer detection based on Gabor-wavelet transform and machine learning methods. International Journal of Machine Learning and Cybernetics. 2019;10(7):1603-12.	Internal validation – Cross validation
220.	Ghosh A. Artificial Intelligence Using Open Source BI-RADS Data Exemplifying Potential Future Use. J. 2019;16(1):64-72.	Internal validation – Cross validation
221.	Gomez-Flores W, Hernandez-Lopez J. Assessment of the invariance and discriminant power of morphological features under geometric transformations for breast tumor classification. Computer Methods and Programs in Biomedicine. 2020;185.	Internal validation – Cross validation
222.	Ha R, Chang P, Karcich J, Mutasa S, Pascual Van Sant E, Liu MZ, et al. Convolutional Neural Network Based Breast Cancer Risk Stratification Using a Mammographic Dataset. Acad Radiol. 2019;26(4):544-9.	Internal validation – Cross validation
223.	Hai J, Tan H, Chen J, Wu M, Qiao K, Xu J, et al. Multi-level features combined end-to-end learning for automated pathological grading of breast cancer on digital mammograms. Computerized Medical Imaging and Graphics. 2019;71:58-66.	Internal validation – Cross validation
224.	Heidari M, Mirniaharikandehei S, Liu W, Hollingsworth AB, Liu H, Zheng B. Development and Assessment of a New Global Mammographic Image Feature Analysis Scheme to Predict Likelihood of Malignant Cases. IEEE Trans Med Imaging. 2020;39(4):1235-44.	Internal validation – Cross validation
225.	Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging (Bellingham). 2016;3(3).	Internal validation – Cross validation

Reference		Main reason for exclusion
226.	Jacomini RS, Nascimento MZ, Dantas RD, Ramos RP. Classification of mass in two views mammograms: Use of analysis of variance (ANOVA) for reduction of the features. Recent Patents on Medical Imaging. 2013;3(1):80-8.	Internal validation – Cross validation
227.	Keles A, Keles A. Extracting fuzzy rules for the diagnosis of breast cancer. Turkish Journal of Electrical Engineering and Computer Sciences. 2013;21(5):1495-503.	Internal validation – Cross validation
228.	Khan S, Hussain M, Aboalsamh H, Mathkour H, Bebis G, Zakariah M. Optimized Gabor features for mass classification in mammography. Applied Soft Computing. 2016;44:267-80.	Internal validation – Cross validation
229.	Khan S, Khan A, Maqsood M, Aadil F, Ghazanfar MA. Optimized Gabor Feature Extraction for Mass Classification Using Cuckoo Search for Big Data E-Healthcare. Journal of Grid Computing. 2019;17(2):239-54.	Internal validation – Cross validation
230.	Kilic N, Gorgel P, Ucan ON, Sertbas A. Mammographic mass detection using wavelets as input to neural networks. J Med Syst. 2010;34(6):1083-8.	Internal validation – Cross validation
231.	Kim DH, Choi JY, Ro YM. Region based stellate features combined with variable selection using AdaBoost learning in mammographic computer-aided detection. Computers in Biology and Medicine. 2015;63:238-50.	Internal validation – Cross validation
232.	Kim DH, Lee SH, Ro YM. Mass type-specific sparse representation for mass classification in computer-aided detection on mammograms. Biomed. 2013;12 Suppl 1:S3.	Internal validation – Cross validation
233.	Kim S. Margin-maximised redundancy-minimised SVM-RFE for diagnostic classification of mammograms. Int J Data Min Bioinform. 2014;10(4):374-90.	Internal validation – Cross validation
234.	Kooi T, van Ginneken B, Karssemeijer N, den Heeten A. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. Med Phys. 2017;44(3):1017-27.	Internal validation – Cross validation
235.	Kozegar E, Soryani M. A cost-sensitive Bayesian combiner for reducing false positives in mammographic mass detection. Biomed Tech (Berl). 2019;64(1):39-52.	Internal validation – Cross validation
236.	Kozegar E, Soryani M, Minaei B, Domingues I. Assessment of a novel mass detection algorithm in mammograms. J Cancer Res Ther. 2013;9(4):592-600.	Internal validation – Cross validation
237.	Kyono T, Gilbert FJ, van der Schaar M. Improving Workflow Efficiency for Mammography Using Machine Learning. J. 2020;17(1 Pt A):56-63.	Internal validation – Cross validation (UK Tommy dataset)

Reference	Main reason for exclusion
238. Lakshmanan R, Shiji TP, Jacob SM, Pratab T, Thomas C, Thomas V. Detection of architectural distortion in mammograms using geometrical properties of thinned edge structures. Intelligent Automation and Soft Computing. 2017;23(1):183-97.	Internal validation – Cross validation
239. Lee J, Nishikawa RM. Detecting mammographically occult cancer in women with dense breasts using deep convolutional neural network and Radon Cumulative Distribution Transform. J Med Imaging (Bellingham). 2019;6(4):044502.	Internal validation – Cross validation
240. Li H, Zhuang S, Li DA, Zhao J, Ma Y. Benign and malignant classification of mammogram images based on deep learning. Biomedical Signal Processing and Control. 2019;51:347-54.	Internal validation – Cross validation
241. Shan LH, Faust O, Yu W. Data mining framework for breast cancer detection in mammograms: A hybrid feature extraction paradigm. Journal of Medical Imaging and Health Informatics. 2014;4(5):756-65.	Internal validation – Cross validation
242. Liu N, Qi ES, Xu M, Gao B, Liu GQ. A novel intelligent classification model for breast cancer diagnosis. Information Processing & Management. 2019;56(3):609-23.	Internal validation – Cross validation
243. Luo ST, Cheng BW. Diagnosing breast masses in digital mammography using feature selection and ensemble methods. J Med Syst. 2012;36(2):569-77.	Internal validation – Cross validation
244. Mednikov Y, Nehemia S, Zheng B, Benzaquen O, Lederman D. Transfer Representation Learning using Inception-V3 for the Detection of Masses in Mammography. Conf Proc IEEE Eng Med Biol Soc. 2018;2018:2587-90.	Internal validation – Cross validation
245. Melendez J, Sanchez CI, van Ginneken B, Karssemeijer N. Improving mass candidate detection in mammograms via feature maxima propagation and local feature selection. Med Phys. 2014;41(8):081904.	Internal validation – Cross validation
246. Milosevic M, Jankovic D, Peulic A. Comparative analysis of breast cancer detection in mammograms and thermograms. Biomed Tech (Berl). 2015;60(1):49-56.	Internal validation – Cross validation
247. Min H, Chandra SS, Crozier S, Bradley AP. Multi-scale sifting for mammographic mass detection and segmentation. Biomedical Physics and Engineering Express. 2019;5(2).	Internal validation – Cross validation
248. Naghibi S, Teshnehlab M, Shoorehdeli MA. Breast cancer classification based on advanced multi dimensional fuzzy neural network. J Med Syst. 2012;36(5):2713-20.	Internal validation – Cross validation
249. Nassif H, Wu Y, Page D, Burnside E. Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women. AMIA Annu Symp Proc. 2012;2012:1330-9.	Internal validation – Cross validation

Reference		Main reason for exclusion
250.	Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L. A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics. 2017;34(4):133-44.	Internal validation – Cross validation
251.	Oliver A, Freixenet J, Marti J, Perez E, Pont J, Denton ER, et al. A review of automatic mass detection and segmentation in mammographic images. Med Image Anal. 2010;14(2):87-110.	Internal validation – Cross validation
252.	Peng J, Bao C, Hu C, Wang X, Jian W, Liu W. Automated mammographic mass detection using deformable convolution and multiscale features. Medical and Biological Engineering and Computing. 2020;58(7):1405-17.	Internal validation – Cross validation
253.	Perez NP, Guevara Lopez MA, Silva A, Ramos I. Improving the Mann-Whitney statistical test for feature selection: an approach in breast cancer diagnosis on mammography. Artif Intell Med. 2015;63(1):19-31.	Internal validation – Cross validation
254.	Qiu Y, Yan S, Gundreddy RR, Wang Y, Cheng S, Liu H, et al. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. Journal of X-Ray Science and Technology. 2017;25(5):751-63.	Internal validation – Cross validation
255.	Ragab DA, Sharkas M, Attallah O. Breast cancer diagnosis using an efficient CAD system based on multiple classifiers. Diagnostics (Basel). 2019;9(4).	Internal validation – Cross validation
256.	Sapate S, Talbar S, Mahajan A, Sable N, Desai S, Thakur M. Breast cancer diagnosis using abnormalities on ipsilateral views of digital mammograms. Biocybernetics and Biomedical Engineering. 2020;40(1):290-305.	Internal validation – Cross validation
257.	Sapate SG, Mahajan A, Talbar SN, Sable N, Desai S, Thakur M. Radiomics based detection and characterization of suspicious lesions on full field digital mammograms. Comput Methods Programs Biomed. 2018;163:1-20.	Internal validation – Cross validation
258.	Suresh A, Udendhran R, Balamurgan M, Varatharajan R. A Novel Internet of Things Framework Integrated with Real Time Monitoring for Intelligent Healthcare Environment. J Med Syst. 2019;43(6):165.	Internal validation – Cross validation
259.	Tan M, Aghaei F, Wang Y, Zheng B. Developing a new case based computer-aided detection scheme and an adaptive cueing method to improve performance in detecting mammographic lesions. Phys Med Biol. 2017;62(2):358-76.	Internal validation – Cross validation

Reference	Main reason for exclusion
260. Tan M, Pu JT, Zheng B. Reduction of false-positive recalls using a computerized mammographic image feature analysis scheme. Physics in Medicine and Biology. 2014;59(15):4357-73.	Internal validation – Cross validation
261. Torabi M, Razavian SM, Vaziri R, Vosoughi-Vahdat B. A Wavelet-packet-based approach for breast cancer classification. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:5100-3.	Internal validation – Cross validation
262. Velikova M, Lucas PJ, Samulski M, Karssemeijer N. A probabilistic framework for image information fusion with an application to mammographic analysis. Med Image Anal. 2012;16(4):865-75.	Internal validation – Cross validation
263. Velikova M, Lucas PJ, Samulski M, Karssemeijer N. On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks. Artif Intell Med. 2013;57(1):73-86.	Internal validation – Cross validation
264. Wang Z, Huang Y, Li M, Zhang H, Li C, Xin J, et al. Breast mass detection and diagnosis using fused features with density. Journal of X-Ray Science and Technology. 2019;27(2):321-42.	Internal validation – Cross validation
265. Wang Z, Yu G, Kang Y, Zhao Y, Qu Q. Breast tumor detection in digital mammography based on extreme learning machine. Neurocomputing. 2014;128:175-84.	Internal validation – Cross validation
266. Xie W, Li Y, Ma Y. Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing. 2016;Part 3. 173:930-41.	Internal validation – Cross validation
267. Yang LY, Xu ZS. Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning. International Journal of Machine Learning and Cybernetics. 2019;10(3):591-601.	Internal validation – Cross validation
268. Zadeh HG, Seryasat OR, Haddadnia J. Assessment of a novel computer aided mass diagnosis system in mammograms. Biomedical Research (India). 2017;28(7):3129-35.	Internal validation – Cross validation
269. Zeng JM, Gimenez F, Burnside ES, Rubin DL, Shachter R. A Probabilistic Model to Support Radiologists' Classification Decisions in Mammography Practice. Med Decis Making. 2019;39(3):208-16.	Internal validation – Cross validation
270. Zhang C, Zhao J, Niu J, Li D. New convolutional neural network model for screening and diagnosis of mammograms. PLoS ONE. 2020;15(8):e0237674.	Internal validation – Cross validation
Internal validation – Leave-one out (n=12)	

Reference	Main reason for exclusion
271. Casti P, Mencattini A, Salmeri M, Rangayyan RM. Analysis of structural similarity in mammograms for detection of bilateral asymmetry. IEEE Trans Med Imaging. 2015;34(2):662-71.	Internal validation – Leave-one out
272. Dhahbi S, Barhoumi W, Zagrouba E. Breast cancer diagnosis in digitized mammograms using curvelet moments. Comput Biol Med. 2015;64:79-90.	Internal validation – Leave-one out
273. Drukker K, Duewer F, Giger ML, Malkov S, Flowers CI, Joe B, et al. Mammographic quantitative image analysis and biologic image composition for breast lesion characterization and classification. Med Phys. 2014;41(3):031915.	Internal validation – Leave-one out
274. Kelder A, Lederman D, Zheng B, Zigel Y. A new computer-aided detection approach based on analysis of local and global mammographic feature asymmetry. Med Phys. 2018;45(4):1459-70.	Internal validation – Leave-one out
275. Kendall EJ, Barnett MG, Chytyk-Praznik K. Automatic detection of anomalies in screening mammograms. BMC med. 2013;13:43.	Internal validation – Leave-one out
276. Kendall EJ, Flynn MT. Automated breast image classification using features from its discrete cosine transform. PLoS ONE. 2014;9(3):e91015.	Internal validation – Leave-one out
277. Liang C, Bian Z, Lv W, Chen S, Zeng D, Ma J. A computer-aided diagnosis scheme of breast lesion classification using GLGLM and shape features: Combined-view and multi-classifiers. Phys Med. 2018;55:61-72.	Internal validation – Leave-one out
278. Muramatsu C, Hara T, Endo T, Fujita H. Breast mass classification on mammograms using radial local ternary patterns. Comput Biol Med. 2016;72:43-53.	Internal validation – Leave-one out
279. Ramos-Pollan R, Guevara-Lopez MA, Suarez-Ortega C, Diaz-Herrero G, Franco-Valiente JM, Rubio-Del-Solar M, et al. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. J Med Syst. 2012;36(4):2259-69.	Internal validation – Leave-one out
280. Rangayyan RM, Nguyen TM, Ayres FJ, Nandi AK. Effect of pixel resolution on texture features of breast masses in mammograms. J Digit Imaging. 2010;23(5):547-53.	Internal validation – Leave-one out
281. Wang X, Lederman D, Tan J, Wang XH, Zheng B. Computerized detection of breast tissue asymmetry depicted on bilateral mammograms: a preliminary study of breast risk stratification. Acad Radiol. 2010;17(10):1234-41.	Internal validation – Leave-one out
282. Wang Y, Aghaei F, Zarafshani A, Qiu Y, Qian W, Zheng B. Computer-aided classification of mammographic masses using visually sensitive image features. Journal of X-Ray Science and Technology. 2017;25(1):171-86.	Internal validation – Leave-one out

Reference		Main reason for exclusion
Internal validation – Split sample (n=49)		
283.	Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening. Clin Cancer Res. 2018;24(23):5902-9.	Internal validation – Split sample
284.	Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, et al. Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms. Radiology. 2019;292(2):331-42.	Internal validation – Split sample
285.	Akselrod-Ballin A, Karlinsky L, Alpert S, Hashoul S, Ben-Ari R, Barkan E. A CNN based method for automatic mass detection and classification in mammograms. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization. 2019;7(3):242-9.	Internal validation – Split sample
286.	Alqudah AM, Algharib AMS, Algharib HMS. Computer aided diagnosis system for automatic two stages classification of breast mass in digital mammogram images. Biomedical Engineering - Applications, Basis and Communications. 2019;31(1).	Internal validation – Split sample
287.	Andreadis, II, Spyrou GM, Nikita KS. A CADx scheme for mammography empowered with topological information from clustered microcalcifications' atlases. IEEE j. 2015;19(1):166-73.	Internal validation – Split sample
288.	Arevalo J, Gonzalez FA, Ramos-Pollan R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Programs Biomed. 2016;127:248-57.	Internal validation – Split sample
289.	Bakkouri I, Afdel K. Multi-scale CNN based on region proposals for efficient breast abnormality recognition. Multimedia Tools and Applications. 2019;78(10):12939-60.	Internal validation – Split sample
290.	Banaem HY, Dehnavi AM, Shahnazi M. Ensemble Supervised Classification Method Using the Regions of Interest and Grey Level Co-Occurrence Matrices Features for Mammograms Data. Iranian Journal of Radiology. 2015;12(3).	Internal validation – Split sample
291.	Bandeira Diniz JO, Bandeira Diniz PH, Azevedo Valente TL, Correa Silva A, de Paiva AC, Gattass M. Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks. Comput Methods Programs Biomed. 2018;156:191-207.	Internal validation – Split sample

Reference		Main reason for exclusion
292.	Barkana BD, Saricicek I. Classification of breast masses in mammograms using 2D homomorphic transform features and supervised classifiers. Journal of Medical Imaging and Health Informatics. 2017;7(7):1566-71.	Internal validation – Split sample
293.	Beura S, Majhi B, Dash R. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing. 2015;154:1-14.	Internal validation – Split sample
294.	Bhardwaj A, Tiwari A. Breast cancer diagnosis using Genetically Optimized Neural Network model. Expert Systems with Applications. 2015;42(10):4611-20.	Internal validation – Split sample
295.	Boumaraf S, Liu X, Ferkous C, Ma X. A New Computer-Aided Diagnosis System with Modified Genetic Feature Selection for BI-RADS Classification of Breast Masses in Mammograms. Biomed Res Int. 2020;2020:7695207.	Internal validation – Split sample
296.	Cha KH, Petrick N, Pezeshk A, Graff CG, Sharma D, Badal A, et al. Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning. J Med Imaging (Bellingham). 2020;7(1).	Internal validation – Split sample
297.	Chinnasamy VA, Shashikumar DR. Breast cancer detection in mammogram image with segmentation of tumour region. International Journal of Medical Engineering and Informatics. 2020;12(1):1-18.	Internal validation – Split sample
298.	Choi JY, Kim DH, Plataniotis KN, Ro YM. Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography. Expert Systems with Applications. 2016;46:106-21.	Internal validation – Split sample
299.	Chu J, Min H, Liu L, Lu W. A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation. Med Phys. 2015;42(7):3859-69.	Internal validation – Split sample
300.	de Nazare Silva J, de Carvalho Filho AO, Correa Silva A, Cardoso de Paiva A, Gattass M. Automatic Detection of Masses in Mammograms Using Quality Threshold Clustering, Correlogram Function, and SVM. J Digit Imaging. 2015;28(3):323-37.	Internal validation – Split sample
301.	de Sampaio WB, Silva AC, de Paiva AC, Gattass M. Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, LBP and SVM. Expert Systems with Applications. 2015;42(22):8911-28.	Internal validation – Split sample

Reference	Main reason for exclusion
302. Dhas AS, Vijikala V. An improved CAD system for abnormal mammogram image classification using SVM with linear kernel. Biomedical Research (India). 2017;28(12):5499-505.	Internal validation – Split sample
303. Duggento A, Aiello M, Cavaliere C, Cascella GL, Cascella D, Conte G, et al. An Ad Hoc Random Initialization Deep Neural Network Architecture for Discriminating Malignant Breast Cancer Lesions in Mammographic Images. Contrast Media Mol Imaging. 2019;2019:5982834.	Internal validation – Split sample
304. Duraisamy S, Emperumal S. Computer-aided mammogram diagnosis system using deep learning convolutional fully complex-valued relaxation neural network classifier. Iet Computer Vision. 2017;11(8):656-62.	Internal validation – Split sample
305. Ferreira P, Fonseca NA, Dutra I, Woods R, Burnside E. Predicting malignancy from mammography findings and image-guided core biopsies. Int J Data Min Bioinform. 2015;11(3):257-76.	Internal validation – Split sample
306. Gao X, Wang Y, Li X, Tao D. On combining morphological component analysis and concentric morphology model for mammographic mass detection. IEEE Trans Inf Technol Biomed. 2010;14(2):266-73.	Internal validation – Split sample
307. Gastounioti A, Oustimov A, Hsieh MK, Pantalone L, Conant EF, Kontos D. Using Convolutional Neural Networks for Enhanced Capture of Breast Parenchymal Complexity Patterns Associated with Breast Cancer Risk. Acad Radiol. 2018;25(8):977-84.	Internal validation – Split sample
308. Hinton B, Ma L, Mahmoudzadeh AP, Malkov S, Fan B, Greenwood H, et al. Deep learning networks find unique mammographic differences in previous negative mammograms between interval and screen-detected cancers: a case-case study. Cancer Imaging. 2019;19(1):41.	Internal validation – Split sample
309. Ibrahim IM, Wahed MA. Visual versus statistical features selection applied to mammography mass detection. Journal of Medical Imaging and Health Informatics. 2014;4(2):237-44.	Internal validation – Split sample
310. Kim EK, Kim HE, Han K, Kang BJ, Sohn YM, Woo OH, et al. Applying Data-driven Imaging Biomarker in Mammography for Breast Cancer Screening: Preliminary Study. Sci. 2018;8(1):2762.	Internal validation – Split sample
311. Kooi T, Karssemeijer N. Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. J Med Imaging (Bellingham). 2017;4(4):044501.	Internal validation – Split sample

Reference		Main reason for exclusion
312.	Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017;35:303-12.	Internal validation – Split sample
313.	Lesniak JM, Hupse R, Blanc R, Karssemeijer N, Szekely G. Comparative evaluation of support vector machine classification for computer aided detection of breast masses in mammography. Phys Med Biol. 2012;57(16):5295-307.	Internal validation – Split sample
314.	Li H, Meng X, Wang T, Tang Y, Yin Y. Breast masses in mammography classification with local contour features. Biomed. 2017;16(1).	Internal validation – Split sample
315.	Liu B, Jiang Y. A multitarget training method for artificial neural network with application to computer-aided diagnosis. Med Phys. 2013;40(1):011908.	Internal validation – Split sample
316.	Mao N, Yin P, Wang Q, Liu M, Dong J, Zhang X, et al. Added Value of Radiomics on Mammography for Breast Cancer Diagnosis: A Feasibility Study. J. 2019;16(4 Pt A):485-91.	Internal validation – Split sample
317.	Memon MH, Li JP, Ul Haq A, Memon MH, Zhou W. Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection. Wireless Communications & Mobile Computing. 2019;2019.	Internal validation – Split sample
318.	Raj JR, Rahman SMK, Anand S. Preliminary evaluation of differentiation of benign and malignant breast tumors using non-invasive diagnostic modalities. Biomedical Research (India). 2016;27(3):596-603.	Internal validation – Split sample
319.	Ramos-Pollan R, Franco JM, Sevilla J, Guevara-Lopez MA, de Posada NG, Loureiro J, et al. Grid infrastructures for developing mammography CAD systems. Conf Proc IEEE Eng Med Biol Soc. 2010;2010:3467-70.	Internal validation – Split sample
320.	Sasikala S, Ezhilarasi M. Fusion of k-Gabor features from medio-lateral-oblique and craniocaudal view mammograms for improved breast cancer diagnosis. J Cancer Res Ther. 2018;14(5):1036-41.	Internal validation – Split sample
321.	Shankar RS, Gupta VM, Murthy KVSS, Rao CS. Breast cancer data classification using machine learning mechanisms. Indian Journal of Public Health Research and Development. 2019;10(5):214-20.	Internal validation – Split sample
322.	Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci. 2019;9(1):12495.	Internal validation – Split sample

Reference		Main reason for exclusion
323.	Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy Detection on Mammography Using Dual Deep Convolutional Neural Networks and Genetically Discovered False Color Input Enhancement. J Digit Imaging. 2017;30(4):499-505.	Internal validation – Split sample
324.	Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning. Sci. 2016;6:27327.	Internal validation – Split sample
325.	Wang Y, Shi H, Ma S. A new approach to the detection of lesions in mammography using fuzzy clustering. Journal of International Medical Research. 2011;39(6):2256-63.	Internal validation – Split sample
326.	Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging. 2020;39(4):1184-94.	Internal validation – Split sample
327.	Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. Radiology. 2019;293(1):38-46.	Internal validation – Split sample
328.	Zadeh Shirazi A, Seyyed Mahdavi Chabok SJ, Mohammadi Z. A novel and reliable computational intelligence system for breast cancer detection. Med Biol Eng Comput. 2018;56(5):721-32.	Internal validation – Split sample
329.	Zeiser FA, da Costa CA, Zonta T, Marques NMC, Roehe AV, Moreno M, et al. Segmentation of Masses on Mammograms Using Data Augmentation and Deep Learning. J Digit Imaging. 2020;23:23.	Internal validation – Split sample
330.	Zhang YD, Pan CC, Chen XQ, Wang FB. Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. Journal of Computational Science. 2018;27:57-68.	Internal validation – Split sample
331.	Zhou L, Ding M, Xu L, Zhou Y, Zhang X. Automated segmentation of malignant mass in mammography using the principal component analysis network based deep learning model. Journal of Medical Imaging and Health Informatics. 2018;8(8):1678-83.	Internal validation – Split sample
Intervention – Detecting subtypes (n=17)		
332.	Bekker AJ, Shalhon M, Greenspan H, Goldberger J. Multi-view probabilistic classification of breast microcalcifications. IEEE Trans Med Imaging. 2016;35(2):645-6536.	Intervention – Detecting subtypes
333.	Berks M, Chen Z, Astley S, Taylor C. Detecting and classifying linear structures in mammograms using random forests. Inf. 2011;22:510-24.	Intervention – Detecting subtypes

Reference	Main reason for exclusion
334. Devisuganya S, Suganthe RC. A wrapper based binary shuffled frog algorithm for efficient classification of mammograms. Current Signal Transduction Therapy. 2016;11(2):105-13.	Intervention – Detecting subtypes
335. Du GM, Dong M, Sun Y, Li SY, Mu XM, Wei HB, et al. A New Method for Detecting Architectural Distortion in Mammograms by NonSubsampled Contourlet Transform and Improved PCNN. Applied Sciences-Basel. 2019;9(22).	Intervention – Detecting subtypes
336. Huang ML, Hung YH, Lee WM, Li RK, Wang TH. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. J Med Syst. 2012;36(2):407-14.	Intervention – Detecting subtypes
337. Jing H, Yang Y, Nishikawa RM. Retrieval boosted computer-aided diagnosis of clustered microcalcifications for breast cancer. Med Phys. 2012;39(2):676-85.	Intervention – Detecting subtypes
338. Kamra A, Jain VK, Singh S, Mittal S. Characterization of Architectural Distortion in Mammograms Based on Texture Analysis Using Support Vector Machine Classifier with Clinical Evaluation. J Digit Imaging. 2016;29(1):104-14.	Intervention – Detecting subtypes
339. Keles A, Keles A, Yavuz U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. Expert Systems with Applications. 2011;38(5):5719-26.	Intervention – Detecting subtypes
340. Magna G, Casti P, Jayaraman SV, Salmeri M, Mencattini A, Martinelli E, et al. Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system. Knowledge-Based Systems. 2016;101:60-70.	Intervention – Detecting subtypes
341. Matsubara T, Ito A, Tsunomori A, Hara T, Muramatsu C, Endo T, et al. An automated method for detecting architectural distortions on mammograms using direction analysis of linear structures. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:2661-4.	Intervention – Detecting subtypes
342. Mordang JJ, Gubern-Merida A, Bria A, Tortorella F, den Heeten G, Karssemeijer N. Improving computer-aided detection assistance in breast cancer screening by removal of obviously false-positive findings. Med Phys. 2017;44(4):1390-401.	Intervention – Detecting subtypes
343. Mordang JJ, Gubern-Merida A, Den Heeten G, Karssemeijer N. Reducing false positives of microcalcification detection systems by removal of breast arterial calcifications. Med Phys. 2016;43(4):1676-87.	Intervention – Detecting subtypes
344. Scaranelo AM, Eiada R, Bukhanov K, Crystal P. Evaluation of breast amorphous calcifications by a computer-aided detection system in full-field digital mammography. Br J Radiol. 2012;85(1013):517-22.	Intervention – Detecting subtypes

Reference		Main reason for exclusion
345.	Shao YZ, Liu LZ, Bie MJ, Li CC, Wu YP, Xie XM, et al. Characterizing the Clustered Microcalcifications on Mammograms to Predict the Pathological Classification and Grading: A Mathematical Modeling Approach. J Digit Imaging. 2011;24(5):764-71.	Intervention – Detecting subtypes
346.	Tiedeu A, Daul C, Kentsop A, Graebling P, Wolf D. Texture-based analysis of clustered microcalcifications detected on mammograms. Digital Signal Processing. 2012;22(1):124-32.	Intervention – Detecting subtypes
347.	Wang X, Li L, Xu W, Liu W, Lederman D, Zheng B. Improving the performance of computer-aided detection of subtle breast masses using an adaptive cueing method. Phys Med Biol. 2012;57(2):561-75.	Intervention – Detecting subtypes
348.	Wang X, Li L, Xu W, Liu W, Lederman D, Zheng B. Improving performance of computer-aided detection of masses by incorporating bilateral mammographic density asymmetry: an assessment. Acad Radiol. 2012;19(3):303-10.	Intervention – Detecting subtypes
Intervention – No detection/classification (n=7)		
349.	Angayarkanni SP, Kamal NB, Thangaiya RJ. Dynamic graph cut based segmentation of mammogram. Springerplus. 2015;4.	Intervention – No detection/classification
350.	Dabagov AR, Gorbunov VA, Filist SA, Malyutina IA, Kondrashov DS. An Automated System for Classification of Radiographs of the Breast. Biomedical Engineering. 2020;53(6):425-8.	Intervention – No detection/classification
351.	James JJ, Giannotti E, Chen Y. Evaluation of a computer-aided detection (CAD)-enhanced 2D synthetic mammogram: comparison with standard synthetic 2D mammograms and conventional 2D digital mammography. Clin Radiol. 2018;73(10):886-92.	Intervention – No detection/classification
352.	Mayo RC, Kent D, Sen LC, Kapoor M, Leung JWT, Watanabe AT. Reduction of False-Positive Markings on Mammograms: a Retrospective Comparison Study Using an Artificial Intelligence-Based CAD. J Digit Imaging. 2019;32(4):618-24.	Intervention – No detection/classification
353.	Patel BC, Sinha GR. Abnormality detection and classification in computer-aided diagnosis (CAD) of breast cancer images. Journal of Medical Imaging and Health Informatics. 2014;4(6):881-885.	Intervention – No detection/classification
354.	Shen R, Yan K, Xiao F, Chang J, Jiang C, Zhou K. Automatic Pectoral Muscle Region Segmentation in Mammograms Using Genetic Algorithm and Morphological Selection. J Digit Imaging. 2018;31(5):680-91.	Intervention – No detection/classification

Reference		Main reason for exclusion
355.	Sujatha K, Shalini Punithavathani D, Mary Sowbaghya P. Model based non-rigid registration framework for high dynamic range mammography. WSEAS Transactions on Biology and Biomedicine. 2014;11(1):126-32.	Intervention – No detection/classification
Intervention – Not AI (“old” CAD) (n=16)		
356.	Bargallo X, Santamaria G, Del Amo M, Arguis P, Rios J, Grau J, et al. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. Eur J Radiol. 2014;83(11):2019-23.	Intervention – Not AI
357.	Bargallo X, Velasco M, Santamaria G, Del Amo M, Arguis P, Sanchez Gomez S. Role of computer-aided detection in very small screening detected invasive breast cancers. J Digit Imaging. 2013;26(3):572-7.	Intervention – Not AI
358.	Cole EB, Zhang Z, Marques HS, Hendrick RE, Yaffe MJ, Pisano ED. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. American Journal of Roentgenology. 2014;203(4):909-16.	Intervention – Not AI
359.	Fenton JJ, Xing G, Elmore JG, Bang H, Chen SL, Lindfors KK, et al. Short-term outcomes of screening mammography using computer-aided detection a population-based study of medicare enrollees. Ann Intern Med. 2013;158(8):580-7.	Intervention – Not AI
360.	Guerriero C, Gillan MG, Cairns J, Wallis MG, Gilbert FJ. Is computer aided detection (CAD) cost effective in screening mammography? A model based on the CADET II study. BMC Health Serv Res. 2011;11:11.	Intervention – Not AI
361.	Hupse R, Samulski M, Lobbes M, den Heeten A, Imhof-Tas MW, Beijerinck D, et al. Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. Eur Radiol. 2013;23(1):93-100.	Intervention – Not AI
362.	Hupse R, Samulski M, Lobbes MB, Mann RM, Mus R, den Heeten GJ, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology. 2013;266(1):123-9.	Intervention – Not AI
363.	Jung NY, Kang BJ, Kim HS, Cha ES, Lee JH, Park CS, et al. Who could benefit the most from using a computer-aided detection system in full-field digital mammography? World J Surg Oncol. 2014;12:168.	Intervention – Not AI
364.	Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med. 2015;175(11):1828-37.	Intervention – Not AI

Reference	Main reason for exclusion
365. Onega T, Aiello Bowles EJ, Miglioretti DL, Carney PA, Geller BM, Yankaskas BC, et al. Radiologists' perceptions of computer aided detection versus double reading for mammography interpretation. Acad Radiol. 2010;17(10):1217-26.	Intervention – Not AI
366. Romero C, Varela C, Munoz E, Almenar A, Pinto JM, Botella M. Impact on breast cancer diagnosis in a multidisciplinary unit after the incorporation of mammography digitalization and computer-aided detection systems. AJR Am J Roentgenol. 2011;197(6):1492-7.	Intervention – Not AI
367. Sato M, Kawai M, Nishino Y, Shibuya D, Ohuchi N, Ishibashi T. Cost-effectiveness analysis for breast cancer screening: Double reading versus single + CAD reading. Breast Cancer. 2014;21(5):532-41.	Intervention – Not AI
368. Singh S, Maxwell J, Baker JA, Nicholas JL, Lo JY. Computer-aided classification of breast masses: performance and interobserver variability of expert radiologists versus residents. Radiology. 2011;258(1):73-80.	Intervention – Not AI
369. Skaane P, Kshirsagar A, Hofvind S, Jahr G, Castellino RA. Mammography screening using independent double reading with consensus: is there a potential benefit for computer-aided detection? Acta Radiol. 2012;53(3):241-8.	Intervention – Not AI
370. Sohns C, Angic BC, Sossalla S, Konietschke F, Obenauer S. CAD in full-field digital mammography-influence of reader experience and application of CAD on interpretation of time. Clin Imaging. 2010;34(6):418-24.	Intervention – Not AI
371. Zheng B, Sumkin JH, Zuley ML, Lederman D, Wang X, Gur D. Computer-aided detection of breast masses depicted on full-field digital mammograms: a performance assessment. Br J Radiol. 2012;85(1014):e153-61.	Intervention – Not AI
Intervention – Prediction of cancer (n=2)	
372. Chen X, Moschidis E, Taylor C, Astley S. Breast cancer risk analysis based on a novel segmentation framework for digital mammograms. Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv. 2014;17(Pt 1):536-43.	Intervention – Prediction of cancer
373. Timmers JM, Verbeek AL, IntHout J, Pijnappel RM, Broeders MJ, den Heeten GJ. Breast cancer risk prediction model: a nomogram based on common mammographic screening findings. Eur Radiol. 2013;23(9):2413-9.	Intervention – Prediction of cancer
Outcomes – No relevant outcomes for Q1 or Q2 (n=24)	

Reference		Main reason for exclusion
374.	Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys. 2017;44(10):5162-71.	No relevant outcomes
375.	Benndorf M. Conditional non-independence of radiographic image features and the derivation of post-test probabilities - A mammography BI-RADS example. Radiography. 2012;18(3):201-5.	No relevant outcomes
376.	Benndorf M, Burnside ES, Herda C, Langer M, Kotter E. External validation of a publicly available computer assisted diagnostic tool for mammographic mass lesions with two high prevalence research datasets. Med Phys. 2015;42(8):4987-96.	No relevant outcomes
377.	Clancy K, Aboutalib S, Mohamed A, Sumkin J, Wu S. Deep Learning Pre-training Strategy for Mammogram Image Classification: an Evaluation Study. J Digit Imaging. 2020;30:30.	No relevant outcomes
378.	Cole EB, Zhang Z, Marques HS, Nishikawa RM, Hendrick RE, Yaffe MJ, et al. Assessing the stand-alone sensitivity of computer-aided detection with cancer cases from the Digital Mammographic Imaging Screening Trial. AJR Am J Roentgenol. 2012;199(3):W392-401.	No relevant outcomes
379.	Li Z, Yu L, Wang X, Yu H, Gao Y, Ren Y, et al. Diagnostic Performance of Mammographic Texture Analysis in the Differential Diagnosis of Benign and Malignant Breast Tumors. Clin Breast Cancer. 2018;18(4):e621-e7.	No relevant outcomes
380.	Lobbes M, Smidt M, Keymeulen K, Girometti R, Zuiani C, Beets-Tan R, et al. Malignant lesions on mammography: accuracy of two different computer-aided detection systems. Clin Imaging. 2013;37(2):283-8.	No relevant outcomes
381.	Mayo RC, Leung JWT. Impact of artificial intelligence on women's imaging: Cost-benefit analysis. American Journal of Roentgenology. 2019;212(5):1172-3.	No relevant outcomes
382.	Mendel K, Li H, Sheth D, Giger M. Transfer Learning From Convolutional Neural Networks for Computer-Aided Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography. Acad Radiol. 2019;26(6):735-43.	No relevant outcomes
383.	Murakami R, Kumita S, Tani H, Yoshida T, Sugizaki K, Kuwako T, et al. Detection of breast cancer with a computer-aided detection applied to full-field digital mammography. J Digit Imaging. 2013;26(4):768-73.	No relevant outcomes
384.	Oliver A, Llado X, Freixenet J, Marti R, Perez E, Pont J, et al. Influence of using manual or automatic breast density information in a mass detection CAD system. Acad Radiol. 2010;17(7):877-83.	No relevant outcomes

Reference		Main reason for exclusion
385.	Park CS, Jung NY, Kim K, Jung HS, Sohn KM, Oh SJ. Detection of breast cancer in asymptomatic and symptomatic groups using computer-aided detection with full-field digital mammography. Journal of Breast Cancer. 2013;16(3):322-8.	No relevant outcomes
386.	Punitha S, Ravi S, Devi MA, Vaishnavi J. Particle swarm optimized computer aided diagnosis system for classification of breast masses. Journal of Intelligent & Fuzzy Systems. 2017;32(4):2819-28.	No relevant outcomes
387.	Sadaf A, Crystal P, Scaranelo A, Helbich T. Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers. Eur J Radiol. 2011;77(3):457-61.	No relevant outcomes
388.	Sohns C, Angic B, Sossalla S, Konietschke F, Obenauer S. Computer-assisted diagnosis in full-field digital mammography--results in dependence of readers experiences. Breast J. 2010;16(5):490-7.	No relevant outcomes
389.	Torrents-Barrena J, Puig D, Melendez J, Valls A. Computer-aided diagnosis of breast cancer via Gabor wavelet bank and binary-class SVM in mammographic images. Journal of Experimental & Theoretical Artificial Intelligence. 2016;28(1-2):295-311.	No relevant outcomes
390.	van den Biggelaar FJ, Kessels AG, van Engelshoven JM, Boetes C, Flobbe K. Computer-aided detection in full-field digital mammography in a clinical population: performance of radiologist and technologists. Breast Cancer Res Treat. 2010;120(2):499-506.	No relevant outcomes
391.	Vedanarayanan V, Nandhitha NM. Advanced image segmentation techniques for accurate isolation of abnormality to enhance breast cancer detection in digital mammographs. Biomedical Research (India). 2017;28(6):2753-7.	No relevant outcomes
392.	Warren LM, Given-Wilson RM, Wallis MG, Cooke J, Halling-Brown MD, Mackenzie A, et al. The effect of image processing on the detection of cancers in digital mammography. AJR Am J Roentgenol. 2014;203(2):387-93.	No relevant outcomes
393.	Warren LM, Halling-Brown MD, Looney PT, Dance DR, Wallis MG, Given-Wilson RM, et al. Image processing can cause some malignant soft-tissue lesions to be missed in digital mammography images. Clin Radiol. 2017;72(9):799.e1-.e8.	No relevant outcomes
394.	Wu Y, Vanness DJ, Burnside ES. Using multidimensional mutual information to prioritize mammographic features for breast cancer diagnosis. AMIA Annu Symp Proc. 2013;2013:1534-43.	No relevant outcomes

Reference	Main reason for exclusion
395. Yang X, Cao A, Song Q, Schaefer G, Su Y. Vicinal support vector classifier using supervised kernel-based clustering. Artif Intell Med. 2014;60(3):189-96.	No relevant outcomes
396. Yu SD, Liu LL, Wang ZY, Dai GZ, Xie YQ. Transferring deep neural networks for the differentiation of mammographic breast lesions. Science China-Technological Sciences. 2019;62(3):441-7.	No relevant outcomes
397. Zheng K, Harris C, Bakic P, Makrogiannis S. Spatially localized sparse representations for breast lesion characterization. Computers in Biology and Medicine. 2020;123 (no pagination).	No relevant outcomes
Study type – Systematic reviews with no relevant outcomes for Q2 (n=7)	
398. Azavedo E, Zackrisson S, Mejare I, Heibert Arnlind M. Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. BMC med. 2012;12:22.	Study type – Systematic reviews with no relevant outcomes for Q2
399. Eadie LH, Taylor P, Gibson AP. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. Eur J Radiol. 2012;81(1):e70-6.	Study type – Systematic reviews with no relevant outcomes for Q2
400. Gruppo di studio G-S, Chersevani R, Ciatto S, Del Favero C, Frigerio A, Giordano L, et al. "CADEAT": considerations on the use of CAD (computer-aided diagnosis) in mammography. Radiol Med (Torino). 2010;115(4):563-70.	Study type – Systematic reviews with no relevant outcomes for Q2
401. Henriksen EL, Carlsen JF, Vejborg IM, Nielsen MB, Lauridsen CA. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. Acta Radiol. 2019;60(1):13-8.	Study type – Systematic reviews with no relevant outcomes for Q2
402. 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 Rev Med Devices. 2019;16(5):351-62.	Study type – Systematic reviews with no relevant outcomes for Q2
403. Sadoughi F, Kazemy Z, Hamedan F, Owji L, Rahmanikatigari M, Azadboni TT. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Breast Cancer (Dove Med Press). 2018;10:219-30.	Study type – Systematic reviews with no relevant outcomes for Q2
404. Yassin NIR, Omran S, El Houby EMF, Allam H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Comput Methods Programs Biomed. 2018;156:25-45.	Study type – Systematic reviews with no relevant outcomes for Q2
Full text not available via Document Supply (n=7)	

Reference		Main reason for exclusion
405.	Bhavani SR, Chilambuchelvan A, Senthilkumar J, Manjula D, Krishnamoorthy R, Kannan A. A secure cloud-based multi-agent intelligent system for mammogram image diagnosis. International Journal of Biomedical Engineering and Technology. 2018;28(2):185-202.	Document Supply cancelled request: no location found.
406.	Grout S, Dheeraj Suryaa SR, Hitesh, Venkatesan DH, Sumanth S, Vishnu Vardhan Reddy M. Anomaly detection in digital mammography using neural networks. Journal of International Pharmaceutical Research. 2019;46(3):750-4.	Document Supply cancelled request: no location found.
407.	Saraswathi D, Srinivasan E. An ensemble approach to diagnose breast cancer using fully complex-valued relaxation neural network classifier. International Journal of Biomedical Engineering and Technology. 2014;15(3):243-60.	Document Supply cancelled request: no location found.
408.	Selvan VP, Suganthi M. Clinical support system for classification of tumor in mammogram images using multiple features and neural network classifier. Journal of Pure and Applied Microbiology. 2015;9(Special Edition):253-61.	Document Supply cancelled request: no location found.
409.	Singh B, Jain VK, Singh S. Mammogram mass classification using support vector machine with texture, shape features and hierarchical centroid method. Journal of Medical Imaging and Health Informatics. 2014;4(5):687-96.	Document Supply cancelled request: no location found.
410.	Srivastava S, Sharma N, Singh SK, Srivastava R. Quantitative analysis of a general framework of a CAD tool for breast cancer detection from mammograms. Journal of Medical Imaging and Health Informatics. 2014;4(5):654-74.	Document Supply cancelled request: no location found.
411.	Zhou L, Ding M, Xu L, Zhou Y, Zhang X. The automatic segmentation of mammographic mass using the end-to-end convolutional network based on dense-prediction. Journal of Medical Imaging and Health Informatics. 2019;9(7):1429-34.	Document Supply cancelled request: no location found.
Other reasons (n=5)		
412.	Becker AS, Marcon M, Ghafoor S, et al. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. Invest Radiol 2017;52(7):434-40. doi: https://dx.doi.org/10.1097/RLI.0000000000000358	Study 1: BCDR database; unclear proportion of screening mammograms Study 2: Temporal validation
413.	da Silva R, de Carvalho A. Automatic classification of breast lesions usingTransfer Learning. Ieee Latin America Transactions. 2019;17(12):1964-9.	Language – Not available in English
414.	Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. The Lancet Digital Health 2020;2(3):e138-e48. doi: **http://dx.doi.org/10.1016/S2589-7500%2820%2930003-0**	Evaluation study: unclear proportion of screening mammograms.

Reference		Main reason for exclusion
		Reader study: In parts temporal validation confirmed by corresponding author via email.
415.	Polat K. Application of Attribute Weighting Method Based on Clustering Centers to Discrimination of Linearly Non-Separable Medical Datasets. J Med Syst. 2012;36(4):2657-73.	Separation of two different mage datasets (liver and breast)
416.	Sechopoulos I, Mann RM. Stand-alone artificial intelligence - The future of breast cancer screening? Breast. 2020;49:254-60.	Narrative review

Table 20 presents sub-studies of the 11 included articles (and the 2 excluded articles using temporal validation) that were excluded from the analysis with reasons for exclusion.

Table 20. Excluded sub-studies (studies / datasets) from review analyses with reasons for exclusion

Reference	**Excluded study / dataset and reason**
11 included articles using geographical validation	
Balta 2020[83]	None
Dembrower 2020[84]	None
Lang 2020[85]	None
McKinney 2020[76]	1) **Retrospective clinical comparison** with original decisions of UK and US readers, respectively, excluded due to internal validation test set (split sample). 2) **Comparison with reader study** excluded due to internal validation test set (split sample). 3) **Simulation study** excluded as it is based on test accuracy estimated obtained using internal validation test sets (split sample).
Pacilè 2020[77]	None
Rodriguez-Ruiz 2018[79]	Excluded **AI as stand-alone reader** due to lack of outcomes such as sensitivity and specificity (only AUC).
Rodriguez-Ruiz 2019[56]	Excluded **AI as stand-alone reader** due to lack of outcomes such as sensitivity and specificity (only AUC).
Rodriguez-Ruiz 2019[78]	Excluded **data sets A and D-H** as <90% screening mammograms or unclear proportion of screening mammograms. Excluded **data set B** as no relevant outcomes reported.
Salim 2020[80]	None
Schaffter 2020[81]	Excluded the **Kaiser Permanente Washington (KPW) dataset** as it was used for training and evaluation (split sample).
Watanabe 2019 [82]	None
2 excluded articles using temporal validation	
Becker 2017[97]	Excluded Study 1 (External test cohort) as the proportion of screening mammograms is unclear for the BCDR database.
Kim 2020[87]	Excluded the development dataset due to unknown proportion of diagnostic mammograms.

Appendix 3 — Summary and appraisal of individual studies

Data Extraction

Table 21. Summary table of study characteristics (geographical validation studies)

Reference and country	Study design	Study type	Population N (cancer prevalence), age in years	AI system	Comparator	AI role (envisaged) in screening pathway	Question
Studies with test accuracy outcomes							
Schaffter 2020,[81] Multinational	Retrospective test accuracy, not enriched, 2008-2012	Accuracy of a read, no pathway evaluated	68,008 women from Swedish screening cohort (1.1% cancer), mean age 53.3 (SD 9.4)	AI-1: Top-performing individual model (Therapixel, Paris, France), categorisation into cancer / no cancer at confidence level between 0 and 1 AI-2: Ensemble model of 8 top-performing individual algorithms, categorisation into cancer / no cancer at confidence level between 0 and 1 where sensitivity is the same as comparator AI-3: AI-2 with radiologists' assessment (recall or no recall) integrated (score=1 if women was recalled or 0 otherwise)	Original reader decision of (1) first reader, (2) double reading + consensus	Stand-alone AI (AI-2), AI + original reader decision (AI-3) (AI to replace reader 2 or all human readers)	Q1 + Q2

Reference and country	Study design	Study type	Population N (cancer prevalence), age in years	AI system	Comparator	AI role (envisaged) in screening pathway	Question
Salim 2020,[80] Sweden	Retrospective test accuracy study (case control), enriched, 2008-2015	Accuracy of a read, no pathway evaluated	8,805 women from Swedish screening cohort (8.4% cancer), median age 54.5 (interquartile range, 47.4-63.5)	Three (AI-1, AI-2, AI-3) commercially available anonymised AI systems yielding a prediction score between 0 and 1 for the suspicion of cancer	Original reader decision (double reading + consensus)	Stand-alone AI (AI to replace reader 2 or all human readers)	Q1 + Q2
McKinney 2020,[76] USA, UK	Retrospective test accuracy study, enriched, 2001-2018	Accuracy of a read, no pathway evaluated	3,097 women from 1 US academic medical centre (11.6% cancer within 12 months, 22.2% within 27 months), <40: 181 (5.9%) 40-49: 1,259 (40.8%) 50-59: 800 (26.1%) 60-69: 598 (19.0%) >=70: 259 (8.2%)	In-house AI system (ensemble of three models) reading full mammograms and classifying women into cancer / no cancer based on the mean cancer score between 0 and 1 of the predictions from the 3 independent models. The binary operating point was set using the validation set where the AI system achieved superiority for both sensitivity and specificity.	Original single reader decision in form of BI-RADS score (scores 0, 4, 5 were treated as positive)	Stand-alone AI (AI to replace reader 2 or all human readers)	Q1
Rodriguez-Ruiz 2019*,[78] Multinational	Enriched test set MRMC laboratory study 2003-2008	Accuracy of a read, no pathway evaluated	199 mammograms from digital screening pilot project conducted in Utrecht, Netherlands (39.7% cancer), age range 50-74	Transpara v 1.4.0 (Screenpoint Medical BV, Nijmegen, the Netherlands) providing a continuous score ranging between 1 and 10 representing the level of suspicion of cancer (threshold 8.26)	Seven single reader decisions	Stand-alone AI (AI for pre-screening or replacing human reader)	Q1

Reference and country	Study design	Study type	Population N (cancer prevalence), age in years	AI system	Comparator	AI role (envisaged) in screening pathway	Question
Rodriguez-Ruiz 2018[79] and 2019[56] Netherlands, USA, Germany	Enriched test set MRMC laboratory study, 2013-2017	Accuracy of a read, no pathway evaluated	240 women from one US and one German centre (41.7% cancer), median age 62 (range 39-89)	14 radiologists reading Transpara (version 1.3.0, ScreenPoint Medical, Nijmegen, the Netherlands) using the interactive decision support which provides a level of suspicion (on a scale of 1 to 100) for the area clicked with AI system integrated into reading workstation (7 radiologists) or AI system installed on a different screen (7 radiologists).	The same 14 radiologists reading without AI support providing a BI-RADS score and probability of malignancy (BI-RADS category 3 used as recall threshold)	AI as reader aid	Q1 + Q2
Pacilè 2020,[77] France, USA	Enriched test set MRMC laboratory study, 2013-2016	Accuracy of a read, no pathway evaluated	240 women from 1 US centre (50% cancer), mean age 59 (range 37-85)	14 radiologists reading MammoScreen V1 (Therapixel, Nice, France) reporting image positions with a related suspicion score for suspicion of breast cancer	The same 14 radiologists without AI	AI as reader aid	Q1 + Q2

Reference and country	Study design	Study type	Population N (cancer prevalence), age in years	AI system	Comparator	AI role (envisaged) in screening pathway	Question
Watanabe 2019,[82] USA	Enriched test set MRMC laboratory study, 2009-2016	Accuracy of a read, no pathway evaluated	122 women from 1 US community health centre (73.8% cancer), mean age 65.4 (range 40-90)	7 radiologists reading cmAssist^{TM} (CureMetrix, Inc., La Jolla, CA, US) which provided markings and their corresponding quantitative scores (neuScore™, scale of 0–100)	The same 7 radiologists without AI making a decision on recall prior to reading with AI	AI as reader aid	Q1 + Q2
Studies without test accuracy outcomes							
Balta 2020,[83] Netherlands, Germany	Retrospective cohort, not enriched, January to November 2018	Not test accuracy (simulation of effect of AI-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload)	18,015 women from 1 German Breast Diagnostic Centre (0.64% cancer), age NR	Transpara 1.6.0 (Screenpoint Medical BV, Nijmegen, Netherlands) providing a continuous score ranging between 1 and 10 (Transpara scores 1 – 10 were used as thresholds)	6 radiologists as single readers / independent double reading by two radiologists with consensus	Stand-alone AI (AI triages to single or double reading)	Q2

Reference and country	Study design	Study type	Population N (cancer prevalence), age in years	AI system	Comparator	AI role (envisaged) in screening pathway	Question
Dembrower 2020,[84] Sweden	Retrospective case control, enriched, 2009-2015	Not test accuracy (simulation of effect of AI-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload)	7,364 women from 1 Swedish screening centre, (7.4% cancer), median age 53.6 years (IQR 47.6-63.0)	Commercial AI system (Lunit, Seoul, South Korea) version 5.5.0.16 which provides a cancer detection score in form of a decimal number between 0 and 1 (each decile was used as threshold)	Screening pathway without AI	Stand-alone AI (AI for pre-screening or post-screen of negatives after radiological assessment)	Q2
Lang 2020,[85] Sweden	Retrospective cohort, not enriched, 2012 to 2015	Not test accuracy (effect of AI system identifying normal screening mammograms on radiologist and cancer detection)	9,581 women from 1 Swedish breast screening centre (0.71% cancer), mean age 57.6 (range 40-74)	Transpara v.1.4.0 (ScreenPoint Medical BV, Nijmegen, Netherlands) providing a continuous score ranging between 1 and 10 (low risk 1-5)	Screening pathway without AI	Stand-alone AI (AI for pre-screening)	Q2

AI artificial intelligence, BI-RADS Breast Imaging-Reporting and Data System, SD standard deviation.

*Only 1/9 datasets in the study (Hupse et al. 2013[96]) met the inclusion criteria (screening mammograms) for this review (see **Table 2**).

Table 22. Summary table of study characteristics (excluded studies using temporal validation)

Reference and country	Study design	Study type	Population N (cancer prevalence), age in years	AI system	Comparator	AI role (envisaged) in screening pathway	Question
Becker 2017,[97] Switzerland	Enriched test set MRMC laboratory study, October - December 2012 (trained on data from January to September)	Accuracy of a read, no pathway evaluated	251 women screened in 1 Swiss hospital (7.2% cancer), Mean age (n=1,146 eligible women) Cancer: 59.6 (SD 11.7, range 35-88) Non-cancer: 56.6 (SD 9.3, range 32-85)	ViDi Suite Version 2.0 (ViDi Systems Inc, Villaz-Saint-Pierre, Switzerland) producing a score from 0 to 1 for the whole image and a heat map overlay with suspicious anomalies highlighted. The threshold for binary operating point was based on the Youden index.	3 radiologists rating the images on a 5-point Likert-type scale for malignancy (single reader)	Stand-alone AI (AI to replace reader 2 or all human readers) or reader aid	Q1 + Q2
Kim 2020b,[87] South Korea	Enriched test set MRMC laboratory study, 2009-2018 *	Accuracy of a read, no pathway evaluated	320 mammograms from two hospitals in South Korea (50% cancer), mean age 53.19 (SD 10.01)	In-house (Lunit, Seoul, South Korea) system which provides a per breast abnormality score between 0 and 1. The binary operating point was set to 0·1 which achieved 90% sensitivity in the tuning dataset.	A reader representative score: a cancer-positive case was deemed correctly detected by readers if more than half of the 14 readers identified it correctly	Stand-alone AI (AI to replace reader 2 or all human readers)	Q1 + Q2
Kim 2020a,[87] South Korea	Enriched test set MRMC laboratory study, 2009-2018*	Accuracy of a read, no pathway evaluated	320 mammograms from two hospitals in South Korea (50% cancer), mean age 53.19 (SD 10.01)	14 radiologists reading in-house (Lunit, Seoul, South Korea) system which provides pixel-level abnormality scores as a heat map and an per breast abnormality score between 0 and 1	The same 14 radiologists (single read) from different institutions than the datasets read mammograms unaided (recall decision) prior to AI-aided read	AI as a reader aid	Q1 + Q2

*Refer to text for an explanation of temporal validation in this study.

Appraisal for quality and risk of bias

Quality assessments of included studies are reported below.

Table 23. Quality assessment of 7 included studies and 2 excluded, temporal validation studies, for question 1.

Study	Risk of bias					Applicability concerns			
	Patient selection	Index test	Comparator	Reference standard	Flow and timing	Patient selection	Index test	Comparator	Reference standard
Geographical validation studies (n=7)									
McKinney 2020 (Geographical validation study)[76]	High	High	Low	High	High	High	High	High	High
Pacilè 2020[77]	High	High	High	High	Unclear	High	High	High	High
Rodriguez-Ruiz 2019a[78] (stand-alone)	High	High	High	Unclear	Unclear	High	High	High	High
Rodriguez-Ruiz 2019b[56] (reader aid)	High	High	High	High	Unclear	High	High	High	High
Salim 2020[80]	High	High (for all 4 index tests)	Low (for both comparators)	High	High	High	High (for all 4 index tests)	High* / Low**	High
Schaffter 2020[81]	Low	High (for all 4 index tests)	Low (for both comparators)	High	High	Unclear	High (for all 4 index tests)	High* / Low**	High
Watanabe 2019[82]	High	High	High	Unclear	Unclear	High	High	High	High
Temporal validation studies (n=2)									
Becker 2017[97]	High	High	High	Unclear	Unclear	High	High	High	High
Kim 2020[87]	High	High (for both index tests)	High	High	Unclear	High	High (for both index tests)	High	High

* Original single reader; ** Original consensus reading.

Appendix 4 – Abstract reporting tables

Question 2

TITLE	
Citation	*Balta 2020*[83]
BACKGROUND	
Study type	*Non-enriched, retrospective cohort study*
Objectives	*We investigated whether a deep learning-based artificial intelligence (AI) system can be used to improve breast cancer screening workflow efficiency by making a pre-selection of likely normal screening mammograms where double-reading could be safely replaced with single-reading.*
Components of the study	*Population:* *18,015 consecutive screening FFDM exams acquired between January and November 2018 at a single German institution (Breast Diagnostic Centre, Munich Reference Centre, Germany).* *Intervention:* *Commercially available AI algorithm (Transpara™ 1.6.0, Screenpoint Medical BV, Nijmegen, Netherlands);* *AI system assigned a 1-10 score to each screening exam denoting the likelihood of cancer.* *Partial double screening strategy: AI pre-selects likely normal screening mammograms where double-reading could be safely replaced with single-reading.* *Comparator:* *Original reading decisions from double reading with consensus.* *6 radiologists from a single centre in Germany.* *Reference standard:* *Cancer: Biopsy-proven, screen-detected cancer* *No cancer: No follow-up of screen-negatives.* *Outcomes:* *Cases sent to consensus, recall rate, cancer detection rate (sensitivity for screen-detected cancers), workload (defined as the number of mammogram readings performed by reader 1 and reader 2) and PPV of screening.*
OUTCOMES	
Outcomes reported	*After evaluating all possible AI score thresholds, it was found that when AI scores 1 to 7 are single read instead of double read, the* ***cancer detection rate*** *would have remained the same (no screen-detected cancers missed – the AI score is low but the single-reader would recall the exam),* ***recall rate***

	*would have decreased by 11.8% (from 5.35% to 4.79%), and **screen reading workload** would have decreased by 32.6%.* *It has been necessary to consult the full text to identify the following relevant data:* *The **PPV of screening** would have increased significantly (p<0.0001) by 10.54% (11.90% to 13.30%), and the number of **cases going through consensus** would have decreased significantly (p<0.0001) by 20.79% (2400 to 1987).*
Conclusions	*In conclusion, using an AI system could improve breast cancer screening efficiency by pre-selecting likely normal exams where double-reading might not be needed.*

TITLE	
Citation	*Dembrower 2020*[84]
BACKGROUND	
Study type	*Retrospective simulation study*
Objectives	*To examine triaging based on two complementary roles for a commercially available AI cancer detector: as the first and only reader to dismiss the majority of normal mammograms (no radiologist work stream), and as the final reader after a negative examination to identify women at highest risk of undetected cancer (enhanced assessment work stream).*
Components of the study	*Population:* *All 547 women diagnosed with breast cancer (200 interval cancers and 347 screen-detected cancers at the latest screening round) and 6,817 randomly chosen healthy women who attended 2 consecutive screening rounds within 2.5 years from the Cohort of Screen-Aged Women;* *Karolinska University Hospital uptake area (Stockholm, Sweden) examined with FFDM (Hologic, Marlborough, MA, USA) between Feb 10, 2009, and Dec 10, 2015.* *The simulated screening population (11-times up-sampling of healthy women) contained 75,534 women, resulting in 0.74% cancer incidence over 1 screening interval.* *Intervention:* *Commercially available AI algorithm (Lunit, Seoul, South Korea, version 5.5.0.16).* *Based on the continuous prediction score from the AI cancer detector algorithm between 0 and 1, various cutoff points for the decision to channel women to the 2 new work streams were examined:* *1) triage certain screening examinations into a* ***no radiologist work stream****, and* *2) then after regular radiologist assessment of the remainder, triage certain screening examinations into an* ***enhanced assessment work stream (e.g. ultrasound, MRI)****.* *Comparator:* *Original double reading and consensus decisions (1 centre, Sweden).* *Reference standard:* *Cancer: NR* *No cancer: ≥ 2 years follow-up, 2 consecutive screening rounds within 2.5 years.* *Outcomes:*

	Cancer detection (interval cancers and screen-detected cancers) and radiologist workload.
OUTCOMES	
Outcomes reported	*Latest screening round:* *When including 60%, 70%, or 80% of women with the lowest AI scores in the <u>no radiologist stream</u>, the proportion of **screen-detected cancers** that would have been missed were 0, 0.3% (95% CI 0.0–4.3), or 2.6% (95% CI 1.1–5.4), respectively.* *Second latest screening round:* *When including 1% or 5% of women with the highest AI scores in the <u>enhanced assessment stream</u>, the potential additional cancer detection was 24 (12%) or 53 (27%) of 200 **subsequent interval cancers**, respectively, and 48 (14%) or 121 (35%) of 347 **next-round screen-detected cancers**, respectively.* *It has been necessary to consult the full text to identify the following relevant data:* ***Potential net change in cancer detection:*** *If no radiologist resources were used for 90% of women with the lowest AI scores and were invested into doing MRI for the top 2% AI scores (that were negative after radiologist double reading of the mammograms), a net of 89 of 547 cancers would potentially have been detected up to 2 years earlier, corresponding to a detection rate of 59 cancers per 1000 supplemental screening examinations.*
Conclusions	*Using a commercial AI cancer detector to triage mammograms into no radiologist assessment and enhanced assessment could potentially reduce radiologist workload by more than half, and pre-emptively detect substantial proportion of cancers otherwise diagnosed later.*

TITLE	
Citation	*Lang 2020[85]*
BACKGROUND	
Study type	*Non-enriched, retrospective cohort study*
Objectives	*The aim of this study was to evaluate the potential of a commercially available AI system to identify normal mammograms in a breast cancer screening population, thereby reducing workload related to the radiologists' screen-reading and false positives. In addition, the characteristics of screen-detected cancers that were missed by the AI system were assessed.*
Components of the study	*Population:* *Consecutive subcohort of the Malmö Breast Tomosynthesis Screening Trial; 9,581 non-pregnant women between 40–74 years attending national breast cancer screening (FFDM) at Skåne University Hospital, Malmö, Sweden;* *68 screen-detected cancers; 187 false-positive recalls.* *Intervention:* *Commercially available AI algorithm (Transpara v.1.4.0, Screenpoint Medical BV, Nijmegen, Netherlands) as pre-screen to identify normal mammograms in a breast cancer screening population that do not need human reading.* *Comparator:* *Human double reading (1 centre, Sweden).* *Reference standard:* *Cancer: Histology of surgical specimen or core-needle biopsies and with a cross-reference to a regional cancer register.* *No cancer: A normal mammogram was defined as free of screen-detected cancer.* *Outcomes:* *Reduction of mammography screening exams that would need human double reading, screen-detected and missed cancers, and false positives avoided for the different thresholds.*
OUTCOMES	
Outcomes reported	*If mammograms scored 1 and 2 were excluded from screen-reading, 1829 (19.1%; 95% CI 18.3–19.9) exams could be removed, including 10 (5.3%; 95% CI 2.1–8.6) false positives but no cancers. In total, 5082 (53.0%; 95% CI 52.0–54.0) exams, including 7 (10.3%; 95% CI 3.1–17.5) cancers and 52 (27.8%; 95% CI 21.4–34.2) false positives, had low-risk scores. All, except one, of the seven screen-detected cancers with low-risk scores were judged to be clearly visible.* *All seven cancers with low-risk scores were invasive, of which three were small (≤ 7 mm), low-grade invasive tubular carcinomas, i.e. tumours with excellent prognosis. On the other hand, three cancers, two ductal and one lobular type,*

	were large (20 mm), one of which was histologic grade 3, i.e. of less-favourable prognosis.
Conclusions	*The evaluated AI system can correctly identify a proportion of a screening population as cancer-free and also reduce false positives. Thus, AI has the potential to improve mammography screening efficiency.*

TITLE	
Citation	*Pacilè 2020*[77]
BACKGROUND	
Study type	*Enriched test set MRMC laboratory study,* *14 reader participants read cases over 2 reading sessions without and with AI as reader-aid separated by a washout period of 4 weeks (counterbalance design).*
Objectives	*To evaluate the benefits of an artificial intelligence (AI)–based tool for two-dimensional mammography in the breast cancer detection process.*
Components of the study	*Population:* *Enriched dataset: 240 women from 1 centre in the USA* *(120 cancer, 120 non-cancer), FFDM acquired between 2013 and 2016.* *Intervention:* *AI (MammoScreen V1, Therapixel, Nice, France) as reader-aid* *used by 14 American Board of Radiology and MQSA certified radiologists.* *Comparator:* *14 American Board of Radiology and MQSA-certified radiologists without AI reader aid.* *Reference standard:* *Cancer: Histopathologic evaluation.* *No cancer: Negative result at follow-up of 18 months.* *Outcomes:* *Reading time.*
OUTCOMES	
Outcomes reported	*Reading time changed dependently to the AI-tool score.* *For low likelihood of malignancy (<2.5%), the time was about the same in the first reading session and slightly decreased in the second reading session.* *For higher likelihood of malignancy, the reading time was on average increased with the use of AI.* *It has been necessary to consult the full text to identify the following relevant data:* *The learning curve observed between the first and the second session, together with the fact that the maximum increment of time did not exceed 15 seconds, suggested that the introduction of this tool into screening programs may not prolong the workflow of the radiologists and possibly even lead to a shorter average reading time.*
Conclusions	*This clinical investigation demonstrated that the concurrent use of this AI tool improved the diagnostic performance of radiologists in the detection of breast cancer without prolonging their workflow.*

TITLE	
Citation	*Rodriguez-Ruiz 2018[79] (preliminary data after 7 readers)* *Rodriguez-Ruiz 2019b[56]*
BACKGROUND	
Study type	*Enriched test set MRMC laboratory study*
Objectives	*To compare breast cancer detection performance of radiologists reading mammographic examinations unaided versus supported by an artificial intelligence (AI) system.*
Components of the study	*Population:* *Screening digital mammographic examinations from 240 women performed between 2013 and 2017 at 2 centres (Center A: USA, Center B: Germany) were included (100 showing cancers, 40 leading to false-positive recalls, 100 normal).* *Intervention:* *Transpara (version 1.3.0, ScreenPoint Medical, Nijmegen, the Netherlands) as reader aid.* *14 MQSA-qualified radiologists with AI support.* *AI support provided radiologists with interactive decision support (clicking on a breast region yields a local cancer likelihood score), traditional lesion markers for computer-detected abnormalities, and an examination-based cancer likelihood score.* *Comparator:* *14 MQSA-qualified radiologists without AI support.* *Reference standard:* *Cancer: Cancers were verified by means of histopathologic evaluation.* *No cancer: False-positive findings were verified with histopathologic evaluation (n = 11) or with negative follow-up findings for at least 1 year (n = 29).* *All normal examinations had at least 1 year of negative follow-up findings.* *Outcomes:* *Reading time.*
OUTCOMES	
Outcomes reported	*Reading time per case was similar (unaided, 146 seconds; supported by AI, 149 seconds; p=0.15).* *It has been necessary to consult the full text to identify the following relevant data:*

	Reading unaided and with AI support differed as a function of the computer Transpara score (p<0.001). For the low-suspicion examinations (score, 1–5), radiologists decreased their average reading time per case by 11% when using the AI system. Conversely, reading time per case was 2% higher with use of AI support for the high-suspicion examinations (score, 6–10). *Given the high workload of screening programs, from a cost-effectiveness point of view the performance benefit of using AI support is further enhanced by the fact that radiologists do not lengthen their reading time when using this system. In fact, in a real screening scenario, the average reading time per case would actually decrease by approximately 4.5%.*
Conclusions	*Radiologists improved their cancer detection at mammography when using an artificial intelligence system for support, without requiring additional reading time.*

TITLE	
Citation	*Salim 2020*[80]
BACKGROUND	
Study type	*Enriched, retrospective case-control study.*
Objectives	*To perform an external evaluation of 3 commercially available artificial intelligence (AI) computer-aided detection algorithms as independent mammography readers and to assess the screening performance when combined with radiologists.*
Components of the study	*Population:* *8,805 women (739 cancer, 8.4%) between 40 and 74 years from Swedish Cohort of Screen-Age Women, screened with FFDM at 1 academic hospital in Stockholm (Sweden) from 2008 to 2015.* *618 (84%) screen-detected cancer cases and 121 (16%) interval cancers within 12 months of the screening examination and a random sample of 8,066 healthy controls.* *Intervention:* *3 commercial AI systems yielding a prediction score for each breast ranging between 0 and 1.* *Comparator:* *Retrospective comparison to original reader decision (double reading with consensus).* *25 different first-reader radiologists and 20 different second-reader radiologists from 1 centre in Sweden.* *Reference standard:* *Cancer: Pathology-confirmed diagnosis at screening or within 12 months (secondary analysis 23 months) thereafter.* *No cancer: Healthy women with at least 2-year cancer-free follow-up.* *Outcomes:* *Types of cancer detected; AUC for interval cancers within 12 months of screening; simulated effect of a combination of AI and readers on abnormal interpretations and cancer detection.*
OUTCOMES	
Outcomes reported	*It has been necessary to consult the full text to identify the following relevant data:* ***Spectrum of disease detected*** *<u>In situ cancers</u> (n=85; 78 screen-detected and 7 interval cancers):* *AI-1: 71 (83.5%) First reader: 76 (89.4%)* *AI-2: 65 (76.5%) Second reader: 76 (89.4%)* *AI-3: 65 (76.5%) First and second reader: 80 (94.1%)*

	<u>*Invasive cancers*</u> *(n=640; 534 screen-detected and 106 interval cancers)* *AI-1: 530 (82.8%) First reader: 491 (76.7%)* *AI-2: 426 (66.6%) Second reader: 510 (79.7%)* *AI-3: 431 (67.3%) First and second reader: 553 (86.4%)* <u>*Invasive cancers (Stage 2 or higher)*</u> *(n=204; 142 screen-detected and 62 interval cancers)* *AI-1: 160 (78.4%) First reader: 139 (68.1%)* *AI-2: 119 (58.3%) Second reader: 139 (68.1%)* *AI-3: 124 (60.8%) First and second reader: 153 (75.0%)* ***Interval cancer within 12 months after negative radiologist assessment:*** *AI-1: AUC 0.810 (95% CI, 0.767-0.852)* *AI-2: AUC 0.728 (95% CI, 0.677-0.779)* *AI-3: AUC 0.744 (95% CI, 0.696-0.792)* *AI-1 achieved an AUC of 0.810, suggesting that there is potential for the AI algorithms to promote earlier cancer detection and that there are suspicious findings present in many of those mammograms.* ***Simulated combination of AI with readers:*** *For the first reader, the relative increase in cancer detection was 15% when adding AI-1 and 12% when adding the second reader; the relative increase in abnormal interpretations was 78% when adding AI-1, and 24% when adding the second reader.*
Conclusions	*To our knowledge, this study is the first independent evaluation of several AI computer-aided detection algorithms for screening mammography. The results of this study indicated that a commercially available AI computer-aided detection algorithm can assess screening mammograms with a sufficient diagnostic performance to be further evaluated as an independent reader in prospective clinical trials. Combining the first readers with the best algorithm identified more cases positive for cancer than combining the first readers with second readers.*

TITLE	
Citation	*Schaffter 2020*[81]
BACKGROUND	
Study type	*Non-enriched, retrospective cohort study*
Objectives	*To evaluate whether AI can overcome human mammography interpretation limitations with a rigorous, unbiased evaluation of machine learning algorithms.*
Components of the study	*Population:* *68,008 screening examinations from Karolinska Institute (Stockholm, Sweden) performed between April 2008 and December 2012.* *780 (1.1%) diagnosed with cancer within 12 months of mammogram.* *Intervention:* *DREAM challenge;* *An ensemble method aggregating top-performing AI algorithms and consensus radiologists' recall assessments (CEM+R).* *Comparison:* *Retrospective comparison to original reader decision (double reading with consensus, Sweden).* *Reference standard:* *Cancer: Diagnosed with breast cancer in the left/right breast (confirmed with tissue diagnosis) within 12 months of the given screening mammography exam.* *No cancer: No known diagnosis of cancer in the left/right breast on review of medical records one or more years after the screening exam.* *Outcomes:* *Types of cancer detected and recall rate.*
OUTCOMES	
Outcomes reported	*It has been necessary to consult the full text to identify the following relevant data.* ***Recall rate*** *Our study suggests that a collaboration between radiologists and an ensemble algorithm may reduce the recall rate from 0.095 to 0.08, an absolute 1.5% reduction. Considering that approximately 40 million women are screened for breast cancer in the United States each year, this would result in more than half a million women annually who would not have to undergo unnecessary diagnostic work-up.*
Conclusions	*While no single AI algorithm outperformed radiologists, an ensemble of AI algorithms combined with radiologist assessment in a single-reader screening environment improved overall accuracy. This study*

	underscores the potential of using machine learning methods for enhancing mammography screening interpretation.

TITLE	
Citation	*Watanabe 2019[82]*
BACKGROUND	
Study type	*Enriched test set MRMC laboratory study;* *7 radiologists read all mammograms first without then with AI aid.*
Objectives	*To determine whether cmAssist™, an artificial intelligence-based computer-aided detection (AI-CAD) algorithm, can be used to* *improve radiologists' sensitivity in breast cancer screening and detection.*
Components of the study	*Population:* *122 women with screening FFDM performed at 1 community healthcare facility in Southern California between February 7, 2008 (earliest) and January 8, 2016 (latest) that were all originally interpreted as negative in conjunction with R2 ImageChecker CAD, version 10.0.* *All 90 false-negative cases that were missed by their original interpreting radiologists for up to 5.8 years, 32 normal cases.* *Intervention:* *Commercially available cmAssist™ (CureMetrix, Inc., La Jolla, CA). AI as reader aid: 7 American Board of Radiology and MQSA certified radiologists were provided with the cmAssist markings and their corresponding quantitative scores (neuScore™, scale of 0–100).* *Comparator:* *7 American Board of Radiology and MQSA certified radiologists without AI support.* *Reference standard:* *Cancer: Biopsy-proven cancer; the earliest actionable prior mammograms were obtained between 0.76 and 5.8 years (mean, 2.1 years) prior to the current mammogram that eventually resulted in recall and workup for breast cancer.* *No cancer: BI-RADS 1 and 2 women with a 2-year follow-up of negative diagnosis.* *Outcomes:* *Additional cancers detected (by type); reader's acceptance of AI.*
OUTCOMES	
Outcomes reported	*It has been necessary to consult the full text to identify the following relevant data:* ***Additional cancers detected (by type)*** *<u>Calcifications</u> (=Microcalcifications as the leading lesion type, n=17)*

	With AI-CAD assistance, the 7 readers recalled an average of 3.4 additional cancerous calcifications but ignored on average 6.1 flagged malignant calcification cases. *Masses (all remaining cases without microcalcifications as the leading lesion type, such as focal asymmetry or mass with micro-calcifications; n=73)* *With AI-CAD assistance, readers recalled an average of 6.4 additional cases of malignant masses but ignored on average 11.4 flagged malignant mass cases.* ***Reader acceptance of AI*** *It is noted that all readers in this study appeared to ignore relatively significant number of flagged actionable lesions that would have improved their sensitivity even further. This suggests that even further improvement in reader accuracy and cancer detection rate could occur as radiologists gain experience in using cmAssist and develop more confidence in its markings and use of the neuScore (quantitative probability of malignancy calculated by cmAssist).*
Conclusions	*With the use of cmAssist^{TM}, there was a substantial and statistically significant improvement in radiologists' accuracy and sensitivity for detection of cancers that were originally missed. The percentage increase in cancer detection rate for the radiologists in the reader panel ranged from 6 to 64% (mean 27%) with the use of cmAssist, with negligible increase in false-positive recalls.*

Appendix 5. QUADAS-2 tailored for AI technologies

First author surname and year of publication:

Name of first reviewer: *Name of second reviewer:*

Phase 1: State the review question:

Question 1) What is the accuracy of AI algorithms to detect breast cancer in women attending screening mammography?

Patients (setting, intended use of index test, presentation, prior testing): *Women attending routine breast screening for digital (full field digital mammography) mammograms. Women may have attended previous screening rounds. Population risk screening only, no high risk screening.*
Index test(s) (including human comparators): *Combination of AI with readers (radiologist or equivalent) in any configuration to make a decision based on the mammograms whether to recall the women for further tests. For example, AI and a single radiologist independently deciding whether to recall for further tests, with arbitration of discordant decisions, or AI independently classifying the mammogram as normal or abnormal.* *Comparator: Screening pathway / read without AI (single or double human readers with or without "old" CAD)*
Reference standard and target condition: *Target condition is breast cancer confirmed by histology on biopsy tissue.*

Phase 2: Draw a flow diagram for the primary study

Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

DOMAIN 1: PATIENT SELECTION

A. Risk of Bias

Describe methods of patient selection:

+ Was a consecutive or random sample of patients enrolled?	Yes/No/Unclear
+ Was a case-control design avoided?	Yes/No/Unclear
+ Did the study avoid inappropriate exclusions?	Yes/No/Unclear
+ Were the women and mammograms included in the study independent of those used to train the AI algorithm?	Yes/No/Unclear

Could the selection of patients have introduced bias? **RISK: LOW/HIGH/UNCLEAR**

(Score HIGH if 'no' to any question.)

B. Concerns regarding applicability

Describe included patients (prior testing, presentation, intended use of index test and setting)**:**

Is there concern that the included patients do not match the review question? **CONCERN: LOW/HIGH/UNCLEAR**

High concerns if:
Not a consecutive or random sample of women attending screening;
Enriched sample / cancer prevalence doesn't match screening context (>3%);
Mammograms not from full-field digital mammography;
Mammograms not from screening (e.g. diagnostic or symptomatic) or only subset such as recalled cases or false-negatives included (cancer might be easier or more difficult to detect);
Women/women's mammograms not representative of UK population (ethnicity, age);
Recall rate in original screening population higher than UK recall rate (3.8%)

Was a consecutive or random sample of patients enrolled? Was a case-control design avoided?

RCTs and cohort studies (prospective or retrospective) with unenriched (consecutive or random) sampling – yes.

If not stated – unclear.

Other studies – no.

Did the study avoid inappropriate exclusions?

Exclusion of more than 10% of the samples for any reason, for example retrospective studies with missing data – no.

Systematic exclusion of types of women / images (e.g. of dense breasts) – no.

Exclusion based on outcomes (e.g. exclusion of cancer types, exclusion of interval cancers, exclusion/inclusion based on recall decision) – no.

Were the women and mammograms included in the study independent of those used to train the AI algorithm?

This question has been added.

For test set studies, this translates as has the test set been clearly described as a geographical validation set?

Any internal validation (e.g. cross-validation), split sample or temporal validation – no.

No details stated about the training set and tuning set - unclear.

Geographical validation (Test set was sample from a different centre; can be in another country or the same country) – yes.

For prospective applied studies in a clinical context:

If the study is located at different centre(s) to those who provided mammograms used to train and tune the AI algorithm – yes.

If not stated – unclear.

If there is any overlap – no.

DOMAIN 2: INDEX TEST(S) **If more than one index test or a human comparator was used, please complete for each test.** **A. Risk of Bias**	
Describe the index test and how it was conducted and interpreted:	
+ Were the index test results interpreted without knowledge of the results of the reference standard? (Requires no repeated application of AI to any of the same cases, or use of the same cases for training)	Yes/No/Unclear
+ Were the index test results interpreted without knowledge of the results of any other index tests?	Yes/No/Unclear
+ If a threshold was used, was it pre- specified?	Yes/No/Unclear
+ Where human readers are part of the test, were their decisions made in a clinical practice context? (i.e. avoidance of the laboratory effect)	Yes/No/Unclear
Could the conduct or interpretation of the index test have introduced bias? **(Score HIGH if 'no' to any question.)**	**RISK: LOW/HIGH/UNCLEAR**
B. Concerns regarding applicability	
Is there concern that the index test(s) or comparator, its conduct, or interpretation differ from the review question? High concerns if: AI system not yet commercially available, e.g. in house systems; Study did not use a pre-specified threshold for AI system; Not compatible with FFDM systems used in the UK; Not a complete testing pathway applicable to UK (for example AI accuracy for single read, but not integrated into screening centre decisions, e.g. arbitration); Human comparator not a complete testing pathway applicable to the UK (human double reading with arbitration at UK threshold); AI system / reader had no access to prior mammograms / not 4 views available	**CONCERN: LOW/HIGH/UNCLEAR**

Were the index test results interpreted without knowledge of the results of the reference standard?

For index tests where a human is involved (either human read comparator, CAD versions of AI, or included otherwise on the AI testing pathway, e.g. arbitration):
Require clear statement of blinding, or clear temporal relationships where the human read occurred before the reference standard – yes.
Otherwise - no.

For index test where AI is used without any human element:
AI system has not previously been trained on these mammograms or learned from these mammograms or other mammograms from the same women – yes.
If any repeat use of the same cases then - no (unless explicit that the AI algorithm was pre-set and did not change upon repeat use, and the study did not select one of several AI systems based on use with the same cases).
If not explicit that there has been no repeat within same or previous studies - unclear.

Were the index test results interpreted without knowledge of the results of any other index tests?

This question has been added.
If human readers were not blinded to AI then - no (unless that AI is specifically part of the same index test).
If AI systems are trained or calibrated using decisions from human readers in same cases - no.

If a threshold was used, was it pre- specified?

If using a commercially available AI system which gives a yes/no result, or threshold clearly pre-specified in methods – yes.
For systems giving a risk score study must explicitly state the pre-specified threshold - yes.
Pathways with AI as a single decision in the pathway, requires pre-specification of AI part.
Using sensitivity / specificity of the reader as benchmark using the same dataset - no
Setting the threshold with the validation set without temporal evidence (e.g. published protocol) that threshold was truly pre-specified - no
Human readers or human CAD combinations – NA.

Where human readers are part of the test, were their decisions made in a clinical practice context? (i.e. avoidance of the laboratory effect)

This question has been added.
If the readers made decisions in the clinical context, and those decisions were used to decide whether to recall women (either prospectively as part of a trial or test accuracy study or retrospective studies using the original decision) – yes.
If readers examined a test set (of any prevalence) outside clinical practice, or any other context likely to result in the laboratory effect[92] – no.

DOMAIN 3: REFERENCE STANDARD

A. Risk of Bias

Describe the reference standard and how it was conducted and interpreted:

+ Is the reference standard likely to correctly classify the target condition? Yes/No/Unclear

+ Were the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear

Could the reference standard, its conduct, or its interpretation have introduced bias? **RISK: LOW/HIGH/UNCLEAR**

(Score HIGH if ‘no’ to any question.)

B. Concerns regarding applicability

Is there concern that the target condition as defined by the reference standard does not match the review question? **CONCERN: LOW/HIGH/UNCLEAR**

Different length of screening rounds than UK screening programme for follow-up / definition of interval cancers;
Classification not by biopsy/follow-up.

Is the reference standard likely to correctly classify the target condition?
If the reference standard is histopathology results from biopsy (cancer present or absent) with at least 2 years follow-up to interval cancers - yes.
If the reference standard is histopathology results from biopsy (cancer present or absent) with no follow-up -no.
It is not possible or ethical to biopsy test negatives, or women recalled for further tests where those test do not clinically indicate the need for biopsy, and follow-up will detect some but not all false negatives from screening.

Were the reference standard results interpreted without knowledge of the results of the index test?
For retrospective studies (if we include the human reader comparator as an index test) – no.
For prospective studies if the investigators did not blind the clinicians undertaking the follow-up tests to which index test examined the mammograms, for example by putting location marks in the same format for AI and human readers - no.
Retrospective studies where readers read mammograms prospectively (enriched test sets) – yes

DOMAIN 4: FLOW AND TIMING **A. Risk of Bias**	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):	
Describe the time interval and any intervention between index tests(s) and reference standard:	
+ Did all patients receive a reference standard?	Yes/No/Unclear
+ Did the study avoid choosing which reference standard based on results of just one of the index tests?	Yes/No/Unclear
+ Were all patients included in the analysis?	Yes/No/Unclear
Could the patient flow have introduced bias? **(Score HIGH if 'no' to any question.)**	**RISK: LOW/HIGH/UNCLEAR**

It is not practical or ethical to give all women the gold standard test, histopathology on biopsy samples. This would involve biopsy of women's breasts without clinical indication, which is clearly unethical. Therefore, all population-based studies will include either partial verification (follow-up tests/biopsy only in test positives) or differential verification (follow-up tests/biopsy in test positives, and follow-up to symptomatic cancer or subsequent screen in test negatives). Cancer, where present, is more likely to be detected in women receiving follow-up tests/biopsy after recall from screening, than women who were not recalled from further test from screening, and simply received follow-up to symptomatic detection (interval cancers). This is because many cancers are very slow growing so would not appear symptomatically for many years, and some would never result in symptoms. Therefore, we have adapted the questions as follows:

Was there an appropriate interval between index test(s) and reference standard?
We removed this signaling question. In this case, the index test refers to the date at which the mammograms were taken not the date at which they were examined. A long interval between taking the mammograms and the reference standard of biopsy could mean that the cancer has developed. It would be extremely unusual for there to be such a gap in retrospective or prospective studies in population screening programmes, which control this interval between mammogram and follow-up tests to be small enough. Where there is differential verification, this becomes an issue because we do not know whether interval cancers were present at the point of screening and missed, or have developed since. However, including follow-up data (differential verification) improves study quality in comparison to not including (partial verification), so we have removed this question to

avoid rating those studies which have included this extra follow-up data to be at higher risk of bias.

Did all patients receive a reference standard?
If there was significant (>10%) loss to follow-up for reference standards of interval cancers or subsequent screening results – no.
If any women who should have received a biopsy or follow-up tests after index test positive results did not receive one or results were unavailable – no.

Did all patients receive the same reference standard? Has been changed to:
Did the study avoid choosing which reference standard based on results of just one of the index tests?
All studies will necessarily have differential verification, because not all women can or should be biopsied. Here we are measuring whether deciding which reference standard is received based on results of just one of the index tests is avoided.
If women were recalled for further tests on the basis of one of the index tests, and not other(s) then this will cause bias because cancer, when present, is more likely to be found if the person receives follow-up tests after recall from screening – no.
If women testing positive in any of the included index tests (AI pathways or comparator human pathways) all receive follow-up tests/biopsy in a prospective study - yes.
For test-treat RCTs randomizing to different test strategies and their associated recall decisions – yes.
In retrospective studies, the decision whether to recall for follow-up tests/biopsy was made on the basis of the human readers' decision. We do not know whether AI positive, human reader negative women are false positive or true positive, and what type of true positive. Follow-up to development of interval cancers will detect some, but not all of these cancers, so reduces, but does not eliminate this bias – no.
For prospective studies where decision to recall is informed by one index test but not all, or is more influenced by one index test than others – no.
Retrospective reader studies (enriched test set studies) in which readers prospectively read retrospective data, the reference standard is not based on any index test but the reference standard is based on the original human reader decision. The reviewers are unclear about the risk of bias – unclear (to be further discussed).

Were all patients included in the analysis?
If there were any exclusions after the point of selecting the cohort, for example intermediate or indeterminate results – yes.

Appendix 6 – UK NSC reporting checklist for evidence summaries

All items on the UK NSC Reporting Checklist for Evidence Summaries have been addressed in this report. A summary of the checklist, along with the page or pages where each item can be found in this report, is presented in **Table 24**.

Table 24. UK NSC reporting checklist for evidence summaries

	Section	Item	Page no.
1.	TITLE AND SUMMARIES		
1.1	Title sheet	Identify the review as a UK NSC evidence summary.	Title page
1.2	Plain English summary	Plain English description of the executive summary.	1-2
1.3	Executive summary	Structured overview of the whole report. To include: the purpose/aim of the review; background; previous recommendations; findings and gaps in the evidence; recommendations on the screening that can or cannot be made on the basis of the review.	3-10
2.	INTRODUCTION AND APPROACH		
2.1	Background and objectives	Background – Current policy context and rationale for the current review – for example, reference to details of previous reviews, basis for current recommendation, recommendations made, gaps identified, drivers for new reviews Objectives – What are the questions the current evidence summary intends to answer? – statement of the key questions for the current evidence summary, criteria they address, and number of studies included per question, description of the overall results of the literature search. Method – briefly outline the rapid review methods used.	11-22

2.2	Eligibility for inclusion in the review	State all criteria for inclusion and exclusion of studies to the review clearly (PICO, dates, language, study type, publication type, publication status etc.) To be decided *a priori*.	24-34
2.3	Appraisal for quality/risk of bias tool	Details of tool/checklist used to assess quality, e.g. QUADAS 2, CASP, SIGN, AMSTAR.	34-35 and Appendix 5
3.	SEARCH STRATEGY AND STUDY SELECTION (FOR EACH KEY QUESTION)		
3.1	Databases/ sources searched	Give details of all databases searched (including platform/interface and coverage dates) and date of final search.	23 and Appendix 1 (Table 12)
3.2	Search strategy and results	Present the full search strategy for at least one database (usually a version of Medline), including limits and search filters if used. Provide details of the total number of (results from each database searched), number of duplicates removed, and the final number of unique records to consider for inclusion.	Appendix 1 (Table 13-Table 17)
3.3	Study selection	State the process for selecting studies – inclusion and exclusion criteria, number of studies screened by title/abstract and full text, number of reviewers, any cross checking carried out.	24-26 and Appendix 2 (Figure 7)
4.	STUDY LEVEL REPORTING OF RESULTS (FOR EACH KEY QUESTION)		
4.1	Study level reporting, results and risk of bias assessment	For each study, produce a table that includes the full citation and a summary of the data relevant to the question (for example, study size, PICO, follow-up period, outcomes reported, statistical analyses etc.). Provide a simple summary of key measures, effect estimates and confidence intervals for each study where available. For each study, present the results of any assessment of quality/risk of bias.	Study level reporting: Question 1: Appendix 3 (Table 21 and Table 22) Question 2: Appendix 4 Quality assessment: Question 1: 38-41 and Appendix 3 (Table 23) Question 2: NA

5.	QUESTION LEVEL SYNTHESIS		
5.1	Description of the evidence	For each question, give numbers of studies screened, assessed for eligibility, and included in the review, with summary reasons for exclusion.	Question 1: 36-37 Question 2: 56
5.2	Combining and presenting the findings	Provide a balanced discussion of the body of evidence which avoids over reliance on one study or set of studies. Consideration of four components should inform the reviewer's judgement on whether the criterion is 'met', 'not met' or 'uncertain': quantity; quality; applicability and consistency.	Question 1: 41-54 Question 2: 57-66
5.3	Summary of findings	Provide a description of the evidence reviewed and included for each question, with reference to their eligibility for inclusion. Summarise the main findings including the quality/risk of bias issues for each question. Have the criteria addressed been 'met', 'not met' or 'uncertain'?	Question 1: 55 Question 2: 67
6.	REVIEW SUMMARY		
6.1	Conclusions and implications for policy	Do findings indicate whether screening should be recommended? Is further work warranted? Are there gaps in the evidence highlighted by the review?	78
6.2	Limitations	Discuss limitations of the available evidence and of the review methodology if relevant.	68-77

References

1. Smittenaar CR, Petersen KA, Stewart K, et al. Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer* 2016;115(9):1147-55. doi: 10.1038/bjc.2016.304 [published Online First: 2016/10/26]

2. Office for National Statistics. Cancer registration statistics, England: 2017, 2019.

3. Januškevičienė I, Petrikaitė V. Heterogeneity of breast cancer: The importance of interaction between different tumor cell populations. *Life Sciences* 2019;239:117009. doi: https://doi.org/10.1016/j.lfs.2019.117009

4. Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 2015;1856(1):73-85. doi: https://doi.org/10.1016/j.bbcan.2015.06.002

5. Anothaisintawee T, Wiratkapun C, Lerdsitthichai P, et al. Risk Factors of Breast Cancer: A Systematic Review and Meta-Analysis. *Asia Pacific Journal of Public Health* 2013;25(5):368-87. doi: 10.1177/1010539513488795

6. Harris LN, Ismaila N, McShane LM, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. *J Clin Oncol* 2016;34(10):1134-50. doi: 10.1200/jco.2015.65.2289 [published Online First: 2016/02/10]

7. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415(6871):530-6. doi: 10.1038/415530a [published Online First: 2002/02/02]

8. Page DL, Kidd TE, Jr., Dupont WD, et al. Lobular neoplasia of the breast: higher risk for subsequent invasive cancer predicted by more extensive disease. *Hum Pathol* 1991;22(12):1232-9. doi: 10.1016/0046-8177(91)90105-x [published Online First: 1991/12/11]

9. Mannu GS, Wang Z, Broggio J, et al. Invasive breast cancer and breast cancer mortality after ductal carcinoma in situ in women attending for breast screening in England, 1988-2014: population based observational cohort study. *BMJ* 2020;369:m1570. doi: 10.1136/bmj.m1570 [published Online First: 2020/05/29]

10. Yen MF, Tabar L, Vitak B, et al. Quantifying the potential problem of overdiagnosis of ductal carcinoma in situ in breast cancer screening. *Eur J Cancer* 2003;39(12):1746-54. doi: 10.1016/s0959-8049(03)00260-0 [published Online First: 2003/07/31]

11. Welch HG, Prorok PC, O'Malley AJ, et al. Breast-Cancer Tumor Size, Overdiagnosis, and Mammography Screening Effectiveness. *N Engl J Med* 2016;375(15):1438-47. doi: 10.1056/NEJMoa1600249 [published Online First: 2016/10/13]

12. Jorgensen KJ, Gotzsche PC, Kalager M, et al. Breast Cancer Screening in Denmark: A Cohort Study of Tumor Size and Overdiagnosis. *Ann Intern Med* 2017;166(5):313-23. doi: 10.7326/M16-0270 [published Online First: 2017/01/24]

13. Thomas ET, Del Mar C, Glasziou P, et al. Prevalence of incidental breast cancer and precursor lesions in autopsy studies: a systematic review and meta-analysis. *BMC Cancer* 2017;17(1):808. doi: 10.1186/s12885-017-3808-1 [published Online First: 2017/12/05]

14. Turashvili G, Brogi E. Tumor Heterogeneity in Breast Cancer. *Front Med (Lausanne)* 2017;4:227-27. doi: 10.3389/fmed.2017.00227

15. Freudenberg JA, Wang Q, Katsumata M, et al. The role of HER2 in early breast cancer metastasis and the origins of resistance to HER2-targeted therapies. *Exp Mol Pathol* 2009;87(1):1-11. doi: 10.1016/j.yexmp.2009.05.001 [published Online First: 2009/05/18]

16. NICE. Early and locally advances breast cancer: diagnoss and management, 2018.

17. Public Health England. NHS Breast Screening Programme: Guidance on who can undertake arbitration. London: Public Health England, 2016.

18. Wilson R, Liston J. Guidelines for Breast Cancer Screening Radiology. NHSBSP Publication. Second edition ed. Sheffield, 2011:1-36.

19. Public Health England. NHS Breast Screening Programme Guidance on collecting, monitoring and reporting technical recall and repeat examinations. London: Public Health England, 2017.

20. Royal College of Radiologists. Guidance on screening and symptomatic breast imaging. Fourth edition. 4 ed. London: Royal College of Radiologists, 2019.

21. Screening & Immunisations Team. Breast Screening Programme: England, 2018-19: NHS Digital, 2020.

22. Public Health Wales. Breast Test Wales Annual Statistical Report 2018-19. Screening Division of Public Health Wales Official Statistics Publication. Cardiff, 2020:1-24.

23. Cancer Research UK. Breast cancer incidence (invasive) statistics [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-Zero accessed 15 December 2020.

24. Cancer Research UK. Breast cancer mortality statistics [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/mortality#heading-Zero accessed 15 December 2020.

25. Cancer Research UK. In situ breast carcinoma incidence statistics [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-in-situ#heading-Zero accessed 15 December 2020.

26. Public Health England. Young person and adult screening KPI data: Public Health England, 2020.

27. Northern Ireland Breast Screening Service. KC62 Data 2017/18.

28. Public Health Scotland. Scottish breast screening programme statistics - Annual update to 31 March 2019 2020 [updated 21 April 2020. Available from: https://beta.isdscotland.org/find-publications-and-data/conditions-and-diseases/cancer/scottish-breast-screening-programme-statistics/ accessed 9 December 2020.

29. Public Health Wales. Breast Test Wales Annual Statistical Report 2018-19. Cardiff: Public Health Wales, 2020.

30. van Seijen M, Lips EH, Thompson AM, et al. Ductal carcinoma in situ: to treat or not to treat, that is the question. *Br J Cancer* 2019;121(4):285-92. doi: 10.1038/s41416-019-0478-6 [published Online First: 2019/07/10]

31. Public Health England. Breast screening: reporting, classification and monitoring of interval cancers and cancers following previous assessment, 2020.

32. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* 2012;380(9855):1778-86. doi: 10.1016/s0140-6736(12)61611-0 [published Online First: 2012/11/03]

33. Screening & Immunisations Team. Breast Screening Programme: England, 2019-20: National Statistics; NHS Digital; 2021 [updated 28 January 2021. Available from: https://files.digital.nhs.uk/F9/98C8E3/breast-screening-programme-eng-2019-20-report.pdf accessed 15 March 2021.

34. Maxwell AJ, Clements K, Hilton B, et al. Risk factors for the development of invasive cancer in unresected ductal carcinoma in situ. *Eur J Surg Oncol* 2018;44(4):429-35. doi: 10.1016/j.ejso.2017.12.007 [published Online First: 2018/02/06]

35. Pinder SE, Shaaban A, Deb R, et al. NHS Breast Screening multidisciplinary working group guidelines for the diagnosis and management of breast lesions of uncertain malignant potential on core biopsy (B3 lesions). *Clin Radiol* 2018;73(8):682-92. doi: 10.1016/j.crad.2018.04.004 [published Online First: 2018/05/19]

36. Forester ND, Lowes S, Mitchell E, et al. High risk (B3) breast lesions: What is the incidence of malignancy for individual lesion subtypes? A systematic review and meta-analysis. *Eur J Surg Oncol* 2019;45(4):519-27. doi: 10.1016/j.ejso.2018.12.008 [published Online First: 2018/12/24]

37. King TA, Pilewskie M, Muhsen S, et al. Lobular Carcinoma in Situ: A 29-Year Longitudinal Experience Evaluating Clinicopathologic Features and Breast Cancer Risk. *J Clin Oncol* 2015;33(33):3945-52. doi: 10.1200/JCO.2015.61.4743 [published Online First: 2015/09/16]

38. Dyrstad SW, Yan Y, Fowler AM, et al. Breast cancer risk associated with benign breast disease: systematic review and meta-analysis. *Breast Cancer Res Treat* 2015;149(3):569-75. doi: 10.1007/s10549-014-3254-6 [published Online First: 2015/02/01]

39. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine* 2015;175(11):1828-37.

40. Perry N, Broeders M, de Wolf C. European guidelines for quality assurance in breast cancer screening and diagnosis. In: Karsa L. von, Holland R, Broeders M, et al., eds. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth ed. Luxembourg: European Commission, Office for Official Publications of the European Union 2013:XIV–XX.

41. Nelson HD, Cantor A, Humphrey L, et al. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Screening for Breast Cancer: A Systematic Review to Update the 2009 US Preventive Services Task Force Recommendation. Rockville (MD): Agency for Healthcare Research and Quality (US) 2016.

42. Ebell MH, Thai TN, Royalty KJ. Cancer screening recommendations: an international comparison of high income countries. *Public Health Rev* 2018;39:7-7. doi: 10.1186/s40985-018-0080-0

43. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *Radiographics* 2017;37(7):2113-31.

44. McCorduck P, Cfe C. Machines who think: A personal inquiry into the history and prospects of artificial intelligence: CRC Press 2004.

45. Google AI. Advancing AI for everyone [Available from: https://ai.google./ accessed 9 December 2020.

46. Faes L, Liu X, Wagner SK, et al. A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. *Transl Vis Sci Technol* 2020;9(2):7. doi: 10.1167/tvst.9.2.7 [published Online First: 2020/07/25]

47. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* 2018;286(3):800-09. doi: 10.1148/radiol.2017171920 [published Online First: 2018/01/09]

48. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal* 2020 doi: 10.1093/ckj/sfaa188

49. Chan H-P, Charles E, Metz P, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms. *Arbor* 1990;1001:48109-0326.

50. Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. *Nat Rev Cancer* 2018;18(8):500-10. doi: 10.1038/s41568-018-0016-5 [published Online First: 2018/05/20]

51. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. *European Journal of Cancer* 2008;44(6):798-807.

52. Chan H-P, Sahiner B, Lo S-C, et al. Computer-aided diagnosis in mammography: Detection of masses by artificial neural network1994.

53. Sahiner B, Chan H-P, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE transactions on Medical Imaging* 1996;15(5):598-610.

54. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436-44.

55. Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks. *Advances in neural information processing systems* 2006;19:153-60.

56. Rodriguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* 2019;290(2):305-14. doi: https://dx.doi.org/10.1148/radiol.2018181371

57. Agarwal R, Diaz O, Yap MH, et al. Deep learning for mass detection in Full Field Digital Mammograms. *Comput Biol Med* 2020;121:103774. doi: 10.1016/j.compbiomed.2020.103774 [published Online First: 2020/04/28]

58. Waymel Q, Badr S, Demondion X, et al. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagnostic and interventional imaging* 2019;100(6):327-36.

59. Xiong Z, Fedorov VV, Fu X, et al. Fully automatic left atrium segmentation from late gadolinium enhanced magnetic resonance imaging using a dual fully convolutional neural network. *IEEE transactions on medical imaging* 2018;38(2):515-24.

60. Bray M-A, Carpenter AE. Quality control for high-throughput imaging experiments using machine learning in CellProfiler. High Content Screening: Springer 2018:89-112.

61. Sevenster M, Buurman J, Liu P, et al. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. *Applied clinical informatics* 2015;6(3):600.

62. Bond M, Pavey T, Welch K, et al. Psychological consequences of false-positive screening mammograms in the UK. *BMJ evidence-based medicine* 2013;18(2):54-61.

63. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. *Annals of internal medicine* 2007;146(7):502-10.

64. Petticrew M, Sowden A, Lister-Sharp D. False-negative results in screening programs. Medical, psychological, and other implications. *International journal of technology assessment in health care* 2001;17(2):164-70.

65. Majid AS, de Paredes ES, Doherty RD, et al. Missed breast carcinoma: pitfalls and pearls. *Radiographics* 2003;23(4):881-95.

66. McDonald RJ, Schwartz KM, Eckel LJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic radiology* 2015;22(9):1191-98.

67. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)* 2017;359

68. Clark S, Reeves PJ. Women's experiences of mammography: a thematic evaluation of the literature. *Radiography* 2015;21(1):84-88.

69. Carter SM, Rogers W, Win KT, et al. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *The Breast* 2020;49:25-32.

70. Castelvecchi D. Can we open the black box of AI? *Nature News* 2016;538(7623):20.

71. McCartney M. Margaret McCartney: AI in medicine must be rigorously tested. *Bmj* 2018;361

72. Pesapane F, Volonté C, Codari M, et al. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights into imaging* 2018;9(5):745-53.

73. Sit C, Srinivasan R, Amlani A, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. *Insights into Imaging* 2020;11(1):14.

74. Topol E. The Topol review: preparing the healthcare workforce to deliver the digital future. *Health Education England* 2019

75. Kotter E, Ranschaert E. Challenges and solutions for introducing artificial intelligence (AI) in daily clinical workflow. *Eur Radiol* 2021;31(1):5-7. doi: 10.1007/s00330-020-07148-2 [published Online First: 2020/08/17]

76. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577(7788):89-94. doi: https://doi.org/10.1038/s41586-019-1799-6)

77. Pacilè S, Lopez J, Chone P, et al. Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. *Radiology: Artificial Intelligence* 2020;2(6):e190208. doi: 10.1148/ryai.2020190208

78. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J Natl Cancer Inst* 2019;111(9):916-22. doi: https://dx.doi.org/10.1093/jnci/djy222

79. Rodriguez-Ruiz A, Mordang JJ, Karssemeijer N, et al. Can radiologists improve their breast cancer detection in mammography when using a deep learning based computer system as decision support? *Proceedings of SPIE* 2018;10718:1071803. doi: 10.1117/12.2317937

80. Salim M, Wahlin E, Dembrower K, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol* 2020;27:27. doi: https://dx.doi.org/10.1001/jamaoncol.2020.3321

81. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA netw* 2020;3(3):e200265. doi: https://dx.doi.org/10.1001/jamanetworkopen.2020.0265

82. Watanabe AT, Lim V, Vu HX, et al. Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. *J Digit Imaging* 2019;32(4):625-37. doi: https://dx.doi.org/10.1007/s10278-019-00192-5

83. Balta C, Rodriguez-Ruiz A, Mieskes C, et al. Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact? *Proceedings of SPIE* 2020;11513:115130D.

84. Dembrower K, Wahlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digital Health* 2020;2(9):E468-E74.

85. Lang K, Dustler M, Dahlblom V, et al. Identifying normal mammograms in a large screening population using artificial intelligence. *European Radiology* 2020;02:02. doi: https://dx.doi.org/10.1007/s00330-020-07165-1

86. Chen X, Moschidis E, Taylor C, et al. Breast cancer risk analysis based on a novel segmentation framework for digital mammograms. *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv* 2014;17(Pt 1):536-43.

87. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2020;2(3):e138-e48. doi: http://dx.doi.org/10.1016/S2589-7500%2820%2930003-0

88. Kyono T, Gilbert FJ, van der Schaar M. Improving Workflow Efficiency for Mammography Using Machine Learning. *J* 2020;17(1 Pt A):56-63. doi: https://dx.doi.org/10.1016/j.jacr.2019.05.012

89. Kyono T, Gilbert FJ, van der Schaar M. MAMMO: A Deep Learning Solution for Facilitating Radiologist-Machine Collaboration in Breast Cancer Diagnosis. 2018 30 October 2018.

90. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *European Radiology* 2019;29(9):4825-32. doi: https://dx.doi.org/10.1007/s00330-019-06186-9

91. Taylor-Phillips S, Seedat F, Kijauskaite G, et al. UK National Screening Committee Approach to Reviewing Evidence on Artificial Intelligence in Breast Cancer Screening. Manuscript in preparation, 2020.

92. 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(1):47-53. doi: 10.1148/radiol.2491072025 [published Online First: 2008/08/07]

93. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155(8):529-36. doi: 10.7326/0003-4819-155-8-201110180-00009 [published Online First: 2011/10/19]

94. Yada B, Whiting PF, Davenport C, et al. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. MEMTAB2020. Leuven, Belgium, 2020.

95. National Institute for Health Research (NIHR). Artificial intelligence funding: National Institute for Health Research (NIHR); [Available from: https://www.nihr.ac.uk/explore-nihr/funding-programmes/ai-award.htm accessed 10 December 2020.

96. Hupse R, Samulski M, Lobbes M, et al. Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *European Radiology* 2013;23(1):93-100. doi: https://dx.doi.org/10.1007/s00330-012-2562-7

97. Becker AS, Marcon M, Ghafoor S, et al. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Invest Radiol* 2017;52(7):434-40. doi: https://dx.doi.org/10.1097/RLI.0000000000000358

98. Gilbert FJ, Tucker L, Gillan MG, et al. The TOMMY trial: a comparison of TOMosynthesis with digital MammographY in the UK NHS Breast Screening Programme--a multicentre retrospective reading study comparing the diagnostic performance of digital breast tomosynthesis and digital mammography with digital mammography alone. *Health Technol Assess* 2015;19(4):i-xxv, 1-136. doi: 10.3310/hta19040 [published Online First: 2015/01/20]

99. Sharma N, Ng AY, James JJ, et al. Large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. *medRxiv* 2021:2021.02.26.21252537. doi: 10.1101/2021.02.26.21252537