Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy

Karoline Freeman, Julia Geppert, Chris Stinton, Daniel Todkill, Samantha Johnson, Aileen Clarke, Sian Taylor-Phillips

Division of Health Sciences, University of Warwick, Coventry, UK

Correspondence to:
S Taylor-Phillips
S.Taylor-Phillips@warwick.ac.uk
(ORCID 0000-0002-1841-4346)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2021;374:n1872
http://dx.doi.org/10.1136/bmj.n1872

Accepted: 21 July 2021

ABSTRACT

OBJECTIVE

To examine the accuracy of artificial intelligence (AI) for the detection of breast cancer in mammography screening practice.

DESIGN

Systematic review of test accuracy studies.

DATA SOURCES

Medline, Embase, Web of Science, and Cochrane Database of Systematic Reviews from 1 January 2010 to 17 May 2021.

ELIGIBILITY CRITERIA

Studies reporting test accuracy of AI algorithms, alone or in combination with radiologists, to detect cancer in women's digital mammograms in screening practice, or in test sets. Reference standard was biopsy with histology or follow-up (for screen negative women). Outcomes included test accuracy and cancer type detected.

STUDY SELECTION AND SYNTHESIS

Two reviewers independently assessed articles for inclusion and assessed the methodological quality of included studies using the QUality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool. A single reviewer extracted data, which were checked by a second reviewer. Narrative data synthesis was performed.

RESULTS

Twelve studies totalling 131 822 screened women were included. No prospective studies measuring test accuracy of AI in screening practice were found. Studies were of poor methodological quality. Three retrospective studies compared AI systems with the clinical decisions of the original radiologist, including 79 910 women, of whom 1878 had screen detected cancer or interval cancer within 12 months of screening. Thirty four (94%) of 36 AI systems evaluated in these studies were less accurate than a single radiologist, and all were less accurate than consensus of two or more radiologists. Five smaller studies (1086 women, 520 cancers) at high risk of bias and low generalisability to the clinical context reported that all five evaluated AI systems (as standalone to replace radiologist or as a reader aid) were more accurate than a single radiologist reading a test set in the laboratory. In three studies, AI used for triage screened out 53%, 45%, and 50% of women at low risk but also 10%, 4%, and 0% of cancers detected by radiologists.

CONCLUSIONS

Current evidence for AI does not yet allow judgement of its accuracy in breast cancer screening programmes, and it is unclear where on the clinical pathway AI might be of most benefit. AI systems are not sufficiently specific to replace radiologist double reading in screening programmes. Promising results in smaller studies are not replicated in larger studies. Prospective studies are required to measure the effect of AI in clinical practice. Such studies will require clear stopping rules to ensure that AI does not reduce programme specificity.

STUDY REGISTRATION

Protocol registered as PROSPERO CRD42020213590.

WHAT IS ALREADY KNOWN ON THIS TOPIC

A recent scoping review of 23 studies on artificial intelligence (AI) for the early detection of breast cancer highlighted evidence gaps and methodological concerns about published studies

Published opinion pieces claim that the replacement of radiologists by AI is imminent

Current mammography screening is repetitive work for radiologists and misses 15-35% of cancers—a prime example of the sort of role we would expect AI to be fulfilling

WHAT THIS STUDY ADDS

This systematic review of test accuracy identified 12 studies, of which only one was included in the previous review

Current evidence on the use of AI systems in breast cancer screening is of insufficient quality and quantity for implementation into clinical practice

In retrospective test accuracy studies, 94% of AI systems were less accurate than the original radiologist, and all were less accurate than original consensus of two radiologists; prospective evaluation is required

Introduction

Breast cancer is a leading cause of death among women worldwide. Approximately 2.4 million women were diagnosed with breast cancer in 2015, and 523 000 women died.[1] Breast cancer is more amenable to treatment when detected early,[2] so many countries have introduced screening programmes. Breast cancer screening requires one or two radiologists to examine women's mammograms for signs of presymptomatic cancer, with the aim of reducing breast cancer related morbidity and mortality. Such screening is also associated with harms, such as overdiagnosis and overtreatment of cancer that would not have become symptomatic within the woman's lifetime. Disagreement exists about the extent of overdiagnosis, from 1% to 54% of screen detected cancers, and about the balance of benefits and harms of screening.[2] The spectrum of disease detected at screening is associated with outcomes. For example, detection of low grade ductal carcinoma in situ is more associated with overdiagnosis,[3 4] whereas detection of grade 3 cancer is

more likely to be associated with fewer deaths.[5] Cancer is detected in between 0.6% and 0.8% of women during screening.[6 7] Breast screening programmes also miss between 15% and 35% of cancers owing either to error or because the cancer is not visible or perceptible to the radiologist. Some of these missed cancers present symptomatically as interval cancers.[8]

Considerable interest has been shown in the use of artificial intelligence (AI) either to complement the work of humans or to replace them. In 2019, 3.8% of all peer reviewed scientific publications worldwide on Scopus related to AI.[9] Claims have been made that image recognition using AI for breast screening is better than experienced radiologists and will deal with some of the limitations of current programmes.[10-13] For instance, fewer cancers might be missed because an AI algorithm is unaffected by fatigue or subjective diagnosis,[14 15] and AI might reduce workload or replace radiologists completely.[11 12]

AI might, however, also exacerbate harm from screening. For example, AI might alter the spectrum of disease detected at breast screening if it differentially detects more microcalcifications, which are associated with lower grade ductal carcinoma in situ. In such a case, AI might increase rates of overdiagnosis and overtreatment and alter the balance of benefits and harms.

Autopsy studies suggest that around 4% of women die with, not because of, breast cancer,[16] so there is a "reservoir" of clinically unimportant disease, including incidental in situ carcinoma, which might be detected by AI. The spectrum of disease is correlated with mammographic features (for example, ductal carcinoma in situ is often associated with microcalcifications). Therefore, the cases on which AI systems were trained, and the structures within the AI system, might considerably affect the spectrum of disease detected. These structures and algorithms within an AI system are not always transparent or explicable, making interpretation a potential problem. Unlike human interpretation, how or why an algorithm has made a decision can be difficult to understand (known as the "black box" problem).[17] Unlike human decision makers, algorithms do not understand the context, mode of collection, or meaning of viewed images, which can lead to the problem of "shortcut" learning,[18] whereby deep neural networks reach a conclusion to a problem through a shortcut, rather than the intended solution. Thus, for example, DeGrave et al[19] have shown how some deep learning systems detect covid-19 by means of confounding factors, rather than pathology, leading to poor generalisability. Although this problem does not preclude the use of deep learning, it highlights the importance of avoiding potential confounders in training data, an understanding of algorithm decision making, and the critical role of rigorous evaluation.

This review was commissioned by the UK National Screening Committee to determine whether there is sufficient evidence to use AI for mammographic image analysis in breast screening practice. Our aim was to assess the accuracy of AI to detect breast cancer when integrated into breast screening programmes, with a focus on the cancer type detected.

Methods

Data sources

Our systematic review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of diagnostic test accuracy (PRISMA-DTA) statement.[20] The review protocol is registered on PROSPERO (international prospective register of systematic reviews).

We conducted literature searches for studies published in English between 1 January 2010 and 9 September 2020 and updated our searches on 17 May 2021. The search comprised four themes: breast cancer, artificial intelligence, mammography, and test accuracy or randomised controlled trials. A number of additional synonyms were identified for each theme. Databases searched were Medline (Ovid), Embase (Ovid); Web of Science, and the Cochrane Database of Systematic Reviews (CENTRAL). Details of the search strategies are shown in supplementary appendix 1. We screened the reference lists of systematic reviews and included additional relevant studies and contacted experts in the field.

Study selection

Two reviewers independently reviewed the titles and abstracts of all retrieved records against the inclusion criteria, and subsequently, all full text publications. Disagreements were resolved by consensus or discussion with a third reviewer.

We applied strict inclusion/exclusion criteria to focus on the evaluation of the integration of AI into a breast cancer screening programme rather than the development of AI systems. Studies were eligible for inclusion if they reported test accuracy of AI algorithms applied to women's digital mammograms to detect breast cancer, as part of a pathway change or a complete read (reading+decision resulting in classification). Eligible study designs were prospective test accuracy studies, randomised controlled trials, retrospective test accuracy studies using geographical validation only, comparative cohort studies, and enriched test set multiple reader multiple case laboratory studies. The enriched test set multiple reader multiple case laboratory studies included retrospective data collection of images and prospective classification by standalone AI or AI assisted radiologists. The reference standard was cancer confirmed by histological analysis of biopsy samples from women referred for further tests at screening and preferably also from symptomatic presentation during follow-up.

All studies will necessarily have differential verification because not all women can or should be biopsied. In prospective test accuracy studies this will not introduce significant bias because those positive on either an index or comparator test will receive follow-up tests. In retrospective studies and enriched test set studies (with prospective readers), the decision

as to whether women receive biopsy or follow-up is based on the decision of the original reader, which introduces bias because cancer, when present, is more likely to be found if the person receives follow-up tests after recall from screening. We assessed this using the QUality Assessment of Diagnostic Accuracy Studies-2 tool (QUADAS-2). When AI is used as a pre-screen to triage which mammograms need to be examined by a radiologist and which do not, we also accepted a definition of a normal mammogram as one free of screen detected cancer based on human consensus reading, as this allows estimation of accuracy in the triage.

We excluded studies that reported the validation of AI systems using internal validation test sets (eg, x-fold cross validation, leave one out method), split validation test sets, and temporal validation test sets as they are prone to overfitting and insufficient to assess the generalisability of the AI system. Furthermore, studies were excluded if less than 90% of included mammograms were complete full field digital mammography screening mammograms. Additionally, studies were excluded if the AI system was used to predict future risk of cancer, if only detection of cancer subtypes was reported, if traditional computer aided detection systems without machine learning were used, or if test accuracy measures were not expressed at any clinically relevant threshold (eg, area under the curve only) or did not characterise the trade-off between false positives and false negative results (eg, sensitivity for cancer positive samples only). Finally, we excluded simulation results of the hypothetical integration of AI with radiologists' decisions as they do not reliably estimate radiologist behaviour when AI is applied.

Data extraction and quality assessment

One reviewer extracted data on a predesigned data collection form. Data extraction sheets were checked by a second reviewer and any disagreements were resolved by discussion. Study quality was assessed independently by two reviewers using QUADAS-2[21] tailored to the review question (supplementary appendix 2).

Data analysis

The unit of analysis was the woman. Data were analysed according to where in the pathway AI was used (for example, standalone AI to replace one or all readers, or reader aid to support decision making by a human reader) and by outcome. The primary outcome was test accuracy. If test accuracy was not reported, we calculated measures of test accuracy where possible. Important secondary outcomes were cancer type and interval cancers. Cancer type (eg, by grade, stage, size, prognosis, nodal involvement) is important in order to estimate the effect of cancer detection on the benefits and harms of screening. Interval cancers are also important because they have worse average prognosis than screen detected cancers,[22] and by definition, are not associated with overdiagnosis at screening. We synthesised studies narratively owing to their small number and extensive heterogeneity. We plotted reported sensitivity and specificity for the AI systems and any comparators in a receiver operating characteristic plot using the package "ggplot2"[23] in R version 3.6.1 (Vienna, Austria).[24]

Patient and public involvement

The review was commissioned on behalf of the UK National Screening Committee (UKNSC), and the scope was determined by the UKNSC adult reference group, which includes lay members. The results were discussed with patient contributors.

Results

Study selection

Database searches yielded 4016 unique results, of which 464 potentially eligible full texts were assessed. Four additional articles were identified: one through screening the reference lists of relevant systematic reviews, one through contact with experts, and two by hand searches. Overall, 13 articles[25-37] reporting 12 studies were included in this review (see supplementary fig 1 for full PRISMA flow diagram). Exclusions on full text are listed in supplementary appendix 3.

Characteristics of included studies

The characteristics of the 12 included studies are presented in table 1, table 2, and table 3 and in supplementary appendix 4, comprising a total of 131 822 screened women. The AI systems in all included studies used deep learning convolutional neural networks. Four studies evaluated datasets from Sweden,[26 27 35 36] three of which had largely overlapping populations,[26 35 36] one from the United States and Germany,[32] one from Germany,[25] one from the Netherlands,[33] one from Spain[31] and four from the US.[28-30 37] Four studies enrolled women consecutively or randomly,[25 27 31 36] while the remaining studies selected cases and controls to enrich the dataset with patients with cancer. Three studies included all patients with cancer and a random sample of those without cancer.[26 29 35] One study included all patients with cancer and controls matched by age and breast density.[28] In two studies, patients and controls were sampled to meet predefined distributions and were reviewed by one radiologist to exclude images not meeting quality standards and images with obvious signs of cancer.[30 32] One study used a range of rules for selection, including by perceived difficulty and mammographic features.[33] Finally, one study included only false negative mammograms.[37] No prospective test accuracy studies in clinical practice were included, only retrospective test accuracy studies[25-27 29 31 35 36] and enriched test set multiple reader multiple case laboratory studies.[28 30 32 33 37] Of these enriched test set laboratory studies, three reported test accuracy for a single AI read as a reader aid.[30 32 37] Another nine studies reported test accuracy for a single AI read as a standalone system in a retrospective test accuracy study [25-27 29 31 35 36] or an enriched test set multiple reader multiple case laboratory study.[28 33]

In studies of standalone systems, the AI algorithms provided a cancer risk score that can be turned into a binary operating point to classify women as high risk (recall) or low risk (no recall). The in-house or commercial standalone AI systems (table 1, table 2, table 3) were evaluated in five studies as a replacement for one or all radiologists. Three studies compared the performance of the AI system with the original decision recorded in the database, based on either a single US radiologist[29] or two radiologists with consensus within the Swedish screening programme.[35 36] Two studies compared the performance of the AI system with the average performance of nine Dutch single radiologists[33] and five US single radiologists,[28] respectively, who read the images under laboratory conditions. Four commercial AI systems were evaluated as a pre-screen to remove normal cases[25-27 31] or were used as a post-screen of negative mammograms after double reading to predict interval and next round screen detected cancers.[26]

In studies of assistive AI, the commercial AI systems provided the radiologist with a level of suspicion for the area clicked. All three studies compared the test accuracy of the AI assisted read with an unassisted read by the same radiologists under laboratory conditions.[30 32 37] The experience of the radiologists in the reader assisted studies ranged from 3 to 25 years (median 9.5 years) in 14 radiologists,[32] from 0 to 25 years (median 8.5 years) in 14 American Board of Radiology and Mammography Quality Standards Act (MQSA) certified radiologists,[30] and from less than 5 to 42 years in 7 MQSA certified radiologists.[37] The role of the AI system in the screening pathway in the 12 studies is summarised in figure 1.

Assessment of risk of bias and applicability

The evidence for the accuracy of AI to detect breast cancer was of low quality and applicability across all studies (fig 2) according to QUADAS-2 (supplementary

Table 1 | Summary of study characteristics for studies using AI as standalone system

Study	Study design	Population	Mammography vendor	Index test	Comparator	Reference standard
Lotter 2021[28]	Enriched test set MRMC laboratory study (accuracy of a read)	285 women from 1 US health system with 4 centres (46.0% screen detected cancer); age and ethnic origin NR	Hologic 100%	In-house AI system (DeepHealth); threshold NR (set to match readers' sensitivity and specificity, respectively)	5 MQSA certified radiologists (US), single reading; threshold of BI-RADS scores 3, 4, and 5 considered recall	Cancer: pathology confirmed cancer within 3 months of screening; confirmed negative: a negative examination followed by an additional BI-RADS score 1 or 2 interpretation at the next screening examination 9-39 months later
McKinney 2020[29]	Retrospective test accuracy study (accuracy of a read)	3097 women from 1 US centre (22.2% cancer within 27 months of screening); age <40, 181 (5.8%); 40-49, 1259 (40.7%); 50-59, 800 (25.8%); 60-69, 598 (19.3%); ≥70, 259 (8.4%)	Hologic / Lorad branded: >99%; Siemens or General Electric: <1%	In-house AI system (Google Health); threshold: to achieve superiority for both sensitivity and specificity compared with original single reading using validation set	Original single radiologist decision (US); threshold: BI-RADS scores 0, 4, 5 were treated as positive	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
Rodriguez-Ruiz 2019[33]	Enriched test set MRMC laboratory study (accuracy of a read)	199 examinations from a Dutch digital screening pilot project (39.7% cancer); age range 50-74	Hologic 100%	Transpara version 1.4.0 (Screenpoint Medical BV, Nijmegen, Netherlands); threshold: 8.26/10, corresponding to the average radiologist's specificity	Nine Dutch radiologists, single reading, as part of a previously completed MRMC study[38]; no threshold	Cancer: histopathology-proven cancer; non-cancer: ≥1 normal follow-up screening examination (2 year screening interval)
Salim 2020[35]	Retrospective test accuracy study (accuracy of a read)	8805 women from a Swedish cohort study (8.4% cancer within 12 months of screening); median age 54.5 (IQR 47.4-63.5)	Hologic 100%	3 commercial AI systems (anonymised: AI-1, AI-2, and AI-3); threshold: corresponding to the specificity of the first reader	Original radiologist decision (Sweden); (1) single reader (R1; R2), (2) consensus reading; no threshold	Cancer: pathology confirmed cancer within 12 months of screening; non-cancer: ≥2 years cancer free follow-up
Schaffter 2020[36]	Retrospective test accuracy study (accuracy of a read)	68 008 consecutive women from 1 Swedish centre (1.1% cancer within 12 months of screening) mean age 53.3 (SD 9.4)	NR	4 in-house AI systems: 1 top performing model submitted to the DREAM challenge, 1 ensemble method of the eight best performing models (CEM), CEM combined with reader decision (single reader or consensus reading); threshold: corresponding to the sensitivity of single and consensus reading, respectively	Original radiologist decision (Sweden); (1) single reader (R1; R2), (2) consensus reading; no threshold	Cancer: tissue diagnosis within 12 months of screening; non-cancer: no cancer diagnosis ≥12 months after screening

AI=artificial intelligence; BI-RADS=breast imaging reporting and data system; CEM=challenge ensemble method; DREAM=Dialogue on Reverse Engineering Assessment and Methods; IQR=interquartile range; MQSA=Mammography Quality Standards Act; MRMC=multiple reader multiple case; NR=not reported; R1=first reader; R2=second reader; SD=standard deviation.

Table 2 | Summary of study characteristics for studies using AI for triage

Study	Study design	Population	Mammography vendor	Index test	Comparator	Reference standard
Balta 2020[25]	Retrospective cohort study (accuracy of classifying into low and high risk categories)	17 895 consecutively acquired screening examinations from 1 centre in Germany (0.64% screen detected cancer), age NR	Siemens 70% Hologic 30%	Transpara version 1.6.0 (Screenpoint Medical BV, Nijmegen, Netherlands); preselection of probably normal mammograms; Transpara risk score of 1-10, different cutoff points evaluated; optimal cutoff point ≤7: low risk	No comparator as human consensus reading decisions used as reference standard for screen negative results	Cancer: biopsy proven screen detected cancers; non-cancer: no information about follow-up for the normal examinations was available. available; for this review, a normal mammogram was defined as free of screen detected cancer based on human consensus reading
Dembrower 2020[26]	Retrospective case-control study (accuracy of classifying into low and high risk categories)	7364 women with screening examinations obtained during 2 consecutive screening rounds in 1 centre in Sweden (7.4% cancer: 347 screen detected in current round, 200 interval cancers within 30 months of previous screening round), median age 53.6 (IQR 47.6-63.0)	Hologic 100%	Lunit (Seoul, South Korea, version 5.5.0.16) (1) AI for preselection of mammograms probably normal, (2) AI as post-screen after negative double reading to recall women at highest risk of undetected cancer; AI risk score: decimal between 0 and 1, different cutoff points evaluated	None	Cancer: diagnosed with breast cancer at current screening round or within ≤30 months of previous screening round; non-cancer: >2 years' follow up
Lång 2021[27]	Retrospective cohort study (accuracy of classifying into low and high risk categories)	9581 women attending screening at 1 centre in Sweden, consecutive subcohort of Malmö Breast Tomosynthesis Screening Trial (0.71% screen detected cancers), mean age 57.6 (range 40-74)	Siemens 100%	Transpara version 1.4.0 (Screenpoint Medical BV, Nijmegen, Netherlands); preselection of mammograms probably normal; Transpara risk score of 1–10, different cutoff points evaluated; chosen cutoff point ≤5: low risk	No comparator as human consensus reading decisions used as reference standard for screen negative results	Cancer: histology of surgical specimen or core needle biopsies with a cross reference to a regional cancer register; non-cancer: a normal mammogram was defined as free of screen detected cancer based on human consensus reading
Raya-Povedano 2021[31]	Retrospective cohort study (accuracy of classifying into low and high risk categories)	15 986 consecutive women from the Córdoba Tomosynthesis Screening Trial, 1 Spanish centre (0.7% cancer: 98 screen detected (FFDM or DBT), 15 interval cancers within 24 months of screening); mean age 58 (SD 6), range 50-69 years	Hologic (Selenia Dimensions) 100%	Transpara, version 1.6.0 (ScreenPoint Medical BV, Nijmegen, Netherlands); preselection of mammograms probably normal; Transpara risk score of 1–10; cutoff point ≤7 low risk (chosen based on previous research by Balta 2020[35])	Original radiologist decision from Córdoba Tomosynthesis Screening Trial (double reading without consensus or arbitration)	Cancer: histopathologic results of biopsy, screen detected via FFDM or DBT and interval cancers within 24 months of screening; non-cancer: normal reading with 2-years' follow-up

AI=artificial intelligence; DBT=digital breast tomosynthesis; FFDM=full field digital mammography; IQR=interquartile range; NR=not reported; SD=standard deviation.

Table 3 | Summary of study characteristics for studies using AI as reader aid

Study	Study design	Population	Mammography vendor	Index test	Comparator	Reference standard
Pacilè 2020[30]	Enriched test set MRMC laboratory study, counterbalance design (accuracy of a read)	240 women from 1 US centre (50.0% cancer), mean age 59 (range 37-85)	NR	14 MQSA certified radiologists (US) with AI support (MammoScreen version 1, Therapixel, Nice, France); threshold: level of suspicion (0-100) >40	14 MQSA certified radiologists (US) without AI support, single reading; threshold: level of suspicion (0-100) >40	Cancer: histopathology; non-cancer: negative biopsy or negative result at follow-up for ≥18 months
Rodriguez-Ruiz 2019[32]	Enriched test set MRMC laboratory study, fully crossed (accuracy of a read)	240 women (120 from 1 US centre and 120 from 1 German centre; 41.7% cancer), median age 62 (range 39-89)	Hologic 50% Siemens 50%	14 MQSA certified radiologists (US) with AI support (Transpara version 1.3.0, Screenpoint Medical BV, Nijmegen, the Netherlands); threshold: BI-RADS score ≥3	14 MQSA certified radiologists (US) without AI support, single reading; threshold: BI-RADS score ≥3	Cancer: histopathology confirmed cancer; false positives: histopathologic evaluation or negative follow-up for ≥1 year; non-cancer: ≥1 year of negative follow-up
Watanabe 2019[37]	Enriched test set MRMC laboratory study, first without AI support, then AI aided (accuracy of a read)	122 women from 1 US centre (73.8% cancer, all false negative mammograms), mean age 65.4 (range 40-90)	NR	7 MQSA certified radiologists (US) with AI support (cmAssist· CureMetrix, Inc., La Jolla, CA); no threshold	7 MQSA certified radiologists (US) without AI support, single reading; no threshold	Cancer: biopsy proven cancer; non-cancer: BI-RADS 1 and 2 women with a 2 year follow-up of negative diagnosis

AI=artificial intelligence; BI-RADS=Breast Imaging-Reporting and Data System; MQSA=Mammography Quality Standards Act; MRMC=multireader multicase; NR=not reported.

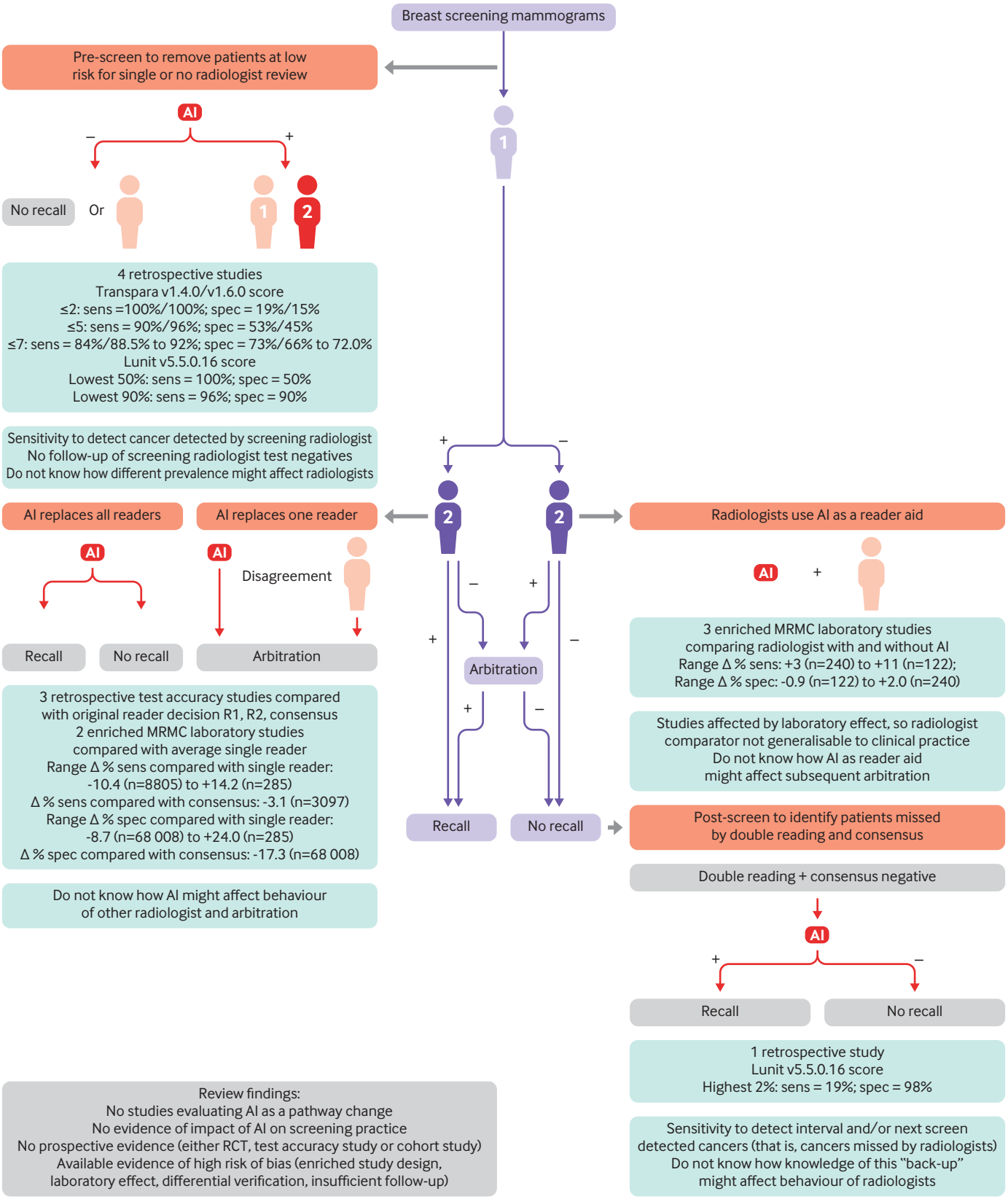

Fig 1 | Overview of published evidence in relation to proposed role in screening pathway. Purple shade=current pathway; orange shade=AI added to pathway; green shade=level of evidence for proposed AI role. AI=artificial intelligence; +/−=high/low risk of breast cancer, person icon=radiologist reading of mammograms as single, first, or second reader; MRMC=multiple reader multiple case; R1, R2=reader 1, reader 2; RCT, randomised controlled trial; sens=sensitivity; spec=specificity

appendix 2). Only four studies (albeit the four largest comprising 85% of all 131 822 women in the review) enrolled women consecutively or randomly, with a cancer prevalence of between 0.64% and 1.1%.[25 27 31 36] The remaining studies used enrichment leading to breast cancer prevalence (ranging from 7.4%[26] to 73.8%[37]), which is atypical of screening populations. Five studies[28 30 32 33 37] used reading under "laboratory" conditions at risk of introducing bias because radiologists read mammograms differently in a retrospective laboratory experiment than in clinical practice.[39] Only one of the studies used a prespecified test threshold which was internal to the AI system to classify mammographic images.[31]

The reference standard was at high (n=8) or unclear (n=3) risk of bias in 11/12 studies. Follow-up of screen negative women was less than two years in seven studies,[25-28 30 32 36] which might have resulted in underestimation of the number of missed cancers and overestimation of test accuracy.

Furthermore, in retrospective studies of routine data the choice of patient management (biopsy or follow-up) to confirm disease status was based on the decision of the original radiologist(s) but not on the decision of the AI system. Women classified as positive by AI who did not receive biopsy based on the original radiologists' decision only, received follow-up to confirm disease status. Therefore, cancers with a lead time from screen to symptomatic detection longer than the follow-up time in these studies will be misclassified as false positives for the AI test, and cancers which would have been overdiagnosed and overtreated after detection by AI would not be identified as such because the type of cancer that can indicate overdiagnosis, is unknown. The direction and magnitude of bias is complex and dependent on the positive and negative concordance between AI and radiologists but is more likely to be in the direction of overestimation of sensitivity and underestimation of specificity.

The applicability to European or UK breast cancer screening programmes was low (fig 2). None of the studies described the accuracy of AI integrated into a clinical breast screening pathway or evaluated the accuracy of AI prospectively in clinical practice in any country. Only two studies compared AI performance with the decision from human consensus reading.[35 36] The studies included only interval cancers within 12 months of screening, which is not typical for screening

Study reference	Risk of bias					Applicability concerns			
	Patient selection	Index test	Comparator test	Reference standard	Flow and timing	Patient selection	Index test	Comparator test	Reference standard
Standalone AI systems (5 studies)									
Lotter 2021[28]	High	High	High	Unclear	Unclear	High	High	High	High
McKinney 2020[29]	High	High	Low	High	High	High	High	High	Low
Rodriguez-Ruiz 2019[33]	High	High	High	Unclear	Unclear	High	High	High	Unclear
Salim 2020[35]	High	High	Low	High	High	High	High	Low* / High*	High
Schaffter 2020[36]	Low	High	Low	High	High	Unclear	High	Low* / High*	High
AI as reader aid (3 studies)									
Pacilè 2020[30]	High	High	High	High	Unclear	High	High	High	High
Rodriguez-Ruiz 2019[32,34]	High	High	High	High	Unclear	High	High	High	High
Watanabe 2019[37]	High	High	High	Unclear	Unclear	High	High	High	Low
AI for triage (4 studies)									
Balta 2020[25]	Low	High	None	High	High	Low	High	None	High
Dembrower 2020[26]	High	High	None	Low† / High†	High	High	High	None	Low† / High†
Lång 2020[27]	Low	High	None	High	High	Low	High	None	High
Raya-Povedano 2021[31]	Low	Low	Low	Low	High	Low	High	High	Low

Fig 2 | Overview of concerns about risk of bias and applicability of included studies. *Low concerns about applicability for consensus reading; high concerns about applicability for single reading as comparator test. †Low concerns about risk of bias and low applicability for the previous screening round (biopsy proven cancer or at least two years' follow-up); high concerns about risk of bias and high applicability for the current screening round (biopsy-proven cancer but no follow-up of test negatives)

Table 4 | Summary of test accuracy outcomes

Study	Index test (manufacturer)/comparator	TP	FP	FN	TN	% Sensitivity (95% CI)	Δ % Sensitivity, P value or (95% CI)	% Specificity (95% CI)	Δ % Specificity, value or (95% CI)
Standalone AI (5 studies):									
Lotter 2021,[28] Index cancer	AI (in-house) at reader's specificity	126	51	5	103	96.2 (91.7 to 99.2)	+14.2, P<0.001	66.9	Set to be equal
	AI (in-house) at reader's sensitivity	107	14	24	140	82.0	Set to be equal	90.9 (84.9 to 96.1)	+24.0, P<0.001
	Comparator: average single reader†	NA	NA	NA	NA	82.0	—	66.9	
McKinney 2020[29]*	AI (in-house)	NR	NR	NR	NR	56.24	+8.1, P<0.001	84.29	+3.46, P=0.02
	Comparator: original single reader	NR	NR	NR	NR	48.1	—	80.83	—
Rodriguez-Ruiz 2019[33]	AI (Transpara version 1.4.0)	63	25	16	95	80 (70 to 90)	+3 (-6.2 to 12.6)	79 (73 to 86)	Set to be equal
	Comparator: average single reader§	NA	NA	NA	NA	77 (70 to 83)	—	79 (73 to 86)	—
Salim 2020[35]†	AI-1 (anonymised)	605	NR	NR	NR	81.9 (78.9 to 84.6)	See below	96.6 (96.5 to 96.7)	Set to be equal
	AI-2 (anonymised)	495	NR	NR	NR	67.0 (63.5 to 70.4)	−14.9 *v* AI-1 (P<0.001)	96.6 (96.5 to 96.7)	Set to be equal
	AI-3 (anonymised)	498	NR	NR	NR	67.4 (63.9 to 70.8)	−14.5 *v* AI-1 (P<0.001)	96.7 (96.6 to 96.8)	Set to be equal
	Comparator: original reader 1	572	NR	NR	NR	77.4 (74.2 to 80.4)	−4.5 *v* AI-1 (P=0.03)	96.6 (96.5 to 96.7)	—
	Comparator: original reader 2	592	NR	NR	NR	80.1 (77.0 to 82.9)	−1.8 *v* AI-1 (P=0.40)	97.2 (97.1 to 97.3)	+0.6 *v* AI-1 (NR)
	Comparator: original consensus reading	628	NR	NR	NR	85.0 (82.2 to 87.5)	+3.1 *v* AI-1 (P=0.11)	98.5 (98.4 to 98.6)	+1.9 *v* AI-1 (NR)
Schaffter 2020[36]‡	Top-performing AI (in-house)	NR	NR	NR	NR	77.1	Set to be equal	88	−8.7 *v* reader 1 (NR)
	Ensemble method (CEM; in-house)	NR	NR	NR	NR	77.1	Set to be equal	92.5	−4.2 *v* reader 1 (NR)
	Comparator: original reader 1	NR	NR	NR	NR	77.1	—	96.7 (96.6 to 96.8)	
Schaffter 2020[36]	Top-performing AI (in-house)	NR	NR	NR	NR	83.9	Set to be equal	81.2	−17.3 *v* consensus (NR)
	Comparator: original consensus reading	NR	NR	NR	NR	83.9	—	98.5	—
AI for triage pre-screen (4 studies):									
Balta 2020[25]	AI as pre-screen (Transpara version 1.6.0):								
	AI score ≤2: ~15% low risk	114	15 028	0	2754	100.0	NA	15.49	NA
	AI score ≤5: ~45% low risk	109	9791	5	7991	95.61	NA	44.94	NA
	AI score ≤7: ~65% low risk	105	6135	9	11647	92.11	NA	65.50	NA
Lång 2020[27]	AI as pre-screen (Transpara version 1.4.0):								
	AI score ≤2: ~19% low risk	68	7684	0	1829	100.0	NA	19.23	NA
	AI score ≤5: ~53% low risk	61	4438	7	5075	89.71	NA	53.35	NA
	AI score ≤7: ~73% low risk	57	2541	11	6972	83.82	NA	73.29	NA
Raya-Povedano 2021[31]	AI as pre-screen (Transpara version 1.6.0); AI score ≤7: ~72% low risk	100	4450	13	11424	88.5 (81.1 to 93.7)	NA	72.0 (71.3 to 72.7)	NA
Dembrower 2020[26]§	AI as pre-screen (Lunit version 5.5.0.16):								
	AI score ≤0.0293: 60% low risk¶	347	29 787	0	45 200	100.0	NA	60.28	NA
	AI score ≤0.0870: 80% low risk¶	338	14 729	9	60 258	97.41	NA	80.36	NA
AI for triage post-screen (1 study):									
Dembrower 2020[26]§	AI as post-screen (Lunit v5.5.0.16); prediction of interval cancers: AI score ≥0.5337: ~2% high risk	32	1413	168	73921	16	NA	98.12	NA
Dembrower 2020[26]§	AI as post-screen (Lunit version 5.5.0.16); prediction of interval and next round screen detected cancers: AI score ≥0.5337: ~2% high risk	103	1342	444	73645	19	NA	98.21	NA
AI as reader aid (3 studies):									
Pacilè 2020[30]	AI support§ (MammoScreen version 1)	NA	NA	NA	NA	69.1 (60.0 to 78.2)	+3.3, P=0.02	73.5 (65.6 to 81.5)	+1.0, P=0.63
	Comparator: average single reader**	NA	NA	NA	NA	65.8 (57.4 to 74.3)	—	72.5 (65.6 to 79.4)	—
Rodriguez-Ruiz 2019[32]	AI support (Transpara version 1.3.0)	86	29	14	111	86 (84 to 88)	+3, P=0.05	79 (77 to 81)	+2, P=0.06
	Comparator: average single reader	83	32	17	108	83 (81 to 85)	—	77 (75 to 79)	—

(*Continued*)

doi: 10.1136/bmj.n1872 | *BMJ* 2021;374:n1872 | the bmj

Table 4 | Continued

Study	Index test (manufacturer)/comparator	TP	FP	FN	TN	% Sensitivity (95% CI)	Δ % Sensitivity, P value or (95% CI)	% Specificity (95% CI)	Δ % Specificity, value or (95% CI)
Watanabe 2019[37]	AI support** (cmAssist)	NA	NA	NA	NA	62 (range 41 to 75)	+11, P=0.03	77.2	−0.9 (NR)
	Comparator: average single reader**	NA	NA	NA	NA	51 (range 25 to 71)	—	78.1	—

AI=artificial intelligence; CEM=challenge ensemble method of eight top performing AIs from DREAM challenge; CI=confidence interval; DREAM=Dialogue on Reverse Engineering Assessment and Methods; FN=false negatives; F=false positives; NA=not applicable; NR=not reported; TN=true negatives; TP=true positives.
*Inverse probability weighting: negative cases were upweighted to account for the spectrum enrichment of the study population. Patients associated with negative biopsies were downweighted by 0.64. Patients who were not biopsied were upweighted by 23.61.
†Applied an inverse probability weighted bootstrapping (1000 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.
‡In addition, the challenge ensemble method prediction was combined with the original radiologist assessment. At the first reader's sensitivity of 77.1%, CEM+reader 1 resulted in a specificity of 98.5% (95% confidence interval 98.4% to 98.6%), higher than the specificity of the first reader alone of 96.7% (95% confidence interval, 96.6% to 96.8%; P<0.001). At the consensus readers' sensitivity of 83.9%, CEM+consensus did not significantly improve the consensus interpretations alone (98.1% *v* 98.5% specificity, respectively). These simulated results of the hypothetical integration of AI with radiologists' decisions were excluded as they did not incorporate radiologist behaviour when AI is applied.
§Applied 11 times upsampling of the 6817 healthy women, resulting in 74 987 healthy women and a total simulated screening population of 75 534.
¶Specificity estimates not based on exact numbers; the numbers were calculated by reviewers from reported proportions applied to 75 334 women (347 screen detected cancers and 74 987 healthy women).
**In enriched test set multiple reader multiple case laboratory studies where the same images, there are considerable problems in summing 2x2 test data across readers.

programmes. No direct evidence is therefore available as to how AI might affect accuracy if integrated into breast screening practice.

Analysis

AI as a standalone system to replace radiologist(s)

No prospective test accuracy studies, randomised controlled trials, or cohort studies examined AI as a standalone system to replace radiologists. Test accuracy of the standalone AI systems and the human comparators from retrospective cohort studies is summarised in table 4. All point estimates of the accuracy of AI systems were inferior to those obtained by consensus of two radiologists in screening practice, with mixed results in comparison with a single radiologist (fig 3). Three studies compared AI accuracy with that of the original radiologist in clinical practice,[29 35 36] of which two were enriched with extra patients with cancer.

The DREAM challenge of 68 008 consecutive women from the Swedish screening programme found the specificity of the top performing AI system (by Therapixel in a competition between 31 AI systems evaluated in the competitive phase on the independent Swedish dataset) was inferior in comparison with the original first radiologist (88% *v* 96.7%) and inferior also in comparison with the original consensus decision (81% *v* 98.5%) when the AI threshold was set to match the first reader's sensitivity and the consensus of readers' sensitivity, respectively.[36] The specificity of an ensemble method of the eight top performing AI systems remained inferior to that of the original first radiologist (92.5% *v* 96.7%, P<0.001), even in the same dataset that was used to choose the top eight.

An enriched Swedish cohort study (which overlapped that of the DREAM challenge, n=8805, 8.4% cancer) used three commercially available AI systems with thresholds set to match the specificity of the original radiologists. The study found that one commercially available AI system had superior sensitivity (81.9%, P=0.03) and two had inferior sensitivity (67%, 67.4%) in comparison with the original first radiologist (77.4%).[35] All had inferior sensitivity in comparison with the original consensus decision (85%, P=0.11 for best AI system *v* consensus). The manufacturer and identity were not reported for any of the three AI systems.

An enriched retrospective cohort from the US (n=3097, 22.2% cancer) found the AI system outperformed the original single radiologist in sensitivity (56% *v* 48%, P<0.001) and specificity (84% *v* 81%, P=0.021), although absolute values for the radiologist were lower than those found in clinical practice in the US and Europe.[29] Two enriched test set multiple case multiple reader laboratory studies reported that AI outperformed an average single radiologist reading in a laboratory setting, but the generalisability to clinical practice is unclear.[28 33]

AI as a standalone system for triage

Four studies used the Transpara versions 1.4.0 and 1.6.0 and Lunit version 5.5.0.16 AI systems,

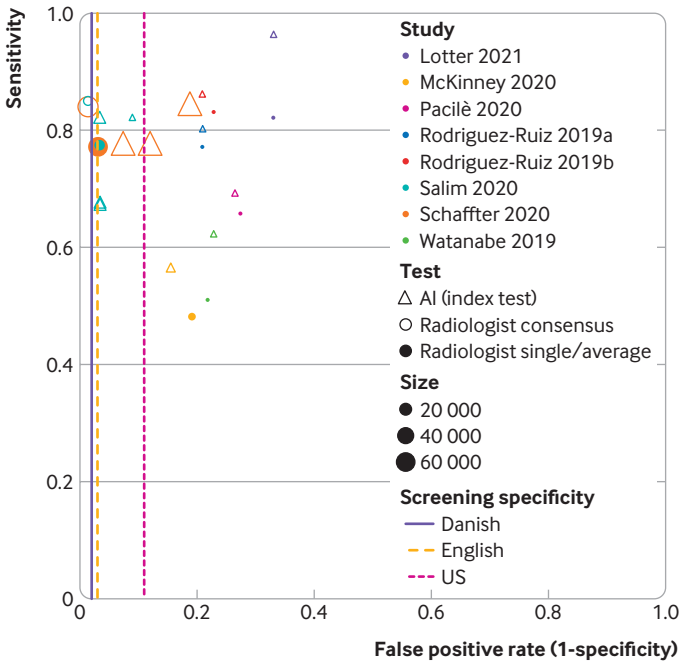

Fig 3 | Study estimates of sensitivity and false positive rate (1–specificity) in receiver operating characteristic space by index test (artificial intelligence) and comparator (radiologist) for eight included studies. Comparators are defined as consensus of two readers and arbitration (radiologist consensus), or single reader decision/average of multiple readers (radiologist single/average). iVertical dashed lines represent specificity for screening programmes for Denmark (2% false positive rate),[61] UK (3% false positive rate),[62 63] and US (11% false positive rate).[64] Retrospective test accuracy studies: Salim et al,[35] Schaffter et al,[36] and McKinney et al.[29] Enriched test set multiple reader multiple case laboratory studies: Pacilè et al,[30] Watanabe et al,[37] Rodriguez-Ruiz et al[33] (Rodriguez-Ruiz 2019a in figure), Lotter 2021,[28] and Rodriguez-Ruiz et al[32] (Rodriguez-Ruiz 2019b in figure)

respectively, as a pre-screen to identify women at low risk whose mammograms required less or no radiological review.[25-27 31] In this use, AI systems require high sensitivity so that few patients with cancer are excluded from radiological review, and only moderate specificity, which determines the radiology case load saved.

In a retrospective consecutive German cohort (n=17 895, 0.64% cancer) the Transpara version 1.6.0 AI system achieved a sensitivity of 92% and a specificity of 66% at the Transpara score 7 to remove patients at low risk from double reading, and 96% sensitivity (45% specificity) at a Transpara score of 5.[25] A Transpara version 1.4.0 score of 5 had 90% sensitivity and 53% specificity in a Swedish cohort (n=9581, 0.71% cancer).[27] Both studies reported 100% sensitivity at a score of 2 (and specificities of 15% and 19%, respectively). The threshold for classification (7[25] and 5[27]) was determined by exploring the full range of Transpara scores from 1 to 10 in the same dataset (fig 4A). In these studies, screen negative women were not followed up, so the sensitivity refers to detection of cancers which were detected by the original radiologists.

One study predefined the Transpara score of 7 to identify women at low risk in a Spanish cohort (n=15 986, 0.7% cancer, including 15 interval cancers within 24 months of follow-up) and achieved 88% sensitivity and 72% specificity.[31]

A Swedish case-control study (n=7364, 7.4% cancer) used a range of thresholds to consider use of the Lunit version 5.5.0.16 AI system as a pre-screen to remove normal patients (fig 4A) and then as a post-screen of patients who were negative after double reading to identify additional cancers (interval cancers and next round screen detected cancers; fig 4B).[26] Using 11 times upsampling of healthy women to simulate a screening population, they reported that use of AI alone with no subsequent radiologist assessment in the 50% and 90% of women with the lowest AI scores had 100% and 96% sensitivity and 50% and 90% specificity, respectively. AI assessment of negative mammograms after double reading detected 103 (19%) of 547 interval and next round screen detected cancers if the 2% women with the highest AI scores were post-screened (with a hypothetical perfect follow-up test).[26]

None of these studies reported any empirical data on the effect on radiologist behaviour of integrating AI into the screening pathway.

AI as a reader aid

No randomised controlled trials, test accuracy studies, or cohort studies evaluated AI as a reader aid in clinical practice. The only three studies of AI as a reader aid reported accuracy of radiologists' reading of an enriched test set in a laboratory environment, with limited generalisability to clinical practice. Sensitivity and specificity were reported as an average of 14,[30] 14,[32] or 7[37] radiologists with and without the AI reader aid. Point estimates of the average sensitivity were higher for radiologists with AI support than for unaided reading (absolute difference +3.0%, P=0.046,[32] +3.3%, P=0.021,[30] and +11%, P=0.030[37]) in all three studies of 240,[30] 240,[32] and 122[37] women. The effect of AI support on average reader specificity in a laboratory setting was small (absolute difference +2.0%, P=0.06,[32] +1.0%, P=0.63,[30] and −0.9%,[37] no P value reported; table 4).

Cancer type

Limited data were reported on types of cancer detected, with some evidence of systematic differences between different AI systems. Of the three retrospective cohort studies investigating AI as a standalone system to replace radiologist(s), only one reported measuring whether there was a difference between AI and radiologists in the type of cancer detected. One anonymised AI system detected more invasive cancers (82.8%) than a radiologist (radiologist 1: 76.7%; radiologist 2: 79.7%, n=640) and less ductal carcinoma in situ (83.5%) than a radiologist (radiologist 1: 89.4%; radiologist 2: 89.4%, n=85), though the grades of ductal carcinoma in situ and invasive cancer were not reported.[35] This same AI system detected more stage 2 or higher invasive cancers (n=204, 78.4% than radiologist 1: 68.1% and radiologist 2: 68.1%).[35] The other two anonymised AI systems detected fewer stage

doi: 10.1136/bmj.n1872 | *BMJ* 2021;374:n1872 | the bmj

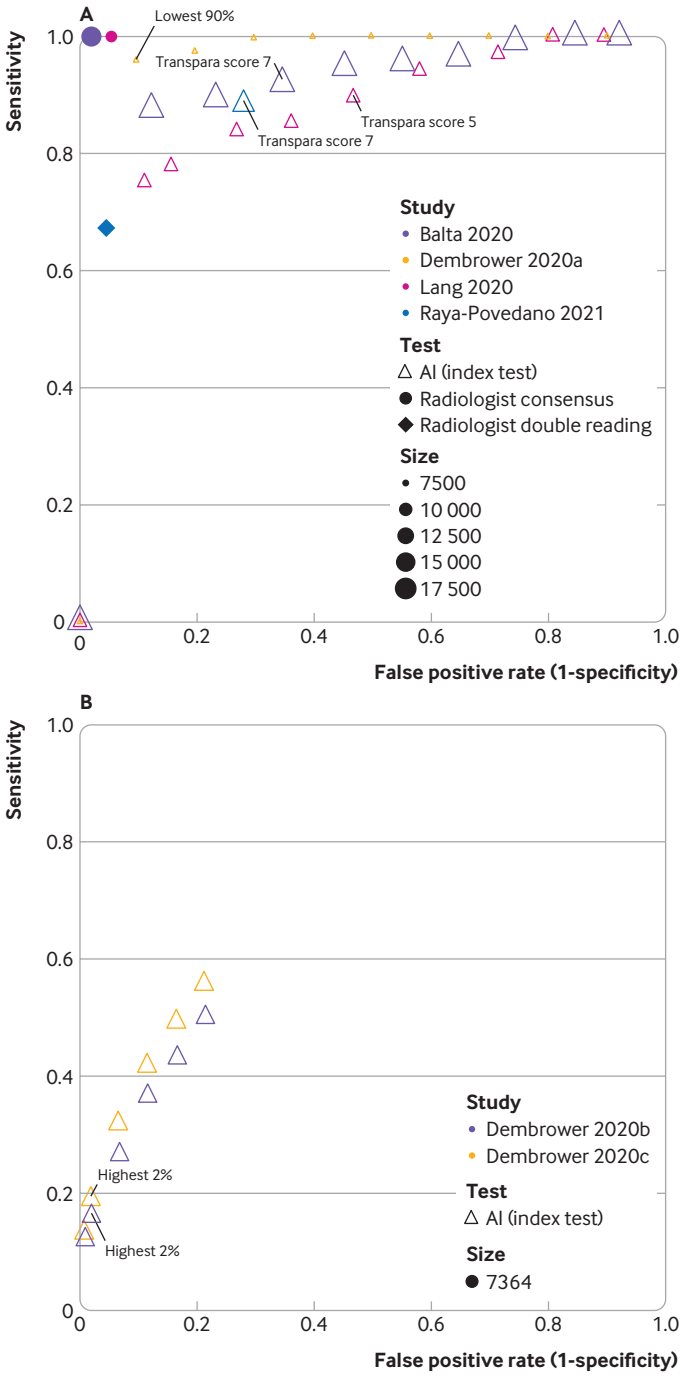

Fig 4 | Study estimates of sensitivity and false positive rate (1–specificity) in receiver operating characteristic space for studies of artificial intelligence (AI) as a pre-screen (A) or post-screen (B). Pre-screen requires very high sensitivity, but can have modest specificity, post-screen requires very high specificity, but can have modest sensitivity. Reference standard for test negatives was double reading not follow-up. (A) Dembrower 2020a: retrospective study using AI (Lunit version 5.5.0.16) for pre-screen (point estimates not based on exact numbers). Reference standard includes only screen detected cancers. No data reported for radiologists.[26] Balta 2020 (Transpara version 1.6.0),[25] Raya-Povedano 2021 (Transpara version 1.6.0),[31] and Lång 2020 (Transpara version 1.4.0)[27]: retrospective studies using AI as pre-screen. Reference standard includes only screen detected cancers. (B) Dembrower 2020b: retrospective study using AI (Lunit version 5.5.0.16) for post-screen detection of interval cancers,[26] Dembrower 2020c: retrospective study using AI (Lunit version 5.5.0.16) for post-screen detection of interval cancers and next round screen detected cancers.[26] Thresholds highlighted represent thresholds specified in studies. Radiologist double reading for this cohort would be 100% specificity and 0% sensitivity as this was only in a cohort of women with screen (true and false) negative mammograms

2 or higher invasive cancers (58.3% and 60.8%) than the radiologists.

In an enriched test set multiple reader multiple case laboratory study, a standalone in-house AI model (DeepHealth Inc.) detected more invasive cancer (+12.7%, 95% confidence interval 8.5 to 16.5) and more ductal carcinoma in situ (+16.3%, 95% confidence interval 10.9 to 22.2) than the average single reader.[28] This trend for higher performance of the AI model was also seen for lesion type, cancer size, and breast density.

In an enriched test set multiple reader multiple case laboratory study, addition of the CureMetrix AI system to assist readers increased detection of microcalcifications (n=17,+20%) preferentially in comparison with other mammographic abnormalities such as masses (n=73,+9%).[37] Microcalcifications are known to be more associated with ductal carcinoma in situ than with invasive cancer, but the spectrum of disease was not directly reported.

Forty seven (87%) of 54 screen detected invasive cancers were classified as high risk using Transpara version 1.4.0 with a threshold of 5, in comparison with 14 (100%) of 14 microcalcifications.[27] Using Transpara version 1.6.0 with a threshold of 7 as pre-screen, four additional cancers were classified as high risk by AI that had been missed by original double reading without consensus (two ductal carcinoma in situ, one low grade invasive ductal cancer, and one high grade invasive ductal cancer).[31] No information on cancer type was reported for the two screen detected cancers that were classed by AI as low risk.

Discussion

Main findings

In this systematic review of AI mammographic systems for image analysis in routine breast screening, we identified 12 studies which evaluated commercially available or in-house convolutional neural network AI systems, of which nine included a comparison with radiologists. One of the studies reported that they followed STARD reporting guidelines.[36] The six smallest studies (total 4183 women) found that AI was more accurate than single radiologists.[28-30 32 33 37] The radiologists in five of six of these studies were examining the mammographic images of 932 women in a laboratory setting, which is not generalisable to clinical practice. In the remaining study, the comparison was with a single reading in the US with an accuracy below that expected in usual clinical practice.[29] Whether this lower accuracy was due to case mix or radiologist expertise is unclear. In two of the largest retrospective cohort studies of AI to replace radiologists in Europe (n=76 813 women),[35 36] all AI systems were less accurate than consensus of two radiologists, and 34 of 36 AI systems were less accurate than a single reader. One unpublished study is in line with these findings.[40] This large retrospective study (n=275 900 women) reported higher sensitivity of AI in comparison with the original first reader decision but lower specificity, and the AI system was less

accurate than consensus reading.[40] Four retrospective studies[25-27 31] indicated that at lower thresholds, AI can achieve high sensitivity so might be suitable for triaging which women should receive radiological review. Further research is required to determine the most appropriate threshold as the only study which prespecified the threshold for triage achieved 88.5% sensitivity.[31] Evidence suggests that the accuracy and spectrum of disease detected between different AI systems is variable.

Considerable heterogeneity in study methodology was found, some of which resulted in high concerns over risk of bias and applicability. Compared with consecutive sampling, case-control studies added bias by selecting cases and controls[41] to achieve an enriched sample. The resulting spectrum effect could not be assessed because studies did not adequately report the distribution of original radiological findings, such as the distribution of the original BI-RADS scores. The effect was likely to be greater, however, when selection was based on image or cancer characteristics rather than if enrichment was achieved by including all available women with cancer and a random sample of those who were negative.

The overlap of populations in three Swedish studies means that they represent only one rather than three separate cohorts.[26 35 36] Performance of the AI system might have been overestimated if the same AI system read the same dataset more than once and, therefore, could have had the opportunity to learn. We could not confirm this as the three AI systems used by Salim et al were anonymised.[35]

The included studies have some variation in reference standard for the definition of normal cases, from simply consensus decision of radiologists at screening, to one to three years of follow-up. This inconsistency means accuracy estimates are comparable within, but not between, studies. Overall, the current evidence is a long way from the quality and quantity required for implementation in clinical practice.

Strengths and limitations

We followed standard methodology for conducting systematic reviews, used stringent inclusion criteria, and tailored the quality assessment tool for included studies. The stringent inclusion criteria meant that we included only geographical validation of test sets in the review—that is, at different centres in the same or different countries, which resulted in exclusion of a large number of studies that used some form of internal validation (where the same dataset is used for training and validation—for example, using cross validation or bootstrapping). Internal validation overestimates accuracy and has limited generalisability,[42] and might also result in overfitting and loss of generalisability as the model fits the trained data extremely well but to the detriment of its ability to perform with new data. The split sample approach similarly does not accurately reflect a model's generalisability.[43]

Temporal validation is regarded as an approach that lies midway between internal and external validation[43] and has been reported by others to be sufficient in meeting the expectations of an external validation set to evaluate the effectiveness of AI.[42] For screening, however, temporal validation could introduce bias because, for instance, the same women might attend repeat screens, and be screened by the same personnel using the same machines. Only geographical validation offers the benefits of external validation and generalisability.[42]

We also excluded computer aided detection for breast screening using systems that were categorised as traditional. The definition was based on expert opinion and the literature.[14] The distinction is not clear cut and this approach might have excluded relevant studies that poorly reported the AI methods or used a combination of methods.

We extracted binary classifications from AI systems, and do not know how other information on a recall to assessment form from a radiologist, such as mammographic characteristics or BI-RADS score/level of suspicion, might affect the provision of follow-up tests. In addition, AI algorithms are short lived and constantly improve. Reported assessments of AI systems might be out of date by the time of study publication, and their assessments might not be applicable to AI systems available at the time.

The exclusion of non-English studies might have excluded relevant evidence. The available methodological evidence suggests that this is unlikely to have biased the results or affected the conclusions of our review.[44 45] Finally, 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.

Strengths and limitations in comparison with previous studies

The findings from our systematic review disagree with the publicity some studies have received and opinions published in various journals, which claim that AI systems outperform humans and might soon be used instead of experienced radiologists.[10-13] Our different conclusion is based on our rigorous and systematic evaluation of study quality. We did not extract the "simulation" parts of studies, which were often used as the headline numbers in the original papers, and often estimated higher accuracy for AI than the empirical data of the studies. In these simulations various assumptions were made about how radiologist arbitrators would behave in combination with AI, without any clinical data on behaviour in practice with AI. Although a great number of studies report the development and internal validation of AI systems for breast screening, our study shows that this high volume of published studies does not reflect commercially available AI systems suitable for integration into screening programmes. Our emphasis on comparisons with the accuracy of radiologists in clinical practice explains why our conclusions are more cautious than many of the included papers.

A recent scoping review with a similar research question, but broader scope, reported a potential role for AI in breast screening but identified evidence gaps that showed a lack of readiness of AI for breast screening programmes.[46] The 23 included studies were mainly small, retrospective, and used publicly available and institutional image datasets, which often overlapped. The evidence included only one study with a consecutive cohort, one study with a commercially available AI system, and five studies that compared AI with radiologists. We found overlap of only one study between the scoping review and our review despite the same search start date, probably because we focused on higher study quality. Our review identified nine additional recent eligible studies, which might suggest that the quality of evidence is improving, but as yet no prospective evaluations of AI have been reported in clinical practice settings.

Possible explanations and implications for clinicians and policy makers

Our systematic review should be considered in the wider context of the increasing proposed use of AI in healthcare and screening. Most of the literature focuses, understandably, on those screening programmes in which image recognition and interpretation are central components, and this is indicated by a number of reviews recently published describing studies of AI and deep learning for diabetic retinopathy screening.[47 48] Beyond conventional screening programmes, the use of deep learning in medicine is increasing, and has been considered in the diagnosis of melanoma,[49] ophthalmic diseases (age-related macular degeneration[50] and glaucoma[51]), and in interpretation of histological,[52] radiological,[53] and electrocardiogram[54] images.

Evidence is insufficient on the accuracy or clinical effect of introducing AI to examine mammograms anywhere on the screening pathway. It is not yet clear where on the clinical pathway AI might be of most benefit, but its use to redesign the pathway with AI complementing rather than competing with radiologists is a potentially promising way forward. Examples of this include using AI to pre-screen easy normal mammograms for no further review, and post-screen for missed cases. Similarly, in diabetic eye screening there is growing evidence that AI can filter which images need to be viewed by a human grader, and which can be reported as normal immediately to the woman.[55 56] Medical decisions made by AI independently of humans might have medicolegal implications.[57 58]

Implications for research

Prospective research is required to measure the effect of AI in clinical practice. Although the retrospective comparative test accuracy studies, which compared AI performance with the original decision of the radiologist, have the advantage of not being biased by the laboratory effect, the readers were "gatekeepers" for biopsy. This means that we do not know the true cancer status of women whose mammograms were AI positive and radiologist negative. Examination of follow-up to interval cancers does not fully resolve this problem of true cancer status, as lead times to symptomatic presentation are often longer than the study follow-up time. Prospective studies can answer this question by recalling for further assessment women whose mammograms test positive by AI or radiologist. Additionally, evidence is needed on the types of cancer detected by AI to allow an assessment of potential changes to the balance of benefits and harms, including potential overdiagnosis. We need evidence for specific subgroups according to age, breast density, prior breast cancer, and breast implants. Evidence is also needed on radiologist views and understanding and on how radiologist arbitrators behave in combination with AI.

Finally, evidence is needed on the direct comparison of different AI systems; the effect of different mammogram machines on the accuracy of AI systems; the effect of differences in screening programmes on cancer detection with AI, or on how the AI system might work within specific breast screening IT systems; and the effect of making available additional information to AI systems for decision making. Commercially available AI systems should not be anonymised in research papers, as this makes the data useless for clinical and policy decision makers. The most applicable evidence to answer this question would come from prospective comparative studies in which the index test is the AI system integrated into the screening pathway, as it would be used in screening practice. These studies would need to report the change to the whole screening pathway when AI is added as a second reader, as the only reader, as a pre-screen, or as a reader aid. No studies of this type or prospective studies of test accuracy in clinical practice were available for this review. We did identify two ongoing randomised controlled trials, however: one investigating AI as pre-screen with the replacement of double reading for women at low risk with single reading (randomising to AI integrated mammography screening *v* conventional mammography screening), and one investigating AI as a post-screen (randomising women with the highest probability of having had a false negative screening mammogram to MRI or standard of care.)[59 60]

Conclusions

Current evidence on the use of AI systems in breast cancer screening is a long way from having the quality and quantity required for its implementation into clinical practice. Well designed comparative test accuracy studies, randomised controlled trials, and cohort studies in large screening populations are needed which evaluate commercially available AI systems in combination with radiologists. Such studies will enable an understanding of potential changes to the performance of breast screening programmes with an integrated AI system. By highlighting the shortcomings, we hope to encourage future users,

commissioners, and other decision makers to press for high quality evidence on test accuracy when considering the future integration of AI into breast cancer screening programmes.

The views expressed are those of the authors and not necessarily those of the UK National Screening Committee, National Institute for Health Research, or Department of Health and Social Care.

Contributors: KF, JG, SJ, and CS undertook the review. SJ devised and managed the search strategy in discussion with the other authors. KF, JG, CS, DT, AC, ST-P contributed to the conception of the work and interpretation of the findings. KF drafted the manuscript. All authors critically revised the manuscript and approved the final version. ST-P takes responsibility for the integrity and accuracy of the data analysis. STP acts as guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: This study was funded by the UK National Screening Committee. The funder had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: CS, ST-P, KF, JG, and AC have received funding from the UK National Screening Committee for the conduct of the review; ST-P is funded by the National Institute for Health Research (NIHR) through a career development fellowship; AC is partly supported by the NIHR Applied Research Collaboration West Midlands; SJ and DT have nothing to declare; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: No additional data available.

The lead author and manuscript's guarantor (ST-P) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Dissemination to participants and related patient and public communities: The results will be discussed with patient contributors.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Fitzmaurice C, Allen C, Barber RM, et al, Global Burden of Disease Cancer Collaboration. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol* 2017;3:524-48. doi:1001/jamaoncol.2016.5688
2 Youlden DR, Cramb SM, Dunn NA, Muller JM, Pyke CM, Baade PD. The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. *Cancer Epidemiol* 2012;36:237-48. doi:1016/j.canep.2012.02.007
3 van Luijt PA, Heijnsdijk EA, Fracheboud J, et al. The distribution of ductal carcinoma in situ (DCIS) grade in 4232 women and its impact on overdiagnosis in breast cancer screening. *Breast Cancer Res* 2016;18:47. doi:1186/s13058-016-0705-5
4 Yen MF, Tabár L, Vitak B, Smith RA, Chen HH, Duffy SW. Quantifying the potential problem of overdiagnosis of ductal carcinoma in situ in breast cancer screening. *Eur J Cancer* 2003;39:1746-54. doi:1016/S0959-8049(03)00260-0
5 Tabar L, Chen TH, Yen AM, et al. Effect of Mammography Screening on Mortality by Histological Grade. *Cancer Epidemiol Biomarkers Prev* 2018;27:154-7. doi:1158/1055-9965.EPI-17-0487
6 Baines CJ, Miller AB, Wall C, et al. Sensitivity and specificity of first screen mammography in the Canadian National Breast Screening Study: a preliminary report from five centers. *Radiology* 1986;160:295-8. doi:1148/radiology.160.2.3523590
7 Houssami N, Macaskill P, Bernardi D, et al. Breast screening using 2D-mammography or integrating digital breast tomosynthesis (3D-mammography) for single-reading or double-reading--evidence to guide future screening strategies. *Eur J Cancer* 2014;50:1799-807. doi:1016/j.ejca.2014.03.017
8 Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer* 2017;3:12. doi:1038/s41523-017-0014-x
9 AI Index Steering Committee. *The AI Index 2021 Annual Report* . Human-Centered AI Institute, Stanford University, 2021.
10 Dustler M. Evaluating AI in breast cancer screening: a complex task. *Lancet Digit Health* 2020;2:e106-7. doi:1016/S2589-7500(20)30019-4
11 Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375:1216-9. doi:1056/NEJMp1606181
12 Chockley K, Emanuel E. The End of Radiology? Three Threats to the Future Practice of Radiology. *J Am Coll Radiol* 2016;13(12 Pt A):1415-20. doi:1016/j.jacr.2016.07.010
13 Stower H. AI for breast-cancer screening. *Nat Med* 2020;26:163. doi:1038/s41591-020-0776-9
14 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018;18:500-10. doi:1038/s41568-018-0016-5
15 McDonald RJ, Schwartz KM, Eckel LJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad Radiol* 2015;22:1191-8. doi:1016/j.acra.2015.05.007
16 Thomas ET, Del Mar C, Glasziou P, Wright G, Barratt A, Bell KJL. Prevalence of incidental breast cancer and precursor lesions in autopsy studies: a systematic review and meta-analysis. *BMC Cancer* 2017;17:808. doi:1186/s12885-017-3808-1
17 Castelvecchi D. Can we open the black box of AI?*Nature* 2016;538:20-3. doi:1038/538020a
18 Geirhos R, Jacobsen J-H, Michaelis C, et al. Shortcut learning in deep neural networks. *Nat Mach Intell* 2020;2:665-73. doi:1038/s42256-020-00257-z
19 DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal.*medRxiv* 2020. doi:10.1101/2020.09.13.20193565
20 McInnes MDF, Moher D, Thombs BD, et al, the PRISMA-DTA Group. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* 2018;319:388-96. doi:1001/jama.2017.19163
21 Whiting PF, Rutjes AW, Westwood ME, et al, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529-36. doi:7326/0003-4819-155-8-201110180-00009
22 Niraula S, Biswanger N, Hu P, Lambert P, Decker K. Incidence, Characteristics, and Outcomes of Interval Breast Cancers Compared With Screening-Detected Breast Cancers. *JAMA Netw Open* 2020;3:e2018179. doi:1001/jamanetworkopen.2020.18179
23 Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag 2016.
24 *R: A language and environment for statistical computing.* [program] R Foundation for Statistical Computing, 2017.
25 Balta C, Rodriguez-Ruiz A, Mieskes C, Karssemeijer N, Heywang-Köbrunner SH. Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact? *Proc SPIE* 2020;11513:115130D
26 Dembrower K, Wåhlin 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 Digit Health* 2020;2:e468-74. doi:1016/S2589-7500(20)30185-0
27 Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol* 2021;31:1687-92. doi:1007/s00330-020-07165-1
28 Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med* 2021;27:244-9. doi:1038/s41591-020-01174-9
29 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89-94. doi:1038/s41586-019-1799-6
30 Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. *Radiol Artif Intell* 2020;2:e190208. doi:1148/ryai.2020190208
31 Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. *Radiology* 2021;300:57-65. doi:1148/radiol.2021203555
32 Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* 2019;290:305-14. doi:1148/radiol.2018181371

33 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J Natl Cancer Inst* 2019;111:916-22. doi:1093/jnci/djy222
34 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?*Proc SPIE* 2018;10718:1071803. doi:1117/12.2317937
35 Salim M, Wåhlin E, Dembrower K, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol* 2020;6:1581-8. doi:1001/jamaoncol.2020.3321
36 Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw* 2020;3:e200265. doi:1001/jamanetworkopen.2020.0265
37 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;32:625-37. doi:1007/s10278-019-00192-5
38 Hupse R, Samulski M, Lobbes M, et al. Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *Eur Radiol* 2013;23:93-100. doi:1007/s00330-012-2562-7
39 Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008;249:47-53. doi:1148/radiol.2491072025
40 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:1101/2021.02.26.21252537
41 Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005;51:1335-41. doi:1373/clinchem.2005.048595
42 Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* 2018;286:800-9. doi:1148/radiol.2017171920
43 Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where?*Clin Kidney J* 2020;14:49-58. doi:1093/ckj/sfaa188
44 Nussbaumer-Streit B, Klerings I, Dobrescu AI, et al. Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. *J Clin Epidemiol* 2020;118:42-54. doi:1016/j.jclinepi.2019.10.011
45 Morrison A, Polisena J, Husereau D, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *Int J Technol Assess Health Care* 2012;28:138-44. doi:1017/S0266462312000086
46 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;16:351-62. doi:1080/17434440.2019.1610387
47 Nielsen KB, Lautrup ML, Andersen JKH, Savarimuthu TR, Grauslund J. Deep Learning-Based Algorithms in Screening of Diabetic Retinopathy: A Systematic Review of Diagnostic Performance. *Ophthalmol Retina* 2019;3:294-304. doi:1016/j.oret.2018.10.014
48 Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. *Eye (Lond)* 2020;34:451-60. doi:1038/s41433-019-0566-0
49 Codella NCF, Nguyen Q, Pankanti S, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM J Res Dev* 2017;61:5:1-5:15. doi:1147/JRD.2017.2708299
50 Lee CS, Baughman DM, Lee AY. Deep learning is effective for the classification of OCT images of normal versus Age-related Macular Degeneration. *Ophthalmol Retina* 2017;1:322-7. doi:1016/j.oret.2016.12.009
51 Xiangyu Chen , Yanwu Xu , Jiang Liu, Damon Wing Kee Wong, Tien Yin Wong. Glaucoma detection based on deep convolutional neural network. *Annu Int Conf IEEE Eng Med Biol Soc* 2015;2015:715-8. doi:1109/embc.2015.7318462
52 Sirinukunwattana K, Ahmed Raza SE, Yee-Wah Tsang , Snead DR, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Trans Med Imaging* 2016;35:1196-206. doi:1109/TMI.2016.2525803
53 McBee MP, Awan OA, Colucci AT, et al. Deep Learning in Radiology. *Acad Radiol* 2018;25:1472-80. doi:1016/j.acra.2018.02.018
54 Pourbabaee B, Roshtkhari MJ, Khorasani K. Deep Convolutional Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillation Patients. *IEEE Trans Syst Man Cybern Syst* 2018;48:2095-104. doi:1109/TSMC.2017.2705582
55 Tufail A, Rudisill C, Egan C, et al. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. *Ophthalmology* 2017;124:343-51. doi:1016/j.ophtha.2016.11.014
56 Heydon P, Egan C, Bolter L, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *Br J Ophthalmol* 2021;105:723-8. doi:1136/bjophthalmol-2020-316594
57 Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast* 2020;49:25-32. doi:1016/j.breast.2019.10.001
58 Gerke S, Minssen T, Cohen G. *Ethical and legal challenges of artificial intelligence-driven healthcare.* Artificial Intelligence in Healthcare, 2020: 295-336. doi:1016/B978-0-12-818438-7.00012-5
59 NCT. Mammography Screening With Artificial Intelligence (MASAI). https://clinicaltrials.gov/ct2/show/NCT04838756. 2021.
60 NCT. Using AI to Select Women for Supplemental MRI in Breast Cancer Screening. https://clinicaltrials.gov/ct2/show/NCT04832594. 2021.
61 Lynge E, Bak M, von Euler-Chelpin M, et al. Outcome of breast cancer screening in Denmark. *BMC Cancer* 2017;17:897. doi:1186/s12885-017-3929-6
62 Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. *Br J Cancer* 2011;104:571-7. doi:1038/bjc.2011.3
63 Screening and Immunisations Team. Breast Screening Programme: England, 2019-20: National Statistics; NHS Digital. 2021. https://files.digital.nhs.uk/F9/98C8E3/breast-screening-programme-eng-2019-20-report.pdf.
64 Breast Cancer Surveillance Consortium. Screening Mammography Sensitivity, Specificity & False Negative Rate. 2017. https://www.bcsc-research.org/statistics/screening-performance-benchmarks/screening-sens-spec-false-negative.

Web appendix: Supplementary appendices