

Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis



Sarah E. Hickman, MBBS • Ramona Woitek, MD, PhD • Elizabeth Phuong Vi Le, MA • Yu Ri Im, MB BChir • Carina Mouritsen Luxboj, MA, MSci • Angelica I. Aviles-Rivero, BSc, MSc, PhD • Gabrielle C. Baxter, PhD • James W. MacKay, MA, MB BChir, PhD • Fiona J. Gilbert, MD

From the Department of Radiology (S.E.H., R.W., G.C.B., J.W.M., F.J.G.) and Department of Medicine (E.P.V.L., Y.R.I., C.M.L.), University of Cambridge School of Clinical Medicine, Box 218, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, England; Department of Radiology, Addenbrooke's Hospital, Cambridge University Hospitals National Health Service Foundation Trust, Cambridge, England (R.W., F.J.G.); Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria (R.W.); Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, England (A.I.A.R.); and Norwich Medical School, University of East Anglia, Norwich, England (J.W.M.). Received February 15, 2021; revision requested April 26; revision received July 14; accepted August 5. **Address correspondence to** F.J.G. (e-mail: fjg28@cam.ac.uk).

Supported by the National Institute for Health Research Cambridge Biomedical Research Centre and Cancer Research UK grant (grant no. C543/A26884). Cancer Research UK funds the PhD studentship for S.E.H. through an Early Detection program grant (grant no. C543/A26884). G.C.B. is funded by a studentship from GE Healthcare. R.W. is funded by the Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre (grant no. C9685/A25177).

Conflicts of interest are listed at the end of this article.

See also the editorial by Whitman and Moseley in this issue.

Radiology 2022; 302:88–104 • <https://doi.org/10.1148/radiol.2021210391> • Content codes:  

Background: Advances in computer processing and improvements in data availability have led to the development of machine learning (ML) techniques for mammographic imaging.

Purpose: To evaluate the reported performance of stand-alone ML applications for screening mammography workflow.

Materials and Methods: Ovid Embase, Ovid Medline, Cochrane Central Register of Controlled Trials, Scopus, and Web of Science literature databases were searched for relevant studies published from January 2012 to September 2020. The study was registered with the PROSPERO International Prospective Register of Systematic Reviews (protocol no. CRD42019156016). Stand-alone technology was defined as a ML algorithm that can be used independently of a human reader. Studies were quality assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 and the Prediction Model Risk of Bias Assessment Tool, and reporting was evaluated using the Checklist for Artificial Intelligence in Medical Imaging. A primary meta-analysis included the top-performing algorithm and corresponding reader performance from which pooled summary estimates for the area under the receiver operating characteristic curve (AUC) were calculated using a bivariate model.

Results: Fourteen articles were included, which detailed 15 studies for stand-alone detection ($n = 8$) and triage ($n = 7$). Triage studies reported that 17%–91% of normal mammograms identified could be read by adapted screening, while “missing” an estimated 0%–7% of cancers. In total, an estimated 185 252 cases from three countries with more than 39 readers were included in the primary meta-analysis. The pooled sensitivity, specificity, and AUC was 75.4% (95% CI: 65.6, 83.2; $P = .11$), 90.6% (95% CI: 82.9, 95.0; $P = .40$), and 0.89 (95% CI: 0.84, 0.98), respectively, for algorithms, and 73.0% (95% CI: 60.7, 82.6), 88.6% (95% CI: 72.4, 95.8), and 0.85 (95% CI: 0.78, 0.97), respectively, for readers.

Conclusion: Machine learning (ML) algorithms that demonstrate a stand-alone application in mammographic screening workflows achieve or even exceed human reader detection performance and improve efficiency. However, this evidence is from a small number of retrospective studies. Therefore, further rigorous independent external prospective testing of ML algorithms to assess performance at preassigned thresholds is required to support these claims.

©RSNA, 2021

Online supplemental material is available for this article.

There are now more than five U.S. Food and Drug Administration–approved algorithms for mammographic interpretation, primarily to be used as clinical decision support systems (1). Research has demonstrated that these machine learning (ML) computer-aided detection (CAD) algorithms can reach and even exceed clinician performance, providing an independent definitive output (ie, case-level decision) on two-dimensional standard-view mammogram (ie, mediolateral oblique and craniocaudal) data (Fig 1) (2,3). This could allow for ML stand-alone CAD and computer-aided diagnosis (CADx), or, when ML algorithms are set at a high sensitivity, for the automated case-based computer-aided triage (CADt) of mammograms within the screen reading workflow (4).

Many countries have implemented breast screening to detect cancer at an earlier stage, albeit with differing screening processes, such as single reading in the United States and double reading in many European countries, with screening starting at varied ages (40–50 years) and differing intervals between screening (annual, biennial, and triennial) (5–8). Mammography remains the most common imaging modality used, although its cost-effectiveness is debated because of false-positive findings, overdiagnosis, and false-negative findings (ie, interval cancers) (9,10). Human readers—for example, radiologists and reporting radiographers in the United Kingdom—are under increasing pressure because of increasing workloads, demands from busy clinics, strict

Abbreviations

AUC = area under the receiver operating characteristic curve, CAD = computer-aided detection, CADt = computer-aided triage, CADx = computer-aided diagnosis, ML = machine learning

Summary

Retrospective studies demonstrate the performance of stand-alone machine learning applications in screening mammography can reach reader performance and can provide a mechanism for case triage, which merits investigation with prospective studies.

Key Results

- Seven retrospective studies suggested that machine learning (ML) could be used to reduce the number of mammography examinations read by radiologists by 17%–91% while “missing” 0%–7% of cancers.
- A meta-analysis of five retrospective mammography breast cancer detection studies with 185 252 cases demonstrated a higher area under the receiver operating characteristic curve (AUC) for ML (AUC = .89) compared to readers (AUC = .85).
- The mean Checklist for Artificial Intelligence in Medical Imaging score was 30 of 42 (71%); ML model explainability methods were underreported.

screening program targets, and staff shortages (11). Alternatives to double reading of mammograms are being sought to further alleviate pressure, including single reading using CAD prompts, stand-alone ML algorithms with a second reader, or CADt with various reader combinations (2).

Studies investigating the use of traditional CAD mammography systems demonstrated no significant improvement in reader performance, and although sensitivity was similar to that of double reading, given the increase in recall rates, these systems were deemed not cost-effective (12,13). Additional limitations of traditional CAD systems include high rates of false-positive prompts, limited reproducibility of prompts, increased reading time, and a CAD preference for calcification detection over soft-tissue masses and architectural distortion (14,15). Traditional CAD systems were trained using handcrafted features extracted from human delineations. The latest ML methods can use pre-trained deep learning networks and automatically delineated cancer regions by means of iterative interactive software to rely upon learned features, and they have the potential to overcome the limitations of traditional CAD systems. However, how these new ML systems should be used in real-time workflows is still unclear. One route could be to improve efficiency of the workflow by operating as stand-alone systems. Although the performance expected by such stand-alone ML applications in a screening workflow is yet to be agreed upon, a system should meet a “clinically relevant threshold” (16). In general, recall rates should not be increased because of the huge impact on workload, thus algorithms with a lower specificity would require human intervention to reduce recalls (16,17). Therefore, making a definitive decision on whether current systems reach the standard required for routine workflow use is challenging.

We conducted a systematic review and meta-analysis to investigate whether or not ML algorithms (ie, CAD and CADx) are as sensitive and specific as radiologists in detecting breast cancer in patients undergoing screening mammography. In addition,

we evaluated the application of stand-alone ML algorithms (ie, CADt) used in breast cancer screening for mammography interpretation and the impact of ML algorithms if adopted into clinical practice. Furthermore, we aimed to identify areas of bias and gaps in the reported evidence. Appendix E1 (online) contains a glossary of terms.

Materials and Methods

This systematic review and meta-analysis was reported in accordance with the Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies guidance (18). The review protocol was registered with the PROSPERO International Prospective Register of Systematic Reviews (protocol no. CRD42019156016) (Appendix E2 [online]). Data generated or analyzed during the study are available from the corresponding author by request.

Literature Search

Digital literature databases, including Ovid Embase, Ovid Medline, Scopus, Web of Science, and the Cochrane Central Register of Controlled Trials, were searched from January 2012 to September 2020, with the final search conducted on September 3, 2020, to include the advancements in ML algorithms for medical image interpretation and increased mammographic data availability (2,19). Hand searches of included article references, a gray literature search of computer science databases (ie, DBLP computer science bibliography, Association for Computing Machinery Digital Library, and Institute of Electrical and Electronics Engineers Xplore Digital Library), and a search of arXiv, a preprint literature database, were also conducted for the same time period. The search strategy is detailed in Appendix E3 (online).

Study Selection

To limit bias, all publication types and all study designs were included, with no language restriction or data set age limit applied. Eligibility criteria included women imaged with mammography for screening or diagnosis of breast cancer and a ML algorithm applied as stand-alone workflow application (ie, CAD and CADx or CADt) with sufficient information reported for the performance of stand-alone ML algorithms and reader performance, or the simulated impact on reader performance and workflow to allow for comparison. Any ground truth (eg, histopathologic findings) was accepted. Because data are available at multiple levels (Fig 1), we included algorithms only if they provided an interpretation at the case or examination level to enable comparison with clinician performance as reported in screening programs.

Two independent reviewers undertook the initial title and abstract screening (S.E.H., a physician with 2 years of experience, and then E.P.V.L., C.M.L., or Y.R.I., all medical students) with discordance arbitration by a third reviewer (E.P.V.L., C.M.L., or Y.R.I.), with independent full text review (S.E.H. and R.W., a radiologist with 11 years' experience) and discordance arbitration by a third reviewer (F.J.G., a senior radiologist with more than 30 years of experience).

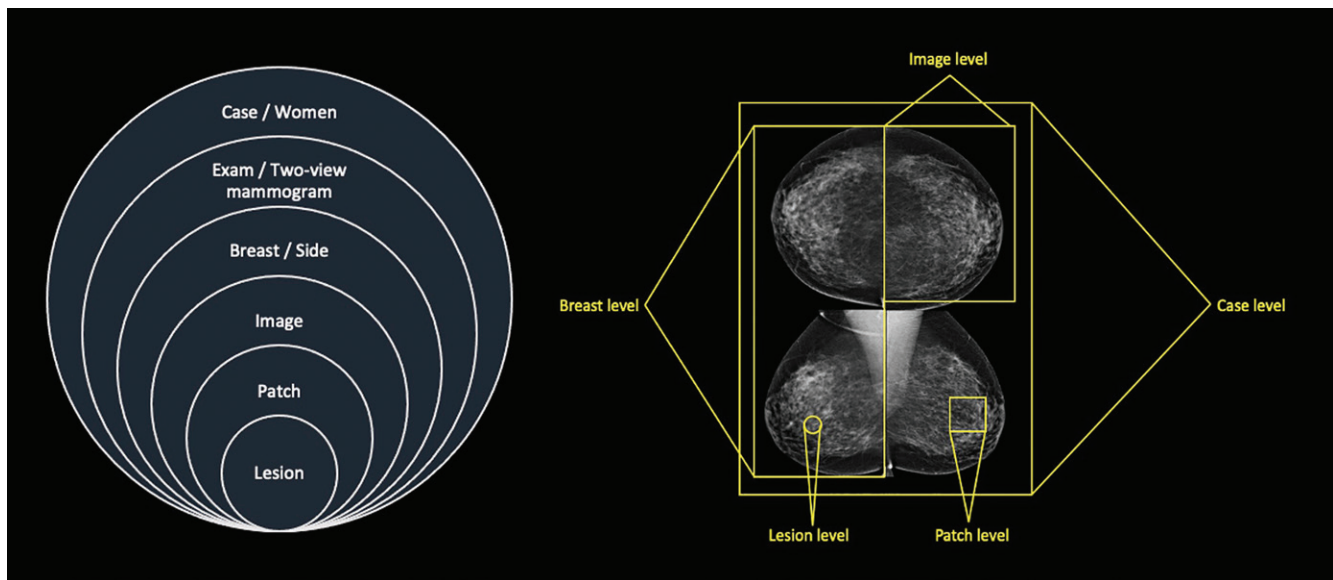


Figure 1: Diagrams show multitime (left) and multiview (right) point data produced with two-dimensional standard view mammography. Data can be analyzed at different levels.

Data Extraction

A predesigned data extraction spreadsheet was used by the reviewers (S.E.H. and R.W.) and checked by a third reviewer (A.I.A.R., a computer scientist with 4 years of experience). Results were only extracted for studies where algorithm performance was compared with readers or the impact on reader workflow and performance was reported. If studies reported multiple stand-alone algorithms, results for all algorithms were extracted (Appendix E4 [online]).

Meta-Analysis Protocol

For the meta-analysis, CAD and CADx algorithm performance was evaluated by adapting the method described in Liu et al (19). The primary meta-analysis compared the best-performing algorithm of each study, at the test stage using screening mammography data, with the performance of readers. Details of the primary meta-analysis study selection are available in Appendix E5 (online). The secondary meta-analysis extended the primary meta-analysis and compared the performance of all reported algorithms and readers in all stand-alone CAD and CADx studies, which used external data sets for addressing the generalization capabilities of the techniques, with no limitations of ground truth.

Quality Assessment

Risk of bias and quality assessment of all included studies took place using Quality Assessment of Diagnostic Accuracy Studies 2 (20,21) and Prediction Model Risk of Bias Assessment Tool (22) by two reviewers (S.E.H. and R.W.), with discussion between reviewers to resolve discordance. Signaling questions for Quality Assessment of Diagnostic Accuracy Studies 2 were adapted for ML studies. Prediction Model Risk of Bias Assessment Tool questions were adapted using the technique in Nagendran et al

(23). However, as our review focused on mammography ML, applicability was assessed in all fields except the predictor field.

The Checklist for Artificial Intelligence in Medical Imaging (24) was used by two reviewers (S.E.H. and A.I.A.R.), with discussion between reviewers to resolve discordance. An overall reporting score for all parameters was generated as well as for eight key fields identified, and common areas underreported were documented.

Statistical Analysis

All statistical analyses were implemented in R software (version 4.0.3, R Project for Statistical Computing) (25) using the “mada” (26) and “boot” (27) packages. Normal and benign examinations were combined, and 2×2 contingency tables were created by calculating true-positive, true-negative, false-positive, and false-negative findings from the reported data set characteristics and sensitivity and specificity provided, ensuring there was an integer, or whole, number of cases. The heterogeneity of the included studies in the quantitative analysis was measured using the I^2 and Cochrane Q tests, where high heterogeneity was defined by I^2 greater than 50% and $P < .05$ for Cochrane Q test. The estimated pooled sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) were calculated for both readers and ML algorithms using a bivariate random effects model by Reitsma et al (28) with 95% CIs. Bootstrapping with 100 iterations was used to generate 95% CIs for the AUC, and a t test was used to compare the ML algorithm and reader sensitivity and specificity, with $P < .05$ indicating a statistically significant difference. Summary receiver operating characteristic plots were constructed for both primary and secondary analyses for pooled reader and ML algorithm performance.

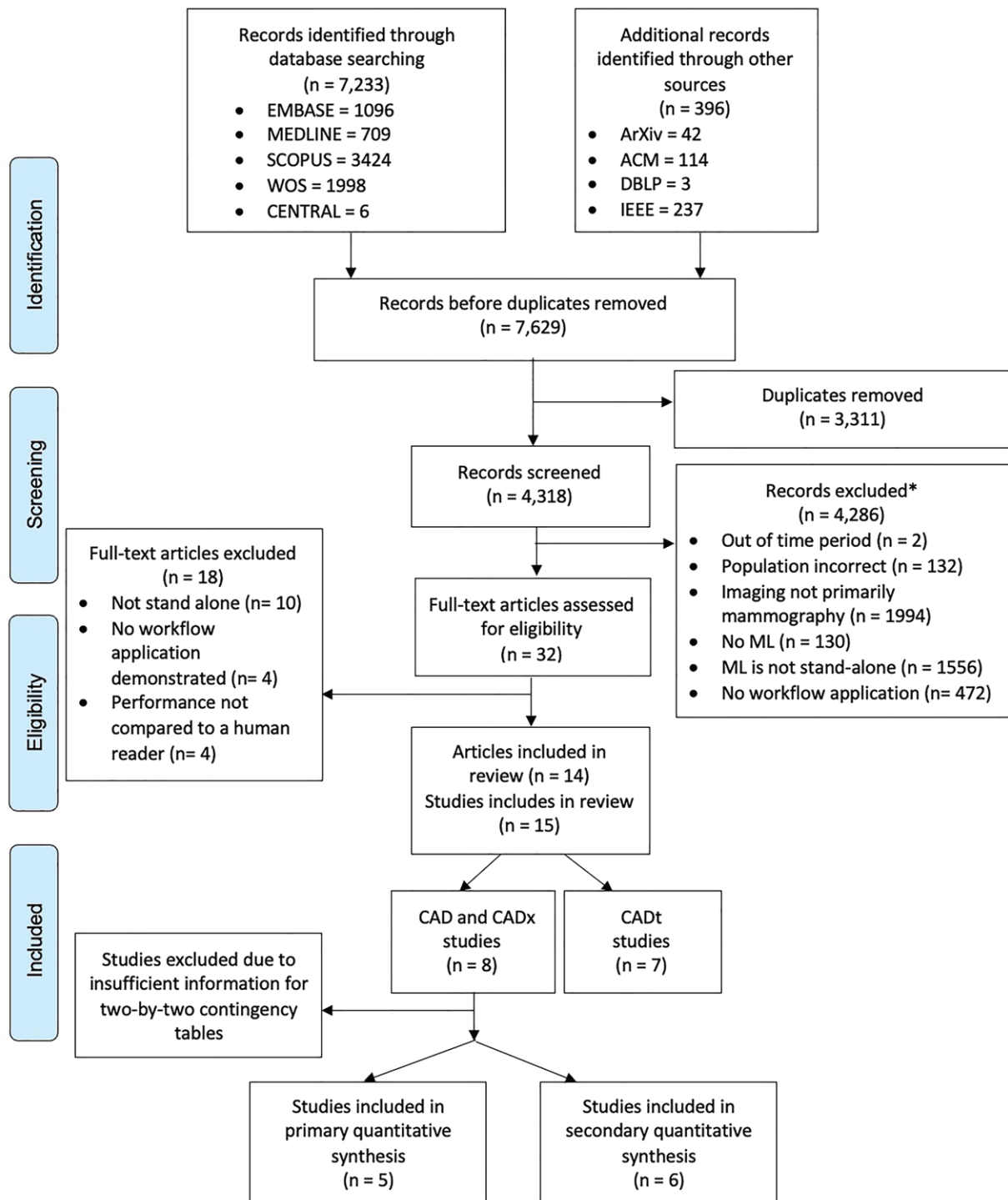


Figure 2: Flowchart of Preferred Reporting Items for Systematic Review and Meta-Analysis for Diagnostic Test Accuracy for studies included in identification, de-duplication, screening, and data-extraction stages of review. ACM = Association for Computing Machinery, CAD = computer-aided detection, CADt = computer-aided triage, CADx = computer-aided diagnosis, IEEE = Institute of Electrical and Electronics Engineers, ML = machine learning, WOS = Web of Science. * = Studies could have been excluded for multiple reasons.

Results

Study Selection and Data Extraction

A Preferred Reporting Items for a Systematic Review and Meta-Analysis diagram (Fig 2) demonstrates the study inclusion process. The search of electronic literature databases and computer

science databases returned 7629 records. Removal of duplicates resulted in 4318 records. After the screening of titles and abstracts, 4286 records were excluded, the remaining 32 full texts were reviewed, and 14 articles were included in the qualitative review. References of included studies can be found in Appendix E6 (online).

Table 1: CADt Algorithm Details and Results

Images Evaluated, Study, and Year	Machine Learning Technique	Triage Task	Sample Size Development Images	Internal or External Testing	Test Threshold	Test No. of Readers and Years of Reader Experience	Test Reader Country and Format	Test Validation Method	Percentage of Normal Cases Triage and Workload Reduction	Result
Screening mam-mograms										
Yala et al, 2019 (39)*	DL: ResNet-18	Triage of normal cases	Total = 238 117 (63 852 cases); training = 212 276 (56 831 cases); validation = 25 841 (7021 cases)	Internal	“Minimum probability score of a radiologist TP assessment on the validation set”	23 readers, 1–31 years of experience	U.S., single	Hold-out method	Normal cases triaged = 19.3%	Missed cancers = 1.1%; sensitivity: 90.1% (172 of 191; 95% CI: 86.0, 94.3); specificity: 94.2% (24,814 of 26,349; 95% CI: 94.0, 94.6)
McKinney et al, 2020 (41)	DL: ensemble ResNet (V2 50 and V1 50), MobileNetV2, and RetinaNet	Triage of normal cases	Training in U.K. = 13 918 cases; validation = 62 866 cases; training in U.S. = 55.0% of 22 225 cases; validation = 15.0% of 22 225 cases	Internal and external†	NPV in U.K. = 99.9%; NPV in the U.S. = 99.9%	U.K., 51 readers with 5 to 20 years of experience; U.S., 1–30 years of experience	U.K., double; U.S., single	Hold-out method	Normal cases triaged: U.K. = 41.0%; U.S. = 35.0%	U.S. reader study ML vs reader: change in AUC (Δ) +0.115 (95% CI: 0.06, 0.18), $P < .001$; U.S. and U.K. data set performance; FP reduction = 5.7% and 1.2%; FN reduction = 9.4% and 2.7%
Balta et al, 2020 (49)‡	DL: unclear, architecture commercial system, Transpara (version 1.6.0)†	Triage of normal cases to single reading†	Unclear, commercial system was directly used†	Internal and external†	7	Six readers	Germany, double	External validation	Workload reduction = 32.6%	Missed cancers = 0.0%; ML decreased recall rate = 11.8% ($P < .001$); PPV = 10.5% ($P < .001$)
Dembrower et al, 2020 (48)§	DL: unclear, architecture commercial system, Lunix (version 5.5.0)†	Triage of normal cases	Training = 170 230 examinations	External	NA	NA	Sweden, double	External validation	Normal cases triaged >60.0%	Missed cancers at 60.0%, 70.0%, 80.0%, 0.0%, 0.3% (95% CI: 0.0, 4.3) and 2.6% (95% CI: 1.1, 5.4)

Screening mam-mograms used from recalled screening cases

Table 1 (continues)

From the included 14 articles, eight studies reported a stand-alone CAD and CADx algorithm performance, and seven studies reported the use of a CADt system. One

article reported on both stand-alone CAD and CADx and CADt. Five studies for stand-alone CAD and CADx provided enough information to be included in the primary

Table 1 (continued): CADt Algorithm Details and Results

Images Evaluated, Study, and Year	Machine Learning Technique	Triage Task	Sample Size Development Images	Internal or External Testing	Test Threshold	Test No. of Readers and Years of Reader Experience	Test Reader Country and Format	Test Validation Method	Percentage of Normal Cases Triaged and Workload Reduction	Result
Kyono et al, 2018 (46)	DL: Inception-ResNetV2 and multi-task learning	Triage for all cases [†]	Training = 90.0% of cases; validation = 10.0% of cases; 100.0% = 7162 cases	Internal	Least patients seen by radiologist without adversely affecting radiologist's FPR or FNR [‡]	Details provided in Kyono et al, 2019 (40) [†]	Single multi-reader [†]	Hold-out method	Workload reduction = 42.8%	Cohen κ = 0.716; F1 statistical test score = 0.757; TP = 120; TN = 803; FP = 41; FN = 36
Kyono et al, 2019 (40)	DL: Inception-ResNetV2 and multitask learning [†]	Triage of normal cases	Unclear, training = 5060 cases plus eight of 10-fold training plus one of 10-fold validation of 2000 cases [†]	Internal	NPV >99.0%	>30 readers with more than 2 years of experience	Single multi-reader [†]	10-fold CV	Normal cases triaged = 34.0% (95% CI: 25.0, 43.0); Low prevalence = 91.0% (95% CI: 88.0, 94.0)	NPV <99.0% [†]
Screening and diagnostic mammograms: Rodriguez-Ruiz et al, 2019 (38) [†]	DL: unclear architecture commercial system, Transpara (version 1.4.0) [†]	Triage of normal cases	Unclear, data partition for training and validation of 189 000 examinations [†]	Internal and external [†]	5, 2	101 readers (U.S. = 52.0% and Europe = 48.0%), further details provided in Rodriguez-Ruiz et al, 2019 (37)	Single multi-reader	External validation	Normal cases triaged threshold of 5 = 47.0%; normal cases triaged threshold of 2 = 17.0%	At the threshold of 5, missed cancers = 7.0; at the threshold of 2, missed cancers = 1.0

Note.—Algorithm performance is compared with reader performance for all included studies. Data partition level is at patient level. All studies are retrospective. AUC = area under the receiver operating characteristic curve, CADt = computer-aided triage, CV = cross-validation, DL = deep learning, FN = false-negative, FNR = false-negative rate, FP = false-positive, FPR = false-positive rate, ML = machine learning, NA = not available, NPV = negative predictive value, PPV = positive predictive value, N = true-negative, TP = true-positive, U.K. = United Kingdom, U.S. = United States.

* Code available at https://github.com/yala/OncoNet_public.

[†] Indicates caveat or another reported format.

[‡] Code available at <https://screenpoint-medical.com/in-practice/>.

[§] Code available at <https://www.lunit.io>.

meta-analysis and six studies for the secondary meta-analysis (algorithm [$n = 17$] and reader [$n = 15$]).

The included articles were published between 2017 and 2020, with three of 14 articles (21%) published on a preprint platform (ie, arXiv). A total of 16 algorithms, including 12 unique algorithms, were included in this review, with two algorithms reported multiple times using different versions.

All included studies were conducted retrospectively. Generalizability was demonstrated in four studies where algorithms were tested on data sets from a different country to the training data set. All data sets used for reader comparison testing were private. Eight of 14 articles (57%) evaluated algorithms on external data sets only, with a further two of 14 articles (14%) using both internal and external datasets. Cancer prevalence within testing data sets varied

Table 2: CAD† Test Set Data Characteristics of All Included Studies

Images Evaluated, Study, and Year	Database Used	Internal or External Data	Country	No. of Centers	Study Years	No. of Images	No. of Cancer Images	Vendor	Screening or Diagnostic Mammograms	Patient Age (y)	Breast Density	Ground Truth
Screening mammograms												
Yala et al, 2019 (39)	Private	Internal	U.S.	1	2009–2016	26 540; 7176 cases	191 examinations; 187 cases, 2.6% of cases	Hologic	Screening	57.8 ± 10.9*	Yes	HP and FU, >1 year
McKinney et al, 2020 (41)	OPTI-MAM (private) and Northwestern Memorial Hospital (private)	Internal	U.K. and U.S.	2 in U.K. and 1 in U.S.	U.K.: 2012–2015; U.S.: 2001–2018	U.K.: 25 856 cases; U.S.: 3097 cases†	U.K.: 414 cases, 1.6% of cases; U.S.: 686 cases, 22.2% of cases†	Hologic, GE, and Siemens	Screening NA	NA	Yes, U.S. only†	HP and FU, >1 year
Balta et al, 2020 (49)	Private	External†	Germany	1	2018	17 895 examinations	114 cases; 0.6% of cases	Hologic and Siemens	Screening NA	NA	NA	HP, no FU
Dem-brower et al, 2020 (48)	CSAW (private)	External	Sweden	1	2009–2015	7364 cases; simulated 75 534 cases	547 cases; 0.7% of cases	Hologic	Screening Range, 40–74; 53.6 (15.4)‡	40–74; 53.6 (15.4)‡	Yes	HP and FU >2 years
Screening mammograms used from recalled screening cases												

Table 2 (continues)

from 0.6% to 50.0%, and the total testing data set size ranged from 240 examinations to 113 663 cases (ie, cohort size was simulated using bootstrapping). The comparator of readers ranged in number (four to 101 readers), experience (1–44 years), and specialization (general or breast

for all studies. The algorithms code was available in nine of 14 articles (64%). Commonly used architectures included ResNet, RetinaNet, and MobileNet, which are all a type of convolutional neural network. This included algorithms that were commercially available in six of 14 articles (43%)

Table 2 (continued): CADt Test Set Data Characteristics of All Included Studies

Images Evaluated, Study, and Year	Database Used	Internal or External Data	Country	No. of Centers	Study Years	No. of Images	No. of Cancer Images	Vendor	Screening or Diagnostic Mammograms	Patient Age (y)	Breast Density	Ground Truth
Kyono et al, 2018 (46)	TOMMY (private)	Internal	U.K.	6	NA	1000 cases	156 of cases; 15.6% of cases	NA	Screening Range, (re-called for assessment and family history)	40–73	Yes	HP and review by three readers of 2D and DBT images
Kyono et al, 2019 (40)	TOMMY (private)	Internal	U.K.	6	NA	Unclear, 1/10-fold of 2000 cases [†]	300 cases; 15.0% of cases	NA	Screening Range, (re-called for assessment and family history)	40–73	Yes	HP and review by three readers of 2D and DBT images
Screening and diagnostic mammograms: Rodriguez-Ruiz et al, 2019 (38)	Private	External	Seven countries, further details provided in Rodriguez-Ruiz et al, 2019 (37)	NA	NA	2654 examinations	653 examinations; 24.6% of examinations	GE, Ho-logic, Sectra, and Siemens	Both (50.0% screening, 50.0% clinical)	Details provided in Rodriguez-Ruiz et al, 2019 (37)	NA	HP and FU ≥ 1 year

Note.—All test data were full-field digital mammograms, and all test data were processed. This information was not available for Balta et al (49) and Dembrower et al (48). CADt = computer-aided triage, CSAW = Cohort of Screen-Aged Women, DBT = digital breast tomosynthesis, FU = follow-up, HP = histopathologic findings, NA = not available, OPTIMAM = OPTIMAM Mammography Image Database, TOMMY = Tomosynthesis with Digital Mammography, 2D = two-dimensional, U.K. = United Kingdom, U.S. = United States.

* Numbers are means ± standard deviations.

[†] Indicates caveat or another reported format.

[‡] Numbers are medians, with interquartile ranges in parentheses.

or where code was available in a public repository in three of 14 articles (21%).

Independent CADt studies reported that between 17% and 91% of normal mammograms could be identified, while missing 0%–7% of cancers (Tables 1, 2). For CAD and CADx tasks, eight studies reported the algorithms' AUCs between 0.69 and 0.96 (Tables 3, 4).

Quality Assessment

The Prediction Model Risk of Bias Assessment Tool and Quality Assessment of Diagnostic Accuracy Studies 2 tools were applied to all included articles in this review, and summary results of assessments are shown in Figure 3 and in Appendix E7 (online). Applying both tools identified a high risk of bias for analysis, as well as high bias and applicability concerns for

Table 3: CAD and CADx Algorithm Details and Results

Images Evaluated, Study, and Year	Machine Learning Technique	No. of Development Images	Internal or External Testing	Test No. of Readers and Years of Reader Experience	Test Reader Country and Format	Test Validation Method	AUC of ML vs Reader	Sensitivity of ML vs Reader (%)	Specificity of ML vs Reader (%)
Screening mammograms									
Geras et al, 2017 (47)*	DL: Custom-trained CNN	721 186 (164 224 examinations); validation = 108 276 (24 552 examinations)	Internal	Four readers	Single multi-reader	Hold-out method	macro AUC = 0.688 vs 0.704	NA	NA
Lotter et al, 2019 (45)	DL: ResNet-50 and RetinaNet	97769 cases	Internal and external [†]	Five readers, 2–15 years of experience	Single multi-reader	External validation and bootstrapping	Test 1 ML: 0.95 (95% CI: 0.92, 0.97); test 2 ML: 0.77 (95% CI: 0.71, 0.82) [‡]	Test 1: +14.2 (95% CI: 9.2, 18.5); $P < .001$; ML vs reader test 2: +17.5 (95% CI: 6.0, 26.2; $P < .001$) ML vs reader [‡]	Test 1: +24.0 (95% CI: 17.4, 30.4); $P < .001$; ML vs reader test 2: +16.2 (95% CI: 7.3, 24.6; $P < .001$) ML vs reader [‡]
Rodriguez-Ruiz et al, 2019 (44) [§]	DL: unclear architecture, commercial system, Transpara (version 1.3.0) [‡]	Unclear data partition for training and validation of 18 000 examinations [†]	Internal and external [†]	14 readers, 11 specialists with 3–25 years of experience	Single multi-reader	External validation	0.89 vs 0.87 ($P = .33$)	83.0 (95% CI: 77.0 (95% CI: 81.0, 85.0), reader only [†]	79.0 (95% CI: 75.0, 83.0), reader only [†]

Table 3 (continues)

the index test, participants, and patient selection (Fig 3). Reasons for high bias and applicability include eight of 14 articles (57%) with cancer-enriched cohorts, five of 14 articles (36%) that tested the algorithm on an internal data set, and three of 14 articles (21%) that did not preset the algorithm threshold in CADt studies. According to the Prediction Model Risk of Bias Assessment Tool assessment, articles were reported to have an overall low (7%), unclear (7%), and high (86%) risk of bias.

Critical appraisal of the reporting quality in the 14 included articles using the 42 parameters of the Checklist for Artificial Intelligence in Medical Imaging, found scores ranging from 22 to 34, with a mean total score of 30 of 42 (71%). The points most commonly underreported included robustness or sensitivity analysis, methods for explainability or interpretability, and protocol registration. Methods for

explainability (eg, saliency maps) to provide transparency of the algorithm's deduction were reported in three articles. Only 50% of articles reported all eight key fields (Fig 4).

Statistical Analysis

Low heterogeneity was found for both algorithms and readers in the primary and secondary analyses ($I^2 = 0.0\%–0.6\%$; Cochrane Q test $P = .45–.78$).

An estimated 185 252 cases from three countries with more than 39 readers were included in the primary meta-analysis. The pooled summary estimates for sensitivity, specificity, and AUC were 75.4% (95% CI: 65.6, 83.2), 90.6% (95% CI: 82.9, 95.0), and 0.89 (95% CI: 0.84, 0.98), respectively, for ML algorithms. For readers, the pooled sensitivity, specificity, and AUC were 73.0% (95% CI: 60.7, 82.6), 88.6% (95% CI: 72.4, 95.8), and

Table 3 (continued): CAD and CADx Algorithm Details and Results

Images Evaluated, Study, and Year	Machine Learning Technique	No. of Development Images	Internal or External Testing	Test No. of Readers and Years of Reader Experience	Test Reader Country and Format	Test Validation Method	AUC of ML vs Reader	Sensitivity of ML vs Reader (%)	Specificity of ML vs Reader (%)
McKinney et al, 2020 (41)	DL: ensemble ResNet (V2 50 and V1 50), MobileNetV2, and RetinaNet	U.K.: training (13 918 cases); validation (62 866 cases); U.S.: training (55.0% of 22 225 cases); validation (15.0% of 22 225 cases)	Internal and external [†]	U.K.: 51 readers, 5 to more than 20 years of experience; U.S.: 1–30 years of experience; U.S. reader study, six readers, 4–15 years of experience	U.K.: double; U.S.: single; reader study: single multi-reader	Hold-out method and external validation	ML U.K.: AUC = 0.89 (95% CI: 0.87, 0.91); U.S. (with training for U.K. and U.S.): AUC = 0.81 (95% CI: 0.79, 0.83); U.K. training only: AUC = 0.76 (95% CI: 0.73, 0.78); [‡] reader study ML vs reader: change in AUC = +0.115 (95% CI: 0.06, 0.18; $P < .001$)	+8.1 ($P < .001$); ML improvement vs reader (minimum and maximum range, 0.0–9.4) [‡]	+3.5 ($P = .02$); ML improvement vs reader (minimum and maximum range, 3.4–5.7) [‡]
Schaffter et al, 2020 (43)	DL: Ensemble CEM (eight networks, including VGG and Faster R-CNN); DL: customized VGG network	KPW (59 923 cases and 100 974 examinations), DDSM, and other data sets (eg, OPTIMA MAM)	External	U.S. screening readers and Sweden screening readers	U.S.: single; Sweden: double (reported single first reader) [†]	Hold-out method and external validation	KPW CEM = .90; top-performing model = 0.86; KI CEM = 0.92; [‡] top-performing model = 0.90	KPW reader sensitivity = 85.9; [†] KI first reader = 77.1; reader consensus = 83.9 ^{†‡}	KPW CEM = 76.1; top-performing model = 66.3 vs 90.5; KI CEM = 92.5; [‡] top-performing model = 88.0 and 81.2 vs first reader = 96.7; ^{†‡} reader consensus = 98.5

Table 3 (continues)

0.85 (95% CI: 0.78, 0.97), respectively (Fig 5). The differences in sensitivity and specificity were not statistically significant ($P = .11$ and $.40$, respectively). Algorithm performance thresholds were set at the reported reader sensitivity and specificity in four studies.

When including all available results from CAD and CADx studies conducted using external data sets that provided a direct comparison between ML algorithms and readers for a secondary meta-analysis, the pooled sensitivity,

Table 3 (continued): CAD and CADx Algorithm Details and Results

Images Evaluated, Study, and Year	Machine Learning Technique	No. of Development Images	Internal or External Testing	Test No. of Readers and Years of Reader Experience	Test Reader Country and Format	Test Validation Method	AUC of ML vs Reader	Sensitivity of ML vs Reader (%)	Specificity of ML vs Reader (%)
Salim et al, 2020 (29)	DL: ResNet-34; MobileNet; unknown	752 000; 239 000; 112 000	External	Sweden screening readers: 25 first readers and 20 second readers	Sweden, double	External validation and bootstrapping	0.96; 0.92; 0.92	ML: 81.9 ($P = .03$); 67.0; 67.4 vs first reader = 77.4; reader consensus = 85.0 [‡]	ML: 96.6, and 96.7 vs first reader = 96.6; reader consensus = 98.5 [‡]
Screening and diagnostic mammograms: Rodriguez-Ruiz et al, 2019 (37) [§]	DL: unclear architecture; commercial system, Transpara (version 1.4.0) [†]	Unclear data partition for training and validation of 189 000 examinations [†]	Internal and external [†]	101 readers: 95 readers for sensitivity and specificity (1–44 years of experience) [†]	Single multi-reader	External validation	0.84 (95% CI: 0.82, 0.86) vs 0.81 (95% CI: 0.79, 0.84) [‡]	75.0–86.0 vs 76.0–84.0 [‡]	49.0–79.0 [‡] , clinician specificity [†]
Mammography and US used for screening: Kim et al, 2019 (42) [#]	DL: ResNet-34; commercial system, Lunit	Total: 166 968 examinations; training: 152 693 examinations; validation: 14 275 examinations	Internal and external [†]	14 readers, seven specialists (>6 months of experience)	Single multi-reader	External validation	0.94 (95% CI: 0.92, 0.97) vs 0.81 (95% CI: 0.77, 0.85); $P < .001$	88.8 vs 75.3 ($P < .001$)	($P < .001$) vs 72.0 ($P = .002$)

Note.—Algorithm performance is compared with reader performance for all included studies. Task is stand-alone artificial intelligence detection and diagnosis. The artificial intelligence decision is made at the per-case level. All testing is retrospective. AUC = area under the receiver operating characteristic curve, CAD = computer-aided detection, CADx = computer-aided diagnosis, CEM = Challenge Ensemble Method, CNN = convolutional neural network, DL = deep learning, DDSM = Digital Database for Screening Mammography, KI = Karolinska Institute, KPW = Kaiser Permanente Washington, NA = not available, OPTIMAM = OPTIMAM Mammography Image Database, U.K. = United Kingdom, U.S. = United States.

* Code available at https://github.com/nyukat/BIRADS_classifier.

[†] Indicates caveat or other reported format.

[‡] Indicates the results of studies included in the primary meta-analysis.

[§] Code available at <https://screenpoint-medical.com/in-practice/>.

^{||} Code available at <https://github.com/Sage-Bionetworks/DigitalMammographyEnsemble>.

[#] Code available at <https://www.lunit.io>.

specificity, and AUC were 80.4% (95% CI: 75.5, 84.6), 82.1% (95% CI: 72.7, 88.8), and 0.86 (95% CI: 0.84, 0.90), respectively, for algorithms. For readers, the pooled sensitivity, specificity, and AUC were 78.5% (95% CI: 73.8, 82.5), 82.6% (95% CI: 69.2, 90.9), and 0.84 (95% CI: 0.81, 0.88), respectively (Fig 5). The differences in sensitivity and specificity were not statistically significant ($P = .70$ and $.73$, respectively). Summary Tables E1–E5 (online) and

additional information are available in Appendixes E8–E11 (online) with associated Figures E1–E4 (online).

Discussion

We found the performance of mammography screening algorithms is reaching equivalence to readers in stand-alone computer-aided detection and computer-aided diagnosis tasks. Comparing our results to two recently published reader per-

Table 4: CAD and CADx Test Set Data Characteristics of All Included Studies

Images Evaluated, Study, and Year	Database Used	Internal or External Data	Country	No. of Centers	Year of Studies	No. of Images	No. of Cancer Images	Vendor	Screening or Diagnostic Mammograms	Patient Age (y)	Breast Density	Ground Truth
Screening mammo-grams												
Geras et al, 2017 (47)	New York University (private)	Internal	U.S.	5	2010–2016	500 examinations	NA	NA	Screening	Range, 19–99; 57.2 ± 11.6*	NA	BI-RADS score: 0, 1, and 2†
Lotter et al, 2019 (45)	Private	External	U.S.	1	2011–2014	Test 1: index cases, 285 examinations; Test 2: pre-index cases 12–24 months previously, 274 examinations	Test 1: 131 examinations, 46.0% of examinations; Test 2: 120 examinations, 43.8% of examinations	NA	Screening	NA	NA	HP and FU >1 year
Rodriguez-Ruiz et al, 2019 (44)	Private	External	U.K. and Europe	1 in U.S. and 1 in Europe	U.S.: 2013–2017; Europe: 2014–2015	240 examinations	100 examinations, 41.7% of examinations	Hologic and Siemens	Screening	Range, 39–89; mean, 61.0	Yes	HP and FU >1 year
McKinney et al, 2020 (41)	OPTIMA MAM (private) and Northwestern Memorial Hospital (private)	Internal and external†	U.S. and U.K.	2 in U.K. and 1 in U.S.	U.K.: 2012–2015; U.S.: 2001–2018	U.K.: 25 856 cases; U.S.: 3097 cases; U.S. reader: 500 cases†	U.K.: 414 cases, 1.6% of cases; U.S.: 686 cases, 22.2% of cases; U.S. reader study: 125 cases, 25.0% of cases†	GE, Hologic, and Siemens	Screening	NA	Yes, U.S. only†	HP and FU >1 year

Table 4 (continues)

Table 4 (continued): CAD and CADx Test Set Data Characteristics of All Included Studies

Images Evaluated, Study, and Year	Database Used	Internal or External Data	Country	No. of Centers	Year of Studies	No. of Images	No. of Cancer Images	Vendor	Screening or Diagnostic Mammograms	Patient Age (y)	Breast Density	Ground Truth
Schaffter et al, 2020 (43)	KPW (private) and KI (private)	Internal and external [†]	U.S. and Sweden	1 and 2 in KI	KPW: NA; KI: 2008–2012	KPW: 25 657 cases; 43 257 examinations; KI: 68 008 cases; 166 578 examinations [†]	KPW: 283 cases, 1.1% of cases; KI: 780 cases, 1.1% of cases	283NA	Screening	KPW: range, 40–74, 58.4 ± 9.7;* KI: range, 40–74, 53.3 ± 9.4*	NA	HP and FU >1 year
Salimet al, 2020 (29)	CSAW (private)	External	Sweden	1	2008–2015	8805 cases, 113 663 examinations simulated [†]	739 cases, 0.7% of simulated cases [†]	Hologic	Screening	Range, 40–74; median, 54.5	Yes	HP and FU >2 years
Screening and diagnostic mammograms: Rodriguez-Ruiz et al, 2019 (37)	Private	External	Sweden, U.K., Netherlands, U.S., Italy, Spain, and Austria	NA	NA	2652 examinations, 2389 examinations for sensitivity and specificity [†]	653 examinations, 24.6% of examinations; 610 examinations, 24.6% of examinations	GE, Hologic, Sectra, and Siemens	Both (some unilateral only) [†]	30–92	Yes	HP and FU >1 year
Mammography and US used for screening: Kim et al, 2019 (42)	Private	External [†]	South Korea	2	2009–2018	320 examinations	160 examinations, 50.0% of examinations	GE and Hologic	Screening (including US) [†]	53.2 ± 10.0*	Yes	Mammography and US detected plus HP [†]

Note.—All test data were full-field digital mammograms, and all test data were processed. This information was not available for Kim et al (42).

BI-RADS = Breast Imaging Reporting and Data System, CSAW = Cohort of Screen-Aged Women, FU = follow-up, HP = histopathologic findings, KI = Karolinska Institute, KPW = Kaiser Permanente Washington, NA = not available, OPTIMAM = OPTIMAM Mammography Image Database, U.K. = United Kingdom, U.S. = United States.

* Numbers are means ± standard deviations.

[†] Indicates caveat or other reported format.

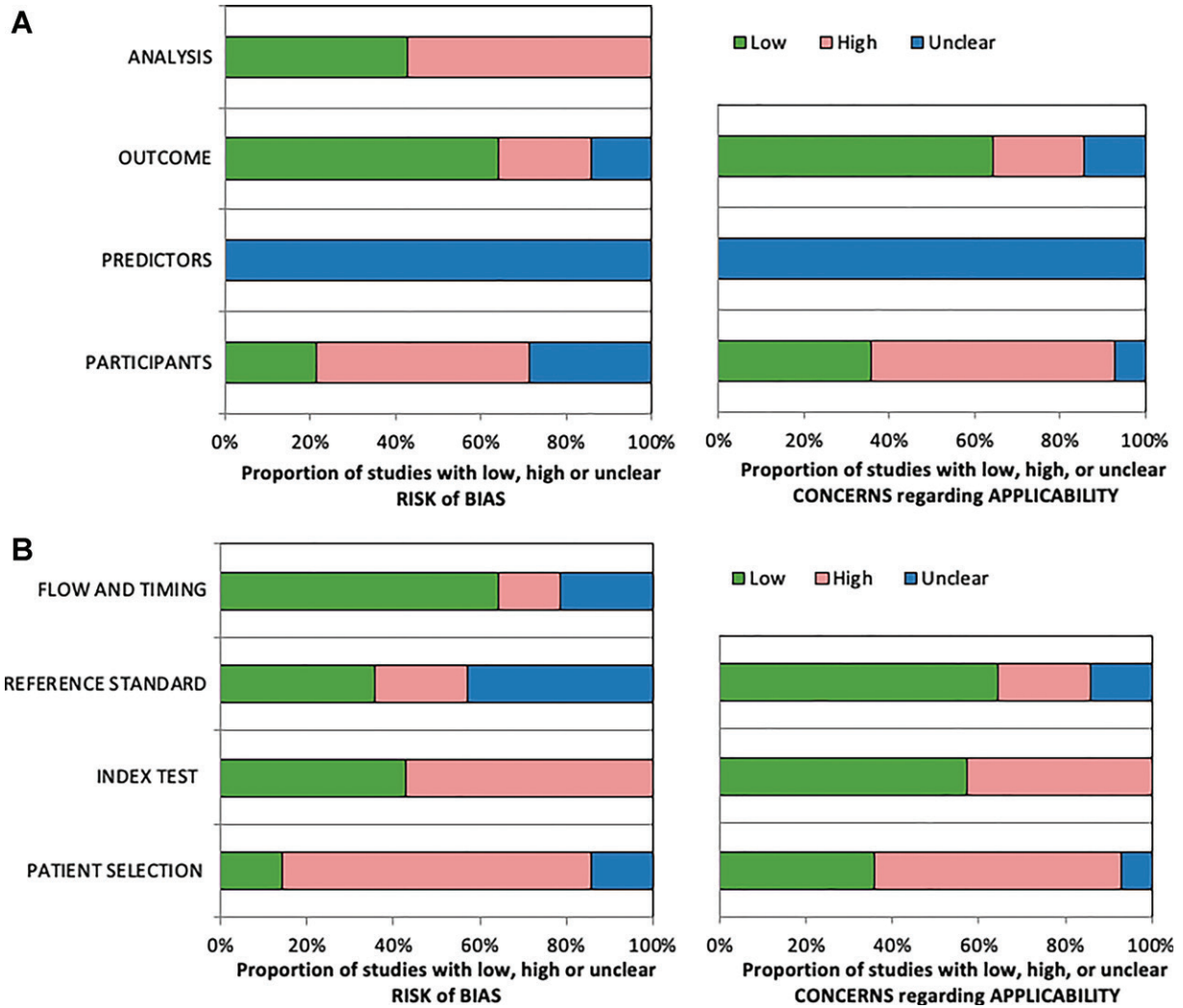


Figure 3: Stacked bar charts show summary results of included articles assessed with (A) Prediction Model Risk of Bias Assessment Tool and (B) Quality Assessment of Diagnostic Accuracy Studies 2 assessment. For 14 included articles, each category is represented as percentage of number of articles that have high, low, or unclear levels of bias and applicability.

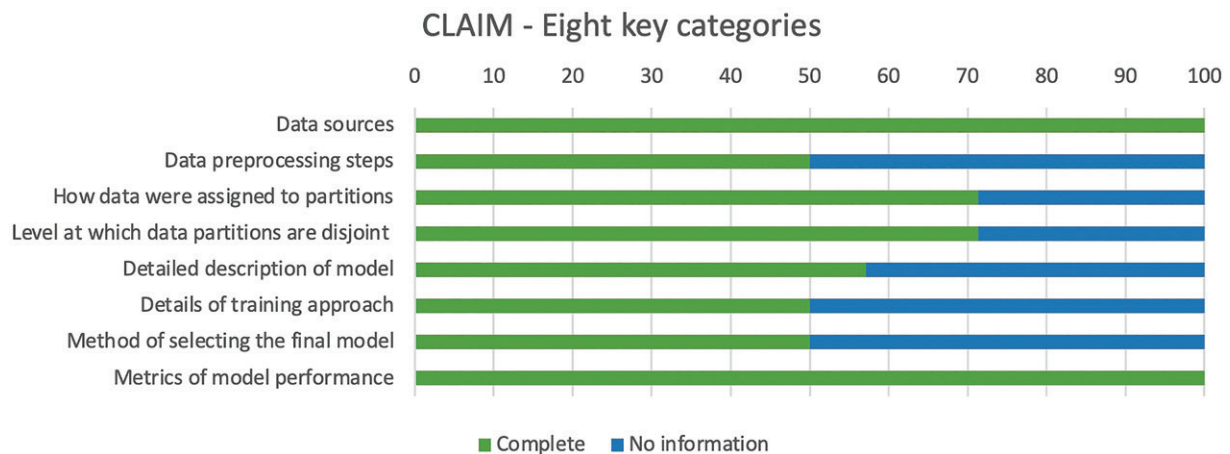


Figure 4: Stacked bar chart of Checklist for Artificial Intelligence in Medical Imaging (CLAIM) assessment. Results for 14 articles included in this review across eight key categories identified from checklist are shown. Score of 1 was given if complete information was provided, and score of 0 was given where no information was provided. X-axis indicates percentage of articles in review that included information about eight key categories detailed in y-axis.

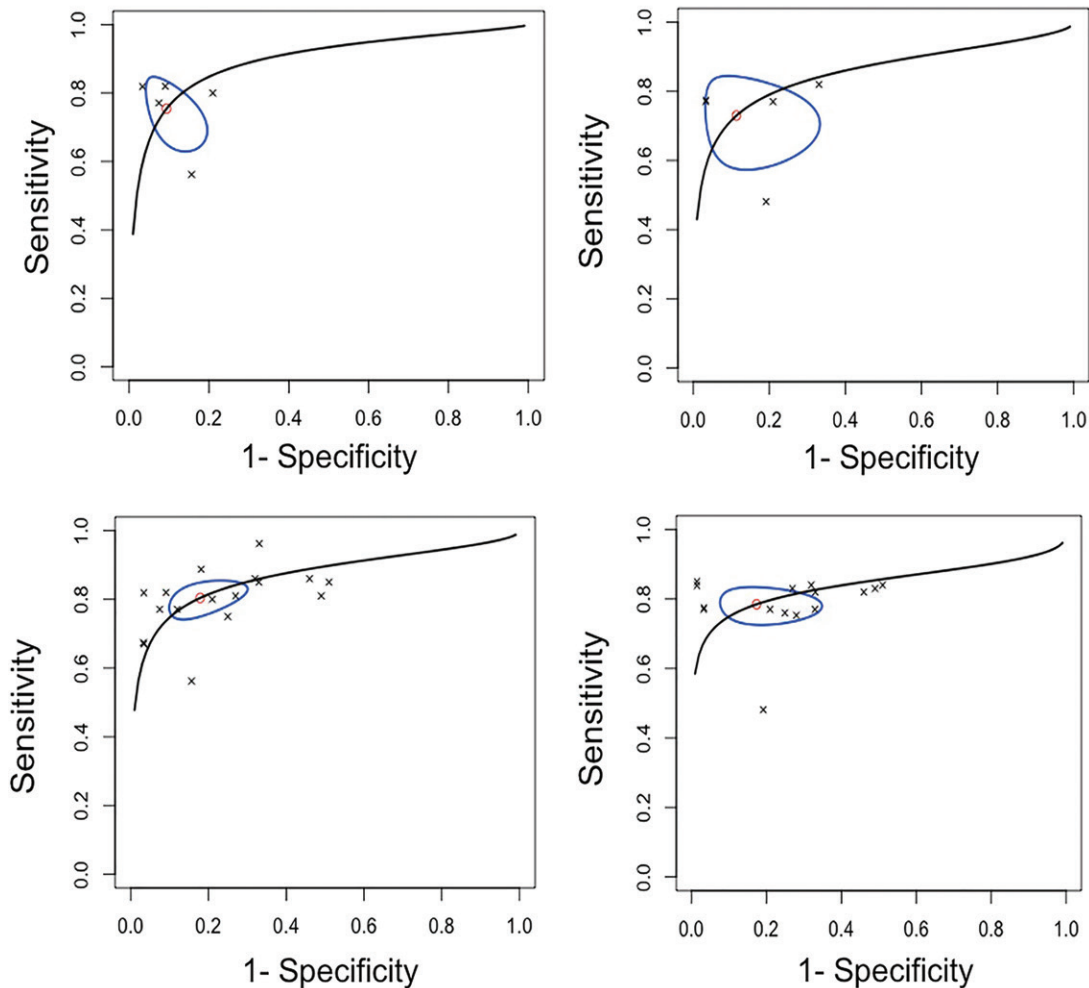


Figure 5: (A, B) Summary receiver operating characteristic (sROC) curves in (A) five studies for included algorithm and (B) reader results reported for top-performing machine learning algorithm tested on external data set, compared with reader performance for computer-aided detection and computer-aided diagnosis applications, with a ground truth of more than 1 year follow-up and histopathologic findings (primary meta-analysis). (C, D) Summary receiver operating characteristic (sROC) curves for (C) 17 algorithm-reported results and (D) 15 reader-reported results from included studies for computer-aided detection and computer-aided diagnosis applications tested externally (secondary meta-analysis). Line represents summary receiver operating characteristic curve, oval represents 95% CIs, circle represents summary estimate, and crosses represent individual results.

formance studies demonstrated that although the pooled sensitivity of algorithms (75.4%) was higher than that of pooled readers (73.0%) and single reading in Sweden (73.0%) (8), it was inferior to both single reading in the United States (86.9%) (6) and double reading with consensus in Sweden (85.0%) (8). The pooled specificity of algorithms (90.6%) was superior to pooled readers (88.6%) and single reading in the United States (88.9%) (6) but was inferior to both single (96.0%) and double reading with consensus in Sweden (98.0%) (8). Therefore, further improvements are needed to make sure machine learning systems meet the “clinically relevant thresholds” of current reader performance and screening program targets. Our findings were similar to a systematic review and meta-analysis comparing deep learning applications across all medical imaging to health care professionals, who came to a similar conclusion and highlighted the importance of continued external testing (19).

Algorithms are also performing tasks not feasible by readers such as high-volume normal case triage, with no detrimental change when reader performance was extrapolated in an adapted screening workflow (ie, using machine-only reading of cases assigned to be normal as an alternative to single or double reading) (2). However, the acceptable “miss” rate for a system, similar to the interval cancer targets, should be agreed upon and specified for machine-only reading of normal mammograms before clinical adoption. The biggest barrier may be public understanding of the concept of acceptable “misses.”

No prospective studies have yet been reported, many studies are still conducted with retrospective internal testing, and few studies are conducted by an independent party where multiple algorithms are cross-compared using external data sets (29). In addition, most of the studies used enriched cancer cohorts for testing, which do not include the class imbalance of cancers to healthy controls in screening. Thus, these data sets may

not provide a realistic representation from which to infer model performance in clinical implementation limiting generalizability, clinical applicability, and feasibility of workflow translation. Our findings highlight the need for well-designed prospective randomized and nonrandomized controlled trials to be conducted across different breast screening programs. These prospective studies should include representative case proportions to replicate the class imbalance in screening, with readers of varying experience interacting with ML algorithm outputs within the clinical workflow. This will allow for performance to be assessed as well as technologic feasibility, reading time, reader acceptability, and effect on reader performance (17). Prospective studies investigating ML applications for mammographic screening are currently underway in the United Kingdom, Norway, Sweden, China, and Russia, with results pending (30–32).

Most articles were from 2019 onward, reflecting the exponential growth in publications since major milestones such as the ImageNet (33) and the Digital Mammography Dialogue for Reverse Engineering Assessment and Methods (3,34) challenges. Although the computer codes were available in 64% of articles, only 21% of code was available on an open-source platform. However, the provision of code alone does not result in a deployable model, including training weights and the threshold at which the algorithm performance was determined, thus limiting reproducibility and transparency (35,36). Large data sets were used for testing, but the majority of these are private, which limits the ability to replicate results.

Two commonly used tools for bias assessment found a high risk of bias due to cancer-enriched cohorts and use of internal data sets as well as the algorithm threshold in triage studies not being preset. Therefore, these results may not be applicable and generalizable to all breast screening populations (21). We applied a specific artificial intelligence medical imaging reporting guideline, the Checklist for Artificial Intelligence in Medical Imaging, to critically appraise artificial intelligence medical imaging literature. It should be noted that the Checklist for Artificial Intelligence in Medical Imaging was published after more than half of the articles in this review were published. Therefore, we have not presented the results of each individual study but have used this as a foundation to find underreported areas within the current literature, as well as to confirm the applicability of the Checklist for Artificial Intelligence in Medical Imaging for ML mammography studies (24).

The meta-analysis was limited by both the small number of eligible studies and because the contingency tables were constructed using reported sensitivity, specificity, total cases, and malignant cases to provide estimated integers, or whole numbers, for calculating true-positive, true-negative, false-positive, and false-negative findings. The primary meta-analysis included studies where reader performance did not reach the level reported in national screening standards; therefore, it is possible that the relative improved performance of ML algorithms is overestimated, and the performance of readers is underestimated as part of this analysis. The primary analysis also used only the highest-performing (ie, based on test performance) algorithm if multiple algorithms were tested and therefore may be slightly biased toward the selected algorithms. The

secondary meta-analysis incorporated multiple algorithms and readers from the same study, in the same population, which could potentially lead to overrepresentation. Therefore, the results from the meta-analysis should be interpreted with caution. Last, for the secondary meta-analysis, both screening and diagnostic mammograms were included in studies, including one study in which women were screened with mammography and US, both of which would have an impact on the expected performance metrics.

There is a growing evidence base that stand-alone machine learning (ML) performance is comparable to reader performance and that ML can undertake triage tasks at a volume and speed not feasible for human readers. Although only retrospective trials have been conducted, the potential for algorithms to perform at the level of or even exceed the performance of a reader within the real-time breast screening workflow is realistic. However, further robust prospective data are critical to understanding where algorithm thresholds are set and are required to examine the interaction between human readers and algorithms, as well as the effect on reader performance and patient outcomes over time.

Acknowledgment: We would like to thank the library team at the University of Cambridge Clinical School for their guidance in developing the search strategy.

Author contributions: Guarantor of integrity of entire study, F.J.G.; study concepts/design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, S.E.H., R.W., E.P.V.L., Y.R.I., C.M.L., F.J.G.; experimental studies, S.E.H., A.I.A.R., R.W.; statistical analysis, S.E.H., G.C.B., J.W.M.; and manuscript editing, all authors

Disclosures of conflicts of interest: S.E.H. research collaborations with Merantix, ScreenPoint, Volpara, and Lunit. R.W. employee of University of Cambridge. E.P.V.L. no relevant relationships. Y.R.I. no relevant relationships. C.M.L. no relevant relationships. A.I.A.R. no relevant relationships. G.C.B. no relevant relationships. J.W.M. employee of Astra Zeneca. F.J.G. funding from Lunit; consultant for Alphabet and Kheiron; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing, or educational events from GE Healthcare; president of European Society of Breast Imaging; equipment, materials, drugs, medical writing, gifts, or other services from GE Healthcare, Bayer, Lunit, and ScreenPoint.

References

1. American College of Radiology Data Science Institute. AI Central. <http://web.archive.org/web/20211018160712/https://aicentral.acrdsi.org/>. Accessed September 10, 2020.
2. Sechopoulos I, Mann RM. Stand-alone artificial intelligence - The future of breast cancer screening? *Breast* 2020;49:254–260.
3. Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. *Clin Radiol* 2019;74(5):357–366.
4. Watanabe L. The power of triage (CADt) in breast imaging. *Applied Radiology*. <http://web.archive.org/web/20211018160942/https://www.appliedradiology.com/communities/Breast-Imaging/the-power-of-triage-cadt-in-breast-imaging>. Accessed November 24, 2020.
5. Schünemann HJ, Lerda D, Quinn C, et al. Breast cancer screening and diagnosis: A synopsis of the European breast guidelines. *Ann Intern Med* 2020;172(1):46–56.
6. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* 2017;283(1):49–58.
7. DeAngelis CD, Fontanarosa PB. US Preventive Services Task Force and breast cancer screening. *JAMA* 2010;303(2):172–173.
8. Salim M, Dembrower K, Eklund M, Lindholm P, Strand F. Range of radiologist performance in a population-based screening cohort of 1 million digital mammography examinations. *Radiology* 2020;297(1):33–39.
9. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013;108(11):2205–2240.

10. Pharoah PDP, Sewell B, Fitzsimmons D, Bennett HS, Pashayan N. Cost effectiveness of the NHS breast screening programme: life table model. *BMJ* 2013;346:f2618 [Published correction appears in *BMJ* 2013;346:f3822].
11. The Royal College of Radiologists. Clinical Radiology UK Workforce Census 2019 Report. <http://web.archive.org/web/20211018161155/https://www.rcr.ac.uk/publication/clinical-radiology-uk-workforce-census-2019-report>. Published 2020. Accessed June 1, 2021.
12. Gilbert FJ, Astley SM, Gillan MGC, et al. Single reading with computer-aided detection for screening mammography. *N Engl J Med* 2008;359(16):1675–1684.
13. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175(11):1828–1837.
14. Kohli A, Jha S. Why CAD failed in mammography. *J Am Coll Radiol* 2018;15(3 Pt B):535–537.
15. Philpotts LE. Can computer-aided detection be detrimental to mammographic interpretation? *Radiology* 2009;253(1):17–22.
16. UK National Screening Committee. Interim guidance on incorporating artificial intelligence into the NHS Breast Screening Programme. <http://web.archive.org/web/20211018161746/https://www.gov.uk/government/publications/artificial-intelligence-in-the-nhs-breast-screening-programme/interim-guidance-on-incorporating-artificial-intelligence-into-the-nhs-breast-screening-programme>. Published 2019. Accessed June 1, 2020.
17. Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *Br J Cancer* 2021;125(1):15–22.
18. McInnes MDF, Moher D, Thoms BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* 2018;319(4):388–396.
19. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1(6):e271–e297.
20. 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):529–536.
21. UK National Screening Committee. Use of artificial intelligence for image analysis in breast cancer screening. Rapid review and evidence map. http://web.archive.org/web/20210922212910/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/987021/AI_in_BSP_Rapid_review_consultation_2021.pdf. Published 2021. Accessed May 19, 2021.
22. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170(1):51–58.
23. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
24. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2020;2(2):e200029.
25. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://web.archive.org/web/20211018074349/https://www.r-project.org/>. Published 2020. Accessed June 1, 2021.
26. Doebler P. mada: meta-analysis of diagnostic accuracy. R package version 0.5.10. <http://web.archive.org/web/20211018162134/https://cran.r-project.org/web/packages/mada/index.html>. Published 2020. Accessed June 1, 2021.
27. Angelo Canty and Brian Ripley. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25. <http://web.archive.org/web/20211018162439/https://cran.r-project.org/web/packages/boot/>. Published 2020. Accessed June 1, 2021.
28. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58(10):982–990.
29. Salim M, Wählin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol* 2020;6(10):1581–1588.
30. NHSX. Mia mammography intelligent assessment. <http://web.archive.org/web/20211009085024/https://www.nhs.uk/ai-lab/explore-all-resources/understand-ai/mia-mammography-intelligent-assessment/>. Accessed October 18, 2021.
31. ClinicalTrials.gov. Development of Artificial Intelligence System for Detection and Diagnosis of Breast Lesion Using Mammography. <http://web.archive.org/web/20211018162549/https://clinicaltrials.gov/ct2/show/NCT03708978>. Accessed October 28, 2020.
32. ClinicalTrials.gov. Experiment on the Use of Innovative Computer Vision Technologies for Analysis of Medical Images in the Moscow Healthcare System. <http://web.archive.org/web/20211018162934/https://clinicaltrials.gov/ct2/show/NCT04489992>. Accessed October 28, 2020.
33. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *NeurIPS Proc*. <http://web.archive.org/web/20211009080513/https://papers.nips.cc/paper/2012/file/c399862d-3b9d6b76c8436e924a68c45b-Paper.pdf>. Published 2012. Accessed June 18, 2020.
34. IBM Research Staff. DREAM Challenge results: Can machine learning help improve accuracy in breast cancer screening? <http://web.archive.org/web/20210922202919/https://www.ibm.com/blogs/research/2017/06/dream-challenge-results/>. Accessed May 22, 2020.
35. Heaven WD. AI is wrestling with a replication crisis. *MIT Technol Rev*. <http://web.archive.org/web/20211018163631/https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/>. Accessed November 24, 2020.
36. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature* 2020;586(7829):E14–E16.
37. 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(9):916–922.
38. Rodriguez-Ruiz A, Lång 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. *Eur Radiol* 2019;29(9):4825–4832.
39. 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.
40. Kyono T, Gilbert FJ, van der Schaar M. Improving workflow efficiency for mammography using machine learning. *J Am Coll Radiol* 2020;17(1 Pt A):56–63.
41. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577(7788):89–94.
42. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multi-reader study. *Lancet Digit Health* 2020;2(3):e138–e148.
43. Schaffter T, Buist DSM, Lee CL, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open* 2020;3(3):e200265.
44. 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(2):305–314.
45. Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using annotation-efficient deep learning approach. *Arxiv [Preprint]*. 2019;1–16. <http://arxiv.org/abs/1912.11027>.
46. Kyono T, Gilbert FJ, van der Schaar M. MAMMO: a deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *Arxiv [Preprint]*. 2018;1–18. <http://arxiv.org/abs/1811.02661>.
47. Geras KJ, Wolfson S, Shen Y, et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *Arxiv [Preprint]*. 2017;1–9. <http://arxiv.org/abs/1703.07047>.
48. Dembrower K, Wählin 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(9):e468–e474.
49. Balta C, Rodríguez-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? In: *Proc SPIE 11513, 15th International Workshop on Breast Imaging (IWBI2020), Conference Location, May 22, 2020, Piscataway, NJ: IEEE, 2020; 115130D*.