Title

Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis.

Journal

Radiology, 307(5)

Authors

Yoon, Jung

Strand, Fredrik

Baltzer, Pascal

et al.

Publication Date

2023-06-01

DOI

10.1148/radiol.222639

Peer reviewed

Radiology

ORIGINAL RESEARCH • EVIDENCE-BASED PRACTICE

Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis

Jung Hyun Yoon, MD, PhD • Fredrik Strand, MD, PhD • Pascal A. T. Baltzer, MD • Emily F. Conant, MD • Fiona J. Gilbert, FRCR, FRCPG, FRCPE, FRSE, FMedSci • Constance D. Lehman, MD, PhD • Elizabeth A. Morris, MD • Lisa A. Mullen, MD • Robert M. Nishikawa, PhD • Nisha Sharma, MBChB, BSc (hons), MRCP, FRCR, MSc • Ilse Vejborg, MD • Linda Moy, MD • Ritse M. Mann, MD, PhD**

From the Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University, College of Medicine, 50 Yonsei-ro, Seodaemun-gu, 03722 Seoul, Korea (J.H.Y.); Department of Oncology and Pathology, Karolinska Institute, Stockholm, Sweden (F.S.); Department of Radiology, Unit of Breast Imaging, Karolinska University Hospital, Stockholm, Sweden (F.S.); Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria (P.A.T.B.); Department of Radiology, University of Pennsylvania, Philadelphia, Pa (E.F.C.); Department of Radiology, University of Cambridge, Cambridge, UK (F.J.G.); Department of Radiology, Harvard Medical School, Massachusetts General Hospital, Boston, Mass (C.D.L.); Department of Radiology, University of California Davis, Davis, Calif (E.A.M.); Department of Radiology, Breast Imaging Division, Johns Hopkins Medicine, Baltimore, Md (L.A.M.); Department of Radiology, University of Pittsburgh, UPMC Magee-Womens Hospital, Pittsburgh, Pa (R.M.N.); Department of Radiology, St James Hospital, Leeds, UK (N.S.); Department of Breast Examinations, Copenhagen University Hospital Herlev-Gentofte, Copenhagen, Denmark (I.V.); Department of Radiology, Laura and Isaac Perlmutter Cancer Center, Center for Biomedical Imaging, Center for Advanced Imaging Innovation and Research, New York University Grossman School of Medicine, New York, NY (L.M.); Department of Medical Imaging, Radboud University Medical Center, Nijmegen, the Netherlands (R.M.M.); and Department of Radiology, Netherlands Cancer Institute, Amsterdam, the Netherlands (R.M.M.).
Received October 14, 2022; revision requested October 25; revision received March 23, 2023; accepted April 3. **Address correspondence to** J.H.Y. (email: *lvjenny@yuhs.ac*).

Supported by the National Institutes of Health at the New York University Grossman School of Medicine (grant P41EB017183).

* L.M. and R.M.M. are co-senior authors.

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

See also the editorial by Scaranelo in this issue.

Radiology 2023; 307(5):e222639 • https://doi.org/10.1148/radiol.222639 • Content code: BR

Background: There is considerable interest in the potential use of artificial intelligence (AI) systems in mammographic screening. However, it is essential to critically evaluate the performance of AI before it can become a modality used for independent mammographic interpretation.

Purpose: To evaluate the reported standalone performances of AI for interpretation of digital mammography and digital breast tomosynthesis (DBT).

Materials and Methods: A systematic search was conducted in PubMed, Google Scholar, Embase (Ovid), and Web of Science databases for studies published from January 2017 to June 2022. Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) values were reviewed. Study quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 and Comparative (QUADAS-2 and QUADAS-C, respectively). A random effects meta-analysis and meta-regression analysis were performed for overall studies and for different study types (reader studies vs historic cohort studies) and imaging techniques (digital mammography vs DBT).

Results: In total, 16 studies that include 1 108 328 examinations in 497 091 women were analyzed (six reader studies, seven historic cohort studies on digital mammography, and four studies on DBT). Pooled AUCs were significantly higher for standalone AI than radiologists in the six reader studies on digital mammography (0.87 vs 0.81, $P = .002$), but not for historic cohort studies (0.89 vs 0.96, $P = .152$). Four studies on DBT showed significantly higher AUCs in AI compared with radiologists (0.90 vs 0.79, $P < .001$). Higher sensitivity and lower specificity were seen for standalone AI compared with radiologists.

Conclusion: Standalone AI for screening digital mammography performed as well as or better than radiologists. Compared with digital mammography, there is an insufficient number of studies to assess the performance of AI systems in the interpretation of DBT screening examinations.

Supplemental material is available for this article.

Digital mammography and digital breast tomosynthesis (DBT) are cornerstones of breast imaging, especially for breast cancer screening. Randomized clinical trials, systematic reviews, and observational studies have demonstrated that screening mammography reduces breast cancer–related mortality by 20%–50% (1–3). Based on this evidence, mammography has become widely implemented for breast cancer screening (4). Although screening with mammography is proven beneficial, it does not detect all cancers. False-negative screening studies, in which cancers are diagnosed symptomatically or are detected by another modality between two screening rounds, represent 20%–33% of cancer cases (5,6), and lack of perception of an abnormality was reported as the most common cause of missed breast cancers (7). In addition, many well-trained dedicated radiologists are required to sustain the screening programs and qualified resources are limited, especially in the double-reading settings that are common in many European countries (8,9). There is an inevitable need for assistance in screening

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, CLAIM = Checklist for Artificial Intelligence for Medical Imaging, DBT = digital breast tomosynthesis, QUADAS-C = Quality Assessment of Diagnostic Accuracy Studies Comparative, QUADAS-2 = QUADAS 2, SE = standard error

Summary

Standalone artificial intelligence (AI) for screening digital mammography performed as well as or better than individual breast radiologists or average reader outcomes; there was insufficient comparative evidence on AI for interpretation of digital breast tomosynthesis.

Key Results

- Pooled areas under the receiver operating characteristic curve (AUCs) were significantly higher for standalone artificial intelligence (AI) than radiologists in the six reader studies on digital mammography (AUC difference, 0.05; 95% CI: 0.2, 0.8; $P = .002$), but not for historic cohort studies (AUC difference, 0.06; 95% CI: 0, 0.14; $P = .152$).
- Four studies on digital breast tomosynthesis showed significantly higher AUCs in AI versus radiologists (AUC difference, 0.11; 95% CI: 0.06, 0.16; $P < .001$).
- Higher sensitivity and lower specificity were seen for standalone AI compared with radiologists, regardless of study type or modality.

interpretations to reduce interobserver variability, while also managing workforce limitations (10,11).

Advances in technology and computer programming enabling automated analysis of medical images may address these issues. Using computer-aided detection for mammographic interpretation, with promising results of increased cancer detection rates (12,13), gained attention in the early 1990s (12). However, this positive effect was not maintained after widespread clinical implementation in the United States (14), mostly due to poor specificity and an increase in false-positive interpretations (15). Further advances in technology have led to a new generation of artificial intelligence (AI) algorithms based on convolutional neural networks and deep learning, which easily surpass the classic feature-based approaches when applied to digital mammograms (16,17). In recent years, numerous mammographic AI systems have been developed and validated and are currently commercially available (18–20). With ongoing advances in AI technology and the increasing need for workflow improvement in breast imaging, such AI systems are rapidly being implemented in the clinical setting. Early studies showed promising results in simulating potential roles of AI to support radiologists in practice (19,21–24). However, AI can only fulfill the goals of improving screening outcomes and/or workload reductions when the independent performance is sufficiently high. Therefore, before we can consider the implementation of AI as a standalone modality for mammographic interpretation, the current systems must be critically evaluated.

Thus, a systematic review and meta-analysis evaluating the standalone performances of AI for interpretation of digital mammography and DBT was conducted, setting the stage for if—and if so, how—AI should be clinically implemented.

Materials and Methods

This systematic review and meta-analysis is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses, or PRISMA, Diagnostic Test Accuracy guidance (25).

Literature Search

An online search in PubMed, Google Scholar (first 1000 hits), Embase (Ovid), and Web of Science for studies that were either in final publication or in the process of electronic publication from January 2017 to June 2022 was conducted using the following query: "artificial intelligence AND (mammography OR tomosynthesis OR DBT) AND detection." This relatively short time span was chosen because the first studies evaluating deep learning–based AI for mammography were published in 2017. References of selected publications were analyzed for additional relevant papers.

Eligibility Criteria

Studies were eligible for inclusion if they reported standalone performance of an AI system applied to either digital mammography or DBT to detect breast cancer and also reported independent reads by radiologists. All study designs were eligible. Exclusion criteria were as follows: studies using AI for triage only, studies only including specific populations or lesion types, studies only reporting a comparison of AI algorithms, and studies aimed at risk prediction rather than cancer detection.

All retrieved entries were evaluated based on the article title and abstract, first separately by two authors (J.H.Y., R.M.) and then in consensus. A third author (L.M.) acted as the arbitrator in case of disagreement.

Data Extraction

Appendix S1 contains a comprehensive list of items for data extraction. Data were extracted using a predefined data extraction sheath agreed upon by the authors before study selection. Items were grouped as follows: general study characteristics; characteristics of the screening program from which cases were obtained; selection of the study cohort; types of images included; reference standards; characteristics of screening radiologists; outcome measures such as the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity of human reading; characteristics of the AI system used; and outcome measures of standalone AI. When possible, performance metrics of AI and radiologists that were not provided were calculated from the available study data.

The task of data extraction was divided among all authors. To facilitate extraction, authors were assigned their own papers whenever possible. After data extraction, authors were assigned a further set of papers to control the extracted data and, for these analyses, the controlling author was not an author of the paper. Disagreements and unclarity were resolved by consensus. In case of persistent unclarity, one of three authors (J.H.Y., R.M.M., L.M.) served as arbitrator.

Assessment of Study Quality

Study quality was assessed using Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) (26) and QUADAS

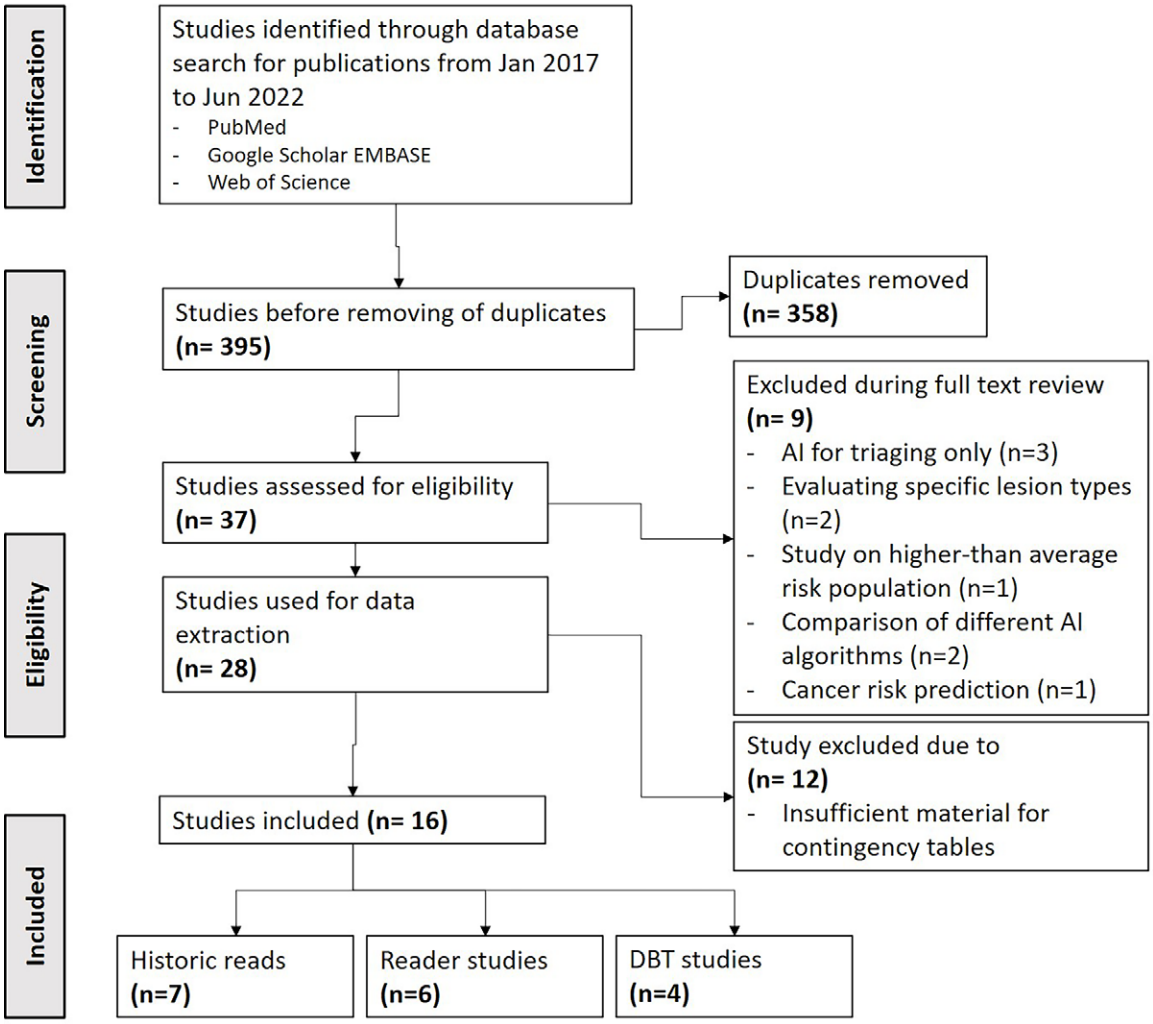

Figure 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram shows study inclusion and exclusion. One of the 16 included studies had data for both interpretation of digital mammography in a historic cohort and interpretation of digital breast tomosynthesis (DBT). AI = artificial intelligence.

Comparative (QUADAS-C) (27) (Table S1). Signaling questions for QUADAS-2 and QUADAS-C were refined for AI studies. Specifically, as the current study aimed to determine whether standalone AI performance is as good as that of radiologists, study characteristics in favor of AI were considered to be biased and those in favor of radiologists were not. Each paper was scored by one of four authors (J.H.Y., F.S., R.M.M., L.M.) and controlled by another one of those authors. In cases of discordance, consensus results were used after discussion among the four authors.

For assessing reporting quality, completeness of the Checklist for Artificial Intelligence for Medical Imaging (CLAIM) (28) was evaluated by the same authors who performed the data extraction for each study. Reported items were scored by the extractors and checked by the controllers, with discussion in cases of discordance. Overall, reporting scores for the total 42 CLAIM items and eight key items were summed for each study. Underreported areas were flagged, analyzed, and documented.

Statistical Analysis

Data were analyzed by one of the authors (P.A.T.B.). A 2 × 2 contingency tables were created by calculating true-positive, false-positive, true-negative, and false-negative findings from the reported data, sensitivity, and specificity provided in the studies. I^2 was calculated to assess the heterogeneity among studies. Values greater than 50% were considered at risk for substantial variability. Pooled estimates with 95% CIs for the performance metrics were calculated for both radiologists and AI using random-effects models. Forest plots were generated to display the performances of studies as a whole and according to study type subgroup. Summary receiver operating characteristic curves for performances of radiologists and AI were constructed using random effects models because bivariate modeling of joint sensitivity and specificity lead numerical overflow during calculation for the overall data set and certain subsets. Multivariable meta-regression analysis was performed by comparing the performance measures of subgroup items according to the following features: study type (ie, reader studies, defined as studies in which radiologists interpreted

mammograms for the study, and historic reads, defined as studies using retrospective interpretation [either single reader or consensus] in clinical practice), definitions for negative examinations, commercially available versus in-house software, the role of the vendor in the study design, and the threshold used for interpretation in the study. Publication bias was assessed separately for radiologists and AI by using the Deeks funnel plot asymmetry test.

Results

Study Selection and Risk of Bias

The PRISMA flow diagram for this meta-analysis is shown in Figure 1. Of 37 studies distributed among the authors for full-text review, nine were excluded after full-text analysis. In total, 28 studies were used for data extraction. During

Table 1: Summary of Characteristics of the 16 Studies Using AI for Mammography

Study and Year	Study Type*	Modality	Reference Standard	Population (% of Cancers)	No. of Readers	AI Algorithm (Version)	Vendor Involved in Study Design?	AI Threshold for Interpretation
Becker et al (33), 2017	R, reader	DM	Cancer: biopsy Benign: negative >2 y FU	251 (7.2%)	3	No information	No	NS
Conant et al (34), 2019	R, reader	DBT	Cancer: biopsy Benign: biopsy or negative >320 d FU	260 (25%)	24	PowerLook Tomo (2.0)	Yes	NS
McKinney et al (22), 2020	R, historic	DM	Cancer: biopsy Benign: negative >3 y FU	UK: 25 856 (1.6%) U.S.: 3097 (22.2%)	6	In-house algorithm	Yes	Set at threshold defined in study
Salim et al (10), 2020	R, historic	DM	Cancer: biopsy Benign: negative >2 y FU	8805 (8.39%)	2	Lunit INSIGHT for MMG (1.1.4.3); anonymous commercial (NR)	No	Set at specificity of radiologist
Sasaki et al (35), 2020	R, reader	DM	Cancer: biopsy Benign: biopsy or negative >1 y FU	310 (22%)	3†	Transpara (1.3.0)	No	Set at threshold defined in study
Pinto et al (36), 2021	R, reader	DBT	Cancer: biopsy Benign: biopsy or negative based on ground truth of previous trial	190 (38.9%)	14	Transpara (1.6.0)	No	Set at threshold defined in study
Shen et al (37), 2019	R, reader	DM	Cancer: biopsy Benign: biopsy or negative at next screening	14 148 (0.21%)	14	In-house algorithm; globally-aware multiple instance classifier (1)	No	Set at sensitivity of radiologist
Dang et al (38), 2022	R, reader	DM	Cancer: biopsy Benign: negative >2 y FU	314 (NR)	12	MammoScreen (1.2)	No	NS
Kim et al (19), 2020	R, reader	DM	Cancer: biopsy Benign: biopsy or negative >2 y FU	320 (50%)	14	Lunit INSIGHT MMG (NR)	Yes	Vendor suggested
Larsen et al (39), 2022	R, historic	DM	Cancer: biopsy Benign: NS	47 877 (1.6%)	24	Transpara (1.7.0)	No	Set at threshold defined in study
Lauritzen et al (40), 2022	R, historic	DM	Cancer: biopsy Benign: NS	114 421 (0.7%)	7	Transpara (1.7.0)	No	Set at threshold defined in study
Lee et al (41), 2022	R, reader	DM	Cancer: biopsy Benign: negative >2 y FU	200 (50%)	10	Lunit INSIGHT for MMG (1.1.1.0)	Yes	Vendor suggested
Leibig et al (42), 2022	R, historic	DM	Cancer: biopsy Benign: biopsy or negative >2 y FU	External test set: 82 851 (3.4%)	NS	Vara (NR)	Yes	Set at specificity of radiologist

Table 1 *(continues)*

Table 1 *(continued)*: Summary of Characteristics of the 16 Studies Using AI for Mammography

Study and Year	Study Type*	Modality	Reference Standard	Population (% of Cancers)	No. of Readers	AI Algorithm (Version)	Vendor Involved in Study Design?	AI Threshold for Interpretation
Romero-Martín et al (29), 2022	R, historic	DM, DBT	Cancer: biopsy Benign: no cancer diagnosis <12 mo, negative for 11–36 mo FU	15 999 (0.7%)	5	Transpara (1.7.0)	No	NS
Sharma et al (43), 2022	R, historic	DM	Cancer: biopsy Benign: no proof of malignancy during <1035 d after negative screening	177 882 (1%)	NS	Mia (2.0.1)	Yes	Vendor suggested
Shoshan et al (44), 2022	R, reader	DBT	Cancer: biopsy Benign: NS	4310 (10.6%)	5	In-house algorithm; IBM-developed AI model (1)	Yes	Set at maximum sensitivity of radiologist

Note.—AI = artificial intelligence, DBT = digital breast tomosynthesis, DM = digital mammography, FU = follow-up, NR = not reported, NS = not specified, R = retrospective, UK = United Kingdom.

* Reader studies are defined as studies in which radiologists interpreted mammograms for the study, and historic reads are defined as studies using retrospective interpretation in clinical practice.

† One radiologist, two technologists.

data extraction and quality assessment, 12 additional studies were excluded due to incomplete data. In total, 16 studies comprising 1 108 328 examinations in 497 091 women were included in this review. Appendix S2 provides a list of the studies included for analysis, and Table 1 summarizes the characteristics of the studies; all used retrospective data sets. Of the 16 studies, nine (56.2%; six on digital mammography and three on DBT) were designed as reader studies, while the remaining seven (43.8%) used historic reads of screening digital mammography. One study (29) used the Córdoba Tomosynthesis Screening Trial, where data for digital mammography, DBT, and standalone AI were extracted. Four (25%) of the 16 studies were on DBT, including three reader studies and one using historic reads (29). For the definition of negative examinations, eight (50%) of the 16 studies used a negative screening after greater than or equal to 2 years of follow-up, another five (31.2%) used a benign histopathologic diagnosis and negative screening at the next screening, and three (18.8%) did not give a precise definition. Of the 16 studies, 12 (75%) used commercially available AI algorithms and seven (43.8%) had vendors involved in the study design. Three (18.8%) of the 16 studies used predefined thresholds provided by the vendor, while five (31.3%) used different thresholds defined specifically for the study.

The QUADAS-2 and QUADAS-C assessment results are displayed in Figure 2. A high risk of bias was seen for patient selection; seven (43.8%) of the 16 studies with selected samples or cancer-enriched cohorts were assessed as having high bias. High concerns for applicability were seen for patient selection (three of 16 studies, 18.8%) and the reference standard (three of 16 studies, 18.8%).

Reporting quality in the 16 articles using CLAIM ranged from 0 to 42 (Fig S1). The mean CLAIM score was 25.1 (range, 14–39) for 42 items (Fig S1). Fifteen (35.7%) of the 42 items were underreported. Three of the eight key items (detailed description of models, details of training approach, and method of selecting the final model) were underreported, with proportions of reporting studies ranging 25%–31%.

Outcome Measures of Standalone AI and Radiologists

Significant heterogeneity was seen in sensitivity and specificity for both radiologists (I^2 = 95.15% and 99.97%, respectively; all $P < .001$) and AI (I^2 = 97.79% and 99.99%, respectively; all $P < .001$) (Fig S2). The heterogeneity was lower in the reader studies than in the historic cohort studies. Pooled estimates for sensitivity were 73.6% (95% CI: 68.7, 78.0) for radiologists and 80.6% (95% CI: 74.3, 85.7) for AI, with a difference of 7% (standard error [SE], 3.8%; $P = .031$) (Table 2). Pooled specificity was 89.6% (95% CI: 82.7, 93.9) and 85.7% (95% CI: 74.1, 92.6) for radiologists and AI, respectively. The difference in specificity (3.9%; SE, 5.5%) was not statistically significant ($P = .221$). Forest plots displaying the pooled estimates of sensitivity and specificity for human readers and AI for all 16 studies are shown in Figure S2. A minor publication bias was observed only for reporting of human reader results ($P = .01$, Fig S3), but not for AI (Fig S4). When analyzed according to study type (ie, reader study vs historic reads for digital mammography), the forest plots showed higher sensitivity and lower specificity for standalone AI compared with that of radiologists (Fig 3).

Within the six reader studies using digital mammography, pooled AUC estimates were 0.81 (SE, 0.014) for radiologists

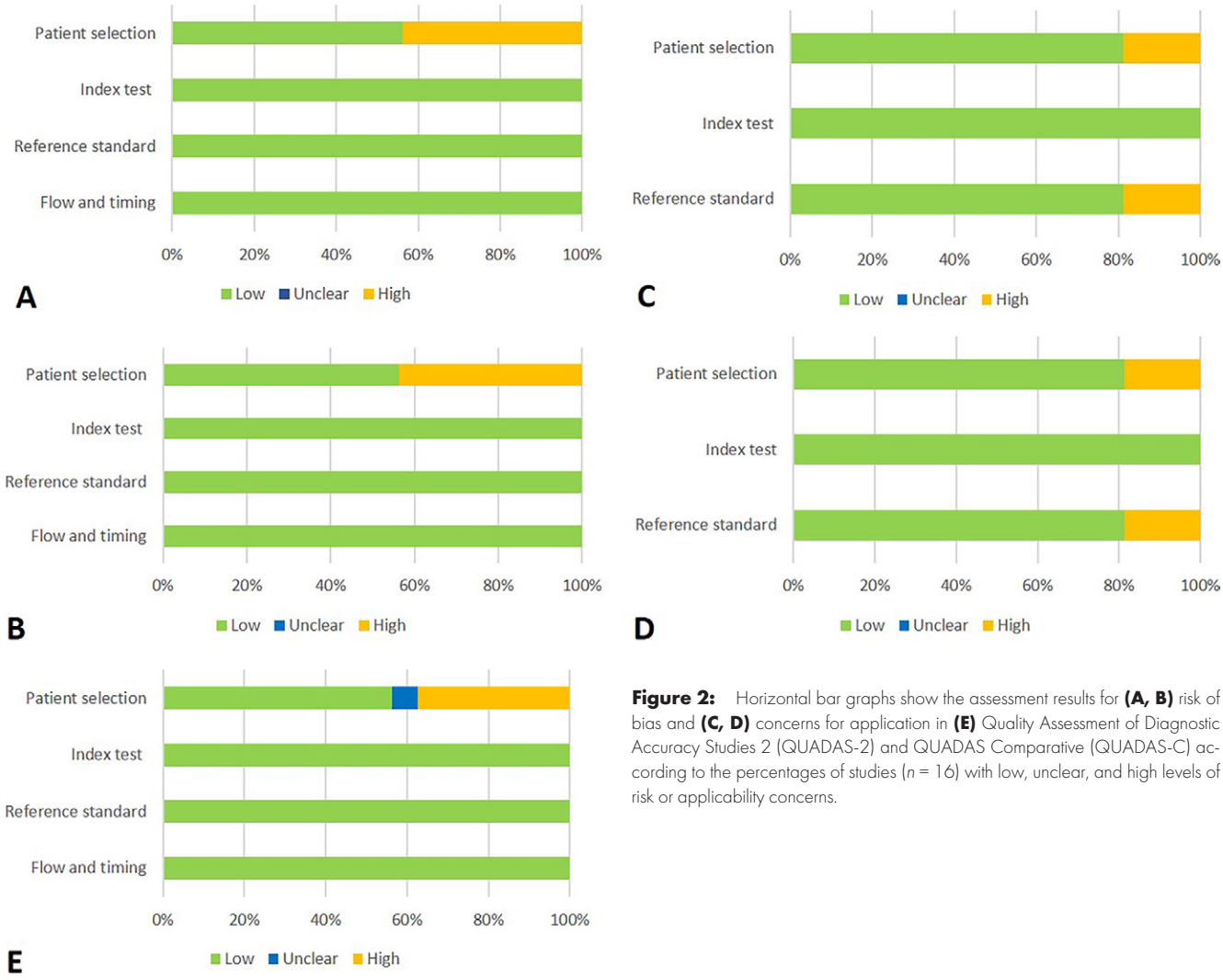

Figure 2: Horizontal bar graphs show the assessment results for **(A, B)** risk of bias and **(C, D)** concerns for application in **(E)** Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) and QUADAS Comparative (QUADAS-C) according to the percentages of studies (n = 16) with low, unclear, and high levels of risk or applicability concerns.

Table 2: Pooled Estimates of Performance Measures for Radiologists and Standalone AI for All Included Studies and Study Type Subgroups

	Sensitivity		Specificity		AUC*	
Variable	Radiologists	AI	Radiologists	AI	Radiologists	AI
All studies (n = 16)	73.6 (68.7, 78.0)	80.6 (74.3, 85.7)	89.6 (82.7, 93.9)	85.7 (74.1, 92.6)	…	…
Reader studies (n = 6)	72.4 (64.1, 79.4)	80.8 (68.0, 89.3)	81.6 (75.7, 86.4)	76.9 (55.2, 90.0)	0.81 [0.014]	0.87 [0.010]
Studies using historic reads (n = 7)†	72.6 (63.7, 80.1)	75.8 (70.2, 80.6)	96.4 (94.9, 97.4)	95.6 (93.7, 96.9)	0.96 [0.022]	0.89 [0.037]
Digital breast tomosynthesis studies (n = 4)†	77.9 (73.1, 82.0)	88.8 (80.2, 94.0)	81.6 (37.8, 97.0)	63.1 (22.1, 91.1)	0.79 [0.020]	0.90 [0.011]

Note.—Except where indicated, data are percentages, with 95% CIs in parentheses. Reader studies are defined as studies in which radiologists interpreted mammograms for the study, and historic reads are defined as studies using retrospective interpretation in clinical practice. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve.

* Data in brackets are standard errors.

† One study included data for historic reads of digital mammography, digital breast tomosynthesis, and standalone AI that were individually extracted.

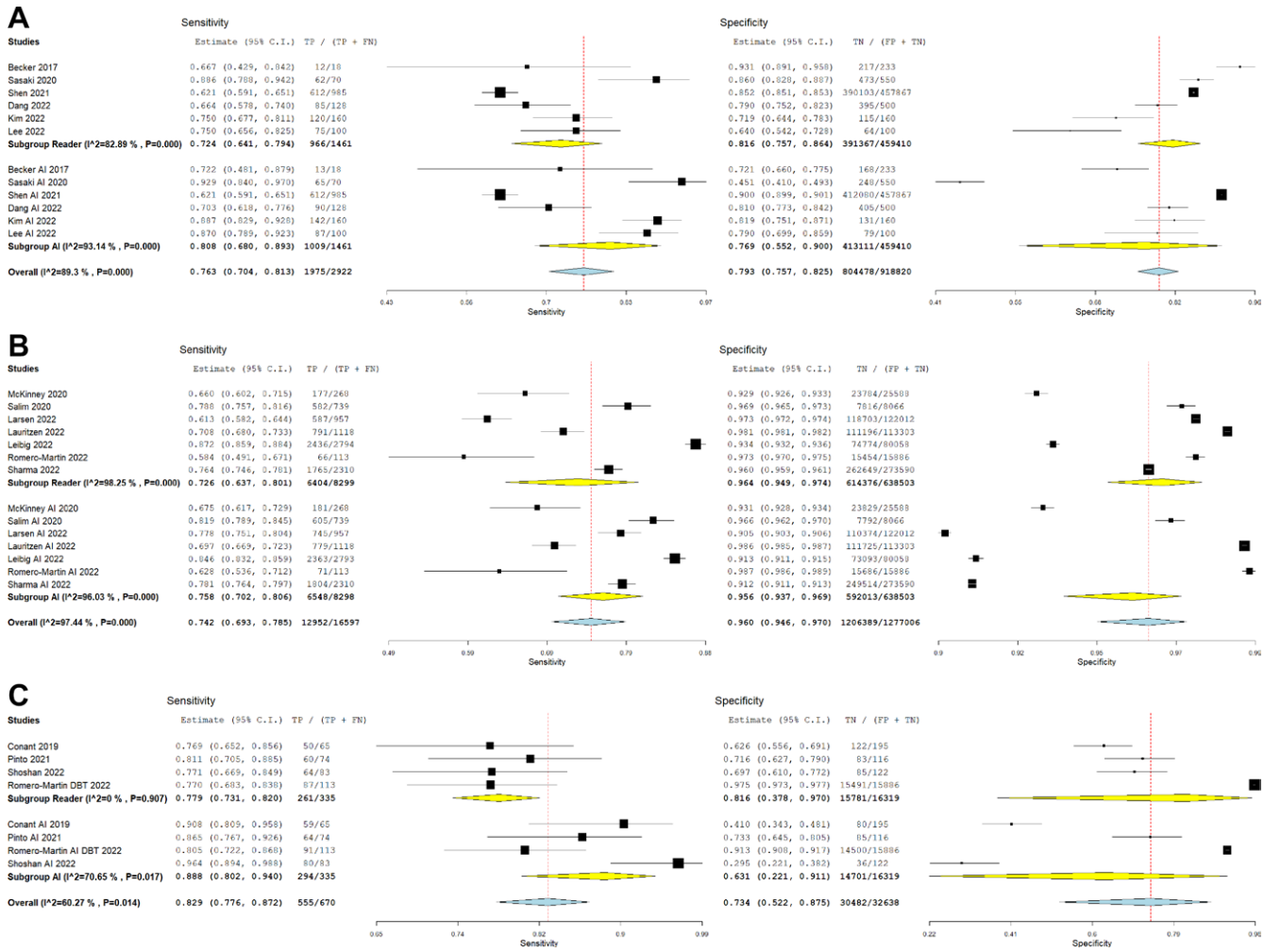

Figure 3: Forest plots show pooled estimates for the **(A)** six reader studies, **(B)** seven studies using historic reads, and **(C)** four digital breast tomosynthesis (DBT) studies. Reader studies are defined as studies in which radiologists interpreted mammograms for the study, and historic reads are defined as studies using retrospective interpretation in clinical practice. AI = artificial intelligence, FN = false negative, FP = false positive, TN = true negative, TP = true positive.

and 0.87 (SE, 0.010) for AI (Fig 4). In the seven studies using historic reads, pooled AUC estimates were 0.96 (SE, 0.022) for radiologists and 0.89 (SE, 0.037) for AI. In the four reader studies using DBT, these metrics were 0.79 (SE, 0.020) and 0.90 (SE, 0.011) for radiologists and AI, respectively. Differences in pooled AUC estimates did not show statistical significance between radiologists and AI for historic reads (AUC difference, 0.06; 95% CI: 0, 0.14; P = .152) but were significantly higher for AI in reader studies (AUC difference, 0.05; 95% CI: 0.2, 0.8; P = .002) and in DBT studies (AUC difference, 0.11; 95% CI: 0.06, 0.16; P < .001).

Meta-regression analysis results investigating the effect of potential confounders on the performances of radiologists and standalone AI are summarized in Tables 3, S2, and S3. In brief, the study type (reader studies vs historic reads) had a significant effect on the diagnostic odds ratios of both human and AI reading results due to a highly significant effect on specificity (P ≤ .001 and P ≤ .002, respectively). Only minor and inconsistent differences in diagnostic performance metrics were observed for the standard of reference for negative findings and whether the AI algorithm was commercially available. The threshold of the AI system had a significant but variable effect on diagnostic performance, depending on the type of threshold used (Table 3). Studies where the vendor was involved in study design showed a lower overall diagnostic performance of radiologists in terms of the diagnostic odds ratio and specificity, while only specificity was significantly affected by this covariate for AI (Tables 3, S2, S3).

Discussion

Our meta-analysis of studies on the standalone performances of artificial intelligence (AI) for interpretation of digital mammography and digital breast tomosynthesis (DBT) shows that current algorithms perform on par with, if not better than, the average performance of breast radiologists. Pooled areas under the receiver operating characteristic curve (AUCs) for standalone AI were statistically superior to those of radiologists in digital mammography reader studies (AUC difference, 0.05; 95% CI: 0.2, 0.8; P = .002) and DBT studies (AUC difference, 0.11; 95% CI: 0.06, 0.16; P < .001). In the seven studies using historic reads, the AUC difference between radiologists and

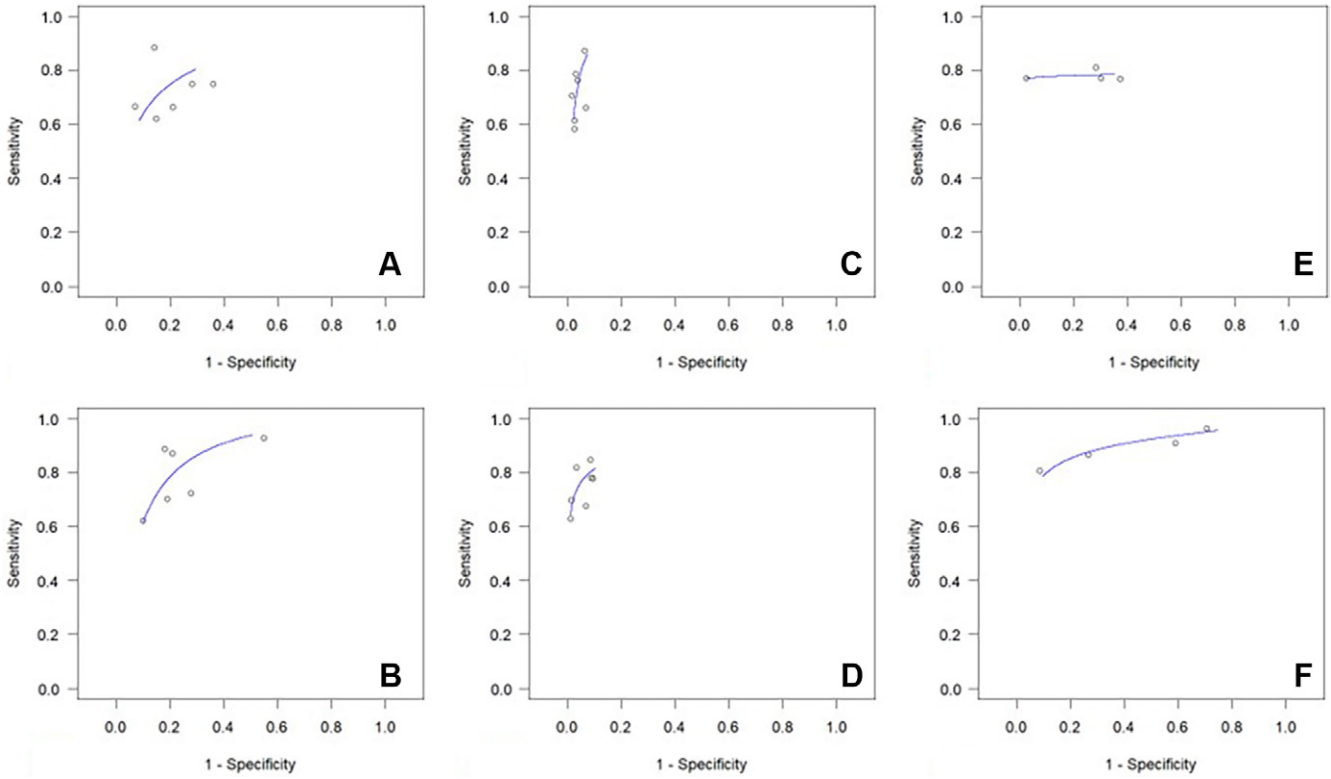

Figure 4: Summary receiver operating characteristic curves show performances of radiologists (top) and artificial intelligence (bottom) according to **(A, B)** reader studies, **(C, D)** historic reads, and **(E, F)** digital breast tomosynthesis (DBT) studies. Lines represent summary receiver operating characteristic curves and circles represent individual study results. Reader studies are defined as studies in which radiologists interpreted mammograms for the study, and historic reads are defined as studies using retrospective interpretation in clinical practice.

standalone AI was 0.06 (95% CI: 0, 0.14; $P = .152$). Although the pooled AUC was significantly higher in AI for DBT studies, there were only four studies evaluating the performance of DBT, and CIs around the estimates are wide. More studies are needed to ascertain the performance of standalone AI for interpretation of DBT.

The combination of sensitivity (alternatively, cancer detection rate) and specificity (alternatively, recall rate) defines the operating point for any given reader—AI or radiologist. Radiologists' operating points in screening settings are usually driven by what is locally regarded as an acceptable recall rate. This may vary from 2% to 3% in northern Europe to up to 12% in the United States, corresponding to a specificity range of approximately 90%–98%. AI thresholds are set differently within the included studies; in general, thresholds are set at a higher sensitivity than we would expect from human readers (81% vs 74%) and a somewhat lower specificity (86% vs 90%) (Table 2). According to the meta-regression analysis results, this seems to explain some of the observed variability between studies (Tables 3, S2, S3). Setting a threshold for AI is an important choice when implementing AI in practice, and the optimum may be different for different use cases. For instance, an AI threshold at high sensitivity might be used when AI is used for triaging, when all recalls are still performed by humans, whereas a threshold for higher specificity may be essential when AI is used as an additional (or independent) reader.

Our meta-analysis included six reader studies on enriched populations. Reader studies have the advantage that readers can be asked to provide a level-of-suspicion score on a continuous scale, allowing for the construction of receiver operating characteristic curves that can be directly compared with the AUC of AI systems. However, it is evident that the performance of radiologists in these studies does not truly represent real-world interpretation. Gur et al (30) showed that radiologists reading mammograms in a laboratory setting not only perform worse, but also demonstrate far more variability than in clinical practice, which is underlined by the observed lower specificity in the reader studies compared with the historical cohort studies in our meta-analysis. Still, although the AUC was lower (0.89 vs 0.96), standalone AI in historic cohort studies did not show significant differences to radiologists, indicating that current AI systems are on par with radiologists.

There are two prior systematic reviews on standalone AI performances. Freeman et al (31) included 12 retrospective studies with 131 822 women screened. The authors found 94% (34 of 36) of AI systems were less accurate than the original radiologist, and all were less accurate than the original consensus of two radiologists. Somewhat different results were seen in the meta-analysis performed by Hickman et al (32). Among the five retrospective studies on breast cancer detection that included 185 252 cases, standalone performances of machine learning systems (AUC, 0.89) exceeded human reader performances (AUC,

Table 3: Meta-Regression Analysis Comparing the Diagnostic Odds Ratio for Radiologists and Standalone AI according to Subgroup Feature

	Radiologists			AI		
Feature	β	SE	*P* Value	β	SE	*P* Value
Study type						
Reader study	…	…	…	…	…	…
Historic reads	0.81 (0.35, 1.28)	0.24	<.001	1.00 (0.37, 1.62)	0.32	.002
Definition for negative examinations						
Histopathologic finding or ≥2 years FU	…	…	…	…	…	…
Histopathologic finding or ≥1 years FU	−0.60 (−0.90, −0.31)	0.15	<.001	−0.41 (−1.12, 0.31)	0.37	.27
≥2 years FU	−1.29 (−2.19, −0.37)	0.46	.01	−0.02 (−1.09, 1.06)	0.55	.98
≥1 year FU	−0.38 (-0.95, 0.19)	0.29	.19	−0.78 (−1.08, −0.47)	0.16	<.001
Not specified	−2.56 (−3.49, 1.63)	0.474	<.001	−1.83 (−2.93, −0.73)	0.56	.001
AI algorithm						
Commercially available	…	…	…	…	…	…
In-house software	0.77 (−0.38, 1.92)	0.59	.19	0.39 (−1.07, 1.85)	0.74	.60
Not specified	1.58 (0.37, 2.80)	0.62	.01	0.17 (−1.10, 1.43)	0.64	.80
Role of vendor						
Involved in study design	…	…	…	…	…	…
Not involved in study design	1.63 (0.75, 2.51)	0.45	<.001	0.96 (−0.09, 2.02)	0.54	.07
AI threshold						
Vendor suggested	…	…	…	…	…	…
Set at specificity of radiologist	−0.25 (−0.53, 0.04)	0.14	.09	−0.33 (−0.62, −0.03)	0.15	.03
Set at maximum sensitivity of radiologist (rule out)	−0.76 (−1.74, 0.21)	0.50	.13	−1.16 (−2.32, 0.01)	0.60	.05
Defined for the study	−1.19 (−1.51, −0.86)	0.17	<.001	−0.73 (−1.08, −0.37)	0.18	<.001
Not specified	−1.74 (−2.69, −0.78)	0.49	<.001	−0.25 (−1.39, 0.86)	0.58	.65

Note.—Data in parentheses are 95% CIs. Reader studies are defined as studies in which radiologists interpreted mammograms for the study, and historic reads are defined as studies using retrospective interpretation in clinical practice. *P* values correspond to comparison to the reference value and $P < .05$ is considered to indicate statistical significance. AI = artificial intelligence, FU = follow-up, SE = standard error.

0.85), which is in line with our results. Both previous studies concluded that rigorous and independent external prospective testing of AI systems is required to estimate the true effect in clinical practice.

Our study had several limitations. First, all included studies are noninterventional (ie, AI assessments did not affect the actual diagnostic workflow) and, therefore, it is unclear whether the assumed cancer detection would have led to actual cancer detection in clinical practice. Second, we did not have access to individual patient data from the included studies and, therefore, only the aggregated data were used for analysis. Third, we analyzed AI systems from multiple vendors and have reported outcomes as an average of AI performance across vendors and not of a specific AI system per se. Lastly, whether different AI systems affect performances could not be evaluated in a meta-regression analysis with the available studies.

In conclusion, when used for interpretation of screening mammograms, standalone artificial intelligence (AI) for digital mammography performed as well as or better than individual breast radiologists or average reader outcomes. There is still an insufficient number of studies to assess the standalone performance of AI systems for interpretation of digital breast tomosynthesis. Future efforts should focus on different implementation strategies and continuous quality control to ensure that the retrospective results lead to both improved cancer detection and optimization of screening programs in a prospective setting.

Author contributions: Guarantors of integrity of entire study, **L.M., R.M.M.**; study concepts/study 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, **J.H.Y., F.S., E.F.C., C.D.L., L.A.M., R.M.N., I.V., L.M., R.M.M.**; clinical studies, **E.F.C., E.A.M., L.M.**; experimental studies, **F.J.G., L.M.**; statistical analysis, **P.A.T.B., L.M., R.M.M.**; and manuscript editing, **J.H.Y., F.S., P.A.T.B., E.F.C., C.D.L., E.A.M., L.A.M., R.M.N., N.S., I.V., L.M., R.M.M.**

Disclosures of conflicts of interest: J.H.Y. No relevant relationships. **F.S.** Grants or contracts from Region Stockholm and Swedish Breast Cancer Association; speaker fees from Lunit. **P.A.T.B.** Secretary general and executive board member of European Society of Breast Imaging (EUSOBI). **E.F.C.** Grants or contracts from iCAD, OM1, National Institutes of Health (NIH), and National Cancer Institute; consulting fees from iCAD; lecture payments from Hologic, MedScape, and Aunt Minnie; meeting and/or travel support from the RSNA; University of Pennsylvania patents planned, issued, or pending; leadership role at Society of Breast Imaging. **F.J.G.** Grants or contracts from Lunit, Bayer, Hologic, and GE Healthcare; consulting fees from Alphabet and Kheiron; lecture payment and receipt of equipment from GE Healthcare; president, EUSOBI. **C.D.L.** Grants or contracts from Breast Cancer Research Foundation, NIH, GE Healthcare, and Hologic; consulting fees from Clairity; meeting travel support from GE Healthcare, Hologic, and Clairity; co-founder of Clairity. **E.A.M.** Grants or contracts from Komen; lecture payments from Bayer and Guerbet; advisory board, Bracco and Kheiron; stockholder, Kheiron and Reveal Pharma. **L.A.M.** Research funding from IBM and Cepheid;

consulting fees and/or honoraria from Hologic; chair of Diversity and Inclusion Committee, Maryland Radiological Society. **R.M.N.** Institutional grants or contracts from GE Healthcare, Koios Medical, Hologic, and iCAD; royalties from Hologic; consulting fees from maiData; advisory board, iCAD. **N.S.** Lecture payments from BARD and Hologic. **I.V.** No relevant relationships. **L.M.** Editor of *Radiology*; grants or contracts from Siemens Foundation, Gordon and Betty Moore Foundation, Mary Kay Foundation, and Google; consulting fees from Guerbet and iCAD; meeting and/or travel support from British Society of Breast Radiology and European Society of Breast Imaging; board member of Society of Breast Imaging and International Society for Magnetic Resonance in Medicine; stockholder, Lunit. **R.M.M.** Associate editor of breast imaging for *Radiology*; grants or contracts from Dutch Research Council, Dutch Cancer Society, European Union, Siemens Healthineers, Bayer Healthcare, Beckton-Dickinson, Medtronic, Screenpoint Medical, Lunit, Koning, and PA Imaging; royalties from Elsevier; consulting fees from Bayer, Guerbet, Siemens, Screenpoint Medical, and Beckton-Dickinson; lecture payments from Siemens, Bayer, and Beckton-Dickinson; data safety monitoring board for SMALL trial; executive board member, EUSOBI; research committee member, ESR Advisory; editorial board member for *European Radiology*, advisory board member, Dutch Cancer Society; clinical advisory board member, Oncode Institute.

References

1. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet 2012;380(9855):1778–1786.
2. Gøtzsche PC, Jørgensen KJ. Screening for breast cancer with mammography. Cochrane Database Syst Rev 2013;2013(6):CD001877.
3. Tabár L, Yen AM, Wu WY, et al. Insights from the breast cancer screening trials: how screening affects the natural history of breast cancer and implications for evaluating service screening programs. Breast J 2015;21(1):13–20.
4. International Agency for Research on Cancer. Breast cancer screening. In: IARC handbooks of cancer prevention, Vol 15. IARC Press, 2015.
5. Hoff SR, Abrahamsen AL, Samset JH, Vigeland E, Klepp O, Hofvind S. Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography-- results from a retrospective review. Radiology 2012;264(2):378–386.
6. Hovda T, Tsuruda K, Hoff SR, Sahlberg KK, Hofvind S. Radiological review of prior screening mammograms of screen-detected breast cancer. Eur Radiol 2021;31(4):2568–2579.
7. Lamb LR, Mohallem Fonseca M, Verma R, Seely JM. Missed Breast Cancer: Effects of Subconscious Bias and Lesion Characteristics. RadioGraphics 2020;40(4):941–960.
8. Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. Br J Radiol 2020;93(1106):20190610.
9. Sechopoulos I, Mann RM. Stand-alone artificial intelligence - The future of breast cancer screening? Breast 2020;49:254–260.
10. 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(10):1581–1588.
11. Lee CS, Moy L, Hughes D, et al. Radiologist Characteristics Associated with Interpretive Performance of Screening Mammography: A National Mammography Database (NMD) Study. Radiology 2021;300(3):518–528.
12. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001;220(3):781–786.
13. Gilbert FJ, Astley SM, McGee MA, et al. Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program. Radiology 2006;241(1):47–53.
14. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175(11):1828–1837.
15. Nishikawa RM, Bae KT. Importance of Better Human-Computer Interaction in the Era of Deep Learning: Mammography Computer-Aided Diagnosis as a Use Case. J Am Coll Radiol 2018;15(1 Pt A):49–52.
16. Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. Clin Radiol 2019;74(5):357–366.
17. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
18. 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.
19. Kim H-E, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020;2(3):e138–e148.
20. 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(6):e190208.
21. 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.
22. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94. [Published correction appears in Nature 2020;586(7829):E19.]
23. 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.
24. 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(3):1687–1692.
25. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020;370:m2632.
26. 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.
27. Yang B, Mallett S, Takwoingi Y, et al. QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. Ann Intern Med 2021;174(11):1592–1599.
28. 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.
29. Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Stand-Alone Use of Artificial Intelligence for Digital Mammography and Digital Breast Tomosynthesis Screening: A Retrospective Evaluation. Radiology 2022;302(3):535–542.
30. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology 2008;249(1):47–53.
31. Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021;374:n1872.
32. Hickman SE, Woitek R, Le EPV, et al. Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. Radiology 2022;302(1):88–104.
33. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. Invest Radiol 2017;52(7):434–440.
34. Conant EF, Toledano AY, Periaswamy S, et al. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. Radiol Artif Intell 2019;1(4):e180096.
35. Sasaki M, Tozaki M, Rodríguez-Ruiz A, et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer 2020;27(4):642–651.
36. Pinto MC, Rodriguez-Ruiz A, Pedersen K, et al. Impact of Artificial Intelligence Decision Support Using Deep Learning on Breast Cancer Screening Interpretation with Single-View Wide-Angle Digital Breast Tomosynthesis. Radiology 2021;300(3):529–536.
37. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep 2019;9(1):12495.
38. Dang LA, Chazard E, Poncelet E, et al. Impact of artificial intelligence in breast cancer screening with mammography. Breast Cancer 2022;29(6):967–977.
39. Larsen M, Aglen CF, Lee CI, et al. Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program. Radiology 2022;303(3):502–511.
40. Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, et al. An Artificial Intelligence-based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. Radiology 2022;304(1):41–49.
41. Lee JH, Kim KH, Lee EH, et al. Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study. Korean J Radiol 2022;23(5):505–516.
42. Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet Digit Health 2022;4(7):e507–e519.
43. Sharma N, Ng AY, James JJ, et al. Retrospective large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. medRxiv 2021. Revision posted June 10, 2022.
44. Shoshan Y, Bakalo R, Gilboa-Solomon F, et al. Artificial Intelligence for Reducing Workload in Breast Cancer Screening with Digital Breast Tomosynthesis. Radiology 2022;303(1):69–77.