Title

Independent External Validation of Artificial Intelligence Algorithms for Automated Interpretation of Screening Mammography: A Systematic Review.

Journal

Journal of the American College of Radiology, 19(2 Pt A)

Authors

Anderson, Anna
Marinovich, M
Houssami, Nehmat
et al.

Publication Date

2022-02-01

DOI

10.1016/j.jacr.2021.11.008

Peer reviewed

HHS Public Access

Author manuscript

J Am Coll Radiol. Author manuscript; available in PMC 2023 February 01.

Published in final edited form as:

J Am Coll Radiol. 2022 February ; 19(2 Pt A): 259–273. doi:10.1016/j.jacr.2021.11.008.

Independent External Validation of Artificial Intelligence Algorithms for Automated Interpretation of Screening Mammography: A Systematic Review

Anna W. Anderson, MD[1,#], **M. Luke Marinovich, PhD, MPH**[2,#], **Nehmat Houssami, MBBS, PhD**[3], **Kathryn P. Lowry, MD**[1], **Joann G. Elmore, MD, MPH**[5], **Diana S.M. Buist, PhD, MPH**[4], **Solveig Hofvind, PhD**[6], **Christoph I. Lee, MD, MS**[1]

[1]Department of Radiology, University of Washington School of Medicine, Seattle, WA

[2]Curtin School of Population Health, Curtin University, Bentley, Western Australia, Australia

[3]The Daffodil Centre, the University of Sydney, a joint venture with Cancer Council NSW, Sydney, New South Wales, Australia

[4]Kaiser Permanente Washington Health Research Institute, Seattle, WA

[5]David Geffen School of Medicine at University of California at Los Angeles, Los Angeles, CA

[6]Cancer Registry of Norway, Oslo, Norway

Abstract

Purpose: To describe the current state of science regarding independent external validation of artificial intelligence (AI) technologies for screening mammography.

Corresponding author: Christoph Lee, MD, MS, 825 Eastlake Avenue East, G2-600, Seattle, WA 98109, Phone: 206-606-6783; stophlee@uw.edu, @christophleemd .

[#]A.W.A. and M.L.M. contributed equally to this work as co-first authors.

ICMJE statement:

Substantial contributions to conception or design of the work: all authors.
Acquisition, analysis, or interpretation of data for the work: all authors.
Drafting of the manuscript: CIL, AWA.
Revising the manuscript critically for important intellectual content: all authors.

Leadership roles:

All authors report being employed at non-profit institutions. Dr. Lee is Director of the Northwest Screening and Cancer Outcomes Research Enterprise at the University of Washington and Deputy Editor of *JACR*. Dr. Houssami is NBCF Chair in Breast Cancer Prevention at the University of Sydney and Co-Editor of *The Breast*. Dr. Elmore is the Director of UCLA's National Clinician Scholars Program and Editor-in-Chief of Adult Primary Care at *Up-To-Date*. Dr. Buist is Director of Research and Strategic Partnerships at Kaiser Permanente Washington Health Research Institute. Dr. Hofvind is Section Head of Breast Cancer Screening at the Cancer Registry of Norway.

Data statement:

The authors declare that they had full access to all of the data in this study and the authors take complete responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of interest disclosures:

CIL and DSMB are co-authors of one paper included in this systematic review. CIL receives personal fees from GRAIL, Inc. for service on a data safety monitoring board and from the American College of Radiology for journal editorial board work.

Materials/Methods: We performed a systematic review across five databases (Embase, PubMed, IEEE Explore, Engineer Village, and Arxiv) through December 10, 2020. Studies that used screening exams from real-world settings to externally validate AI algorithms for mammographic cancer detection were included. The main outcome was diagnostic accuracy defined by area under the receiver operating characteristic curve (AUC). Performance was also compared between radiologists and either standalone AI or combined radiologist and AI interpretation. Study quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool.

Results: After data extraction, 13 studies met inclusion criteria (148,361 total patients). Most (77%10/13) studies evaluated commercially available AI algorithms. Studies included retrospective reader studies (46%,6/13), retrospective simulation studies (38%,5/13), or both (15%,2/13). Across 5 studies comparing standalone AI to radiologists, 60% (3/5) demonstrated improved accuracy with AI (AUC improvement range, 0.02–0.13). All 5 studies comparing combined radiologist and AI interpretation to radiologists alone demonstrated improved accuracy with AI (AUC improvement range, 0.028–0.115). Most studies had risk of bias or applicability concerns for patient selection (69%,9/13) and the reference standard (69%,9/13). Only two studies obtained ground truth cancer outcomes through regional cancer registry linkage.

Conclusions: To date, external validation efforts for AI screening mammography technologies suggest small potential diagnostic accuracy improvements but have been retrospective in nature and suffer from risk of bias and applicability concerns.

Summary Sentence

Independent AI algorithm validation for automated mammography screening interpretation would benefit from development of large, diverse, real-world screening cohorts with linkage to regional cancer registries for robust ground truth.

Introduction

Emerging artificial intelligence (AI) technologies in health care hold promise for improving clinical efficiency and patient outcomes[1, 2]. In medical imaging, developing and incorporating AI algorithms for automated mammography screening interpretation has become a primary use case[3]. Mammograms are highly amenable to AI algorithm development and training due to their standardized imaging positions and projections, large amounts of available digital data, and binary outcome of cancer or no cancer[4]. Multiple recent publications have shown promise for AI-driven screening mammography interpretation both as a standalone tool and as an adjunct tool for interpreting radiologists[5–7].

An earlier scoping review of AI's potential in mammography screening identified methodologic limitations of AI performance assessment including use of non-representative imaging data for model training, limited independent external validation, and potential training data bias[8]. Thus far, most AI algorithm validation studies for mammography screening have used internal validation with a subset of exams from the training cohort to test algorithm performance, which can inflate AI performance due to model overfitting[9]. To truly demonstrate AI algorithm generalizability, external validation is needed using independent target populations not used in training[10]. Moreover, many commonly used

publicly available datasets (such as the Optimam Mammography Database and Digital Database for Screening Mammography (DDSM)[11]) are heavily used in AI algorithm training and therefore are not appropriate sources for independent, external validation[12]. Ideally, AI algorithm performance should be externally validated using real-world screening data to demonstrate generalizability and to inform clinical adoption[13].

We aimed to summarize the current state of the science regarding independent external validation of artificial intelligence (AI) technologies for screening mammography using real-world clinical data and whether the evidence is of high enough quality for widespread clinical adoption. To meet this objective, we performed a comprehensive systematic literature review of studies using real-world screening data for independent, external validation of promising AI algorithms for automated mammography interpretation.

Methods

Our systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement[14]. Our review protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO CRD42021230390). The study was exempt from Institutional Review Board approval as only publicly available data were collected and assessed.

Data Sources and Searches

With the assistance of a professional medical reference librarian, we searched Embase, PubMed, IEEE Explore, Engineer Village, and Arxiv databases from inception through December 10, 2020; the latter three were included given the nature of AI algorithm development and testing involving data scientists and engineers outside the traditional medical research community. In each database, subject headings and free text terms were used to search across three broad concepts: artificial intelligence, mammography, and diagnostic accuracy. The detailed search strategy including search terms for each of the five databases is available in the online Supplement (eTable 1). We only searched for studies with an English translation and where a full manuscript was available.

Study Selection

Studies that used screening mammography exams from real-world clinical settings to independently evaluate AI algorithm cancer detection accuracy were included. We excluded AI algorithm evaluation studies that only used publicly available datasets (e.g., DDSM, mini-Mammographic Image Analysis Society (MIAS) database, Optimam Mammography Database), as these datasets have been heavily used in training and developing AI algorithms. Similarly, we excluded studies that involved internal AI algorithm validation using a subset of the original cohort used to train and develop the algorithms, as such exercises are known to suffer from model overfitting[9]. If studies performed both internal and external validation, we only recorded findings from the external validation portion.

Studies were included if they validated AI alone or in combination with radiologists. Studies validating a single AI algorithm or ensemble models combining multiple algorithms were eligible. We excluded studies that: detailed model training only; involved AI algorithms

developed to only detect specific (restricted) imaging features and not all breast cancers on mammography (e.g., mass detection only, calcification detection only); only involved AI for future cancer risk prediction rather than cancer detection on images; and studies that focused on AI use for improving radiologist workflow (e.g., triage of negative mammograms, decreased interpretation time) rather than automated cancer detection.

Outcomes

The main outcomes of interest were overall accuracy as defined by the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity of AI algorithms for breast cancer detection on real-world screening mammography cohorts. In the few studies that included double-reading by radiologists, only first-reader radiologist performance outcomes with and without AI are presented in our review to ensure comparability across all studies.

Data Extraction

Two authors (AA and CL) independently reviewed all titles and abstracts resulting from the literature search for inclusion and exclusion criteria, with conflicts resolved by a third author (LM). For studies published in online archives before medical journals, only the most recent published medical journal manuscript was included in our study.

We developed a standardized data extraction tool to collect study characteristics (eTable 2). Two reviewers (AA and CL) independently extracted data from each study at the time of full manuscript review. Any data extraction parameter disagreements were resolved by consensus. Data systematically collected for each manuscript include titles, author names, publication date, AI algorithm type (e.g., convolutional neural network), AI algorithm commercial availability, and if AI algorithms were originally trained on public and/or private datasets.

Detailed data collected regarding external validation datasets include a description of the clinical cohort (e.g., clinical setting), screening program interval (e.g., annual, biennial), exam years, imaging modality (digital mammography (DM), digital breast tomosynthesis (DBT), or both), total number of screening exams evaluated, total number of cancer-positive exams evaluated, and the follow-up period for determining interval cancers (e.g., 12-months). We did not require specific reference standards for outcome measures, but did record how studies determined cancer ground truth (e.g., biopsy results, cancer registry linkage).

Data Synthesis and Analysis

We performed a narrative synthesis of the published literature on external validation of promising AI algorithms using real-world clinical mammography exams. Due to methodological heterogeneity across study populations, including enriched reader study exam sets and differences in reported comparator groups and outcome measures, we adopted a descriptive approach for our primary analyses. Estimates of test accuracy (AUC, sensitivity, specificity) for each study were tabulated. For studies reporting AUCs for AI compared with radiologists, descriptive plots of study level differences in the AUC were

generated. For studies that compared the sensitivity and specificity of AI versus radiologists, scatterplots of study-specific (sensitivity, 1-specificity) pairs with exact (Clopper-Pearson) uncertainty regions were plotted in ROC space, and points were joined to highlight within-study comparisons. Scatterplots were examined for inconsistency (statistical heterogeneity) in the direction of the differences in AI and radiologist estimates. When inconsistency was observed, no further analysis was undertaken. When studies were consistent in the direction of the differences, meta-analysis of accuracy was undertaken using the hierarchical summary receiver operating characteristics (HSROC) model proposed by Rutter and Gatsonis[15] to estimate summary AUCs. Because the threshold used to define test positivity varied between studies, summary estimates of sensitivity and specificity were not derived. Fitted sROC curves were overlaid on the scatterplots of study-specific estimates and were restricted to the range of data points. Analyses were undertaken using the 'mada' package[16] in R 4.0.4 (R Project for Statistical Computing, Vienna, Austria).

Quality Assessment

Overall methodological quality of the studies was assessed independently by two reviewers (AA and CL) using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool[17]. The QUADAS-2 quality review emphasizes the risk for bias and concerns for applicability of the primary diagnostic accuracy studies (see Table 4). The QUADAS-2 guiding questions for evidence quality assessment were: 1) Patient selection: Could selection of patients (e.g., sampling design) have introduced bias, or is there concern that the patient population is not representative of the true target population? 2) Index test: Could interpretation of the index test (e.g., AI score, cut-off) have introduced bias, or is there concern that the index test could be interpreted subjectively? 3) Reference standard: Could the reference standard used (e.g., cancer ground truth) have introduced bias or does the target condition defined by the reference standard not match the review question? 4) Flow and timing: Could patient care flow (e.g., AI use timing) have introduced bias?

Results

Our literature search identified 5,072 citations (Figure 1); 1,376 citation titles were reviewed after deduplication, which, yielded 160 citations for dual, independent abstract review. After full abstract review, 66 manuscripts were identified as meeting major inclusion and exclusion criteria and were reviewed by two investigators; 6 were discussed with a third investigator for different review results between the initial two investigators. We extracted data on 25 manuscripts, of which 52% (13/25) were found to use real-world screening exams for true, independent external validation. The remaining 48% (12/25) of studies were found to use the same clinical exam datasets split into training and validation datasets (e.g., internal validation only) and thus were excluded.

Study Characteristics

All 13 external validation studies included in our systematic review evaluated convolutional neural network (CNN) deep learning models[5–7, 18–27] (Table 1). Most (77%, 10/13) evaluated commercially available AI algorithms[7, 18–20, 23–27], and were trained and internally validated using at least some private, proprietary datasets (92%, 12/13)[5–7, 19–27].

All studies were published between 2019–2020 and were retrospective reader studies (46%, 6/13)[18, 20, 22, 24, 26, 27], retrospective simulation studies (38%, 5/13)[5, 7, 19, 23, 25], or both (15%, 2/13)[6, 21]. We identified no published prospective, population-based evaluation studies through 2020.

External Validation Dataset Characteristics

Reader studies were performed using small screening exam collections in the US[6, 18, 20–22, 24, 26, 27], Germany[24], South Korea[20], and Japan[26] (Table 2). Almost all (88%, 7/8) reader studies involved multiple years of DM while only one involved DBT exams[18]. Exam numbers included in these enriched reader studies ranged between 122–500 total exams (66–160 exams were cancer-positive). These enriched case sets were interpreted by 5–24 individual radiologists. All 8 reader studies determined ground truth based on breast biopsy results or negative subsequent screening exams. The cancer follow-up periods were reported by 3 of the reader studies (38%, 3/8), with reported follow-up periods of 12–27 months[6, 22, 24].

Retrospective simulation studies were performed using screening cohorts from the US[5, 6], UK[6], Sweden[5, 7, 19], and China[21] (Table 2). Three studies drew their Swedish screening cohort from the same institution[5, 7, 19]. Two studies, McKinney et al.[6] and Schaffter et al.[5], used their entire screening populations and created their study cohorts from consecutive screening exams (or randomly sampled from consecutive screening exams) over multiple years (total exams 28,953 and 93,665, respectively). The other five studies[7, 19, 21, 23, 25] used case-cohort samples or convenience samples enriched with cancer-positive cases (total exams ranged from 1,633–8,805). While Rodriguez-Ruiz et al. reported exams interpreted by >100 radiologists in two of their studies[23, 25], these samples were comprised of convenience samples of exams collected from multiple prior reader studies. Of the retrospective simulation studies, 2 studies[5, 7] (29%, 2/7) had ground truth determined through robust linkage to regional cancer registries. The remaining simulation studies either did not report how they determined ground truth[19, 21] (29%, 2/7) or used breast biopsy results and subsequent negative screening to determine cancer status[6, 23, 25] (43%, 3/7). McKinney et al.[6] reported cancer follow-up periods of 27 months for their US evaluation cohort and 39 months for their UK evaluation cohort. All other simulation studies reported 1–2 year cancer follow-up[5, 7, 19, 25] (57%, 4/7) or did not report cancer follow-up periods[21, 23] (29%, 2/7).

Diagnostic Accuracy

The most common outcome measure reported was AUC, representing overall diagnostic accuracy. Across reader studies, five[18, 20, 22, 24, 27] provided comparisons of radiologist interpretive performance with vs. without AI while five[6, 20, 24, 26, 27] provided comparisons of radiologist performance vs. standalone AI (Table 3, Figure 2). Another reader study[21] provided AI standalone performance without a comparison group. Reported AUC for radiologists ranged from 0.62–0.87, AUC values for standalone AI ranged from 0.66–0.94, and reported AUC of radiologist+AI performance ranged from 0.80–0.89. All 5 reader studies[18, 20, 22, 24, 27] comparing combined radiologist+AI vs. radiologist alone demonstrated statistically significant improved AUC for radiologist+AI (Figure 2). However,

in the reader studies comparing radiologists vs. standalone AI, there were mixed results with three studies[6, 20, 24] showing superior AI performance and two studies[26, 27] demonstrating significantly worse AUC for standalone AI (Table 3).

For the simulation cohorts (9 different cohorts across 7 retrospective simulation studies), AUC was again the most frequently reported AI performance measure but most studies also reported sensitivity and/or specificity of radiologists and/or AI (Table 3). AUC values for standalone AI ranged between 0.81–0.97. Only one simulation study[25] reported AUC values for radiologists vs. standalone AI, with standalone AI demonstrating non-inferior performance defined as the lower 95% CI of the difference in AUCs not less than a margin of −0.05 (difference in AUC = 0.03) (Figure 2).

Comparisons of sensitivity and specificity for both study types (Figures 3a and 3b) reflected patterns of results for AUCs (Figure 2). Studies comparing radiologists vs. standalone AI (7 cohorts across 5 studies) showed inconsistent results, with AI exhibiting either higher or lower accuracy, or lower specificity with relatively smaller gains in sensitivity (Figure 3a). Studies comparing radiologists versus radiologists+AI (7 cohorts across 5 studies) consistently showed an increase in accuracy for radiologists+AI (squares above and/or to the left of diamonds in Figure 3b), with fitted sROC curves showing a small increase in the AUC that is consistent with the magnitudes of study-level differences observed in Figure 2.

Two simulation studies assessed ensemble model performance, which combined multiple individual models. Schaffter et al.[5] developed ensemble models using 8 top performing models in the Digital Mammography Dialogue on Reverse Engineering Assessment and Methods (DREAM) Challenge. The ensemble model outperformed the top performing algorithm in the Swedish (AUC 0.92 vs. 0.90) and the US evaluation cohorts (AUC 0.90 vs. 0.86). Performance improved further when radiologist performance was added to the ensemble model in both the Swedish and US evaluation cohorts (AUC 0.94 for both ensemble + radiologist models). Salim et al.[7] also explored performance of an ensemble of three commercial AI algorithms together. They found that their ensemble AI model performed as well as combined radiologist and any of the 3 single commercial AI algorithms (Table 3).

Quality Assessment

All studies had high risk or unclear risk of bias or applicability concerns (eFigure 1). Most studies[18, 20–27] (69%, 9/13) had high or unclear risk of bias or applicability concerns in patient selection due to their sampling designs or not obtaining consecutive exams from screening settings (Table 4). All studies suffered some risk of bias or applicability in the index test due to arbitrary AI score cut-offs and/or artificial combination with independent radiologist interpretations in simulation studies (e.g., no actual radiologist-AI interface). Most studies[18–21, 23–27] (69%, 9/13) had high or unclear risk of bias in the reference standard due to the lack of a robust cancer ground truth through linkage with regional cancer registries or long-term cancer follow-up beyond two years to define true negative and false negative screening exams.

Discussion

In our comprehensive systematic review of external validation studies for AI mammography technologies using real-world screening exams, we found that most studies of standalone AI or combined radiologist and AI interpretation demonstrated incremental diagnostic accuracy improvements over radiologist interpretation alone. However, all 13 studies published through 2020 were either retrospective reader or simulation studies with no prospective observational studies or clinical trials. Overall, there was some high or unclear risk of bias or applicability for all included studies. Only two studies linked to regional cancer registries to define cancer ground truth, leading to concerns for the reference standard and cancer determination in all other studies.

The most rigorous and largest simulation studies included study samples built from sequential screening exams from one or more institutions and compared or combined multiple AI algorithms. Salim et al.[7] found the best performance when combining three different commercial AI algorithms. Similarly, Schaffter et al.[5] created an ensemble model comprised of 8 top performing individual algorithms, and observed the highest diagnostic accuracy when the ensemble model was combined with radiologist performance. These studies suggest that ensemble models that aggregate predictions across multiple algorithms can improve performance over individual algorithms[5]. Nevertheless, the applicability and feasibility of ensemble modeling in clinical settings is questionable, especially using multiple commercial AI algorithms together. Both Salim et al.[7] and Schaffter et al.[5] also compared multiple AI algorithms, which is desirable for evaluation studies as some available AI algorithms may perform better than others within distinct screening populations[28].

Our systematic review highlights the urgent need for higher quality external validation studies of AI algorithms for mammography before widespread clinical adoption[9, 12], especially as multiple AI algorithms have gained regulatory approval and are now becoming commercially available both in the US and internationally[29]. The studies included in our systematic review were performed in rather homogeneous populations (distinct European/Caucasian or Asian populations). Future external validation studies need to demonstrate generalizability of AI algorithms across large, diverse screening populations in terms of race/ethnicity, breast cancer risk factors, imaging vendors, and imaging modalities (e.g., DBT vs. DM) and in screening settings where AI would be used (e.g., specific screening age ranges and screening intervals). The largest barriers to conducting such rigorous evaluations are both the availability of population-based screening data with clinical, demographic, and risk factor metadata, and linkage to cancer outcomes determined by regional cancer registries to ensure gold-standard ground truth[12]. More effort is required to collect and curate rich, population-based mammography datasets linked to both clinical metadata and long-term cancer outcomes for AI algorithm external validation purposes.

One other systematic review on diagnostic accuracy of AI on mammography was recently published[30], demonstrating consistent results with our study. However, in addition to diagnostic accuracy, our review adds to the existing literature by collecting more detailed information on specific methodological features important to external validation of AI algorithms. Our review also was more comprehensive, searching the literature from

inception and across five large databases that include engineering and data science databases outside of traditional medical science databases. By including these databases, we searched conference proceedings and online archives not found in traditional medical literature searches but frequented by AI algorithm developers in order to account for all publicly available external validation studies. Our study did have limitations. Our review was limited to English language publications and there may have been more recent publications not included in this systematic review as AI for breast imaging is a fast-moving field. We did not examine the impact of AI algorithms on future breast cancer risk prediction or improved radiologist workflow efficiency (e.g., triaging negative screening exams, less interpretation time), as these were beyond the scope of our review question. We examined AI from an exam-level perspective and not a breast-level or lesion-level perspective. None of the included studies detailed the impact of the presentation format of AI outputs on radiologist interpretation, and none of the studies discussed acceptability of AI from the medical-legal or ethical perspectives. Finally, to provide comparability across studies, we focused on AI as a standalone tool or AI as an adjunct tool for radiologists in single reading settings or first readers in double-reading settings (and not second readers). Future systematic reviews could focus on these additional important aspects of breast cancer screening and the impact of incorporating AI into routine screening practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We thank Teresa E. Jewell, MLIS, for her assistance in designing and executing the comprehensive literature search.

Funding:

CIL and JGE are funded by the National Cancer Institute (R37 CA240403). CIL and DSMB are supported by the National Cancer Institute (P01CA154292). MLM is funded by a National Breast Cancer Foundation Investigator Initiated Research Scheme grant (IIRS-20-011). NH is funded via NBCF Chair in Breast Cancer Prevention grant (EC-21-001) and NHMRC Investigator Leader grant (#1194410). KPL is funded by an American Cancer Society grant (CSDG-21-078-01-CPSH).

References

1. Matheny ME, Whicher D, Thadaney Israni S. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. JAMA. Feb 11 2020;323(6):509–510. doi:10.1001/jama.2019.21579 [PubMed: 31845963]
2. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? JAMA. Jun 18 2019;321(23):2281–2282. doi:10.1001/jama.2019.4914 [PubMed: 31107500]
3. Houssami N, Lee CI, Buist DSM, Tao D. Artificial intelligence for breast cancer screening: Opportunity or hype? Breast. Dec 2017;36:31–33. doi:10.1016/j.breast.2017.09.003 [PubMed: 28938172]
4. Trister AD, Buist DSM, Lee CI. Will Machine Learning Tip the Balance in Breast Cancer Screening? JAMA Oncol. Nov 1 2017;3(11):1463–1464. doi:10.1001/jamaoncol.2017.0473 [PubMed: 28472204]
5. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Netw Open. Mar 2 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265 [PubMed: 32119094]

6. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. Jan 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6 [PubMed: 31894144]
7. Salim M, Wahlin E, Dembrower K, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol. Oct 1 2020;6(10):1581–1588. doi:10.1001/jamaoncol.2020.3321 [PubMed: 32852536]
8. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. Expert Rev Med Devices. May 2019;16(5):351–362. doi:10.1080/17434440.2019.1610387 [PubMed: 30999781]
9. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology. Mar 2018;286(3):800–809. doi:10.1148/radiol.2017171920 [PubMed: 29309734]
10. England JR, Cheng PM. Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers. AJR Am J Roentgenol. Mar 2019;212(3):513–519. doi:10.2214/AJR.18.20490 [PubMed: 30557049]
11. Horsch A, Hapfelmeier A, Elter M. Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies. Int J Comput Assist Radiol Surg. Nov 2011;6(6):749–67. doi:10.1007/s11548-011-0553-9 [PubMed: 21448711]
12. Lee CI, Houssami N, Elmore JG, Buist DSM. Pathways to breast cancer screening artificial intelligence algorithm validation. Breast. Aug 2020;52:146–149. doi:10.1016/j.breast.2019.09.005
13. Thrall JH, Fessell D, Pandharipande PV. Rethinking the Approach to Artificial Intelligence for Medical Image Analysis: The Case for Precision Diagnosis. J Am Coll Radiol. Jan 2021;18(1 Pt B):174–179. doi:10.1016/j.jacr.2020.07.010 [PubMed: 33413896]
14. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. Jan 23 2018;319(4):388–396. doi:10.1001/jama.2017.19163 [PubMed: 29362800]
15. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. Oct 15 2001;20(19):2865–84. doi:10.1002/sim.942 [PubMed: 11568945]
16. Doebler PHH. Meta-analysis of diagnostic accuracy with mada. Available at: https://cran.r-project.org/web/packages/mada/index.html; Accessed October 18, 2021.
17. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. Oct 18 2011;155(8):529–36. doi:10.7326/0003-4819-155-8-201110180-00009 [PubMed: 22007046]
18. 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. Jul 31 2019;1(4):e180096. doi:10.1148/ryai.2019180096 [PubMed: 32076660]
19. Dembrower K, Wahlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health. Sep 2020;2(9):e468–e474. doi:10.1016/S2589-7500(20)30185-0
20. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health. Mar 2020;2(3):e138–e148. doi:10.1016/S2589-7500(20)30003-0 [PubMed: 33334578]
21. Lotter WDA, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using annotation-efficient deep learning approach. arXIV. 2019. https://arxiv.org/abs/1912.11027
22. Pacile 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. Nov 2020;2(6):e190208. doi:10.1148/ryai.2020190208 [PubMed: 33937844]
23. Rodriguez-Ruiz A, Mordang J, Karssemeijer N, Sechopoulos I, Mann RM. Can radiologists improve their breast cancer detection in mammography when using a deep learning based computer system as decision support? 2018:

24. Rodriguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology. Feb 2019;290(2):305–314. doi:10.1148/radiol.2018181371 [PubMed: 30457482]

25. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. Sep 1 2019;111(9):916–922. doi:10.1093/jnci/djy222 [PubMed: 30834436]

26. Sasaki M, Tozaki M, Rodriguez-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. Jul 2020;27(4):642–651. doi:10.1007/s12282-020-01061-8 [PubMed: 32052311]

27. Watanabe AT, Lim V, Vu HX, et al. Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. J Digit Imaging. Aug 2019;32(4):625–637. doi:10.1007/s10278-019-00192-5

28. Geras KJ, Mann RM, Moy L. Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. Radiology. Nov 2019;293(2):246–259. doi:10.1148/radiol.2019182627 [PubMed: 31549948]

29. Elmore JG, Lee CI. Data Quality, Data Sharing, and Moving Artificial Intelligence Forward. JAMA Netw Open. Aug 2 2021;4(8):e2119345. doi:10.1001/jamanetworkopen.2021.19345 [PubMed: 34398208]

30. 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 Sep 1;374:n1872. doi: 10.1136/bmj.n1872. [PubMed: 34470740]

Take-Home Points

- Independent, external validation of promising AI algorithms for automated mammography interpretation using real-world screening cohorts is currently in a nascent state, with only retrospective reader studies or simulation studies documented in the published literature.
- External validation studies for AI in mammography have thus far almost exclusively involved digital mammography exams rather than digital breast tomosynthesis exams.
- Most published external validation studies for AI in mammography suffer from patient selection bias, with use of either enriched reader study cohorts or convenience samples rather than sampling from consecutive screening examinations in real-world settings.
- Most published external validation studies for AI in mammography lack robust cancer registry linkage to determine cancer ground truth, creating a high risk of bias in the reference standard.

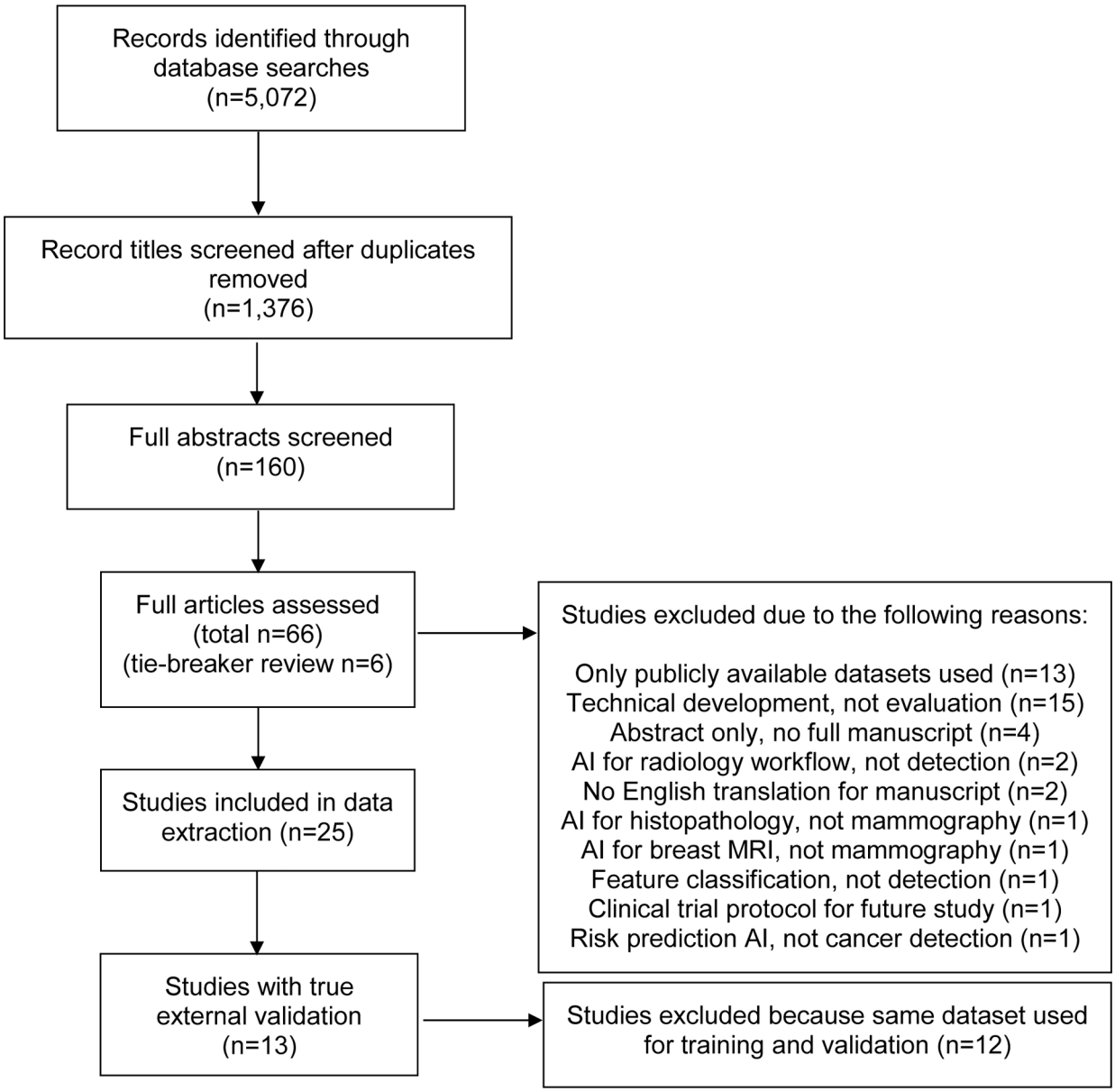

Figure 1.
PRISMA Flowchart

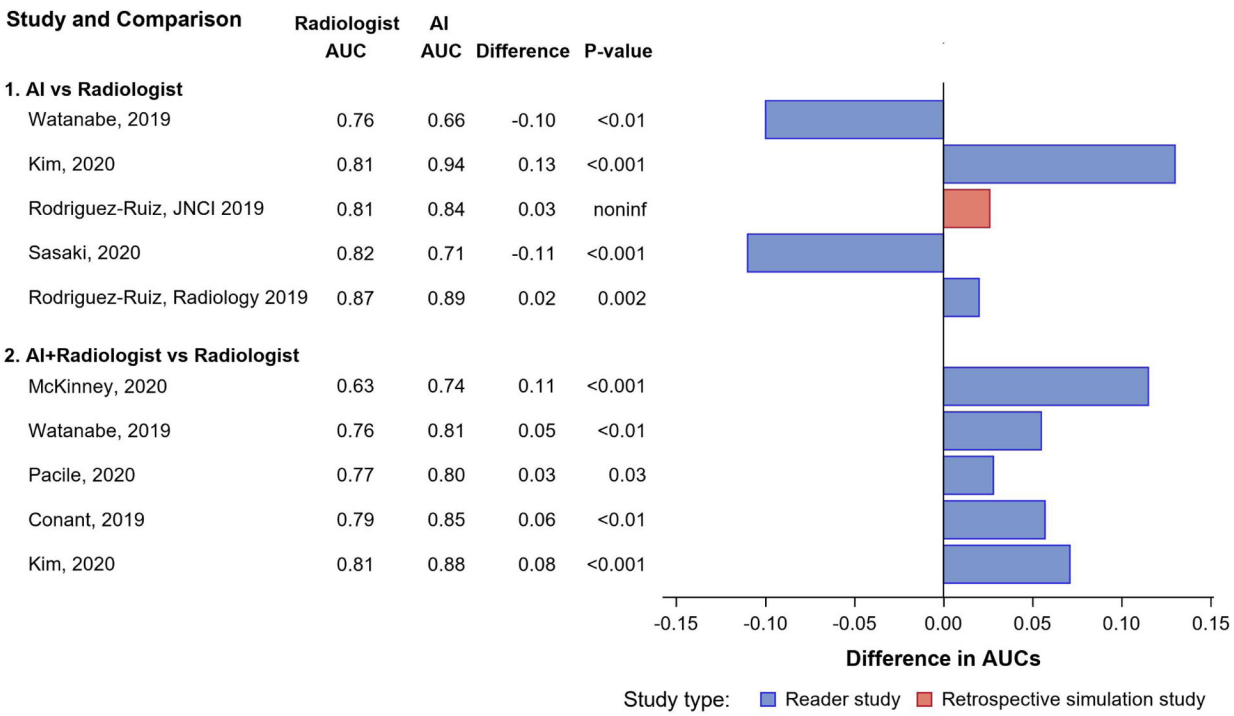

Figure 2. AI Diagnostic Accuracy in Studies with External Validation as Stand-Alone Tools and as Second Readers

Difference in the area under the receiver operating characteristic curve (AUC) between radiologist vs. either AI alone or AI and radiologist combined performance. All AUC values rounded to nearest hundredth. Noninf = non-inferiority.

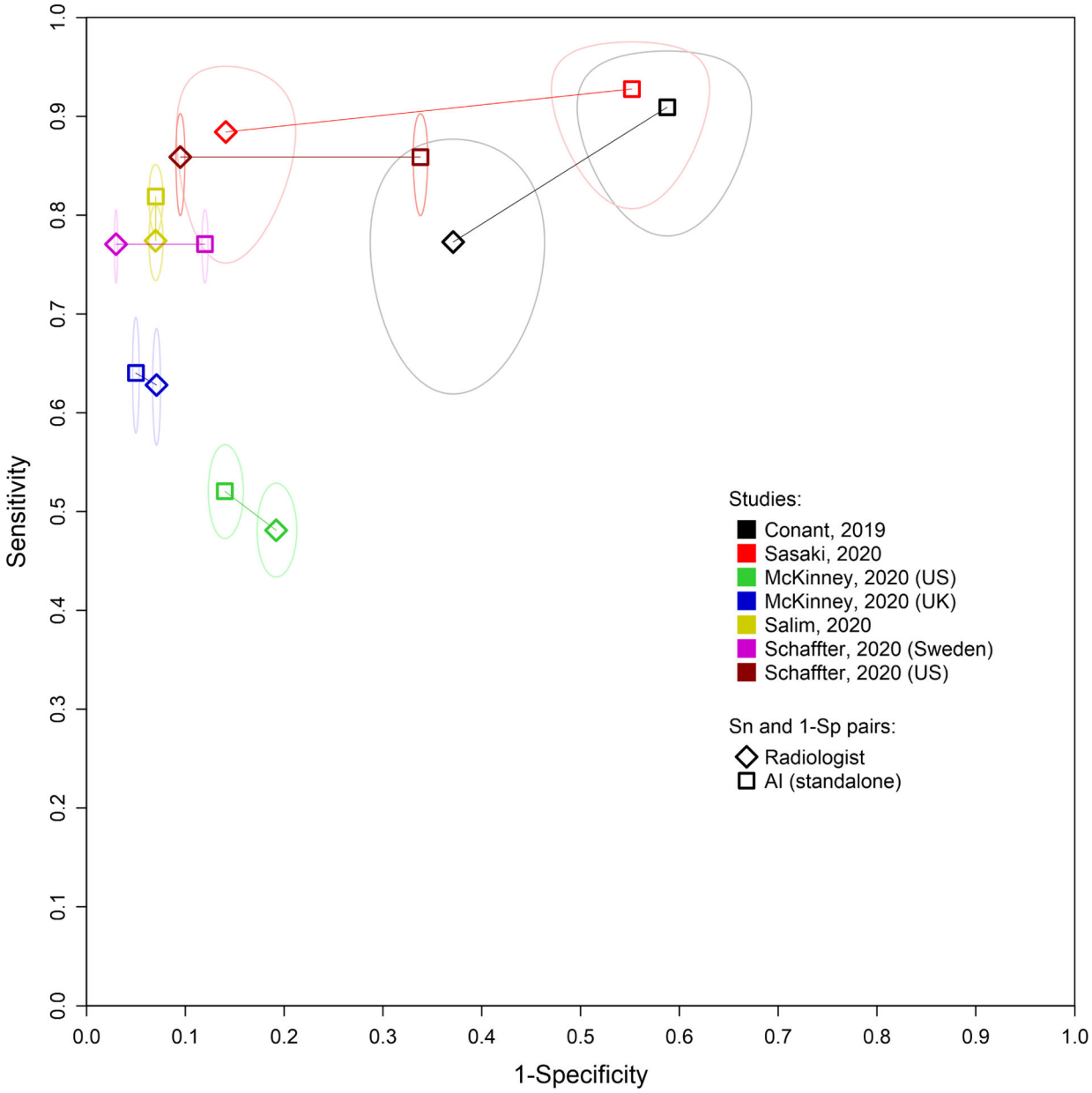

Figure 3a. Summary Receiver Operating Characteristics (sROC) Curves

Study-specific estimates of sensitivity (Sn) and specificity (Sp) for radiologists versus AI standalone.

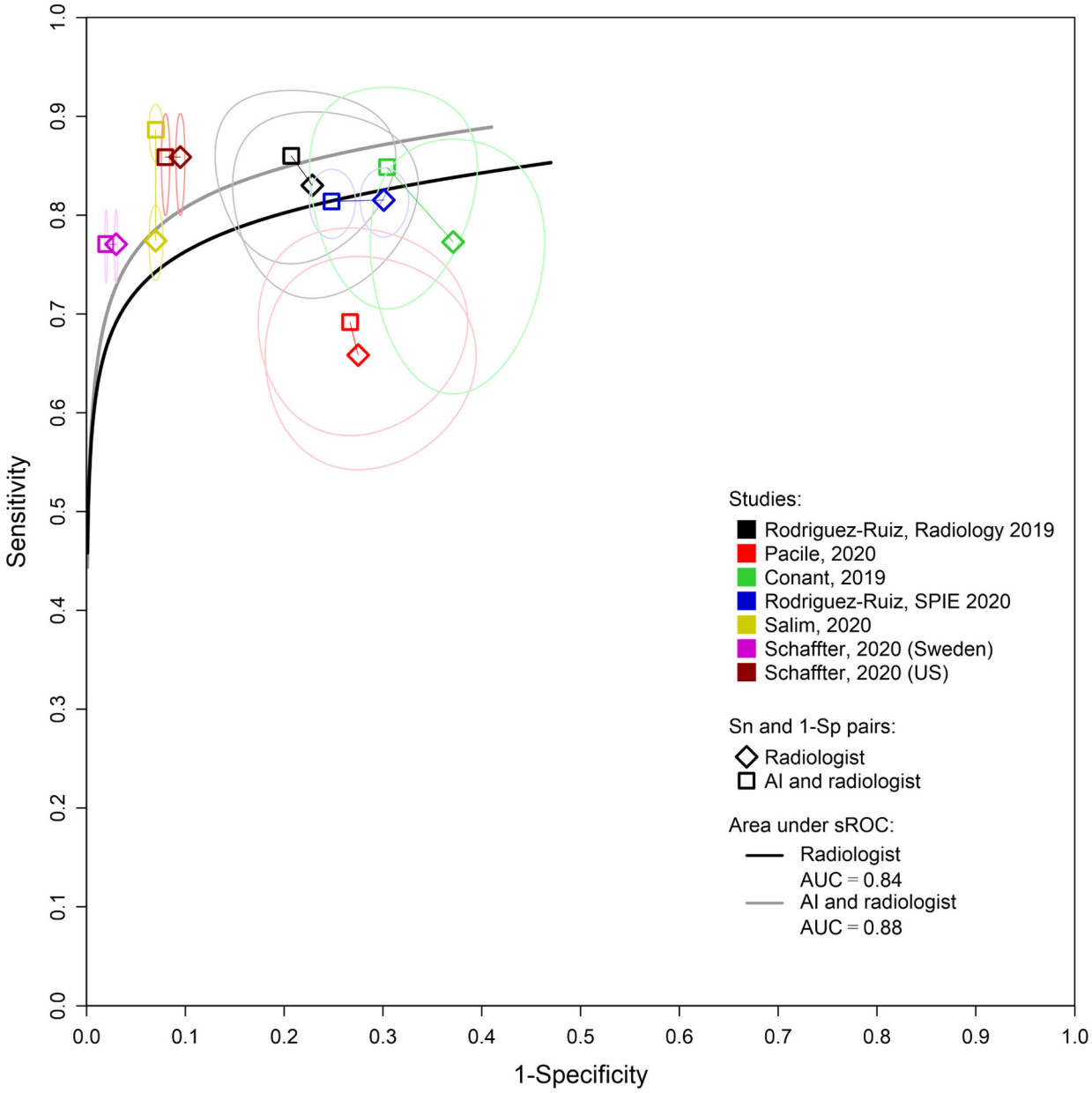

Figure 3b. Summary Receiver Operating Characteristics (sROC) Curves

Study-specific estimates of sensitivity (Sn) and specificity (Sp) for radiologists versus AI and radiologists combined.

Table 1.

General Characteristics of External Validation AI Algorithms

Study	Type of AI algorithm(s)	Commercially available?	Code publicly available	Type of training and internal validation datasets - public or private datasets	Study type
Conant, 2019	CNN model	Yes	No	Unspecified	Reader study
Lotter, 2019	CNN model	No *	No	Public and private datasets	Retrospective simulation study and reader study
Rodriguez-Ruiz, JNCI 2019	CNN model	Yes	No	Private dataset	Retrospective simulation study
Rodriguez-Ruiz, Radiology 2019	CNN model	Yes	No	Private dataset	Reader study
Watanabe, 2019	CNN model	Yes	No	Private dataset	Reader study
Dembrower, 2020	CNN model	Yes	No	Public and private dataset	Retrospective simulation study
Kim, 2020	CNN model	Yes	No	Public and private dataset	Reader study
McKinney, 2020	Ensemble of 3 CNN models	No	No	Public and private datasets	Retrospective simulation study and reader study
Pacile, 2020	CNN model	Yes	No	Private dataset	Reader study
Rodriguez-Ruiz, SPIE 2020	CNN model	Yes	No	Private dataset	Retrospective simulation study
Salim, 2020	3 CNN models	Yes	No	Public and private dataset	Retrospective simulation study
Sasaki, 2020	CNN model	Yes	No	Private dataset	Reader study
Schaffter, 2020	Multiple CNN models	No	Yes	Public and private dataset	Retrospective simulation study

* While not commercially available at time of evaluation, this AI tool is now undergoing the FDA approval process.

Table 2.

External Validation Real-World Dataset Characteristics

Study	Country and Setting	Exam type	Exam year(s)	Number of total exams	Number of cancers	Number of interpreting radiologists	Cancer follow-up period	Screening program interval	Pathology ground truth source
Reader Studies									
Conant, 2019	US: 7 different facilities	DBT	2012–2017	260	66	24	NR	1- or 2-year	Biopsy results
Lotter, 2019 (reader study portion)	US: single health center	DM	2011–2014	405	131	5	NR	1- or 2-year	Biopsy results or subsequent negative screening
Rodriguez-Ruiz, Radiology 2019	US: one center; Germany: one center	DM	2013–2017	240	100	14	1 year	NR	Biopsy results or subsequent negative screening
Watanabe, 2019	US: one community healthcare facility	DM	2008–2016	122	90	7	NR	NR	Biopsy results
Kim, 2020	South Korea: two institutions	DM	2009–2018	320	160	14	NR	1-year	Biopsy results
McKinney, 2020 (reader study portion)	US: single academic center	DM	2001–2018	500	125	6	27 months	1- or 2-year	Biopsy results or subsequent negative screening
Pacile, 2020	Not reported	DM	2013–2016	240	120	14	18 months	NR	Biopsy results
Sasaki, 2020	Japan: one outpatient center	DM	2018	310	69	NR	NR	NR	Biopsy results
Retrospective Simulation Studies									
Lotter, 2019 (simulation study portion)	China: single hospital	DM	2012–2017	1633	533	NR	NR	NR	NR
Rodriguez-Ruiz, JNCI 2019	Variable countries: nine datasets collected from prior reader studies	DM	Not reported	2,652	653	101	1 year	NR	Biopsy results or subsequent negative screening
Dembrower, 2020	Sweden: one major medical center	DM	2009–2015	7,364	547	NR	2 years	2-year	Unclear
McKinney, 2020 (simulation study portion)	US: single academic center; UK: two screening centers	DM	2001–2018	28,953	1,100	US: NR UK: 51	US: 27 months; UK: 39 months	US: 1- or 2- year; UK: 3-year	Biopsy results or subsequent negative screening
Rodriguez-Ruiz, SPIE 2020	Variable countries: data from 10 prior study cohorts	DM	Not reported	2,892	752	115	NR	NR	Biopsy results

Study	Country and Setting	Exam type	Exam year(s)	Number of total exams	Number of cancers	Number of interpreting radiologists	Cancer follow-up period	Screening program interval	Pathology ground truth source
Salim, 2020	Sweden: one major medical center	DM	2008–2015	8,805	739	25 (1st readers); 20 (2nd readers)	2 years	2008–2011: 2-year; 2012–2015: 1-year for age 40–49	Regional cancer registry linkage
Schaffter, 2020	US: one integrated health system; Sweden: one major medical center	DM	2016–2017	93,665	1,063	NR	US: 1 year; Sweden: 18–24 months	1- or 2-years	Regional cancer registry linkage

DM = digital mammography; DBT = digital breast tomosynthesis; US = United States; UK = United Kingdom; NR = not reported.

Table 3.

Diagnostic Performance of AI Mammography Technologies on External Validation Studies

Study	Radiologist performance Accuracy (95%CI if reported)	AI standalone performance Accuracy (95%CI if reported)	Combined AI and radiologist performance Accuracy (95%CI if reported)	Ensemble AI model performance Accuracy (95%CI if reported)
Reader Studies				
Conant, 2019	AUC: 0.80 Sensitivity: 0.77 Specificity: 0.63	AUC: NR Sensitivity: 0.91 (0.81–0.96) Specificity: 0.41 (0.34–0.48)	AUC: 0.85 * Sensitivity: 0.85 * Specificity: 0.70 *	N/A
Lotter, 2019 (reader study portion)	AUC: NR Sensitivity: NR Specificity: NR	AUC: 0.94 (0.92–0.97) Sensitivity: 0.96 (0.92–0.99) Specificity: 0.91 (0.85–0.96)	AUC: NR Sensitivity: NR Specificity: NR	N/A
Rodriguez-Ruiz, Radiology 2019	AUC: 0.87 (0.83–0.90) Sensitivity: 0.83 (0.81–0.85) Specificity: 0.77 (0.75–0.79)	AUC: 0.89 Sensitivity: NR Specificity: NR	AUC: 0.89 (0.85–0.92) * Sensitivity: 0.86 (0.84–0.88) * Specificity: 0.79 (0.77–0.81)	N/A
Watanabe, 2019	AUC: 0.76 Sensitivity: NR Specificity: NR	AUC: 0.66 Sensitivity: NR Specificity: NR	AUC: 0.81 * Sensitivity: NR Specificity: NR	N/A
Kim, 2020	AUC: 0.81 (0.77–0.85) Sensitivity: NR Specificity: NR	AUC: 0.94 (0.92–0.96) * Sensitivity: NR Specificity: NR	AUC: 0.88 (0.85–0.91) * Sensitivity: NR Specificity: NR	N/A
McKinney, 2020 (reader study portion)	AUC: 0.62 Sensitivity: NR Specificity: NR	AUC: 0.74 (0.70–0.79) * Sensitivity: NR Specificity: NR	AUC: NR Sensitivity: NR Specificity: NR	N/A
Pacile, 2020	AUC: 0.77 (0.72–0.81) Sensitivity: 0.66 (0.55–0.74) Specificity: 0.72 (0.66–0.79)	AUC: NR Sensitivity: NR Specificity: NR	AUC: 0.78 (0.75–0.84) * Sensitivity: 0.69 (0.60–0.78) * Specificity: 0.74 (0.66–0.82)	N/A
Sasaki, 2020	AUC: 0.82 Sensitivity: 0.89 Specificity: 0.86	AUC: 0.71 * Sensitivity: 0.93 Specificity: 0.45	AUC: NR Sensitivity: NR Specificity: NR	N/A
Retrospective Simulation Studies				
Lotter, 2019 (China evaluation study)	AUC: NR Sensitivity: NR Specificity: NR	AUC: 0.97 Sensitivity: NR Specificity: NR	AUC: NR Sensitivity: NR Specificity: NR	N/A
Rodriguez-Ruiz, JNCI 2019	AUC: 0.81 (0.79–0.84) Sensitivity: 0.76–0.84 (range) Specificity: 0.49–0.79 (range)	AUC: 0.84 (0.82–0.86) * Sensitivity: NR Specificity: NR	AUC: NR Sensitivity: NR Specificity: NR	N/A

Study	Radiologist performance Accuracy (95%CI if reported)	AI standalone performance Accuracy (95%CI if reported)	Combined AI and radiologist performance Accuracy (95%CI if reported)	Ensemble AI model performance Accuracy (95%CI if reported)
Dembrower, 2020	AUC: NR Sensitivity: NR Specificity: NR	AUC: NR Sensitivity: 0.96–1.00 (range depending on AI cut-off)	AUC: NR Sensitivity: NR Specificity: NR	N/A
McKinney, 2020 (US simulation study portion)	AUC: NR Sensitivity: 0.48 Specificity: 0.81	AUC: 0.81 (0.79–0.83) Sensitivity: 0.52 * Specificity: 0.86 *	AUC: NR Sensitivity: NR Specificity: NR	N/A
McKinney, 2020 (UK simulation study, first reader)	AUC: NR Sensitivity: 0.63 Specificity: 0.93	AUC: 0.89 (0.87–0.91) Sensitivity: 0.64 * Specificity: 0.95 *	AUC: NR Sensitivity: NR Specificity: NR	N/A
Rodriguez-Ruiz, SPIE 2020	AUC: NR Sensitivity: 0.82 (0.76–0.87) Specificity: 0.70 (0.68–0.72)	AUC: NR Sensitivity: NR Specificity: NR	AUC: NR Sensitivity: 0.81 (0.75–0.87) Specificity: 0.75 (0.74–0.77) *	N/A
Salim, 2020 (3 AI algorithms vs. first reader)	AUC: NR Sensitivity: 0.77 (0.74–0.80) Specificity: NR	AUC: 0.92–0.96 (range) Sensitivity: 0.67–0.82 (range) Specificity: fixed at radiologist specificity	AUC: NR Sensitivity: 0.84–0.89 (range) * Specificity: 0.93 (for all 3 AI algorithms)	AUC: NR Sensitivity: 0.87 (0.84–0.89) Specificity: 0.92 (0.92–0.93)
Schaffter, 2020 (Sweden simulation study, first reader)	AUC: NR Sensitivity: 0.77 Specificity: 0.97 (0.97–0.97)	AUC: 0.90 (top model) Sensitivity: fixed at radiologists' sensitivity 0.77 Specificity: 0.88	AUC: 0.94 (ensemble + radiologist) Sensitivity: fixed at radiologists' sensitivity 0.77 Specificity: 0.98 (0.98–0.98) *	AUC: 0.92
Schaffter, 2020 (US simulation study)	AUC: NR Sensitivity: 0.86 Specificity: 0.90 (0.90–0.91)	AUC: 0.86 (top model) Sensitivity: fixed at radiologists' sensitivity 0.86 Specificity: 0.66	AUC: 0.94 (ensemble + radiologist) Sensitivity: fixed at radiologists' sensitivity 0.86 Specificity: 0.92 (0.92–0.92)	AUC: 0.90

*= Performance improvement over radiologist alone was reported as statistically significant. AUC, sensitivity, and specificity values were rounded to nearest hundredth to allow for ease of comparisons. For double-reading environments, only single (first) reader performance values are provided to allow for direct comparisons across studies.

Table 4.

QUADAS-2 Quality of Evidence Assessment

Study	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Conant, 2019	Unclear	Unclear	High	Low	Unclear	Unclear	Low
Lotter, 2019 (reader study)	Unclear	Unclear	High	Low	Low	Low	Low
Lotter, 2019 (evaluation study)	High	Unclear	High	High	High	Unclear	Low
McKinney, 2019 (reader study)	Low	Unclear	Low	Low	Low	Low	Low
McKinney, 2019 (evaluation study)	Low	Unclear	Low	Low	Low	Low	Low
Rodriguez-Ruiz, JNCI 2019	High	High	Unclear	Unclear	High	Low	Low
Rodriguez-Ruiz, Radiology 2019	High	High	Unclear	Unclear	High	High	Low
Watanabe, 2019	High	Unclear	High	Low	High	Unclear	Low
Dembrower, 2020	Low	High	Unclear	Low	Low	High	Low
Kim, 2020	High	Unclear	Unclear	Low	High	Unclear	Low
Pacile, 2020	High	High	Low	Low	High	Unclear	Low
Rodriguez-Ruiz, SPIE 2020	High	Unclear	Unclear	Unclear	High	Unclear	Low
Salim, 2020	Low	Unclear	Low	Low	Low	Low	Low
Sasaki, 2020	High	Unclear	High	Unclear	High	Unclear	Low
Schaffter, 2020	Low	Unclear	Low	Low	Low	Low	Low

Low = low risk of bias or applicability concerns. High = high risk of bias or applicability concerns. Unclear = insufficient data to categorize as high or low risk of bias or applicability concerns.