European Journal of Radiology 204 (2026) 113136

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Diagnostic performance of FDA-cleared and CE-marked AI systems for mammography screening: a systematic review and *meta*-analysis

Putu Irma Wulandari [a,b,*], Ziba Gandomkar [a], Mary Rickard [a], Sahand Hooshmand [a], Mo'ayyad Suleiman [a], Patrick Brennan [a]

[a] Faculty of Medicine and Health, The University of Sydney, Camperdown 2050, NSW, Australia
[b] Akademi Teknik Radiodiagnostik dan Radioterapi Bali, Denpasar 80225, Bali, Indonesia

ARTICLE INFO

Keywords:
Artificial Intelligence
Mammography
Breast screening

ABSTRACT

Objective: To evaluate the diagnostic performance of FDA/CE-cleared AI systems for mammography screening and assess their roles across clinical workflows.
Methods: A systematic search was conducted in Scopus, Embase, and PubMed for full-text articles published from January 2018 to July 2024 evaluating commercially available FDA/CE-cleared AI for breast cancer detection using digital mammography (DM) and/or digital breast tomosynthesis (DBT). Included studies specified the AI system and reported diagnostic performance metrics (e.g. sensitivity, specificity, AUC) and/or clinical impact (e.g. cancer detection rate [CDR], recall rate, workload). Study quality was assessed using a modified QUADAS-2 tool. A random-effects *meta*-analysis pooled standalone AI performance (sensitivity, specificity, and AUC); other workflow roles were synthesised narratively.
Results: In total, 42 studies were included (39 retrospective; 3 prospective): 33 evaluated DM, 6 DBT, and 3 both DM and DBT. Risk of bias was frequently high, particularly for flow and timing. In DM, pooled standalone AI AUC was 0.89 (95% CI: 0.85–0.92), with sensitivity of 76.3% (95% CI: 66.2–84.2%) and specificity of 89.6% (95% CI: 84.5–93.2%). Prospective evaluations of AI-integrated workflows, including triage and second-reader replacement, showed non-inferior cancer detection with significant workload reduction compared with standard double reading. A single prospective study of AI as an additional reviewer reported additional cancer detection with minimal increase in recalls.
Conclusions: FDA/CE-cleared AI systems demonstrate promising diagnostic performance in breast screening. The strongest evidence supports workflow configurations that reduce radiologist workload, particularly triage and independent/supporting reader models, while prospective evidence supporting complete replacement of radiologists remains lacking.

1. Introduction

Breast cancer is the most common cancer and the leading cause of cancer-related death for women [1]. Mammography screening improves early detection; however, significant variation in radiologists' performance when interpreting mammography images have been reported [2]. European Guidelines therefore recommend double reading with consensus or arbitration over single reading practice to improve diagnostic accuracy [3,4], that would normally result in high reading volumes and radiologist fatigue [5]. Additionally, double reading practice might not be practical in countries facing a shortage of radiologists. To address these challenges, artificial intelligence (AI) has emerged as a potential solution.

AI tools in breast imaging support cancer detection, decision making, treatment response assessment, risk prediction, density quantification, triage, and image quality evaluation [6–8]. While some systems remain investigational, others are commercially available, including AI-based computer aided detection (AI-CAD) [6], which can improve

Abbreviations: AI, Artificial Intelligence; FDA, Food and Drug Administration; CE, Conformité Européenne; DM, digital mammography; DBT, digital breast tomosynthesis; ROC, receiver operating characteristic curve; AUC, area under the receiver operating characteristic curve; CDR, cancer detection rate; AIR, abnormal interpretation rate; PPV, positive predictive value.

* Corresponding author.
E-mail address: pwul4550@uni.sydney.edu.au (P.I. Wulandari).

https://doi.org/10.1016/j.ejrad.2026.113136
Received 11 June 2026; Received in revised form 29 July 2026; Accepted 4 August 2026
Available online 5 August 2026

consistency in image interpretation.

The optimal implementation of AI-CAD remains debated. Some studies support AI use for triaging likely negative mammograms without radiologists' review [9], while others propose partial replacement of human readers [10–12], though standalone use is not yet supported [10,11]. Although the most effective use of AI is still unclear, there is promising potential for AI to complement rather than compete with radiologists [13]. European guidelines recommend double reading with AI assistance, with consensus for discordant reads, over double reading without AI [4]. AI-supported single reading is not recommended since evidence regarding AI accuracy and its long-term impact on breast cancer mortality remains limited [4].

Although several studies report AI performance is not inferior to that of radiologists [14–17], results vary across systems and vendors [18]. Nevertheless, the accuracy of AI-CAD is continually improving, and many devices have been authorised by the U.S. Food Drug Administration (FDA) in the USA and Conformité Européenne (CE) in Europe. As a result, there has been rapid growth in large-scale prospective and retrospective studies.

Recent systematic reviews have evaluated the performance of Artificial Intelligence in breast screening [13,19–22] as well as its potential impacts on workload [23]. However, they have not focused on FDA- or CE-approved systems or incorporated recent prospective studies of real-world clinical implementation [24–27]. This review therefore synthesizes evidence on FDA-cleared and/or CE-marked AI systems for mammography screening, with particular focus on diagnostic performance and roles within clinical workflows, including standalone reading, reader assistance, triage, and reader-replacement models.

2. Methods

This systematic review and *meta*-analysis is reported in accordance with the Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA). This review analyses the diagnostic performance of artificial intelligence or machine learning (AI/ML) based medical devices specifically developed for breast cancer detection in mammography, including 2D-mammography and digital breast tomosynthesis (DBT). Only commercially available AI devices that had received FDA clearance and/or CE marking were included in this study.

2.1. Identifying FDA/CE-Cleared AI products

FDA-cleared AI devices were identified through a publicly available FDA database [28], with data retrieved in June 2024. Extracted information included device name, manufacturer, submission details, and regulatory classification.

CE-marked devices were identified using the European Database on Medical Devices (EUDAMED) [29]. Due to incomplete and inconsistent entries, additional sources, including prior publications [30,31], manufacturer websites, and targeted online searches (e.g., https://radiology.healthairegister.com/) were used. Identified devices were then cross-checked against EUDAMED.

Based on unique device names (irrespective of versions or multiple regulatory entries), 12 FDA-cleared and 15 CE-marked AI systems were included. Detailed identification procedures are provided in Supplementary Methods (Appendix A).

2.2. Literature search strategy and study selection

After accumulating the data of FDA-cleared and CE-marked AI products for lesion detection in mammography, this information was utilised for a literature search. We searched Scopus, Embase, and PubMed up to 4 July 2024 using terms related to AI product names, manufacturers, and mammography screening. Searches were limited to English full-text articles (2018–2024). Conference proceeding and reviews were excluded. Detailed search strings in each database are provided in Appendix B.

The screening phase was completed by 3 researchers (P.I.W., Z.G., and S.H.) and documented using a PRISMA flowchart (Fig. 1). We included prospective and retrospective studies evaluating FDA-cleared and/or CE-marked AI systems for mammography screening. Studies were required to report the name of evaluated AI systems and quantitative performance metrics (e.g., sensitivity, specificity, and area under the receiver operating characteristic curve [AUC]) and/or clinical impact measures (e.g., cancer detection rate, recall rate, workload). Only studies using 2D-mammography or DBT were included. Studies on other imaging modalities or non-approved/self-developed AI systems were excluded.

2.3. Data extraction and quality assessments

Extracted data included study characteristics (author, year, design, setting), AI specifications and role, sample size, reference standard, comparator, and outcomes.

The risk of bias and applicability was assessed using Quality Assessments of Diagnostic Accuracy Studies 2 (QUADAS-2) modified for AI studies. The modified template was developed based on previous studies [13,22,32] with further refinements to reflect the aims of this review. The full template is provided in Appendix C.

2.4. Study Clustering and data analysis

Studies were grouped by AI role in the screening workflow:

1) Standalone reader (AI-only interpretation),
2) Reader aid (decision support),
3) Triage (case prioritisation or exclusion),
4) Independent reader (replacement of one reader in double reading), and
5) Secondary review (additional review of non-recall cases)

Studies were classified according to their primary study objective, although some studies evaluated AI in multiple roles as part of additional analyses. A pooled analysis was considered for all study groups (AI roles), however due to substantial heterogeneity, *meta*-analysis was limited to studies reporting standalone AI performance (sensitivity, specificity, and AUC). Studies in other categories were included if standalone metrics were available; otherwise, results were synthesized narratively.

Quantitative analyses were conducted in R (v4.4.2) using the *metafor* and *mada* packages. Diagnostic accuracy was synthesised using a bivariate random-effects (Reitsma) model, jointly pooling sensitivity and specificity. The resulting hierarchical receiver operating characteristic (HSROC) curve was plotted with 95% confidence and prediction regions. Heterogeneity was assessed using I^2 statistics and prediction intervals.

Separately, a *meta*-analysis of study-reported AUC values was also conducted for subset that provide an AUC (this subset was not necessarily identical to that used in the diagnostic accuracy analysis). This analysis was conducted using a random-effects model on the logit scale with restricted maximum likelihood (REML) estimation.

To further explore potential sources of heterogeneity, exploratory subgroup and *meta*-regression analyses were conducted according to screening setting, vendor group, cancer definition, and threshold category where applicable. Sensitivity and leave-one-out analyses were performed to assess the robustness of pooled estimates. Detailed analytical methods and results are provided in Appendix G.

A GRADE-style certainty assessment was performed for each AI role category, which was judged according to study design, risk of bias, consistency of findings, directness of evidence, and overall strength of evidence base. The assessment was intended to provide an overall indication of confidence in the available evidence rather than a formal

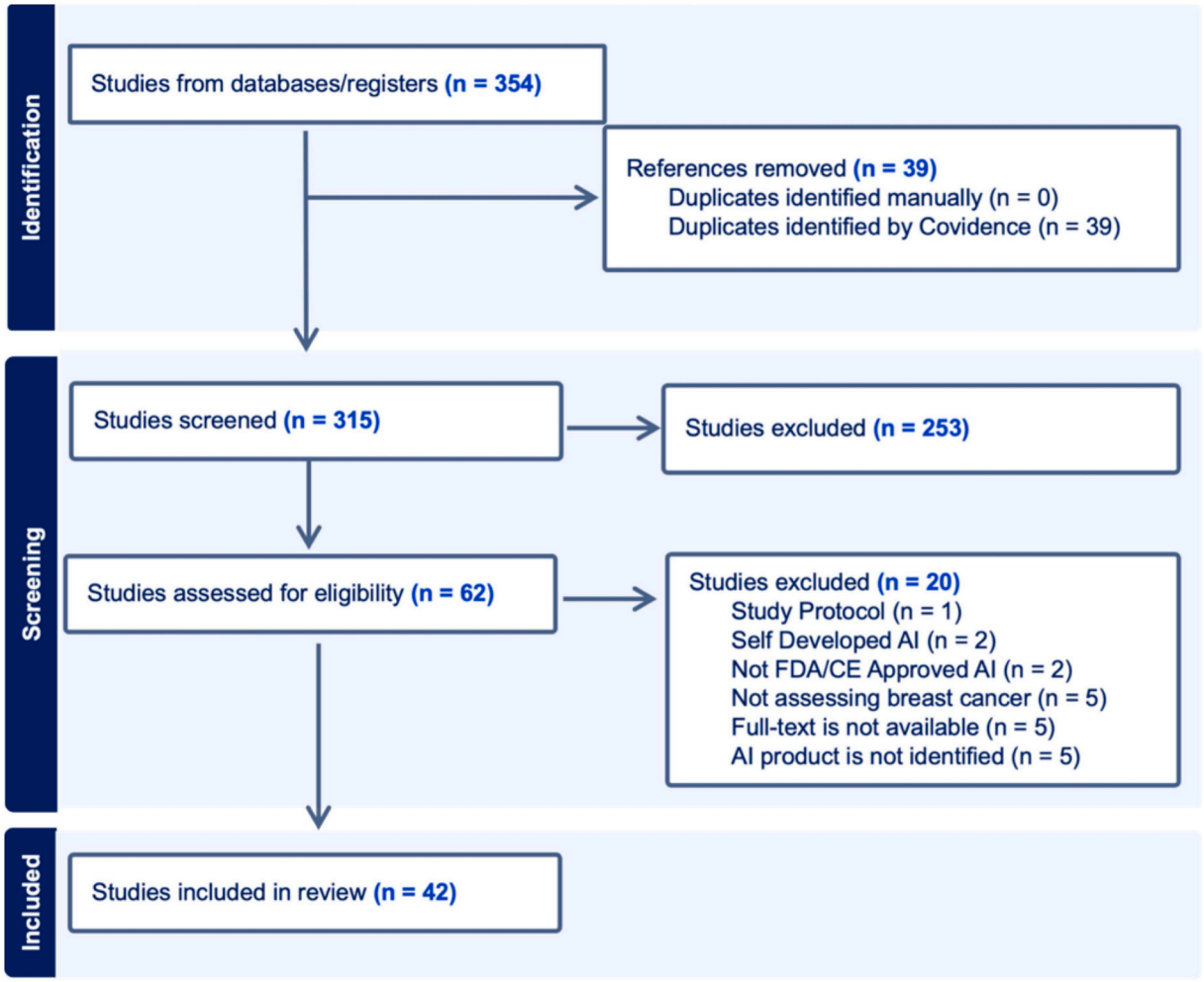

Fig. 1. PRISMA flow diagram of included study.

outcome-level GRADE evaluation.

3. Results

Database search identified 354 records. After duplicate removal, 315 studies were eligible for screening. Following title and abstract screening, 62 studies underwent full-text review, of which 20 were excluded for several reasons, including unclear identification of the AI product, unavailable full-text, or not assessing AI performance in breast cancer detection. As a result, 42 studies were included in the quantitative and qualitative synthesis (Fig. 1).

3.1. Characteristics of included studies

The characteristics of the 42 included studies are summarised in Table 1-5 and in Appendix D. Studies were conducted across Europe, the UK, the US, and Asia. Most evaluated AI performance using 2D mammography (DM, n = 33), while six studies focused on digital breast tomosynthesis (DBT), and three studies evaluated both DM and DBT datasets.

The majority of studies were retrospective (n = 39). Of these, 28 were historical simulation studies using archived datasets or prior clinical reads, eight were laboratory-based reader studies, and three were observational cohort studies comparing screening outcomes before and after AI implementation. The remaining three studies were prospective, including one randomised controlled trial evaluating AI as a triage and concurrent decision support versus standard double reading [24], and two reader studies assessing AI as an independent reader [26] or as an additional review step for non-recalled cases [25].

Across the included studies, FDA-cleared and/or CE-marked AI systems from eight vendors were evaluated. The most frequently assessed systems were Transpara (ScreenPoint Medical BV, Netherlands; n = 21), Lunit (Lunit Inc, South Korea; n = 8), and Mia (Kheiron Medical Technologies, UK; n = 6). Other AI systems included Profound AI (iCAD Inc, USA; n = 3), Vara (Merantix Healthcare, Germany; n = 1), Mammoscreen (Therapixel, France; n = 1), Saige-Dx (DeepHealth, USA; n = 1), and cmTriage (CureMetrix Inc, USA; n = 1).

AI score thresholds or cutoff points varied across studies (see Appendix D). The majority of studies utilised vendor-default thresholds, whereas others applied site-specific thresholds based on validation datasets [33,34] or predefined reader performance benchmarks [35–37].

3.2. Risk of bias assessments

The overall risk of bias and applicability assessments based on QUADAS-2 led to a proportion of studies being classified as high risk, particularly in patient selection, flow and timing, and applicability of the index test (Fig. 2). Comprehensive assessments of 42 studies are provided in Appendix E.

For patient selection, 22 studies (52%) had high or unclear risk of bias, mainly due to non-random sampling, case-control design, or enriched datasets. Furthermore, six of these studies used highly selected populations, including cancer-only cases [38], interval cancers

Table 1
Summary of studies using AI as standalone reader.

Study	Study Design	AI system	Modality/ Population/ Country
Rodriguez-Ruiz et al., 2019 [14]	Retrospective*	Transpara v.1.4.0	Modality: DM. N = 2,652 cases, n cancer = 653 (24.6%). Countries: Spain, Netherlands, Austria, Sweden, UK, Italy, and U.S.
Sasaki et al, 2020 [43]	Retrospective†	Transpara v.1.3.0	Modality: DM. N = 310 patients, with 69 patients (22%, 69/310) had 70 malignant lesions. Country: Japan
Lång et al, 2021 [39]	Retrospective*	Transpara v.1.5.0	Modality: DM. n = 429 women with IC. Country: Sweden
Dahlblom et al, 2021 [44]	Retrospective	Transpara v.1.7.0	Modality: DM. N = 14,640 normal, 136 SDC and 22 IC. Country: Sweden
Byng et al, 2022 [33]	Retrospective	Vara v.1.0.7	Modality: DM. N = 2,396 IC with 374 false negative (FN), 468 minimal signs (MN), 1280 true interval, 274 occult. Country: Germany
Romero-Martín et al, 2022 [45]	Retrospective*	Transpara v.1.7.0	Modality: DM and DBT. N = 15,999 exams including 98 SDC and 15 IC. Country: Spain
Kerschke et al., 2022 [42]	Retrospective*	Transpara v.1.5.0	Modality: DM. N = 2,257 recalled examinations (1,964 normal, 288 SDC and 5 ICs). Country: Germany
Oberije et al, 2023 [38]	Retrospective*	MIA v.2.0	Modality: DM. N = 2,011 examinations with breast cancers (1718 SDC and 293 IC). Country: Hungary
Çelik et al, 2023 [46]	Retrospective*	Transpara v.1.6. and Transpara v.1.7.0	Modality: DM. N normal = 441 women, n IC = 323 women. Country: Turkey
Weigel et al, 2023 [41]	Retrospective	Transpara v.1.7.0	Modality: DM. N = 634 women with 644 calcification-related histologic lesions (195cancerpositivecalcifications). Country: Germany.
Yoen and Chang, 2023 [47]	Retrospective	Lunit INSIGHT MMG v.1.1.4	Modality: DM. N = 238 women with 253 breast cancers and 160 women without breast cancers. Country: South Korea.
Tan et al, 2023 [35]	Retrospective†	Transpara v.1.7.0 and QVCAD v.3.4	Modality: DM and ABUS. N = 430 cases (42malignant,114benign, and274normal) for cohort study, and n = 152 for observer study. Country: China
De Vries et al, 2023 [34]	Retrospective*	MIA v.2.0.1	Modality: DM. N = 45,444 cases with 303 SDC. Country: U.K.
Nguyen et al, 2024 [40]	Retrospective*	Profound AI v.3.0	Modality: DBT. N = 4,855 patients with normal DBT. Country: N/A
Bergan et al, 2024 [48]	Retrospective*	Transpara v.1.7.0	Modality: DM. N = 99,489 screening exams from 74,941 women with 883 breast cancer (688 SDC and 195 IC). Country: Norway
Lee et al, 2024 [49]	Retrospective	Lunit INSIGHT MMG v.1.1.1.0 to 1.1.7.1	Modality: DM. N = 656 breasts from 599 patients that have abnormal AI scores. Country: South Korea
Kwon et al, 2024 [50]	Retrospective*	Lunit INSIGHT MMG v.1.1.7.2.	Modality: DM. N = 89,855 women, n cancers = 143 women (0.16%). Country: South Korea

Note: *= historical study, †= reader study, DM = digital mammography, DBT = digital breast tomosynthesis, n = population/sample, SDC = screen detected cancer, IC = interval cancer.

[33,39,40], normal mammograms [40], calcification cases [41], or recalled screening cases [42].

Reference standard was at high or unclear risk of bias in 19 studies (45%) primarily because they relied on histopathology without adequate follow-up of screen-negative women, or used follow-up periods shorter than two years. This potentially led to underestimation of missed cancers.

Flow and timing had the highest risk of bias, with 40 studies (95%) rated high or unclear. This largely reflected retrospective designs in which follow up testing or biopsy was triggered by historical human reader assessments. As AI-positive but human-negative cases were generally not biopsied, their definite disease status remained uncertain, creating potential verification bias despite follow-up examinations.

Concerns regarding applicability of the index test existed in 31 studies (74%), largely because study workflows deviated from real-world clinical practice, especially in laboratory-based reader studies and simulated AI implementation using historical datasets.

3.3. AI as a standalone reader

No prospective study evaluated AI as a full replacement for radiologists in routine clinical practice. Seventeen studies [14,33–35,38–50] were specifically designed to assess AI as a standalone reader (see Table 1 and Appendix D – Table D.1), while 10 additional studies [36,51–59] primarily evaluating other AI roles (e.g. reader aid, triage, etc) also reported standalone performance. Performance was assessed separately for DM and DBT.

3.4. Standalone AI in digital mammography

For DM, 21 studies reported standalone AI performance across various datasets and AI systems. Seventeen studies used mixed cancer and normal datasets and were included in pooled analyses (see Appendix F), whereas four studies focused on selected populations, including interval cancer [33,39], calcifications [41], or cancer-only datasets [38]. The evaluated systems included Transpara, Lunit, MIA, cmTriage, Vara, and Profound AI.

a. Pooled AUC

Across the 17 mixed-dataset DM studies, standalone AI was evaluated in 1,611,296 DM examinations. Twelve studies reported AUC and 16 studies reported sensitivity and specificity. The pooled AUC of standalone AI across 12 DM studies (n = 976,256 exams) was 0.89 (95% CI: 0.85–0.92), indicating high overall discriminatory performance. However, substantial heterogeneity was observed (I^2 = 97.3%), with a wide prediction interval (0.657–0.971), suggesting that AI performance may vary considerably across different study populations and clinical settings (Fig. 3).

Compared with human reader AUCs reported in three studies, the pooled AI AUC (0.89) was broadly comparable or higher; however, individual-study comparisons were mixed, with two studies reporting higher AI AUC than radiologists [14,58] and one study reporting lower AI AUC [43]. Direct comparison was limited by differences in study design, populations, and reading workflow.

b. Pooled sensitivity and specificity

The pooled sensitivity and specificity of standalone AI in DM across the 16 studies (n = 1,608,644 exams) were 76.3% (95%CI: 66.2–84.2%) and 89.6% (95%CI: 84.5–93.2%), respectively. Considerable heterogeneity was observed for both outcomes (I^2 = 99.2% and 99.98%, respectively), indicating substantial variation in performance across study populations, AI systems, and operating thresholds.

Compared with pooled human benchmarks, standalone AI showed sensitivity comparable to double reading and higher than single reading,

Table 2
Summary of studies using AI as a reader support.

Study	Study Design	AI system	Modality/ Population/ Country
Rodríguez-Ruiz, 2019 [60]	Laboratory study§, MRMC	Transpara v.1.3.0	Modality: DM. N = 240 cases (100cancer,140normal). N radiologists = 14. Country: U.S.A and Europe
van Winkel et al., 2021 [51]	Laboratory study§, MRMC	Transpara v.1.6.0	Modality: DBT. N = 240 cases (65malignant,65benigncases,110normal). N radiologists = 18. Country: U.S.
Pinto et al., 2021 [52]	Laboratory study§, MRMC	Transpara v.1.6.0	Modality: DBT. N = 190 cases (75malignant,26benign,and90normal). N radiologists = 14. Country: Norway
Lee et al, 2022 [61]	Laboratory study§, MRMC	Lunit INSIGHT MMG v.1.1.1.0	Modality: DM. N = 200 cases (100 malignant, 100 normal). N radiologists = 10 (5 breast specialist radiologist/BSR and 5 general radiologists/GR). Country: Korea
Dang et al., 2022 [62]	Laboratory study§, MRMC	Mammoscreen v.1.2	Modality: DM. N = 314 cases (628breasts) with 128 cancers (72 cancers on left breasts, 56 cancers on right breasts). N radiologists = 12. Country: France
Letter et al., 2023 [63]	Retrospective*	Profound AI v2.0	Modality: DBT. N = 5,883 cases with AI, n = 7002 cases without AI. N radiologists = 5. Country: U.S.A.
Kim et al., 2023 [54]	Retrospective‡	Lunit INSIGHT MMG v.1.1.1.0	Modality: DM. N = 1,579 DM without AI, and 1,579 DM with AI. N radiologists = 12. Country: Korea
Kim et al, 2024 [53]	Laboratory study§, MRMC	Saige-Dx v.2.0.0	Modality: DBT. N = 240 cases (100cancerand140normal). N radiologists = 18. Country: U.S.A
Elías-Cabot et al., 2024 [64]	Retrospective‡	Transpara v.1.7.0.	Modality: DM and DBT. N = 11,998 in each control and intervention group (before and after using AI). Country: Spain

Note: §= Laboratory study with enriched test set, MRMC = Multi reader multi case laboratory study, *= historical study, ‡= cohort study (with and without AI), DM = digital mammography, DBT = digital breast tomosynthesis, n = population/sample, SDC = screen detected cancer, IC = interval cancer.

Table 3
Summary of studies using AI as triage tool.

Study	Study Design	AI system	Modality/ Population/ Country
Lång et al, 2021 [66]	Retrospective*	Transpara v.1.4.0	Modality: DM. N = 9581 women with 68 SDC cancers. Country: Sweden
Raya-Povedano et al, 2021 [16]	Retrospective*	Transpara v.1.6.0	Modality: DM and DBT. N = 15,986 women (113 cancers: 98 SDC and 15 IC). Country: Spain
Retson et al, 2022 [55]	Retrospective	cmTriage v.1.0.14	Modality: DM. N = 1,255 cases (400cancerand855normal). Country: U.S.A.
Lauritzen et al, 2022 [17]	Retrospective*	Transpara v.1.7.0	Modality: DM. N = 114,421 women (791 SDC, 327 IC, and 1473 long-term cancers). Country: Denmark
Lång et al, 2023 [24]**	Prospective, RCT	Transpara v.1.7.0.	Modality: DM. N without AI (control group) 40,024 women (203 SDC), n with AI (intervention group) = 39,996 (203 SDC). Country: Sweden
Dahlblom et al, 2023 [59]	Retrospective*	Transpara v.1.7.0	Modality: DBT. N = 14,772 women with 157 cancers (135 SDC and 22 IC). Country: Sweden
Lauritzen et al, 2024 [65]	Retrospective‡	Transpara v.1.7.1	Modality: DM. N before AI = 60,751 women, n with AI = 58,246. Country: Denmark
Larsen et al, 2024 [58]	Retrospective	Lunit INSIGHT MMG v.1.1.7.2.	Modality: DM. Population: 661,695 mammograms (3,807 SDC and 1,110 IC). Country: Norway

Note: **= primarily used AI as triage tool, but also used as concurrent reader support, *= historical study, ‡= cohort study (with and without AI), DM = digital mammography, DBT = digital breast tomosynthesis, n = population/ sample, SDC = screen detected cancer, IC = interval cancer, RCT = randomised controlled trial.

Table 4
Summary of studies using AI as an independent reader.

Study	Study Design	AI system	Modality/ Population/ Country
Sharma, 2021 [56]	Retrospective*	MIA v.2.0.1	Modality: DM. N = 275,900 (2,792cancercases). Countries: U.K and Hungary
Dembrower et al, 2023 [26]	Prospective†	Lunit INSIGHT MMG v.1.1.6	Modality: DM. N = 55,581 women (296womenwithcancer). Country: Sweden
Elhakim et al, 2023 [36]	Retrospective*	Transpara v.1.7.0	Modality: DM. N = 257,671 mammograms (2,014cancers). Country: Denmark
Heywang-Köbrunner et al, 2023 [57]	Retrospective*	Profound AI for 2D mammography V2	Modality: DM. N = 17,884 cases (114 SDC and 17 ICs). Country: Germany
Ng et al, 2023 [67]	Retrospective*	MIA v.2.0	Modality: DM. N = 280,594 cases (2,783 cancers: 2,397 SDC and 386 IC). The dataset consists of UK dataset (n = 194145) and Hungary dataset (n = 86449). Countries: Hungary and U.K.
Dembrower et al, 2023 [37]	Retrospective*	Lunit INSIGHT MMG v.1.1.6	Modality: DM. N = 6,708 women (1,684womenwithcancers). Country: Sweden
Sharma et al, 2023 [15]	Retrospective*	MIA v.2.0	Modality: DM. N = 275,900 cases (2,683cancercases). Countries: U.K and Hungary

Note: *= historical study, †= reader study, DM = digital mammography, n = population/sample, SDC = screen detected cancer, IC = interval cancer.

Table 5
Summary of studies using AI-assisted additional reviewer.

Study and Year	Study Design	AI system	Modality/ Population/ Country
Ng, 2023 [25]	Prospective†	MIA v.2.0	Modality: DM. N initial pilot = 3,746 women, n extended study = 9,112 women, n live use phase = 15,953 women. Country: Hungary

Note: †= reader study, DM = digital mammography, n = population/sample.

with specificity broadly similar to double reading but lower than single reading. These comparisons are shown in the forest plots for sensitivity and specificity (Fig. 4), with overall diagnostic performance illustrated in the HSROC curve (Fig. 5).

c. Sources of heterogeneity and robustness analyses

Exploratory subgroup, *meta*-regression, sensitivity, and leave-one-out analyses were conducted to investigate potential sources of heterogeneity. Although vendor group, screening setting, and cancer

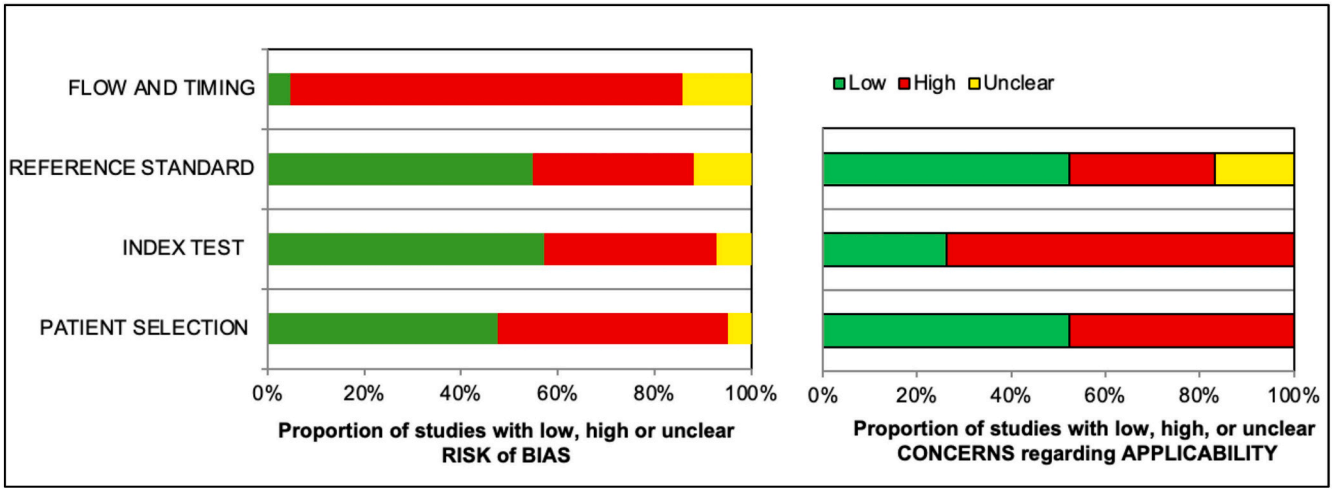

Fig. 2. Stacked bar charts show the summary of included articles assessed by Qualitative assessments of Diagnostic Accuracy Studies 2 (QUADAS-2) modified for AI studies. Each category is represented as percentage of number of articles that have high, low, or unclear levels of bias and applicability.

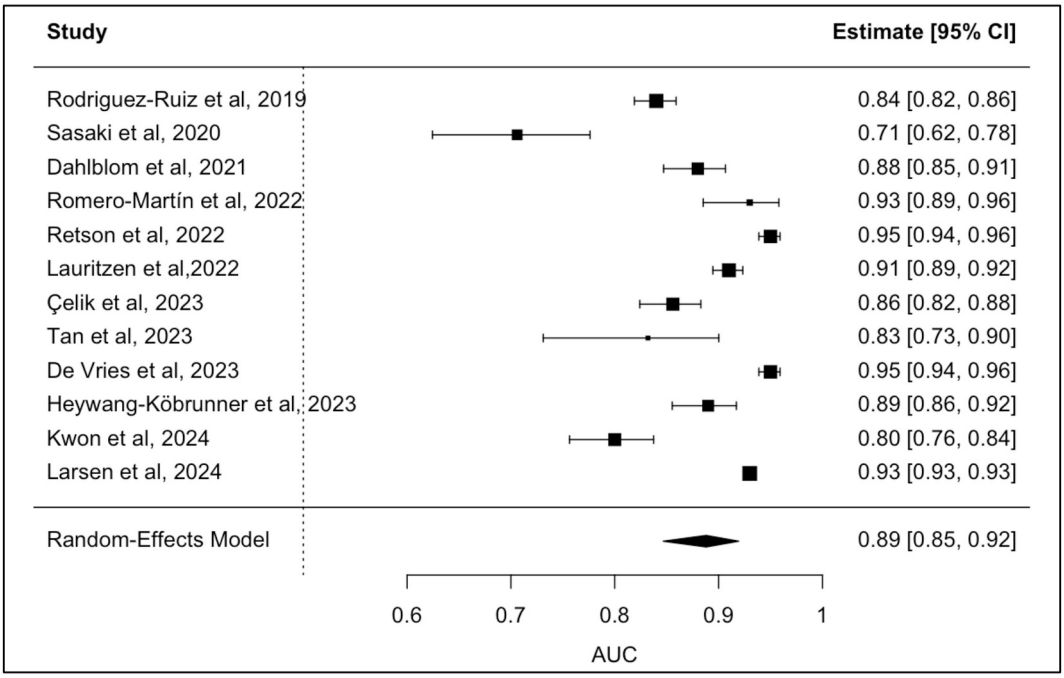

Fig. 3. Forest plot of standalone AI performance in Digital Mammography (DM) for cancer detection. Summary AUC was 0.89 (95%CI: 0.85–0.92) based on 12 studies.

definition explained a small proportion of between-study variability, none were statistically significant moderators. Among the evaluated moderators, vendor grouping explained the largest proportion of between-study variability, suggesting that differences between AI systems may contribute to performance variation, although this finding should be interpreted with caution. Substantial residual heterogeneity remained after these analyses, indicating that performance differences were likely multifactorial. Sensitivity analyses also produced similar pooled estimates, and no individual study materially influenced the overall results, supporting the robustness of findings. Detailed analyses are provided in Appendix G.

d. Selected population studies

Two studies assessed standalone AI in interval cancer cohorts. At a threshold corresponding to 99% specificity (1% recall rate), AI detected a subset of cancers missed or showing minimal signs on prior screening mammograms, suggesting potential value for identifying retrospectively visible missed cancers [33,39]. However, these findings were based on selected interval cancer populations and should not be interpreted as routine screening performance.

3.5. *Standalone AI in digital breast tomosynthesis*

For DBT, the standalone AI performance was evaluated in two large screening populations and three enriched reader studies. In the two screening cohorts, AI achieved high AUCs and sensitivity broadly comparable to single or double reading [45,59], although one study reported a substantially higher recall rate with AI (16.7% vs. 4.4% vs. 3% for AI, double and single reading, respectively) [45]. In enriched DBT reader

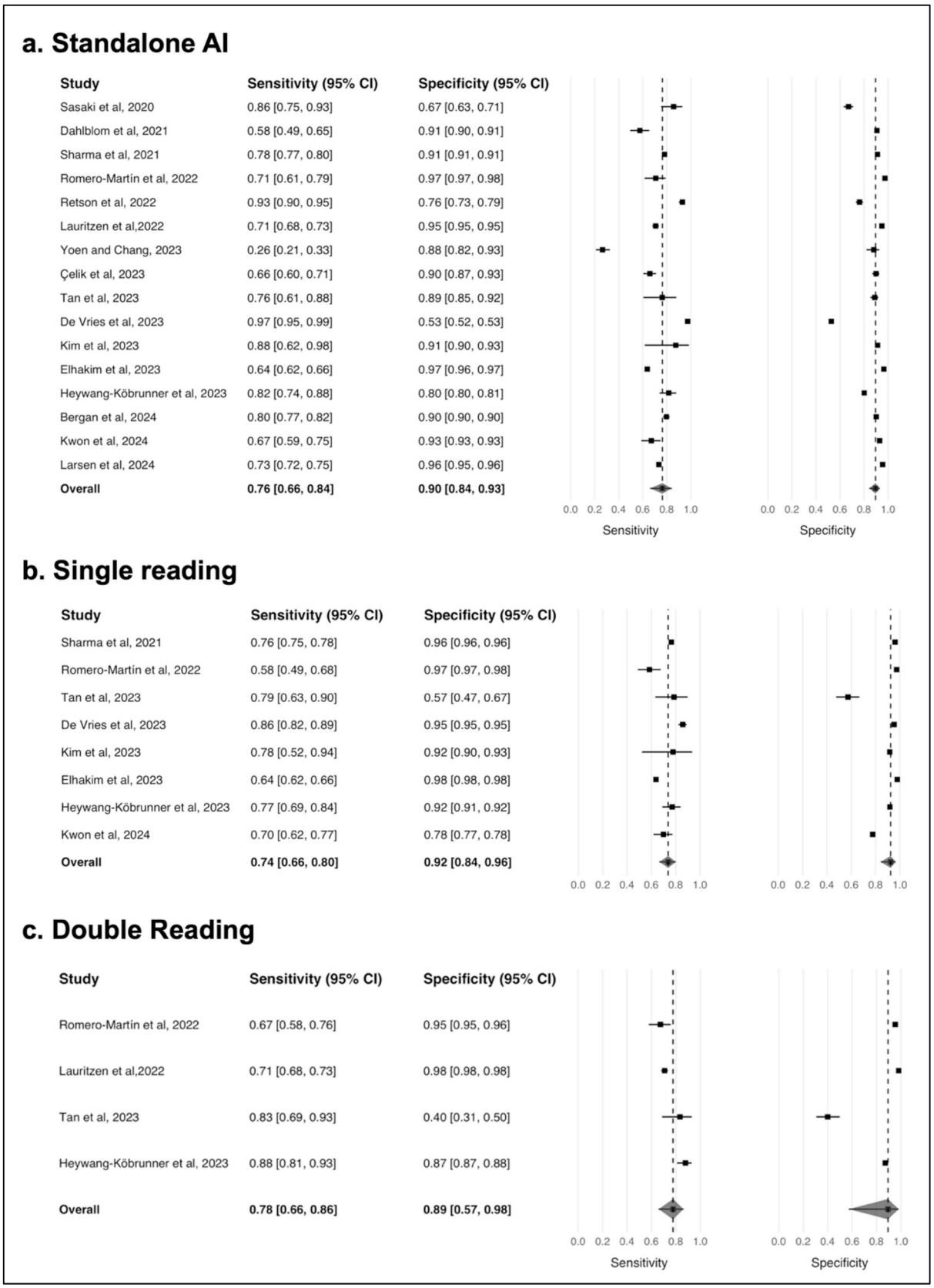

Fig. 4. Forest plots showing joint sensitivity and specificity estimates for: (a) standalone AI system across 16 studies, (b) single reading across 8 studies, and (c) double reading across 4 studies in Digital Mammography (DM). Each plot presents individual study estimates with 95% confidence intervals.

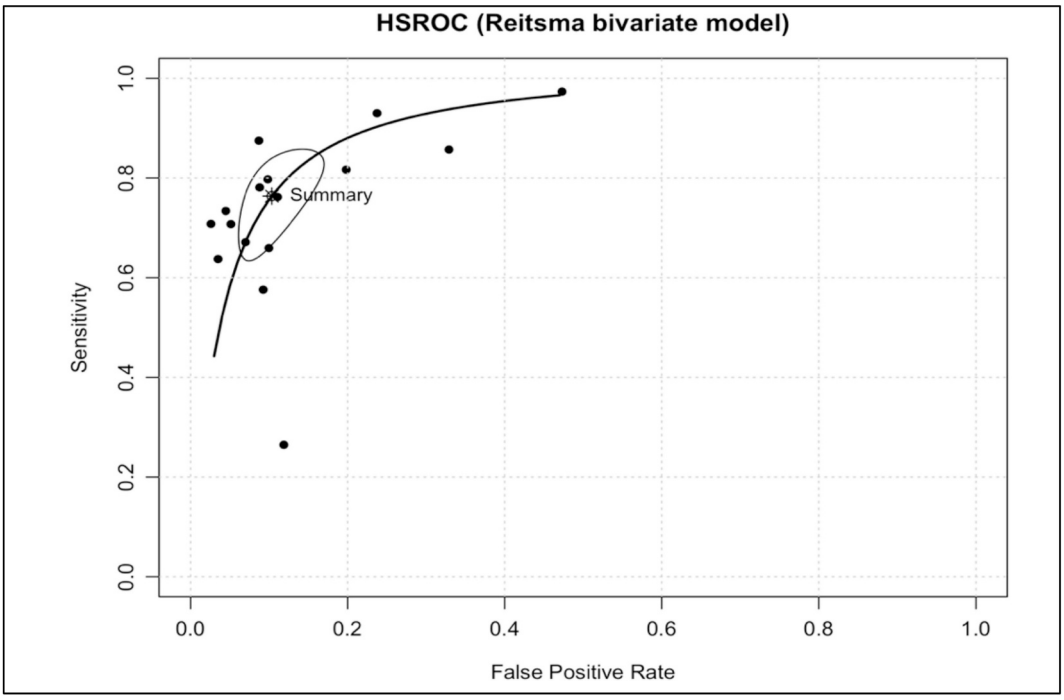

Fig. 5. Hierarchical summary ROC (HSROC) from the bivariate random-effects (Reitsma) model. Each point represents the operating position of an individual study (sensitivity vs. false-positive rate). The star indicates the summary operating point, corresponding to pooled sensitivity of 76.3% and specificity of 89.6%.

studies [51–53] pooled standalone AI AUC was 0.90 (95%CI: 0.85–0.95), compared with 0.85 (95%CI: 0.82–0.87) for radiologists (Fig. 6). Across DBT studies, AI accuracy appeared to vary more across studies than radiologist performance. This suggests that while AI may achieve higher accuracy than radiologists in certain settings, its performance is highly dependent on the dataset characteristics, modality, and the specific AI product.

Overall, standalone AI showed diagnostic performance broadly comparable to radiologists in several study settings, particularly for AUC-based measures. However, findings vary across studies and should be interpreted in the light of the absence of prospective replacement studies, substantial heterogeneity, and use of retrospective or enriched datasets.

3.6. AI as a reader aid

Nine studies [51–54,60–64] evaluated AI as a reader aid supporting radiologists in decision making (Table 2 and Appendix D – Table D.2). Of these, three assessed real-world implementations in screening programs using DM, DBT, and combined modalities, while six were laboratory-based reader studies using enriched datasets. The MASAI trial [24], although primarily a triage study, was also included as it provided concurrent AI decision support during screen reading.

In real-world settings, the impact of AI as reader aid varied across studies. In single-reading workflows, AI integration was generally associated with no significant change in the cancer detection rate or sensitivity, but improvements in specificity (91.6% to 96%) and positive predictive value (PPV1: 9.7% to 18.2%), indicating enhanced diagnostic precision [54]. Similarly, another study reported no significant differences in cancer detection (CDR: 5.9 vs. 7.3 per 1,000), although trend towards improved detection were observed [63]. In contrast, implementation within double reading workflows was associated with increased CDR (5.8 vs. 9.0 per 1,000; $p < 0.001$) and PPV (10.6% vs. 14.6%: $p < 0.001$), with a modest increase in recall rate (5.4% vs 6.1%;

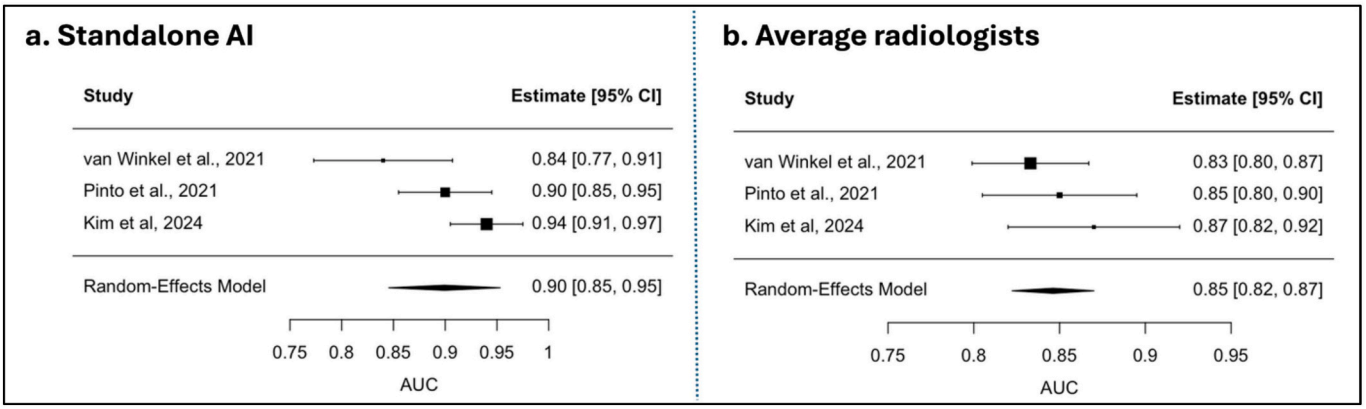

Fig. 6. Forest plots and random effects *meta*-analysis of AUC estimates for standalone AI (a) and radiologists (b) performance in cancer detection using Digital Breast Tomosynthesis (DBT). Diamonds represent pooled estimates under a random-effects model. Horizontal lines indicate 95% confidence intervals. Radiologist performance reflects average single-reader accuracy without AI assistance.

$p < 0.001$), and more consistent benefits observed in DBT than DM [64]. The MASAI trial similarly reported maintained cancer detection (6.1 vs. 5.1; $p = 0.052$), higher PPV (28.3% vs. 24.8%) and a slightly increased recall rate (2.2% vs. 2.0%) compared with standard double reading [24].

Overall, real-world evidence suggests that AI as a reader aid may improve diagnostic precision and, in some settings, cancer detection, particularly within double-reading workflows; however, effects vary by modality and implementation context.

Six laboratory-based reader studies (3 DM [60–62] and 3 DBT [51–53], consistently showed improved radiologist performance with AI assistance. In DM, AI support increased average AUC from 0.79 (range: 0.74–0.87) to 0.84 (range: 0.77–0.89) and average sensitivity from 70.6% (range: 62.9–83.0%) to 79.7% (range: 70.0–86.0%). Specificity improved more subtly from an average of 78% (range: 77.0–79.0%) to 80% (range: 79–81%) [60,62], although one study reported a minor decrease from 68.7% to 67.2% [61]. In DBT, similar improvements were observed, with average AUC increasing from 0.85 (range: 0.83–0.87) to 0.89 (range: 0.86–0.93), and average sensitivity from 78.8% (range: 74.6–81.0%) to 84.9% (range: 79.2–89.6%), while average specificity slightly increase from 73.3% (range: 71.6–75.1%) to 74.7% (range: 73.3–76%) [51–53]. Effects on reading time were inconsistent, with some studies reporting small reductions and others slight increases depending on reader experience and case complexity [51,52,61].

Overall, AI as reader aid generally improves radiologist diagnostic performance in controlled settings and may enhance screening outcomes in clinical practice, although findings are heterogeneous and influenced by study design, dataset characteristics, and workflow implementation.

3.7. AI as a triage tool

Eight studies [16,17,24,55,58,59,65,66] evaluated AI as a triage system using DM, DBT, or combined datasets (see Table 3 and Appendix D – Table D.3). Two studies assessed real-world implementation in screening practice, while six retrospectively simulated AI triage using historical screening data.

In real-world settings, AI triage was associated with substantial workload reduction while maintaining or improving screening performance. In the MASAI randomised trial in Sweden, AI-supported screening reduced radiologists' workload by 44.3% and was associated with a higher cancer detection rate compared with standard double reading (6.1 vs. 5.1 per 1,000; $p = 0.052$) [24]. Similarly, implementation in Denmark reduced radiologists' workload by 33.5% and increased CDR (8.2 vs. 7.0 per 1,000; $p < 0.01$) [65].

Retrospective studies consistently demonstrated that AI triage can substantially reduce reading workload, typically by 50–70%, while maintaining non-inferior sensitivity and specificity. For example, excluding low-risk examinations from human reading based on AI scores reduced workload by over 60% with non-inferior sensitivity (69.7% vs. 70.8%) and higher specificity (98.6% vs 98.1%) compared with double reading [17]. In a large-scale analysis, up to 70% reduction was achieved while retaining detection of over 99% of screen-detected cancers and 85% of interval cancers [58]. Similar findings were reported across other studies [16,55,66].

Overall, AI triage substantially reduces radiologist workload while maintaining screening performance in both real-world and simulated settings, although findings are based largely on retrospective designs and may vary depending on threshold selection and implementation strategy.

3.8. AI as an independent reader in Human-AI screening workflow

Seven studies [15,26,36,37,56,57,67] evaluated workflows where AI replaces or augments radiologists in double- or triple- reading systems, including one prospective trial in Sweden and six retrospective simulation studies using large, unenriched datasets (see Table 4 and Appendix D – Table D.4).

The prospective ScreenTrustCAD trial in Sweden [26] evaluated AI as an independent reader for DM (n = 55,581 women) within a double-reading workflow. Double reading with AI plus one radiologist achieved a non-inferior and slightly higher cancer detection rate (CDR) than standard double reading by two radiologists (0.47% vs. 0.45%, relative proportion 1.04, 95%CI: 1.00–1.09; $p = 0.017$), with a lower post-consensus recall rate (2.80% vs. 2.93%). AI-only reading was also non-inferior (CDR: 0.44%), although not superior to standard double reading. Triple reading (two radiologists and AI) yielded the highest CDR (0.48%) but at the cost of the highest recall rate (3.09%) and consensus workload. Replacing one radiologist with AI reduced initial reading workload by approximately 50% [26].

Retrospective studies consistently showed that replacing one radiologist with AI in double-reading workflows maintains screening performance while reducing workload. Most studies paired AI outputs with archived radiologist reads and applied the original arbitration decisions when discrepancies occurred, although one study [57] recalled cases if either reader (AI or human) flagged them as abnormal without arbitration. Across large multicentre analyses [15,56,67], AI plus one radiologist achieved non-inferior sensitivity and CDR compared with standard double reading, with higher specificity and PPV. Workload reduction typically ranged from 30 to 45% in standard simulation, and up to 87% in modified configurations, although arbitration rates increased from 3.3% to 12.3%. Similar findings were reported across multiple AI systems, with performance influenced by threshold selection and associated trade-offs between recall and specificity [36,37,57]. In DBT, substituting the second reader with AI detected approximately 95% of cancers identified by human double reading while halving the overall reading workload [59].

Overall, these findings indicate that AI can serve as an independent or supporting reader, sustaining or modestly improving screening accuracy while substantially reducing radiologist workload, although some configurations increase arbitration frequency and findings are primarily based on retrospective simulation.

3.9. AI as an additional reader for Non-Recalled cases

A prospective study in Hungary evaluated AI as an additional reviewer within a standard double-reading workflow [25]. AI analysed cases marked as "no-recall", and AI-flagged cases were subsequently reviewed by an additional radiologist with access to AI outputs.

Across a phased implementation (pilot to multi-site rollout), AI increased cancer detection rate by 0.7–1.6 per 1,000 population, with a small increase in recall rate (0.16%-0.30%) and minimal increases in unnecessary recalls (0%-0.23%). PPV improved by up to 1.9%, and most additional cancers detected were invasive (83.3%) and small ($\leq$1cm in 47% of cases). In higher-specificity simulation, the proportion of AI-flagged cases decreased to 2.4%-3.0% while still detecting 4–10% more cancers, indicating that threshold tuning can balance workload and detection gains [25].

3.10. Certainty of evidence

A GRADE-style assessment suggested moderate certainty for triage-related workload reduction, low certainty across most other roles including reader-aid, independent reader, and secondary review applications, and very low certainty for full autonomous replacement of radiologists, reflecting the limited availability of prospective, real-world evidence for most roles (Appendix H).

4. Discussion

This systematic review synthesised evidence from 42 studies evaluating FDA-cleared and/or CE-marked AI systems for breast cancer

screening across multiple clinical roles. Most studies were retrospective and heterogeneous in design, whereas the limited prospective evidence was largely restricted to workflow-based applications rather than standalone AI. Pooled analyses suggest that AI can achieve diagnostic performance comparable to radiologists in digital mammography, while emerging prospective evidence indicates that AI-supported workflow models can reduce workload while maintaining screening performance and, in some settings, improving screening outcomes.

Standalone AI in DM achieved radiologist-level accuracy, consistent with previous *meta*-analyses [19,20]. However, substantial heterogeneity and wide prediction intervals suggested that performance is unlikely to be uniform across screening settings. Exploratory analyses indicated that vendor-related differences contributed more to performance variability than screening setting or cancer definition, although no individual factor sufficiently explained the observed heterogeneity. Interestingly, standalone AI performance remained similar when analyses were restricted to unenriched screening cohorts, suggesting that the overall findings were not solely driven by enriched datasets or case-control designs commonly used in early AI validation studies. Together, these findings suggest that AI performance is context-dependent and reinforce the need for local validation before implementation.

Importantly, standalone AI performance and workflow performance should not be considered interchangeable. Standalone studies primarily evaluate diagnostic accuracy using metrics such as AUC, sensitivity, and specificity, whereas workflow studies assess the impact of AI on screening outcomes including cancer detection rate, recall rate, PPV, workload, and arbitration requirements. Consequently, strong standalone performance does not necessarily translate into improved screening programme performance, which depends on workflow design, radiologist-AI interaction, and implementation strategy. Notably, prospective evidence to date has primarily evaluated AI within screening workflows rather than as a fully autonomous screening system. Although the ScreenTrustCAD trial [26] reported non-inferior cancer detection for standalone AI compared with double reading, AI did not determine recall or diagnostic pathways and detected fewer cancers than standard double reading or AI-plus-one-radiologist approaches [26]. Therefore, current evidence does not support replacing radiologists with standalone AI in routine practice.

AI-assisted reading improved radiologists' performance in laboratory reader studies, particularly AUC and sensitivity. However, these findings derive from controlled settings with enriched datasets, which may overestimate benefits in routine practice. Real-world evidence of AI as a reader aid remains limited and mixed: single-reading workflows showed improvements in diagnostic precision but inconsistent effects on cancer detection, while double-reading workflows showed greater potential for improving CDR and PPV, with small increases in recall. The MASAI trial suggests that concurrent AI decision support can be integrated within hybrid triage workflows, although its independent effect cannot be separated from AI-based triage. Emerging evidence, including AI-STREAM trial [68] and one retrospective evaluation study [64], also suggest potential diagnostic benefit, but generalisability remains limited. Overall, the impact of AI as a reader aid is context-dependent and influenced by workflow design and implementation strategy.

Evidence is also emerging for AI replacing one radiologist in standard double-reading systems. The ScreenTrustCAD trial [26] showed that AI can act as an independent reader, achieving noninferior cancer detection and lower recall compared with standard double reading, consistent with retrospective simulations. One should acknowledge however that human-AI disagreement may increase arbitration workload and workflow complexity [26,37], requiring careful implementation and resource planning.

The strongest operational evidence supports AI as a triage tool, where examinations are allocated to single or double reading based on AI-generated risk scores. This approach may help address workforce shortages and workload pressures. Evidence from Denmark [65] and the MASAI trial in Sweden [24,27] have demonstrated substantial workload reductions while maintaining stable recall rates and improving cancer detection. Updated MASAI results confirmed a 28% increase in cancer detection and 44.2% reduction in workload [27].

Further refinement of AI-triage strategies may improve screening performance by optimising allocation of examinations to different reading pathways. One simulation study [17] explored automatic classification of low-risk cases as normal and high-risk cases as recall, with only intermediate-risk cases double read. While conceptually efficient, removing human oversight for both low- and high-risk groups raises safety, ethical, and medicolegal concerns [69], particularly given the limited prospective evidence for standalone AI decision-making.

The implications of AI implementation may differ between digital mammography (DM) and digital breast tomosynthesis (DBT). The evidence base remains substantially stronger for DM, reflecting both the larger number of studies available and the availability of prospective implementation data. Although AI performance in DBT appeared broadly comparable to DM, the DBT evidence base remains comparatively limited and largely derived from retrospective and reader studies. AI may provide particular value in DBT screening because of the greater burden associated with larger image volumes. Consistent with this, real-world evidence suggests that AI-supported reading can improve DBT screening performance [64]. However, further prospective implementation studies in DBT are required before firm conclusions can be drawn. Accordingly, current evidence supports implementation decisions more strongly in DM than DBT.

Our findings extend Freeman et al. [13], which concluded that evidence was insufficient to determine optimal clinical roles for AI and that AI was not yet accurate enough to replace radiologists. Since then, prospective and real-world evaluations have emerged across multiple AI roles. However, consistent with previous reviews [13,70], standalone AI still lacks prospective validation as a full replacement for radiologists.

Beyond diagnostic performance, safe implementation also depends on the regulatory and surveillance environment in which AI is deployed. Regulatory approval represents an important step towards implementation but does not guarantee consistent long-term performance across all screening settings, since AI systems are typically validated under specific conditions and performance may vary as populations, equipment, protocols, and software versions evolve. Regulatory frameworks are evolving to address this: the FDA's 2024 Predetermined Change Control Plan guidance allows pre-specified software updates without full resubmission, and the EU AI Act extends oversight of high-risk AI-enabled devices beyond initial CE-marking, though requirements still differ across jurisdictions. Consequently, screening programmes adopting AI should incorporate ongoing post-market surveillance, including monitoring of cancer detection, recall, interval cancer rates, PPV, arbitration rates, and workload impact, particularly given the substantial heterogeneity observed across studies.

Alongside this regulatory uncertainty, certainty of evidence was highest, although still only moderate, for triage-related workload reduction, and low or very low for roles involving greater autonomy, including independent reading and full replacement (Appendix H). This gradient, in which certainty declines as the AI's role becomes more autonomous, supports a staged approach to implementation in which higher autonomy-roles warrant more conservative adoption and closer monitoring. Translating this evidence into practice requires an implementation pathway that accounts for both this heterogeneity and the need for ongoing surveillance. We therefore propose a practical, stepwise framework for implementing and governing AI in breast screening (Fig. 7), progressing from defining intended use and local validation, through workflow selection and workforce preparation, to silent-mode piloting and controlled rollout, each separated by explicit go/no-go decision gates rather than a single deployment decision. The framework closes with continuous monitoring and revalidation triggered by performance drift, software updates, or population change, embedding local validation, workflow integration, and workforce training, the elements most consistently supported by current evidence, within an

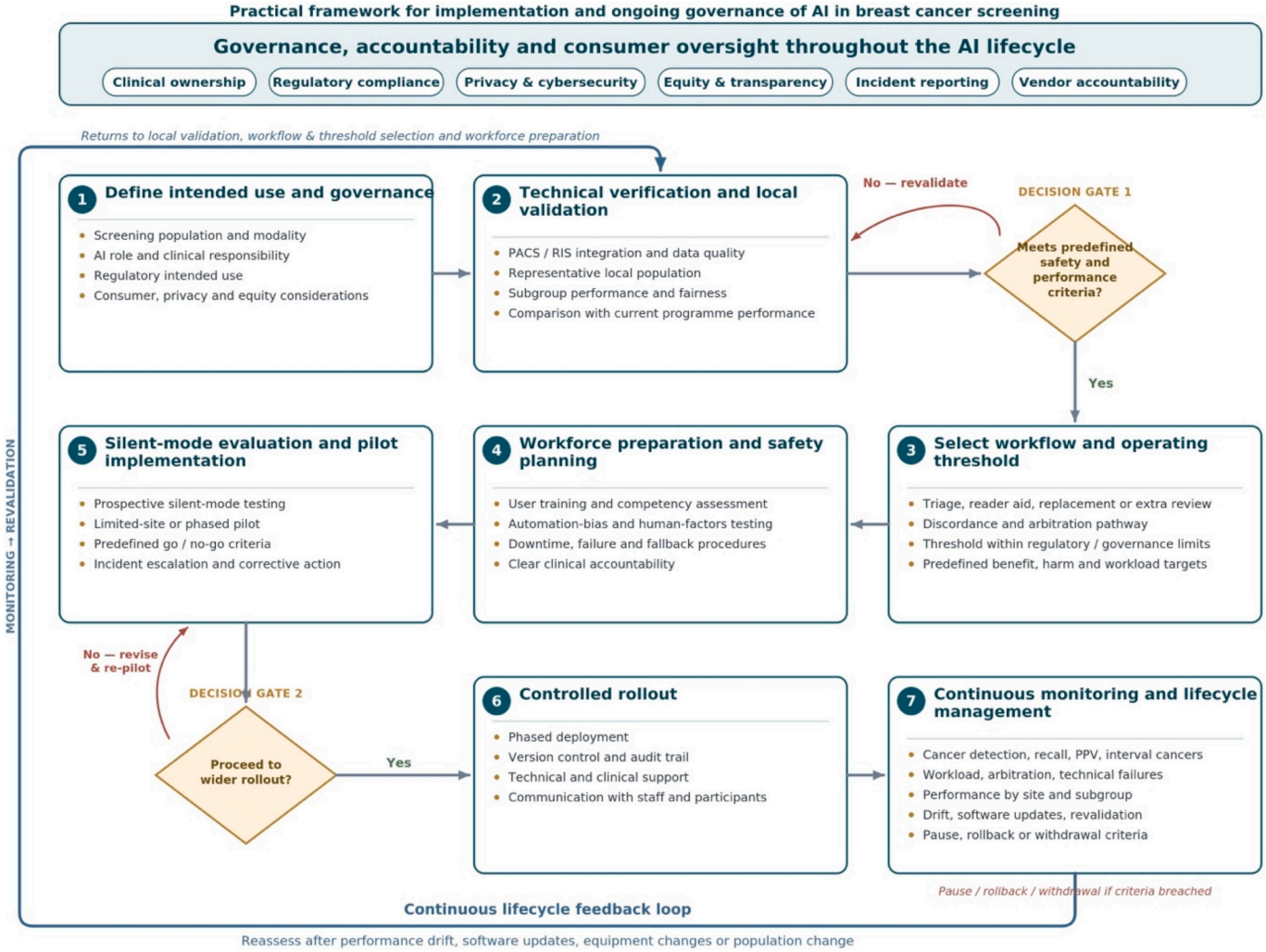

Fig. 7. Proposed framework for the implementation and ongoing governance of AI in breast cancer screening. The framework progresses through seven sequential stages, gated by two explicit safety/performance and pilot go/no-go decision points, and closes with a continuous monitoring and revalidation loop triggered by performance drift, software updates, equipment changes, or population change. All stages are underpinned by governance and accountability structures spanning clinical ownership, regulatory compliance, privacy and cybersecurity, equity and transparency, incident reporting, and vendor accountability.

ongoing governance structure spanning clinical, regulatory, and technical accountability.

Despite emerging prospective evidence, most included studies remain retrospective and often use enriched datasets, limited follow-up, or designs prone to verification bias, reducing generalisability. These factors likely lead to overestimation of AI performance, and simulation studies may also fail to capture real radiologist behaviour when AI influences recall decisions. Consequently, although retrospective studies consistently demonstrated promising diagnostic performance, implementation recommendations currently rely more heavily on the limited prospective evidence evaluating AI within screening workflows. Geographic diversity is also limited, with most studies conducted in Europe. Future evaluations should include more diverse populations, particularly from Asia, Asia-Pacific, and other underrepresented regions, and incorporate prior mammograms and clinical context. This review is further limited by rapid developments since the search period and heterogeneity across included studies.

Overall, commercially available FDA-cleared and/or CE-marked AI systems demonstrate promising diagnostic performance in retrospective studies and early prospective evaluations. The strongest evidence supports workflow models that reduce radiologist workload, particularly triage and independent/supporting reader configurations. However, prospective evidence for complete radiologist replacement remains lacking, and safe implementation will require site-specific adaptation and further evaluations across diverse screening settings.

CRediT authorship contribution statement

Putu Irma Wulandari: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ziba Gandomkar:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mary Rickard:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Conceptualization. **Sahand Hooshmand:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mo'ayyad Suleiman:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Patrick Brennan:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used artificial intelligence tools to assist with language editing, refinement of expression, and improvement of manuscript structure. The authors reviewed and edited all content as needed and take full responsibility for the final version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2026.113136.

References

[1] F. Bray, M. Laversanne, H. Sung, et al., Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 74 (3) (2024) 229–263, https://doi.org/10.3322/caac.21834.

[2] M. Salim, K. Dembrower, M. Eklund, P. Lindholm, F. Strand, Range of radiologist performance in a population-based screening cohort of 1 million digital mammography examinations, Radiology 297 (1) (2020) 33–39, https://doi.org/10.1148/radiol.2020192212.

[3] Y. Chen, J.J. James, E. Michalopoulou, I.T. Darker, J. Jenkins, The relationship between missed breast cancers on mammography in a test-set based assessment scheme and real-life performance in a National Breast Screening Programme, Eur. J. Radiol. 142 (2021) 109881, https://doi.org/10.1016/j.ejrad.2021.109881.

[4] European Commission. European guidelines on breast cancer screening and diagnosis. Available at: https://cancer-screening-and-care.jrc.ec.europa.eu/en/ecibc/european-breast-cancer-guidelines?topic=64&usertype=60#guideline. Accessed 21 August 2024.

[5] S.R. Hoff, T. Myklebust, C.I. Lee, S. Hofvind, Influence of mammography volume on radiologists' performance: results from BreastScreen Norway, Radiology 292 (2) (2019) 289–296, https://doi.org/10.1148/radiol.2019182684.

[6] Taylor CR, Monga N, Johnson C, Hawley JR, Patel M. Artificial Intelligence Applications in Breast Imaging: Current Status and Future Directions. Diagnostics (Basel). 2023; 13(12). doi: 10.3390/diagnostics13122041.

[7] S. Batchu, F. Liu, A. Amireh, J. Waller, M. Umair, A review of applications of machine learning in mammography and future challenges, Oncology 99 (8) (2021) 483–490, https://doi.org/10.1159/000515698.

[8] F. Pesapane, C. Trentin, M. Montesano, et al., Mammography in 2022, from computer-aided detection to artificial intelligence applications, CEOG 49 (11) (2022), https://doi.org/10.31083/j.ceog4911237.

[9] N. Hendrix, B. Hauber, C.I. Lee, A. Bansal, D.L. Veenstra, Artificial intelligence in breast cancer screening: primary care provider preferences, J. Am. Med. Inform. Assoc. 28 (6) (2021) 1117–1124, https://doi.org/10.1093/jamia/ocaa292.

[10] C.F. de Vries, S.J. Colosimo, M. Boyle, et al., AI in breast screening mammography: breast screening readers' perspectives, Insights into Imaging 13 (1) (2022), https://doi.org/10.1186/s13244-022-01322-4.

[11] N. Lennox-Chhugani, Y. Chen, V. Pearson, B. Trzcinski, J. James, Women's attitudes to the use of AI image readers: a case study from a national breast screening programme, BMJ Health Care Inform 28 (1) (2021), https://doi.org/10.1136/bmjhci-2020-100293.

[12] Å.S. Holen, M.A. Martiniussen, M.B. Bergan, N. Moshina, T. Hovda, S. Hofvind, Women's attitudes and perspectives on the use of artificial intelligence in the assessment of screening mammograms, Eur. J. Radiol. 175 (2024) 111431, https://doi.org/10.1016/j.ejrad.2024.111431.

[13] K. Freeman, J. Geppert, C. Stinton, et al., Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy, BMJ (2021), https://doi.org/10.1136/bmj.n1872.

[14] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, et al., Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists, J. Natl Cancer Inst. 111 (9) (2019) 916–922, https://doi.org/10.1093/JNCI/DJY222.

[15] N. Sharma, A.Y. Ng, J.J. James, et al., Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms, BMC Cancer 23 (1) (2023) 460, https://doi.org/10.1186/s12885-023-10890-7.

[16] J.L. Raya-Povedano, S. Romero-Martín, E. Elías-Cabot, A. Gubern-Mérida, A. Rodríguez-Ruiz, M. Álvarez-Benito, AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation, Radiology 300 (1) (2021) 57–65, https://doi.org/10.1148/radiol.2021203555.

[17] A.D. Lauritzen, A. Rodriguez-Ruiz, M.C. von Euler-Chelpin, et al., An artificial intelligence-based mammography screening protocol for breast cancer: outcome and radiologist workload, Radiology 304 (1) (2022) 41–49, https://doi.org/10.1148/radiol.210948.

[18] M. Bahl, Artificial intelligence in clinical practice: implementation considerations and barriers, J. Breast Imaging. 4 (6) (2022) 632–639, https://doi.org/10.1093/jbi/wbac065.

[19] J.H. Yoon, F. Strand, P.A.T. Baltzer, et al., Standalone AI for breast cancer detection at screening digital mammography and digital breast tomosynthesis: a systematic review and meta-analysis, Radiology 307 (5) (2023), https://doi.org/10.1148/radiol.222639.

[20] S.E. Hickman, R. Woitek, E.P.V. Le, et al., Machine learning for workflow applications in screening mammography: systematic review and meta-analysis, Radiology 302 (1) (2022) 88–104, https://doi.org/10.1148/radiol.2021210391.

[21] J. Liu, J. Lei, Y. Ou, et al., Mammography diagnosis of breast cancer screening through machine learning: a systematic review and meta-analysis, Clin. Exp. Med. 23 (6) (2022) 2341–2356, https://doi.org/10.1007/s10238-022-00895-0.

[22] A. Zeng, N. Houssami, N. Noguchi, B. Nickel, M.L. Marinovich, Frequency and characteristics of errors by artificial intelligence (AI) in reading screening mammography: a systematic review, Breast Cancer Res. Treat. 207 (1) (2024) 1–13, https://doi.org/10.1007/s10549-024-07353-3.

[23] D. Xavier, I. Miyawaki, C.A. Campello Jorge, et al., Artificial intelligence for triaging of breast cancer screening mammograms and workload reduction: a meta-analysis of a deep learning software, J. Med. Screen. (2023), https://doi.org/10.1177/09691413231219952.

[24] K. Lång, V. Josefsson, A.M. Larsson, et al., Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study, Lancet Oncol. 24 (8) (2023) 936–944, https://doi.org/10.1016/S1470-2045(23)00298-X.

[25] A.Y. Ng, C.J.G. Oberije, É. Ambrózay, et al., Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer, Nat. Med. 29 (12) (2023) 3044–3049, https://doi.org/10.1038/s41591-023-02625-9.

[26] K. Dembrower, A. Crippa, E. Colón, M. Eklund, F. Strand, Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study, Lancet Digit Heal. 5 (10) (2023) e703–e711, https://doi.org/10.1016/S2589-7500(23)00153-X.

[27] V. Hernström, V. Josefsson, H. Sartor, et al., Screening performance and characteristics of breast cancer detected in the mammography screening with artificial intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study, Lancet Digit Heal. (2025), https://doi.org/10.1016/s2589-7500(24)00267-x.

[28] U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed 19 June 2024.

[29] European Commission. EUDAMED - European Database on Medical Devices. Available at: https://ec.europa.eu/tools/eudamed/#/screen/home. Accessed 30 June 2024.

[30] U.J. Muehlematter, P. Daniore, K.N. Vokinger, Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis, Lancet Digit Heal. 3 (3) (2021) e195–e203, https://doi.org/10.1016/s2589-7500(20)30292-2.

[31] K.G. van Leeuwen, S. Schalekamp, M.J.C.M. Rutten, B. van Ginneken, M. de Rooij, Artificial intelligence in radiology: 100 commercially available products and their scientific evidence, Eur. Radiol. 31 (6) (2021) 3797–3804, https://doi.org/10.1007/s00330-021-07892-z.

[32] V. Sounderajah, H. Ashrafian, S. Rose, et al., A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI, Nat. Med. 27 (10) (2021) 1663–1665, https://doi.org/10.1038/s41591-021-01517-0.

[33] D. Byng, B. Strauch, L. Gnas, et al., AI-based prevention of interval cancers in a national mammography screening program, Eur. J. Radiol. 152 (2022), https://doi.org/10.1016/j.ejrad.2022.110321.

[34] C.F. de Vries, S.J. Colosimo, R.T. Staff, et al., Impact of different mammography systems on artificial intelligence performance in breast cancer screening, Radiol. Artif. Intell. 5 (3) (2023), https://doi.org/10.1148/ryai.220146.

[35] T. Tan, A. Rodriguez-Ruiz, T. Zhang, et al., Multi-modal artificial intelligence for the combination of automated 3D breast ultrasound and mammograms in a population of women with predominantly dense breasts, Insights into Imaging 14 (1) (2023), https://doi.org/10.1186/s13244-022-01352-y.

[36] M.T. Elhakim, S.W. Stougaard, O. Graumann, et al., Breast cancer detection accuracy of AI in an entire screening population: a retrospective, multicentre study, Cancer Imaging 23 (1) (2023) 127, https://doi.org/10.1186/s40644-023-00643-x.

[37] K. Dembrower, M. Salim, M. Eklund, P. Lindholm, F. Strand, Implications for downstream workload based on calibrating an artificial intelligence detection algorithm by standalone-reader or combined-reader sensitivity matching, J. Med. Imaging 10 (2) (2023) S22405, https://doi.org/10.1117/1.JMI.10.S2.S22405.

[38] C.J.G. Oberije, N. Sharma, J.J. James, A.Y. Ng, J. Nash, P.D. Kecskemethy, Comparing prognostic factors of cancers identified by artificial intelligence (AI) and human readers in breast cancer screening, Cancers 15 (12) (2023), https://doi.org/10.3390/cancers15123069.

[39] K. Lång, S. Hofvind, A. Rodríguez-Ruiz, I. Andersson, Can artificial intelligence reduce the interval cancer rate in mammography screening? Eur. Radiol. 31 (8) (2021) 5940–5947, https://doi.org/10.1007/s00330-021-07686-3.

[40] D.L. Nguyen, Y. Ren, T.M. Jones, S.M. Thomas, J.Y. Lo, L.J. Grimm, Patient characteristics impact performance of AI algorithm in interpreting negative

screening digital breast tomosynthesis studies, Radiology 311 (2) (2024), https://doi.org/10.1148/radiol.232286.
[41] S. Weigel, A.K. Brehl, W. Heindel, L. Kerschke, Artificial intelligence for indication of invasive assessment of calcifications in mammography screening, RoFo Fortschritte Auf Dem Gebiet Der Rontgenstrahlen Und Der Bildgebenden Verfahren. 195 (1) (2023) 38–46, https://doi.org/10.1055/a-1967-1443.
[42] L. Kerschke, S. Weigel, A. Rodriguez-Ruiz, N. Karssemeijer, W. Heindel, Using deep learning to assist readers during the arbitration process: a lesion-based retrospective evaluation of breast cancer screening performance, Eur. Radiol. 32 (2) (2022) 842–852, https://doi.org/10.1007/s00330-021-08217-w.
[43] M. Sasaki, M. Tozaki, A. Rodriguez-Ruiz, 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 27 (4) (2020) 642–651, https://doi.org/10.1007/s12282-020-01061-8.
[44] V. Dahlblom, I. Andersson, K. Lång, A. Tingberg, S. Zackrisson, M. Dustler, Artificial intelligence detection of missed cancers at digital mammography that were detected at digital breast tomosynthesis. Radiology, Artif. Intell. 3 (6) (2021), https://doi.org/10.1148/ryai.2021200299.
[45] S. Romero-Martin, E. Elias-Cabot, J.L. Raya-Povedano, A. Gubern-Merida, A. Rodriguez-Ruiz, M. Alvarez-Benito, Stand-alone use of artificial intelligence for digital mammography and digital breast tomosynthesis screening: a retrospective evaluation, Radiology 302 (3) (2022) 535–542, https://doi.org/10.1148/radiol.211590.
[46] L. Celik, D.C. Guner, O. Ozcaglayan, R. Cubuk, M.E. Aribal, Diagnostic performance of two versions of an artificial intelligence system in interval breast cancer detection, Acta Radiol. 64 (11) (2023) 2891–2897, https://doi.org/10.1177/02841851231200785.
[47] H. Yoen, J.M. Chang, Artificial intelligence improves detection of supplemental screening ultrasound-detected breast cancers in mammography, J. Breast Cancer 26 (5) (2023) 504–513, https://doi.org/10.4048/jbc.2023.26.e39.
[48] M.B. Bergan, M. Larsen, N. Moshina, et al., AI performance by mammographic density in a retrospective cohort study of 99,489 participants in BreastScreen Norway, Eur. Radiol. (2024), https://doi.org/10.1007/s00330-024-10681-z.
[49] S.E. Lee, H. Hong, E.K. Kim, Positive predictive values of abnormality scores from a commercial artificial intelligence-based computer-aided diagnosis for mammography, Korean J. Radiol. 25 (2) (2024), https://doi.org/10.3348/kjr.2023.0907.
[50] M.R. Kwon, Y. Chang, S.Y. Ham, et al., Screening mammography performance according to breast density: a comparison between radiologists versus standalone intelligence detection, Breast Cancer Res. 26 (1) (2024), https://doi.org/10.1186/s13058-024-01821-w.
[51] S.L. van Winkel, A. Rodríguez-Ruiz, L. Appelman, et al., Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study, Eur. Radiol. 31 (11) (2021) 8682–8691, https://doi.org/10.1007/s00330-021-07992-w.
[52] M.C. Pinto, A. Rodriguez-Ruiz, K. Pedersen, 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 300 (3) (2021) 529–536, https://doi.org/10.1148/radiol.2021204432.
[53] J.G. Kim, B. Haslam, A.R. Diab, et al., Impact of a categorical AI system for digital breast tomosynthesis on breast cancer interpretation by both general radiologists and breast imaging specialists, Radiol. Artif. Intell. 6 (2) (2024), https://doi.org/10.1148/ryai.230137.
[54] H. Kim, J.S. Choi, K. Kim, E.S. Ko, E.Y. Ko, B.K. Han, Effect of artificial intelligence-based computer-aided diagnosis on the screening outcomes of digital mammography: a matched cohort study, Eur. Radiol. 33 (10) (2023) 7186–7198, https://doi.org/10.1007/s00330-023-09692-z.
[55] T.A. Retson, A.T. Watanabe, H. Vu, C.Y. Chim, Multicenter, multivendor validation of an FDA-approved algorithm for mammography triage, J Breast Imaging. 4 (5) (2022) 488–495, https://doi.org/10.1093/jbi/wbac046.
[56] N. Sharma, A.Y. Ng, J.J. James, et al., Retrospective large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening, medRxiv (2021), https://doi.org/10.1101/2021.02.26.21252537.
[57] S.H. Heywang-Kobrunner, A. Hacker, A. Jansch, et al., Use of novel artificial intelligence computer-assisted detection (AI-CAD) for screening mammography: an analysis of 17,884 consecutive two-view full-field digital mammography screening exams, Acta Radiol. 64 (10) (2023) 2697–2703, https://doi.org/10.1177/02841851231187382.
[58] M. Larsen, C.F. Olstad, C.I. Lee, et al., Performance of an artificial intelligence system for breast cancer detection on screening mammograms from BreastScreen Norway, Radiol. Artif. Intell. 6 (3) (2024), https://doi.org/10.1148/ryai.230375.
[59] V. Dahlblom, M. Dustler, A. Tingberg, S. Zackrisson, Breast cancer screening with digital breast tomosynthesis: comparison of different reading strategies implementing artificial intelligence, Eur. Radiol. 33 (5) (2023) 3754–3765, https://doi.org/10.1007/s00330-022-09316-y.
[60] A. Rodríguez-Ruiz, E. Krupinski, J.J. Mordang, et al., Detection of breast cancer with mammography: effect of an artificial intelligence support system, Radiology 290 (3) (2019) 305–314, https://doi.org/10.1148/radiol.2018181371.
[61] J.H. Lee, K.H. Kim, E.H. Lee, et al., Improving the performance of radiologists using artificial intelligence-based detection support software for mammography: a multi-reader study, Korean J. Radiol. 23 (5) (2022) 505–516, https://doi.org/10.3348/kjr.2021.0476.
[62] L.A. Dang, E. Chazard, E. Poncelet, et al., Impact of artificial intelligence in breast cancer screening with mammography, Breast Cancer 29 (6) (2022) 967–977, https://doi.org/10.1007/s12282-022-01375-9.
[63] H. Letter, M. Peratikos, A. Toledano, et al., Use of artificial intelligence for digital breast tomosynthesis screening: a preliminary real-world experience, J Breast Imaging. 5 (3) (2023) 258–266, https://doi.org/10.1093/jbi/wbad015.
[64] E. Elías-Cabot, S. Romero-Martín, J.L. Raya-Povedano, A.K. Brehl, M. Álvarez-Benito, Impact of real-life use of artificial intelligence as support for human reading in a population-based breast cancer screening program with mammography and tomosynthesis, Eur. Radiol. 34 (6) (2024) 3958–3966, https://doi.org/10.1007/s00330-023-10426-4.
[65] A.D. Lauritzen, M. Lillholm, E. Lynge, M. Nielsen, N. Karssemeijer, I. Vejborg, Early indicators of the impact of using AI in mammography screening for breast cancer, Radiology 311 (3) (2024), https://doi.org/10.1148/radiol.232479.
[66] K. Lang, M. Dustler, V. Dahlblom, A. Akesson, I. Andersson, S. Zackrisson, Identifying normal mammograms in a large screening population using artificial intelligence, Eur. Radiol. 31 (3) (2021) 1687–1692, https://doi.org/10.1007/s00330-020-07165-1.
[67] A.Y. Ng, B. Glocker, C. Oberije, et al., Artificial intelligence as supporting reader in breast screening: a novel workflow to preserve quality and reduce workload, J. Breast Imaging. 5 (3) (2023) 267–276, https://doi.org/10.1093/jbi/wbad010.
[68] Y.-W. Chang, J.K. Ryu, J.K. An, et al., Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study, Nat. Commun. 16 (1) (2025), https://doi.org/10.1038/s41467-025-57469-3.
[69] S.M. Carter, W. Rogers, K.T. Win, H. Frazer, B. Richards, N. Houssami, The ethical, legal and social implications of using artificial intelligence systems in breast cancer care, Breast 49 (2020) 25–32, https://doi.org/10.1016/j.breast.2019.10.001.
[70] O. Díaz, A. Rodríguez-Ruíz, I. Sechopoulos, Artificial Intelligence for breast cancer detection: technology, challenges, and prospects, Eur. J. Radiol. 175 (2024), https://doi.org/10.1016/j.ejrad.2024.111457.