INPLASY

Deep Learning, Radiologist, and Hybrid Reading Strategies in Screening Digital Mammography: a Systematic Review and Network Meta-analysis

INPLASY202620007
doi: 10.37766/inplasy2026.2.0007
Received: 2 February 2026
Published: 2 February 2026

Corresponding author:
Ting Wei Wang

eltonwang1@gmail.com

Author Affiliation:
Taipei Veteran General Hospital.

Wang, TW; Yang, S; Wang, YC; Tsai, YF; Tu, YK; Wu, YT; Tseng, LM; Lai, YC.

ADMINISTRATIVE INFORMATION

Support - NR.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - The authors declare no financial or personal conflicts of interest related to this work. None of the authors is employed by, receives consulting fees from, or holds equity in vendors of deep learning mammography systems evaluated in the included studies. Any vendor involvement reported in primary studies will be extracted and analyzed as a study-level moderator but does not constitute an author conflict.

INPLASY registration number: INPLASY202620007

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 2 February 2026 and was last updated on 2 February 2026.

INTRODUCTION

Review question / Objective Using screening digital mammography (DM), what is the comparative diagnostic accuracy (sensitivity and specificity) of (1) radiologist-only reading strategies, (2) stand-alone deep learning (DL) systems, and (3) hybrid human–DL workflows?

PICOS:
Population: Asymptomatic adults undergoing screening DM.
Index/Interventions: Stand-alone DL and hybrid workflows integrating DL with human readers (e.g., decision support, DL as independent reader, triage/arbitration variants).
Comparators: Single radiologist reading and radiologists' consensus/double-reading strategies (with arbitration/consensus).
Outcomes: Sensitivity and specificity derived from 2×2 tables (TP/FP/TN/FN), using pathology and/or program follow-up as reference standards.
Study type: Head-to-head comparative diagnostic accuracy studies.

Condition being studied Breast cancer detection in population screening using digital mammography. The review focuses on diagnostic accuracy of competing reading strategies (radiologist-only, stand-alone DL, and hybrid human–DL workflows) for identifying breast cancer in screening settings. Reference standards will include pathology confirmation for screen-detected cancers and program/registry-linked

follow-up to verify negative examinations and capture interval cancers, consistent with screening evaluation practice.

METHODS

Participant or population Adults undergoing screening digital mammography (DM) in real-world screening programs or screening-like cohorts. Studies must report comparative (head-to-head) performance of DL vs radiologists and/or hybrid workflows in a screening context. Studies limited to diagnostic workup populations, enriched case-control test sets without screening denominators, pediatric populations, or non-screening indications will be excluded.

Intervention Stand-alone deep learning systems for screening mammography interpretation and hybrid human–DL reading strategies integrating DL outputs into radiologist workflows (e.g., decision support/OR-rule, DL as independent second reader, triage, arbitration/consensus integration), including commercial/regulatory-approved and research-stage DL systems when evaluated head-to-head.

Comparator Radiologist-only reading strategies: single radiologist reading and radiologists' consensus/double reading with arbitration/consensus (without DL). When available, additional workflow comparators within the same cohort (e.g., standard double reading) will be included as separate strategy nodes.

Study designs to be included Head-to-head comparative diagnostic accuracy studies in screening digital mammography, including prospective paired-reader trials, prospective cohorts, and retrospective cohort evaluations (including simulated workflow analyses) that permit reconstruction of TP/FP/TN/FN for at least two competing strategies within the same study.

Eligibility criteria Inclusion:Screening DM studies comparing DL vs radiologists and/or hybrid workflows head-to-head.Adult screening population with case-level outcomes.Reference standard: pathology and/or program/registry follow-up adequate to classify negatives and interval cancers (per study definition).Sufficient data to reconstruct 2×2 tables (TP/FP/TN/FN) for at least two strategies.Exclusion:Non-comparative studies or no head-to-head data.Non-screening settings or diagnostic-only populations.Non-DM primary modality (unless DM results are separable).Outcomes reported only at breast/image/lesion/patch level without exam-level 2×2.Reviews, editorials, protocols.Duplicate/overlapping cohorts: when overlap is likely, only the most complete/appropriate dataset will be retained for quantitative synthesis.

Information sources Electronic databases: PubMed, Embase, Web of Science, IEEE. Additional sources: backward citation searching of included studies and relevant systematic reviews; checking related articles where applicable. No trial registries or grey literature are required, but conference abstracts will be screened if they provide extractable head-to-head 2×2 data and sufficient methodological detail (otherwise excluded).

Main outcome(s) Primary outcomes:

Sensitivity and specificity for breast cancer detection on screening DM for each prespecified strategy node, derived from TP/FP/TN/FN (exam-level).
Effect measures: pooled absolute sensitivity/specificity and relative sensitivity/specificity versus a prespecified reference (single radiologist). 95% CIs and prediction intervals will be reported. Timing aligns with the study's reference standard (pathology and/or follow-up window).

Quality assessment / Risk of bias analysis Two reviewers will independently assess risk of bias and applicability using QUADAS-2, comparative bias using QUADAS-C, and AI/prediction-model related risks using PROBAST-AI. Disagreements will be resolved by consensus with a third author. Certainty of evidence will be summarized using GRADE-DTA, considering risk of bias, indirectness, inconsistency, imprecision, and publication bias.

Strategy of data synthesis We will conduct an arm-based network meta-analysis of diagnostic accuracy. For each study arm, TP and TN will be modeled with binomial likelihoods; sensitivity and specificity will be jointly modeled on the logit scale using a bivariate random-effects (variance-component) framework. Strategy nodes will include: single radiologist, stand-alone DL, radiologists' consensus, single radiologist + DL, and radiologists' consensus + DL. Relative sensitivity/specificity will be estimated versus single radiologist. Network geometry will be summarized by node size (sample size) and edge thickness (number of direct comparisons). Global inconsistency will be tested using design-by-treatment interaction; local inconsistency via loop-specific methods. Analyses will be implemented in Stata (e.g., metadta with abnetwork). We will report

pooled estimates with 95% CIs and prediction intervals.

Subgroup analysis Prespecified meta-regression/ subgroup moderators include:

Thresholding strategy (vendor-suggested, study-defined, matched specificity/sensitivity, rule-out)

Negative-case definition and follow-up duration (e.g., ≥1 year vs ≥2 years; registry linkage)

Vendor involvement (yes/no)

Region/economic setting

Study design (prospective vs retrospective; simulated workflow vs real workflow)
Where data allow, additional subgroup summaries may include breast density and age strata.

Sensitivity analysis Sensitivity analyses will include:

Excluding studies at high/unclear risk of bias in key domains (threshold prespecification, flow/timing, reference standard).

Excluding simulated-workflow studies (retaining only real-world workflow evaluations).

Alternative handling of potential cohort overlap (keeping only the largest/most recent dataset).

Restricting to studies with ≥1-year follow-up (or separately to ≥2-year follow-up) to test robustness to negative-case definitions.

Restricting to commercial/regulatory-approved systems vs including research-stage models, where feasible.

Country(ies) involved Taiwan.

Keywords Digital mammography; breast cancer screening; deep learning; artificial intelligence; radiologist; hybrid workflow; diagnostic accuracy; network meta-analysis.

Contributions of each author
Author 1 - Ting-Wei Wang.
Author 2 - Sheng Yang.
Author 3 - Yun-Chu Wang.
Author 4 - Yi-Fang Tsai.
Author 5 - Yu-Kang Tu.
Author 6 - Yu-Te Wu.
Author 7 - Hong-Jen Chiou.
Author 8 - Ling-Ming Tseng.
Author 9 - Yi-Chen Lai.