INPLASY

Regulatory approval and real-world performance of deep learning systems for breast cancer screening

INPLASY202650006
doi: 10.37766/inplasy2026.5.0006
Received: 1 May 2026
Published: 1 May 2026

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

Corresponding author:
Ting Wei Wang

eltonwang1@gmail.com

Author Affiliation:
Taipei Veteran General Hospital.

ADMINISTRATIVE INFORMATION

Support - NR.

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

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202650006

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 1 May 2026 and was last updated on 1 May 2026.

INTRODUCTION

Review question / Objective This systematic review and meta-analysis aims to evaluate the diagnostic accuracy and real-world performance of regulator-approved deep learning systems for breast cancer screening using digital mammography and digital breast tomosynthesis. The review will quantify pooled stand-alone AI sensitivity and specificity and explore methodological, population-level, imaging-related, regulatory, and implementation-related factors associated with heterogeneity in diagnostic performance.

Condition being studied The condition being studied is breast cancer detected through screening mammography. Breast cancer is the most frequently diagnosed malignancy among women worldwide and remains a major cause of cancer-related mortality. Population-based mammography screening aims to identify invasive breast cancer and ductal carcinoma in situ at an earlier and more treatable stage. This review focuses on breast cancer detection using digital mammography and digital breast tomosynthesis screening pathways.

METHODS

Participant or population The population will include women undergoing breast cancer screening or screening-related mammographic evaluation using digital mammography or digital breast tomosynthesis. Eligible populations may include population-based screening cohorts, opportunistic screening cohorts, clinically enriched screening-related datasets, or real-world implementation cohorts, provided that the AI system is evaluated for breast cancer detection or recall classification.

Intervention The intervention or index test is a regulator-approved deep learning system for breast cancer screening applied to digital mammography or digital breast tomosynthesis. Eligible systems must have regulatory authorization, clearance, certification, or approval,

such as FDA clearance, CE marking, or other national regulatory approval, and must provide stand-alone AI diagnostic performance.

Comparator Comparators may include radiologist interpretation, standard screening pathways, single reading, double reading, human–AI assisted workflows, or alternative AI systems when reported. Studies without a human comparator will also be eligible if they provide sufficient data to evaluate stand-alone AI diagnostic accuracy.

Study designs to be included Prospective and retrospective diagnostic accuracy studies, screening cohort studies, reader studies, paired comparative studies, randomized trials, and real-world implementation studies reporting stand-alone AI diagnostic performance will be included.

Eligibility criteria Eligible studies must evaluate a regulator-approved deep learning system for breast cancer screening using digital mammography or digital breast tomosynthesis. Studies must report stand-alone AI performance as an independent recall/no-recall or cancer/no-cancer classification and use pathology, screening-program follow-up, cancer-registry linkage, or a comparable clinical reference standard. Studies must provide sufficient information to extract or reconstruct true positives, false positives, true negatives, and false negatives. Studies will be excluded if they evaluate non-regulator-approved research prototypes, non-deep-learning systems, non-mammographic imaging, risk-prediction models without a screening-detection endpoint, interval-cancer-only cohorts, conference abstracts or protocols without usable full-text data, studies reporting only AUC without threshold-specific counts, reports lacking sufficient data to reconstruct diagnostic contingency tables, or overlapping cohorts when a more complete or recent report is available.

Information sources Information sources will include PubMed/MEDLINE, Embase, and Web of Science. Additional records will be identified by manually screening reference lists of eligible studies and relevant systematic reviews. Regulatory status may be verified using publicly available regulatory sources, including FDA device databases, CE-marking information, manufacturer documentation, and other national regulatory databases when available.

Main outcome(s) The main outcomes are pooled sensitivity and specificity of regulator-approved deep learning systems for breast cancer detection or recall classification in screening mammography. Outcomes will be calculated from reconstructed 2 × 2 diagnostic contingency tables, including true positives, false positives, true negatives, and false negatives. Digital mammography and digital breast tomosynthesis will be analyzed separately.

Quality assessment / Risk of bias analysis Methodological quality will be assessed using the QUADAS-3 framework for diagnostic test accuracy studies. Two reviewers will independently evaluate risk of bias and applicability concerns across patient selection, index test conduct and interpretation, reference standard appropriateness and independence, and flow and timing. Additional AI-specific concerns, including data leakage, post hoc threshold selection, unclear algorithm lock status, incomplete reporting of model version, and handling of ungradable images, will also be considered. Disagreements will be resolved by consensus or third-reviewer adjudication.

Strategy of data synthesis Diagnostic 2 × 2 tables will be reconstructed for each eligible dataset. Digital mammography and digital breast tomosynthesis datasets will be analyzed separately to avoid conflating modality-specific performance. Pooled sensitivity and specificity will be estimated using hierarchical bivariate random-effects models, which jointly model sensitivity and specificity while accounting for within-study variability, between-study heterogeneity, and their correlation. Summary receiver operating characteristic plots, forest plots, confidence regions, and prediction regions will be generated where appropriate. Meta-regression will be used to explore prespecified sources of heterogeneity. Analyses will be performed using Stata.

Subgroup analysis Subgroup analyses will examine diagnostic performance according to imaging modality, algorithm group, threshold provenance, geographic region, country income status, study design, screening setting, single-vendor versus multi-vendor imaging acquisition, vendor involvement, breast density composition, follow-up duration, and interval-cancer ascertainment. These analyses will be interpreted as exploratory because some subgroups may include limited numbers of datasets.

Sensitivity analysis Sensitivity analyses will assess the robustness of findings by excluding studies at high risk of bias, excluding studies with post hoc threshold selection, excluding studies without adequate follow-up or interval-cancer ascertainment, restricting analyses to peer-reviewed full-text articles, removing clinically enriched datasets, and comparing alternative

model specifications where appropriate. Additional analyses may be performed to assess the influence of overlapping cohorts or influential datasets.

Country(ies) involved Taiwan.

Keywords artificial intelligence; deep learning; breast cancer screening; mammography; digital breast tomosynthesis.

Contributions of each author
Author 1 - Ting Wei Wang.
Author 2 - Yu-Chuan Chen.
Author 3 - Sheng Yang.
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 - Ting-Wei Wang.