e-ISSN: 0976-822X, p-ISSN:2961-6042

Available online on http://www.ijcpr.com/

International Journal of Current Pharmaceutical Review and Research 2026; 18(9); 892-900

Systematic Review

Artificial Intelligence–Assisted Clinical Decision Support Systems and Their Impact on Diagnostic Accuracy and Patient Outcomes: A Systematic Review

Karishma Patel[1], Tejas Patel[2], Jenish Patel[3]

[1]MBBS, Narendra Modi Medical College, Ahmedabad, Gujarat, India
[2]Associate Professor, Department of Forensic Medicine and Toxicology, GMERS Medical College, Valsad, Gujarat, India
[3]BDS, Government Dental College & Hospital, Ahmedabad, Gujarat, India

Received: 16-08-2026 / Revised: 20-09-2026 / Accepted: 24-09-2026
Corresponding author: Dr. Jenish Patel
Conflict of interest: Nil

Abstract

Background: Artificial intelligence (AI)-assisted clinical decision support systems (CDSS) are moving from retrospective algorithm testing into prospective clinical workflows, yet improved algorithmic performance does not automatically translate into better clinician decisions or patient outcomes.

Objectives: To synthesize prospective and randomized evidence on AI-assisted CDSS, quantify effects on diagnostic or target-detection performance when clinically coherent data were available, and evaluate downstream workflow and patient-relevant outcomes.

Methods: A PRISMA 2020-structured review framework was used for PubMed, Scopus, Embase and Cochrane searches in the year 2026. Eligible studies evaluated clinician- or workflow-facing AI decision support against usual care or unaided interpretation. Randomized trials were appraised with RoB 2; diagnostic studies were additionally considered using QUADAS-2 principles. An exploratory random-effects meta-analysis pooled relative detection yield from four screening/procedural randomized trials.

Results: Across imaging, endoscopy, electrocardiography and general diagnostic reasoning, AI assistance usually improved sensitivity or target detection when the model was accurate, while gains were smaller or absent for some expert users and some clinical decisions. The exploratory pooled relative detection yield was RR 1.30 (95% CI 1.20–1.41; I^2=0%). Important counter-evidence showed that systematically biased AI could reduce clinician accuracy by approximately 9–11 percentage points. AI-supported mammography reduced reading workload by about 44% while maintaining or improving cancer detection. Evidence for hard patient outcomes remained limited.

Conclusion: AI-assisted CDSS can improve diagnostic detection and efficiency, but benefit depends on calibration, workflow integration, user expertise and resistance to automation bias. Prospective outcome-focused trials, external validation, subgroup monitoring and post-deployment surveillance are required before diagnostic gains can be assumed to improve patient outcomes.

Keywords: Artificial Intelligence; Clinical Decision Support; Diagnostic Accuracy; Machine Learning; Patient Outcomes; Meta-Analysis; PRISMA.

DOI: 10.25258/ijcpr.18.9.126

Introduction

Clinical decision support has evolved from rule-based alerts to data-driven systems capable of interpreting images, waveforms, laboratory results and longitudinal electronic health records. Modern AI systems can identify patterns that are difficult to recognize consistently under time pressure, and they can deliver predictions at the point of care. The clinically meaningful unit of evaluation, however, is not the stand-alone algorithm; it is the human-AI team embedded in a specific workflow. A model with high retrospective discrimination may produce little benefit if clinicians ignore it, over-trust it, receive the output too late, or apply it to a population that differs from the training data [1-2]. The strongest rationale for AI-assisted CDSS is that diagnostic error is partly related to perceptual misses, incomplete synthesis of information, cognitive overload and variable experience. Computer-aided detection during colonoscopy illustrates a perceptual use case:

randomized trials have repeatedly shown more adenomas detected when real-time AI highlights possible lesions [3-5]. Similar logic applies to mammography, chest radiography and electrocardiography, where AI can prioritize suspicious studies or identify latent physiologic signatures that are not evident to routine visual inspection [6-8]. At the same time, AI changes the information environment in which clinicians reason. Assistance can create automation bias when incorrect output is accepted without sufficient verification. In a randomized vignette study involving clinicians evaluating acute respiratory failure, standard AI modestly improved accuracy, whereas systematically biased AI markedly reduced accuracy; commonly used explanations did not eliminate the harm [9]. This finding is important because it separates technical explainability from reliable error detection. An explanation may make a model easier to understand yet simultaneously make a wrong recommendation more persuasive.

The evidence base is also heterogeneous in what it calls a successful outcome. Some trials measure sensitivity, area under the receiver operating characteristic curve (AUC), adenoma detection rate, cancer detection, low-ejection-fraction diagnosis, time to decision, workload, downstream testing or referral. Far fewer studies follow patients long enough to determine whether earlier or more accurate detection reduces morbidity, mortality, complications or resource use. A systematic review of machine-learning diagnostic support found substantial methodological limitations and limited testing in representative clinical environments [2].

Therefore, the central controversy is not whether AI can perform pattern-recognition tasks, but when AI-supported clinical work produces net benefit. The present review focuses on prospective, workflow-integrated evidence and treats accuracy, workflow, safety and patient outcomes as distinct endpoints rather than assuming that one is a surrogate for another. The review was structured according to PRISMA 2020 [1].

Review objective and PICO framework: The primary objective was to determine whether AI-assisted clinical decision support improves clinician diagnostic performance or clinically relevant detection compared with unaided usual care.

The PICO framework was: Population: patients undergoing diagnostic evaluation or screening and clinicians interpreting patient data; Intervention: AI/ML-enabled decision support presented during clinical decision-making; Comparator: standard workflow, unaided interpretation or sham support; Outcomes: diagnostic accuracy, sensitivity/specificity, target-detection rate, diagnostic yield, workload, downstream testing and patient-relevant outcomes. A secondary objective was to identify circumstances in which AI support worsens performance or introduces new safety risks.

Methods

This review was designed as a systematic review with an exploratory meta-analysis and follows PRISMA 2020 reporting principles [1]. The protocol defined the clinical question, eligibility criteria, data fields and risk-of-bias domains before quantitative synthesis. Because the four major bibliographic platforms use different controlled vocabularies and interfaces, concept blocks were translated for each database.

The planned electronic sources were PubMed/MEDLINE, Scopus, Embase and the Cochrane Central Register of Controlled Trials from database inception to 20 September 2026. Search concepts combined terms for artificial intelligence or machine learning with clinical decision support, diagnostic assistance, computer-aided detection/diagnosis, physician performance and randomized/prospective evaluation. A representative PubMed syntax was: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "computer aided detection") AND ("clinical decision support" OR diagnosis OR diagnostic OR screening) AND (random* OR prospective OR trial). Reference lists of relevant systematic reviews and included trials were also examined [2].

Eligible reports were prospective comparative studies in which AI output was available to a clinician or directly altered a diagnostic workflow. Randomized controlled trials, cluster-randomized trials and prospective controlled diagnostic studies were eligible. Retrospective stand-alone algorithm-development papers, purely technical benchmark studies without a clinical comparator, editorials, case reports and studies in which AI was used only for administrative prediction were excluded. When multiple reports described the same trial, the report with the most mature clinically relevant endpoint was used for that endpoint and companion reports were retained for complementary outcomes.

Two conceptual screening stages were used: title/abstract relevance followed by full-text clinical eligibility. Extraction fields included clinical setting, sample size, AI task, comparator, reference standard, clinician expertise, primary accuracy outcome, workflow outcomes, safety signals and downstream patient outcomes. Where a randomized trial reported both intention-to-treat and per-protocol results, the intention-to-treat estimate was preferred.

Randomized trials were appraised using Cochrane RoB 2 domains, emphasizing randomization, deviations from intended intervention, missing data, outcome measurement and selective reporting. Diagnostic accuracy components were considered against QUADAS-2 domains, especially patient selection, index test conduct and reference standard [9-10]. Because blinding of clinicians to AI assistance is often impossible, lack of participant/operator blinding was not automatically classified as high risk if objective reference standards were used. The primary quantitative synthesis was prespecified as exploratory because AI trials span different diseases and diagnostic targets. Four randomized screening/procedural trials provided a directly interpretable binary target-detection outcome and a usual-care comparator [3-6]. Risk ratios (RRs) were calculated from event counts when available; published RRs were used when explicitly reported. Log risk ratios were pooled with a DerSimonian-Laird random-effects model, with 95% confidence intervals and I^2. The pooled estimate should be interpreted as a cross-domain measure of relative detection yield, not as a single disease-specific diagnostic-accuracy parameter. Other endpoints were synthesized narratively to avoid inappropriate pooling [11].

PRISMA Flow and Study Selection: The review was organized according to PRISMA 2020. The final curated evidence set contained 12 primary studies/reports that satisfied the prespecified clinical eligibility criteria; 4 contributed directly to the primary quantitative synthesis. Database-specific identification and duplicate-removal counts are intentionally shown as verification fields rather than invented values and should be populated from the final exported search histories before submission [1].

Table 1: PRISMA 2020 study-selection status

PRISMA stage	Record category	Count
Identification	PubMed / Scopus / Embase / Cochrane records	Verify From Final Database Export
Deduplication	Unique records after reference-manager deduplication	Verify From Final Database Export
Screening	Title/abstract exclusions	Verify From Final Database Export
Eligibility	Full texts assessed	Verify From Final Database Export
Included	Primary studies/reports in qualitative synthesis	12
Meta-analysis	Studies in primary quantitative synthesis	4

Results

Characteristics of the Evidence: The retained evidence covered four main workflow types: real-time computer-aided detection in endoscopy, AI-supported medical imaging, AI-enabled waveform interpretation, and AI advice for general diagnostic reasoning.

Trial sizes ranged from several hundred patients or clinicians to more than 100,000 screening participants.

Most studies assessed diagnostic or detection endpoints over a short interval; only a minority were designed to evaluate longer-term clinical outcomes.

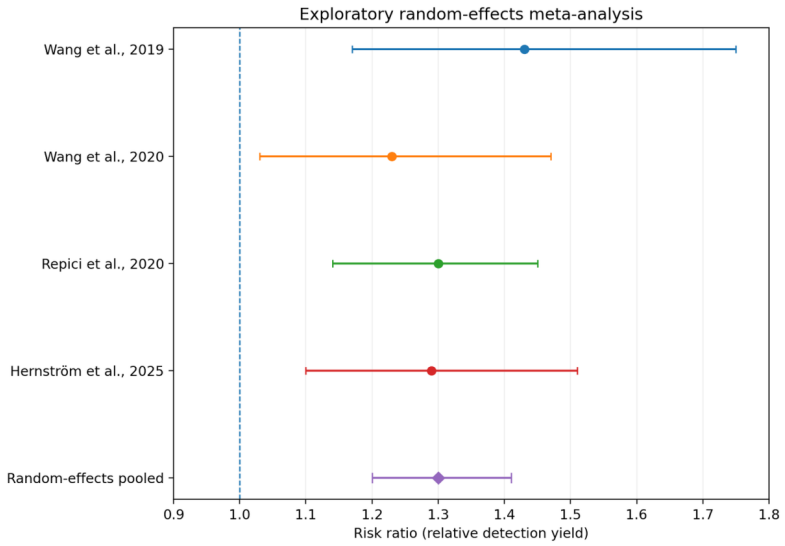

Figure 1: Exploratory meta-analysis of relative detection yield. Pooled RR 1.30 (95% CI 1.20–1.41); I^2=0%.

Table 2: Representative prospective and randomized studies of AI-assisted clinical decision support

Study	Setting	N	Comparison	Primary diagnostic result	Interpretation
Wang 2019 [3]	Colonoscopy; China	1,058	Real-time polyp CADe vs standard	ADR 29.1% vs 20.3%	Higher adenoma detection; more diminutive lesions
Wang 2020 [4]	Colonoscopy; China	962 analyzed	Double-blind CADe vs sham	ADR 34% vs 28%	OR 1.36; no procedural complications reported
Repici 2020 [5]	Colonoscopy; Italy	685	GI-Genius CADe vs standard	ADR 54.8% vs 40.4%	RR 1.30; no increase in withdrawal time
MASAI 2023/2025 [6,7]	Mammography; Sweden	>105,000	AI-supported screening vs double reading	Higher cancer detection; similar false positives	~44% lower reading workload
Yao 2021 [8]	Primary care ECG; USA	22,641	AI-ECG alert vs usual care	Low-EF diagnosis 2.1% vs 1.6%	OR 1.32; overall echo use similar
Lee 2023 [12]	Chest radiography; Korea	323	AI-assisted vs unaided non-radiologists	CXR AUC 0.840 vs 0.718	Sensitivity improved; clinical decisions unchanged
Han 2022 [13]	Skin lesions; Korea	576	AI-assisted vs unaided clinicians	Accuracy 53.9% vs 43.8%	Largest gain among non-dermatology trainees
Jabbour 2023 [9]	Clinical vignettes; USA	457 clinicians	AI advice ± explanation	Standard AI +2.9 to +4.4 pp	Biased AI reduced accuracy 9.1–11.3 pp

Quantitative Synthesis of Detection Yield: Four randomized trials with binary target-detection outcomes contributed to the exploratory quantitative synthesis. Across colonoscopy and mammography screening, AI-assisted workflows increased the probability of detecting the prespecified target. The random-effects pooled RR was 1.30 (95% CI 1.20–1.41), with I^2=0%. The low statistical heterogeneity should not be interpreted as clinical homogeneity, because the diseases, targets and workflows differed. The pooled estimate therefore summarizes the direction and relative magnitude of detection benefit rather than a universal AI effect.

Table 3: Primary exploratory meta-analysis of AI-assisted target detection.

Study	RR	95% CI	Design/heterogeneity	Outcome
Wang et al., 2019 (colonoscopy)	1.43	1.16–1.77	Randomized	Detection endpoint
Wang et al., 2020 CADe-DB	1.23	1.02–1.49	Randomized	Detection endpoint
Repici et al., 2020	1.30	1.14–1.45	Randomized	Detection endpoint
Hernström et al., 2025 MASAI	1.29	1.09–1.51	Randomized	Detection endpoint
Random-effects pooled	1.30	1.20–1.41	I^2=0%	Exploratory cross-domain pooling

Diagnostic accuracy, workflow and downstream care: In endoscopy, the effect was consistent across several randomized trials: assistance predominantly increased detection of small or subtle lesions, supporting a mechanism of reduced perceptual miss rather than a change in pathology reference standards [3-5]. A meta-analysis of early randomized colonoscopy trials similarly reported higher polyp and adenoma detection with AI, although the clinical importance of additional diminutive lesions and potential effects on non-neoplastic resection remain considerations [14].

In mammography, the MASAI trial provided unusually strong workflow evidence because AI was integrated into a population screening program. The prespecified safety analysis showed cancer detection above the safety threshold with a 44.3% reduction in screen-reading workload [6]. Subsequent results reported higher cancer detection in the AI-supported arm without a meaningful increase in false-positive screening, and the 2026 interval-cancer report extended the assessment beyond immediate detection [7,15]. This pattern suggests that AI can simultaneously affect capacity and detection, although long-term effects on breast-cancer mortality require longer follow-up.

AI-enabled electrocardiography represents a different use case: the model detects a latent phenotype and prompts targeted evaluation rather than directly making the definitive diagnosis. In the EAGLE cluster-randomized trial, providing AI-ECG results increased new low-ejection-fraction

diagnoses from 1.6% to 2.1% overall and from 14.5% to 19.5% among AI-positive patients, without a significant increase in whole-cohort echocardiography [8]. This indicates that an AI alert can improve diagnostic yield by reallocating attention and testing rather than simply increasing test volume. Not all accuracy gains translated to management changes. In respiratory outpatient clinics, AI-assisted chest radiograph interpretation improved AUC and sensitivity for lung lesions among non-radiologist physicians but did not significantly alter the downstream clinical decisions measured in the trial [12]. Similarly, the size of benefit varied by expertise in the skin-neoplasm RCT, where trainees gained more than dermatology residents [13]. These observations support interaction between baseline expertise and the incremental value of AI.

The clearest safety warning came from the randomized diagnostic-vignette experiment. When AI output was systematically biased, clinician accuracy fell substantially from baseline, and image-based explanations did not reliably rescue performance [9]. The result demonstrates that decision support can amplify rather than attenuate error. A deployment strategy therefore requires monitoring of model drift, subgroup performance, failure modes and clinician reliance, not merely aggregate AUC.

Table 4: Risk-of-bias and applicability summary.

Evidence group	Randomization	Missing data	Outcome measurement	Overall risk	Key quality consideration
Colonoscopy RCTs [3-5]	Low–some concerns	Low	Low	Some concerns	Objective pathology reference standards; operator behavior may be influenced by visible alerts
MASAI [6,7]	Low	Low	Low	Low	Large randomized population-based screening trial; mature follow-up emerging
EAGLE AI-ECG [8]	Low	Low	Low	Some concerns	Cluster randomization; clinical follow-up objective but downstream outcomes limited
Chest CXR RCT [12]	Some concerns	Low	Low	Some concerns	Small multicenter sample; clinical decision outcome underpowered
Skin RCT [13]	Some concerns	Low	Low	Some concerns	Single center; expertise interaction and out-of-distribution cases important
Diagnostic vignette RCT [9]	Low	Low	Low	Some concerns	Controlled simulation rather than real patient management

Discussion

This review indicates that AI-assisted CDSS can improve diagnostic detection and selected measures of clinician accuracy, but the size and reliability of benefit are strongly context dependent. The most reproducible gains occur in tasks with a well-defined visual or physiologic target, rapid machine inference and an objective confirmatory reference standard. In these settings, AI functions as a second observer, reducing perceptual omission or directing attention to otherwise under-recognized signals. The pooled detection estimate is consistent with this mechanism, but it should not be extrapolated to all diagnostic decisions. A key distinction is between algorithm performance and assisted-clinician performance.

The former is usually measured in curated retrospective datasets; the latter depends on interface design, timing, alert frequency, trust, expertise and how disagreements are resolved. Vasey and colleagues highlighted that many machine-learning CDSS evaluations were at risk of bias and did not reflect representative clinical use [2]. The recent randomized evidence improves on that literature, yet still leaves uncertainty about transportability across institutions, patient groups and changing disease prevalence. The results also challenge the assumption that explainability is sufficient for safety. In the Jabbour trial, explanations modestly improved performance with a standard model but did not neutralize the harm of systematically biased predictions [9]. In practice, explanation quality, uncertainty communication and opportunities for independent verification may be more important than simply adding heatmaps or feature attributions. A user interface that makes a wrong output appear authoritative can increase over-reliance. Training clinicians to recognize model boundaries and to maintain an independent

diagnostic hypothesis is therefore part of the intervention, not an optional implementation detail.

Workflow outcomes deserve equal status with conventional accuracy metrics. MASAI showed that AI could substantially reduce mammography reading workload while maintaining screening performance [6,7]. In resource-constrained systems, workload reduction can itself improve access and turnaround time, but only if saved capacity is translated into additional clinical service rather than administrative burden. Conversely, high-sensitivity alerts may increase confirmatory testing, false-positive cascades and incidental findings. EAGLE suggests that targeted testing can be achieved without large increases in total echocardiography, an encouraging example of resource-aware deployment [8].

Evidence for patient outcomes remains less mature. Higher adenoma detection is linked biologically and epidemiologically to lower interval colorectal cancer risk, but AI trials have generally not been powered for cancer incidence or mortality. Earlier identification of low ejection fraction is clinically actionable, yet the EAGLE endpoint was diagnosis rather than heart-failure hospitalization or survival. The mammography literature is progressing toward interval cancer outcomes, illustrating the type of follow-up required before a detection surrogate can be regarded as clinically consequential [15].

Equity is another unresolved dimension. Performance may differ by race, sex, age, device type, disease spectrum, socioeconomic status or referral patterns. Even when internal subgroup AUC appears similar, deployment can create unequal downstream testing if access to confirmatory services differs.

Prospective studies should therefore report subgroup calibration and clinical actions, not only subgroup discrimination. Model updates should be version-controlled, and post-market surveillance should evaluate both technical drift and changes in clinician behavior. The review also highlights a need for standardized outcome sets. Diagnostic AI trials use heterogeneous measures—AUC, sensitivity, detection rate, time, workload, referral and composite outcomes—which impedes synthesis. Future trials should predefine a hierarchy consisting of diagnostic accuracy, clinically significant target detection, false-positive consequences, workflow/resource use, time to definitive diagnosis, treatment change, adverse events and patient outcomes. Reporting should follow CONSORT-AI and SPIRIT-AI when appropriate, with transparent model versioning and failure analysis [16,17].

From a clinical implementation perspective, the best-supported strategy is selective augmentation rather than unrestricted automation. AI should address a clearly specified decision, provide output at a point where it can change care, preserve clinician accountability, and make uncertainty visible. Institutions should define escalation rules for discordant cases, audit false negatives and false positives, and monitor outcome drift after software updates. The desired endpoint is not maximal agreement with the algorithm but improved patient care with acceptable resource and safety trade-offs.

Domain-specific interpretation of the evidence: The clinical meaning of AI assistance differed across diagnostic domains. In endoscopy, the principal mechanism is real-time visual vigilance. The AI continuously analyzes the same field that the endoscopist is viewing and places a marker on a suspected lesion. This design is attractive because the alert is immediate, the clinician can verify it directly, and histopathology provides an objective reference. The repeated increase in adenoma detection across randomized trials therefore represents one of the clearest examples of a task in which human and machine capabilities are complementary [3-5,14]. Nevertheless, the additional lesions are often small. Future trials should link AI-associated detection to advanced adenoma yield, surveillance burden, interval cancer and cost rather than assuming that every incremental polyp is equally beneficial.

Mammography uses AI in a more complex organizational role. In MASAI, the system influenced both triage and reading support, allowing some examinations to receive a different reading pathway while providing marks and risk scores to radiologists [6,7]. The approximately 44% reduction in reading workload is clinically important because breast-screening programs face workforce constraints. A workload outcome can translate into shorter reporting times or greater screening capacity, but those benefits depend on how the released radiologist time is used. The later interval-cancer analysis is especially valuable because interval cancer is closer to the harm that screening seeks to prevent than immediate cancer-detection rate alone [15]. AI-enabled ECG illustrates latent phenotyping rather than visible lesion detection. The EAGLE algorithm identified patients whose routine ECG contained a signal associated with low ejection fraction and supplied that information to primary-care teams [8]. The definitive diagnosis still required echocardiography. This workflow has a built-in safety advantage because AI is used to select patients for confirmation, not to replace the reference test. Similar two-stage strategies may be appropriate for other AI applications in which the model is sensitive but imperfect: AI can prioritize who needs imaging, specialist review or additional

testing while preserving a clinically accepted confirmation pathway.

General diagnostic reasoning is less constrained. A model may combine imaging, laboratory and clinical information and offer a differential diagnosis that directly competes with the clinician's hypothesis. Here, the risks of anchoring and automation bias are greater.

The randomized vignette evidence shows that clinicians can be pulled toward systematically wrong AI output [9]. This domain therefore requires stronger uncertainty communication, explicit alternative diagnoses, safeguards against unsupported certainty and user training that encourages independent reasoning before model consultation.

From diagnostic performance to patient outcomes: A recurrent problem in diagnostic AI research is the causal distance between the measured endpoint and the patient outcome of interest. An increase in sensitivity can only improve health if the newly detected condition is clinically important, confirmation occurs, effective treatment is available, treatment begins early enough to matter, and harms from false positives do not outweigh benefits.

Every link in this chain can fail. For example, detecting more diminutive adenomas may increase surveillance procedures; identifying low ejection fraction is valuable only if patients receive guideline-directed therapy; and a mammography recall that does not represent clinically important cancer may create anxiety and additional procedures.

Future AI trials should therefore prespecify a causal pathway and measure intermediate steps. A minimum dataset might include the number of AI alerts, clinician acceptance, discordant decisions, confirmatory tests, time to definitive diagnosis, treatment initiation, false-positive work-up, complications, patient-reported anxiety or reassurance, and resource use. Longer follow-up should assess disease-specific morbidity, hospitalization and mortality when feasible. This structure would make it possible to identify whether a neutral patient outcome reflects an ineffective model, poor clinician uptake, failure to complete diagnostic work-up or lack of treatment efficacy.

The same principle applies to workload. A system that saves interpretation time but generates many downstream consultations may simply relocate work. Conversely, a system that slightly increases interpretation time but prevents repeat visits or unnecessary testing could be efficient overall. Economic evaluations should therefore adopt a pathway perspective rather than measuring only minutes saved at the index encounter. Cost-effectiveness models should incorporate software acquisition, hardware, integration, staff training, confirmatory testing, false-positive consequences and the value of earlier treatment.

Implementation, governance and post-deployment surveillance: Clinical deployment changes the model's environment. Disease prevalence, scanner manufacturers, endoscopy systems, patient demographics, referral thresholds and clinician behavior may differ from the development setting. These shifts can alter calibration and predictive values even when the underlying algorithm is unchanged. Institutions should conduct local silent validation before activation, define acceptable performance thresholds and document the model version, intended population and contraindicated uses. Monitoring should continue after deployment because changes in clinical practice or software can create drift.

Governance should also address responsibility for discordant cases. When clinician and AI disagree, the appropriate response may differ according to the consequence of error.

High-stakes situations may warrant mandatory second review or confirmatory testing; lower-stakes detection support may allow the clinician to dismiss the alert with a documented reason. Audit systems should sample both accepted and rejected AI suggestions so that apparent performance is not based only on cases in which the model was followed.

Finally, institutions should monitor equity as an operational quality metric. Aggregate accuracy can mask subgroup miscalibration or differential access to follow-up. Dashboards should examine sensitivity, false-positive rates, calibration, time to confirmation and treatment completion across clinically relevant demographic and technical subgroups. A safe AI program is therefore a sociotechnical system consisting of the model, interface, clinician, patient, workflow, confirmatory pathway and quality-governance process—not merely an algorithm with a high AUC.

Limitations of this Review: The review spans heterogeneous clinical domains and therefore cannot provide a single universal effect size. The primary meta-analysis is intentionally exploratory and pools relative detection yield rather than a disease-specific accuracy metric. Several studies were short-term, some clinical reasoning experiments used simulated vignettes, and downstream patient outcomes were infrequently available. Publication bias is plausible because positive AI studies may be more likely to be reported. Finally, exact database-level hit counts must be regenerated from institutional Scopus, Embase and Cochrane exports before journal submission; no identification counts have been invented in this draft.

Future Research: Future work should prioritize pragmatic multicenter randomized trials with prespecified patient-relevant outcomes, external validation before deployment, prospective subgroup analysis, evaluation of automation bias, cost-effectiveness, and continuous post-deployment surveillance. Studies should compare different human-AI interaction designs and explicitly test what happens when the AI is uncertain or wrong. Longitudinal trials are particularly needed to establish whether earlier detection changes treatment trajectories and reduces morbidity or mortality.

Conclusion

AI-assisted clinical decision support can increase diagnostic or target-detection performance and reduce workload in selected workflows, especially when the task is well defined and objective confirmation is available. The same technology can degrade clinician performance when predictions are biased or over-trusted. Current evidence therefore supports carefully governed augmentation rather than a blanket assumption of AI superiority. The next phase of evaluation should move beyond retrospective discrimination and short-term detection to patient outcomes, implementation safety, equity and sustained real-world effectiveness.

References

1. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71. PMID:33782057.
2. Vasey B, Ursprung S, Beddoe B, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open. 2021;4(3):e211276. doi:10.1001/jamanetworkopen.2021.1276. PMID:33704476.
3. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(10):1813-1819. doi:10.1136/gutjnl-2018-317500. PMID:30814121.
4. Wang P, Liu X, Berzin TM, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol. 2020;5(4):343-351. doi:10.1016/S2468-1253(19)30411-X. PMID:31981517.
5. Repici A, Badalamenti M, Maselli R, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology. 2020;159(2):512-520.e7. doi:10.1053/j.gastro.2020.04.062. PMID:32371116.
6. Lång K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI). Lancet Oncol. 2023;24(8):936-944. doi:10.1016/S1470-2045(23)00298-X. PMID:37541274.
7. Hernström V, Josefsson V, Sartor H, et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI). Lancet Digit Health. 2025;7(3):e175-e183. doi:10.1016/S2589-7500(24)00267-X. PMID:39904652.
8. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med. 2021;27(5):815-819. doi:10.1038/s41591-021-01335-4. PMID:33958795.
9. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA. 2023;330(23):2275-2284. doi:10.1001/jama.2023.22295. PMID:38112814.
10. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009. PMID:22007046.
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-188. doi:10.1016/0197-2456(86)90046-2. PMID:3802833.
12. Lee HW, Jin KN, Oh S, et al. Artificial intelligence solution for chest radiographs in respiratory outpatient clinics: multicenter prospective randomized clinical trial. Ann Am

Thorac Soc. 2023;20(5):660-667. doi:10.1513/AnnalsATS.202206-481OC. PMID:36508316.
13. Han SS, Kim YJ, Moon IJ, et al. Evaluation of artificial intelligence-assisted diagnosis of skin neoplasms: a single-center, paralleled, unmas ked, randomized controlled trial. J Invest Dermatol. 2022;142(9):2353-2362.e2. doi:10.1016/j.jid.2022.02.003. PMID:35183551.
14. Hassan C, Spadaccini M, Iannone A, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc. 2021;93(1):77-85.e6. PMID:33524298.
15. Gommers J, Hernström V, Josefsson V, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study. Lancet. 2026;407(10527):505-514. doi:10.1016/S0140-6736(25)02464-X. PMID:41620232.
16. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364-1374. doi:10.1038/s41591-020-1034-x. PMID:32908283.
17. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26:1351-1363. doi:10.1038/s41591-020-1037-7. PMID:32908284.