



Human oversight of erroneous clinical AI and decision support a systematic review and meta analysis



Di Zeng, Jianyong Lei & Zhihui Li

Artificial intelligence (AI) increasingly supports clinical decisions, but human oversight is protective only if clinicians recognize and resist erroneous output. We systematically reviewed controlled and real-world studies of clinician responses to incorrect automated advice. Forty-six reports represented 41 independent cohort clusters, including 40 core cohorts. Across three adjusted cohorts, clinician correctness was substantially higher with correct than incorrect automated advice (pooled OR 10.19, 95% CI 7.43–13.98; $I^2 = 0\%$); a ≥ 2020 sensitivity analysis was similar but imprecise (OR 9.22, 95% CI 2.06–41.13). In four heterogeneous real-world computerized ECG studies, 26.6% of erroneous interpretations were not corrected or recognized (95% CI 6.4–65.6%; $I^2 = 95.8\%$). Contemporary AI and large language model studies also documented harmful reliance but were too heterogeneous for pooling. Patient-level adverse events were reported in one retrospective cohort. Human oversight should be evaluated as an error-interception capability rather than assumed to ensure safety.

Artificial intelligence is increasingly evaluated and deployed as a clinical collaborator rather than a standalone diagnostic competitor. The intended advantage is complementary performance: algorithms can process complex data streams consistently, while clinicians contribute context, examination findings, patient preferences, and accountability^{1,2}. Yet comparisons between an algorithm and clinicians in isolation do not establish whether the combined human–AI system will be safe in practice. Earlier evaluations of medical AI were often retrospective, used small clinician comparator groups, and made claims that exceeded the strength of the study design³.

Recent reviews have broadened the evidence base for human–AI collaboration. A meta-analysis of clinician–large language model collaboration found heterogeneous and context-dependent effects, while two large scoping reviews showed that empirical studies remain concentrated in short-term diagnostic interpretation and rarely measure patient or system outcomes^{4–6}. These reviews answer whether assistance improves performance on average. Clinical safety, however, depends on a more conditional question: what happens when the automated recommendation is wrong?

Automation bias describes inappropriate reliance on automated cues. It can produce commission errors when a user follows an incorrect recommendation and omission errors when a false-negative or absent alert suppresses an action that should have occurred⁷. Verification complexity is central because human oversight is meaningful only when the reviewer can independently check the basis of the recommendation within the available time, information, and expertise⁸. A clinician may formally retain final

authority yet still function as a weak safety barrier when the automated output is plausible, fluent, difficult to verify, or embedded in a high-throughput workflow⁹.

Governance frameworks emphasize human autonomy, transparency, independent review, monitoring, and lifecycle risk management^{10–13}. Reporting standards for clinical AI similarly call for explicit description of human–computer interaction, failure modes, workflow, and post-deployment monitoring^{14–16}. Nevertheless, the empirical evidence on whether clinicians detect, reject, or propagate incorrect automated advice has not been synthesized around a unified clinical safety model that connects controlled behavioral effects, real-world non-correction, direction-specific decision transitions, and downstream patient consequences.

We therefore conducted a systematic review and meta-analysis of erroneous AI and decision-support output in clinician decision-making. The primary comparative outcome was clinician correctness under correct versus incorrect automated advice. A co-primary implementation outcome was failure to correct or recognize erroneous output in real-world care. We separately evaluated correct-to-wrong switching, wrong-to-correct switching, erroneous-advice adoption, specialty and workflow moderators, mitigation strategies, and documented clinical consequences. Our objective was not to estimate a universal automation-bias rate, but to determine when human oversight succeeds, when it fails, and how those failures should be measured in future clinical AI evaluation.

Results

Study selection and evidence architecture

The original review identified 160 records, of which 10 were removed as duplicates and 150 underwent title/abstract screening. Seventy-one were excluded at this stage, leaving 79 reports sought for retrieval. Sixteen could not be retrieved, and 63 full-text reports were assessed for eligibility; 25 were excluded and 38 were included. During revision, 11 additional reviewer-suggested or citation-linked reports were assessed, three were excluded, and eight were included. These eight reports represented seven new core cohorts because the Nagendran and Fester reports described linked analyses of the same high-fidelity ICU simulation cohort. The revised evidence base therefore comprised 46 eligible reports mapped to 41 independent cohort clusters, including 40 core cohorts and one adjacent mitigation cohort. Figure 1 summarizes the revised evidence architecture, Supplementary Fig. 2 presents the PRISMA 2020 flow, and Table 1 lists the 40 core cohorts. Report-to-cohort linkage and overlap decisions are detailed in Supplementary Table 1.

The technologies ranged from rules-based decision support and computerized ECG interpretation to convolutional neural networks, image-classification systems, selective prediction, AI-generated draft reports, and conversational or multimodal large language models. Among the 40 core cohorts, 10 involved ECG interpretation and 26 were published from 2020 onward; 35 cohorts were controlled or experimental and five were real-world observational. Earlier studies were disproportionately concentrated in computerized ECG interpretation and conventional clinical decision support, whereas the contemporary evidence increasingly included image-based AI, selective prediction and abstention, AI-generated draft reports, and conversational or multimodal large language models. Importantly, the real-world implementation evidence contributing to the pooled non-correction estimate was confined to deployed computerized ECG interpretation, whereas contemporary AI and LLM studies were predominantly conducted in controlled or simulated settings.

The evidence base contained three distinct forms of error provenance: deliberately manipulated erroneous advice, naturally occurring model errors evaluated under controlled conditions, and naturally occurring errors observed in deployed systems. Provenance was retained as a study-level

characteristic throughout the synthesis. Formal subgroup pooling by provenance was not performed because the available comparable studies within individual strata were too few to yield meaningful pooled estimates. The real-world non-correction synthesis was more homogeneous in this respect: all four contributing studies evaluated naturally occurring errors in deployed computerized ECG systems.

Primary meta-analysis: correct versus incorrect automated advice

Correct automated advice was associated with higher odds of a correct clinician decision in all three cohorts, with a pooled OR of 10.19 (95% CI 7.43–13.98; $I^2 = 0\%$; Fig. 2)^{17–21}. The corresponding absolute risk reductions were 40.1 percentage points for radiologists and 37.5 percentage points for internal/emergency medicine physicians in Gaube et al., 36.5 percentage points in Kücking et al., and 43.2 and 59.0 percentage points for cardiology and non-cardiology physicians, respectively, in the Novotny/Bond cohort. Because baseline risks and study structures differed across cohorts, these study-specific ARRs were not pooled.

The three primary cohorts differed in their handling of repeated observations. Gaube et al. accounted for repeated clinician-level observations using multilevel modeling, while Kücking et al. used a generalized linear mixed-effects model with a participant random intercept. The published Novotny/Bond logistic model did not clearly document adjustment for repeated physician and case observations, raising the possibility that its reported confidence interval was narrower than would be expected under full clustering adjustment. We therefore retained the published point estimate but interpreted its precision cautiously. The three cohorts were combined because each estimated the same within-study contrast in clinician correctness under correct versus incorrect automated advice; the pooled estimate summarizes the association across eligible contexts and should not be interpreted as demonstrating equivalence between experimentally manipulated and naturally occurring errors. Excluding the Novotny/Bond cohort and restricting the synthesis to the two studies published from 2020 onward yielded a similar pooled effect (OR 9.22, 95% CI 2.06–41.13; $k = 2$), although substantially less precise. Study-level repeated-measures structures, clustering approaches, adjusted estimates, and

Evidence architecture and conditional safety pathway

The evidence base separates susceptibility to erroneous advice from implementation-level error interception.

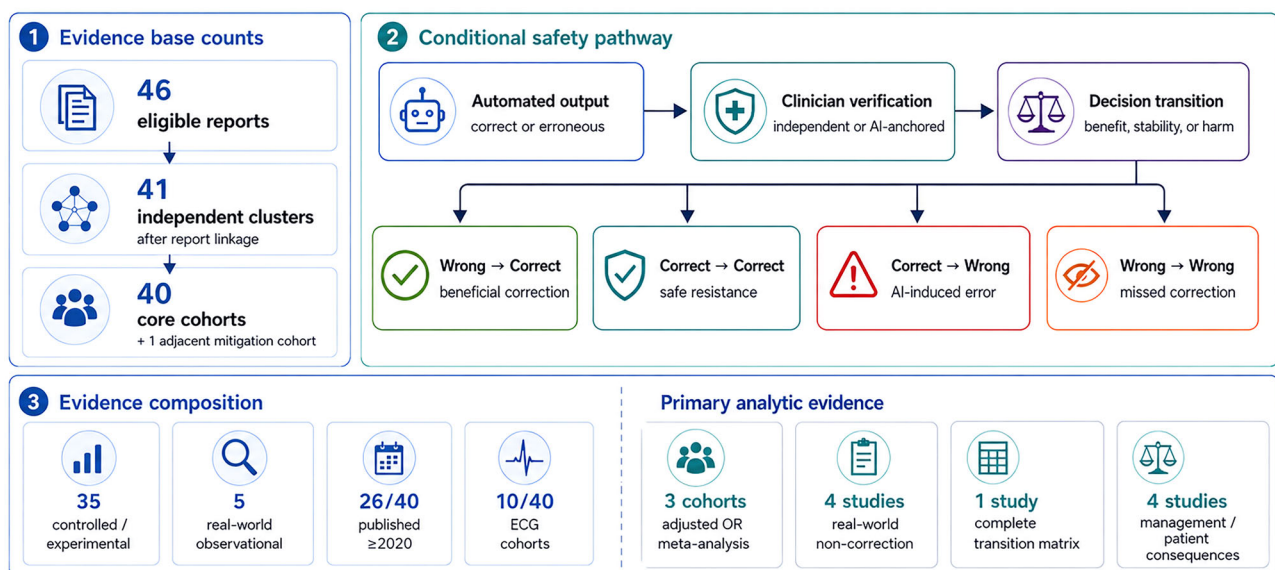


Fig. 1 | Evidence architecture and conditional safety pathway of clinical automation. The figure summarizes the revised evidence base (46 eligible reports, 41 independent cohort clusters, 40 core cohorts and one adjacent mitigation cohort),

the distribution of core cohorts by study design, publication era, and ECG focus, the conditional pathway from automated output through clinician verification to decision transition, and the evidence contributing to the principal analyses.

Table 1 | Characteristics of the 40 independent core cohorts

Study	Year	Clinical domain	Participants/records	Clinical task/material	AI or decision-support system	Design	Erroneous-output exposure	Primary safety outcome
Hillson et al. ²²	1995	ECG interpretation/primary care	40 family physicians and general internists	Ten clinical vignettes with 12-lead ECGs	Computer-assisted ECG interpretation report	Randomized vignette experiment	Computer reports included correct and erroneous diagnoses	Diagnostic agreement, accuracy, and reading time
Morisbak and Gjesdal ²³	1999	ECG interpretation/first-line care	20 first-line physicians	ECG cases with correct or incorrect computerized rhythm diagnoses	Computer-based ECG rhythm interpretation	Controlled physician experiment	Correct and intentionally incorrect computerized interpretations	Accuracy and influence of incorrect interpretations
Goodacre et al. ²⁵	2001	Emergency ECG interpretation	10 accident-and-emergency senior house officers	Fifty ECGs per physician; 25 with and 25 without a computer report	Computer-generated ECG report	Randomized within-reader comparison	Reports could contain major or minor interpretation errors	Major interpretation error and completely correct interpretation
Tsai et al. ²⁴	2003	ECG interpretation/internal medicine	30 internal-medicine residents	ECG interpretation tasks; 1620 subject-finding observations	Computerized diagnostic findings	Controlled crossover experiment	Incorrect computerized findings shown or withheld	Agreement with incorrect findings and diagnostic accuracy
Dreisetti and Binder ⁴⁹	2005	Dermatology decision support	52 dermatologists	Twenty-five pigmented skin lesions per dermatologist	Computerized malignancy estimate and recommendation	Paired pre/post decision-support experiment	CDSS recommendation contradicted the physician's initial decision	Complete wrong-to-correct and correct-to-wrong transition matrix
Anh et al. ⁴⁵	2006	Real-world ECG overreading	2298 ECGs from 1085 patients; 382 false-AF calls with primary-reader response	Routine ECG interpretation and cardiologist overreading	Commercial computerized ECG interpretation	Retrospective expert-adjudicated cohort	Computer falsely labeled ECGs as atrial fibrillation	Erroneous-output non-correction and specialty differences
Southern and Arnsten ⁵²	2009	ECG interpretation/management	105 internal-medicine residents	ECG vignette involving suspected acute ischemia	Computerized ECG interpretation	Randomized survey experiment	Erroneous interpretation suggested acute ischemia	Diagnostic interpretation and urgent-revascularization recommendation
Bae et al. ⁴⁶	2012	Real-world ECG interpretation	10,279 ECGs; 217 AF/AFL-related computer errors	Routine ECG interpretation and clinical follow-up	Computerized ECG analysis	Retrospective expert-adjudicated cohort	Ninety-eight overcalls and 119 missed or misinterpreted AF/AFL cases	Non-correction and downstream treatment consequences
Goddard et al. ³²	2014	Prescribing	26 general practitioners	20 prescribing scenarios (520 decisions)	Simulated prescribing CDSS	Within-participant pre/post experiment	Incorrect advice in 6 scenarios	Correct-to-incorrect switching (negative consultation)
Bouaud et al. ⁵⁰	2015	Breast-cancer multidisciplinary management	394 decisions across three hospitals	Guideline-based breast-cancer management decisions	OncoDoc2 guideline-based CDSS	Prospective decision audit	Eighty-four decisions received entirely non-guideline advice after erroneous input	Conditional wrong-advice adoption
Tsai et al. ³³	2016	Imaging-order safety	66 physicians	3802 contrast-enhanced imaging orders	Naive renal-risk alert CDSS	Cluster randomized trial	1,560 false-positive alerts	False-alert adoption and inappropriate order cancellation
Lyell et al. ³⁴	2017	Electronic prescribing	120 senior medical students	Nine prescribing scenarios per participant	Electronic-prescribing CDS	Randomized repeated-measures simulation	False-negative omissions and false-positive alerts	Omission and commission errors
Novotny et al. ^{17/Bond et al.¹⁸}	2017	ECG interpretation	30 cardiology and non-cardiology fellows; 9000 interpretations	One hundred ECGs across three interpretation phases	Computerized diagnostic proposals	Three-phase repeated-measures experiment	Single proposals were correct or intentionally incorrect; multiple-proposal condition also tested	Adjusted correct-vs-incorrect advice effect, accuracy, agreement, and confidence

Table 1 (continued) | Characteristics of the 40 independent core cohorts

Study	Year	Clinical domain	Participants/records	Clinical task/material	AI or decision-support system	Design	Erroneous-output exposure	Primary safety outcome
Lindow et al. ⁴⁷	2019	Primary-care ECG interpretation	988 patients; 89 erroneous calls; physician interpretation available for 77	Routine primary-care ECG interpretation	Computerized AF/AFL interpretation	Regional retrospective expert re-analysis	Computer falsely labeled atrial fibrillation or atrial flutter	Non-correction and inappropriate anticoagulation
Gaube et al. ¹⁹	2021	Chest radiography	138 radiologists and 127 IM/EM physicians	Eight chest-radiograph cases per participant	Diagnostic advice labeled as AI or human expert	2 x 2 mixed-factorial experiment	Two inaccurate-advice cases per participant	Accuracy decrement and susceptibility to wrong advice
Harada et al. ^{40,51}	2021	General diagnostic reasoning	22 physicians	16 clinical vignettes	AI history-taking system with differential-diagnosis list	Open-label randomized study	Lists variably included the correct diagnosis	Diagnostic accuracy, omission errors, and commission errors
Reverberi et al. ⁴⁶	2022	Colonoscopy	21 endoscopists	504 colorectal-lesion videos	Real-time CADx optical-diagnosis support	Multicenter repeated-measures experiment	Naturally correct or incorrect AI advice	Adoption of correct and incorrect AI advice
Bernstein et al. ²⁶	2023	Chest radiography	Six enrolled radiologists; five analyzed; 90 images across four sessions	Chest-radiograph interpretation	Sham AI findings with optional localization boxes	Fully crossed repeated-reader experiment	Four false-negative and eight false-positive AI outputs per assisted session	Wrong-AI penalty and interface mitigation
Dratsch et al. ²⁷	2023	Mammography	27 radiologists	50 mammograms	Purported AI BI-RADS suggestions	Prospective repeated-measures reader study	Incorrect BI-RADS suggestion in 12 test mammograms	Accuracy and magnitude of automation bias by experience
Jabbour et al. ⁵³	2023	Hospital medicine	457 clinicians	Acute respiratory-failure vignettes with chest radiographs	Diagnostic AI with optional image explanations	Multicenter randomized vignette trial	Systematically biased AI linked to age, BMI, or image blur	Diagnostic harm from biased AI and explanation mitigation
Rezazade Mehrizi et al. ⁵⁴	2023	Mammography	92 radiologists	15 mammography examinations; 2760 decisions	Mammography algorithmic suggestions	Two controlled reader experiments	Correct and intentionally incorrect suggestions	Following wrong suggestions and mitigation by explanation or priming
Wang et al. ⁵¹	2023	Musculoskeletal MRI	40 clinicians	200 anterior cruciate ligament MRI cases; 8000 interactions	Deep-learning MRI diagnostic aid	Randomized crossover laboratory study	Naturally correcting or misleading AI diagnoses	Correcting events, misleading events, and AI suppression
Acosta et al. ⁴²	2024	Radiology reporting	3 radiologists; multi-case reader study	Radiology cases with standard or AI-draft reporting	GPT-4 simulated draft reports	Paired multi-reader multi-case workflow study	1–3 deliberately inserted draft-report errors in half the cases	Residual clinically significant report errors and reporting time
Prinster et al. ⁵⁵	2024	Chest radiography	220 physicians	Eight radiographs per participant	Simulated AI advice with local or global explanations	Multicenter randomized experiment	Advice correctness and confidence were manipulated	Accuracy under incorrect AI and explanation effects
Sakamoto et al. ⁵⁷	2024	General diagnostic reasoning	20 physicians	20 clinical cases	AI history-taking system with top-10 differential list	Quasi-experimental study	Differential lists with variable reliability	Trust calibration and diagnostic accuracy
Yu et al. ⁵⁸	2024	Chest radiography	140 radiologists	Multi-reader evaluations across 15 pathologies	Chest-radiograph AI probabilities	Secondary analysis of randomized multi-reader studies	Naturally inaccurate predictions across tasks	Heterogeneity and individual harm from AI assistance
Biro et al. ⁴⁴	2025	Patient portal messaging/clinical communication	20 practicing primary-care physicians; 18 messages per clinician	18 simulated patient portal messages; four erroneous AI drafts	ChatGPT-4-generated draft replies	Cross-sectional randomized-order within-clinician simulation	Four naturally occurring draft errors identified by two PCP reviewers (two objective inaccuracies; two potentially harmful omissions)	Error insufficiently corrected after clinician review; completely unedited erroneous drafts
Escabi-Mendoza et al. ⁴⁸	2025	Real-world ECG interpretation	5000 ECGs; 238 provider-response records	Routine ECG interpretation and management	Computerized AF analysis	Retrospective blinded cardiologist adjudication	Computer overcalled or missed atrial fibrillation	Non-correction, specialty effects, and clinical consequences

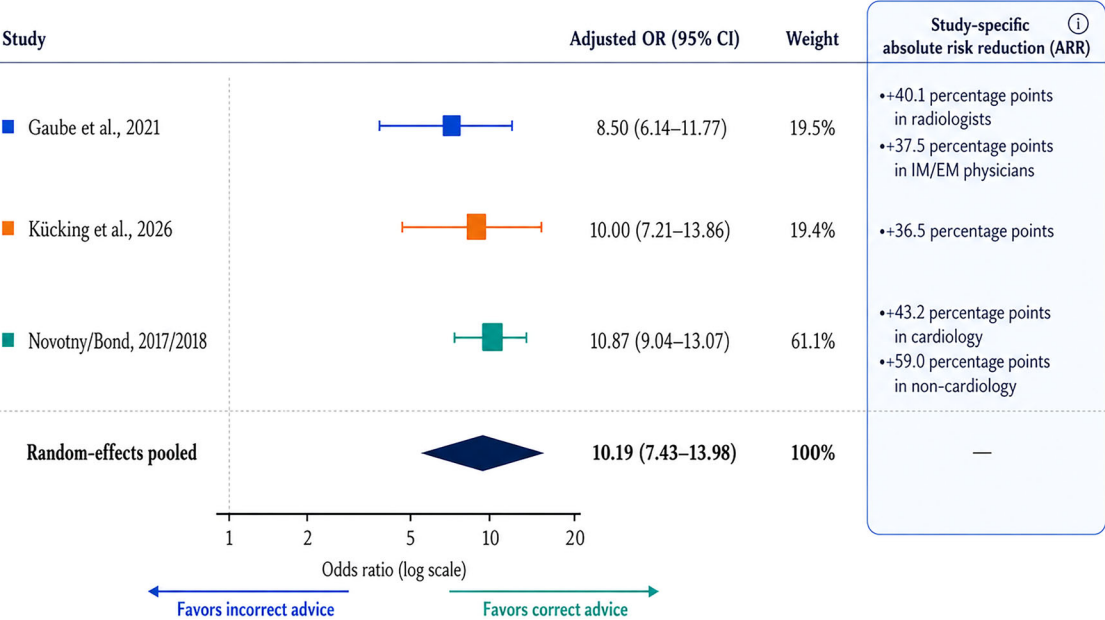
Table 1 (continued) | Characteristics of the 40 independent core cohorts

Study	Year	Clinical domain	Participants/records	Clinical task/material	AI or decision-support system	Design	Erroneous-output exposure	Primary safety outcome
Frei et al. ²⁶	2025	Computational pathology	18 medical students, 23 EMMP participants, and experience-stratified pathologists	Ten colorectal-cancer regions of interest	TCFCAD tumor-cell-fraction estimator	Repeated unaided and AI-assisted scoring	Inaccurate CAD estimate on selected regions	Anchoring toward inaccurate tumor-cell-fraction output
Jabbour et al. ⁴¹	2025	Hospital medicine diagnosis and treatment	259 clinicians	Clinical vignette diagnosis and treatment decisions	Selective AI prediction/abstention	Randomized clinical decision study	Inaccurate AI predictions shown or selectively withheld	Accuracy under inaccurate AI and selective prediction; changed error patterns
Kashiwagi et al. ⁴⁸	2025	Ophthalmic imaging	45 ophthalmologists	120 fundus photographs in two sessions	Glaucoma image-classification AI	Repeated-measures online simulation	Thirty percent of second-session AI judgments intentionally wrong	Accuracy and decision time under correct versus incorrect AI
Nagendran/Festor et al. ³⁷	2025	Intensive care treatment decisions	38 ICU physicians; eye-tracking subset $n = 19$	Six high-fidelity sepsis scenarios	Synthetic AI fluid/vasopressor recommender with XAI	High-fidelity repeated-measures simulation	Four safe and two deliberately unsafe recommendations per physician	Unsafe-advice rejection, escalation, dose shift, and attention
Qazi et al. ³⁹	2025	General diagnostic reasoning/LLM	44 licensed physicians; 264 vignette decisions	Six clinical vignettes	ChatGPT-4o diagnostic recommendations	Randomized single-blind clinical reasoning trial	Three of six LLM recommendations contained deliberate clinically significant errors	Diagnostic reasoning and top-choice accuracy under flawed vs error-free LLM advice
Xu et al. ⁴³	2025	Dermatology/primary care diagnosis	153 primary care physicians	Open-ended differential diagnosis of skin disease images	Image AI with traditional and multimodal-LLM explanations	Controlled human-AI experiment	AI predictions included correct and incorrect outputs; explanation and ordering varied	Accuracy under erroneous AI, explanation effects, and decision-order effects
Kücking et al. ²¹	2026	Wound assessment	223 physicians and nurses	1338 wound-image decisions	CNN wound-maceration CDSS	Simulated pre/intervention study with GLMM	Correct and incorrect AI recommendations	Correct decision, error adoption, and provider-level moderators
Ong et al. ³⁶	2026	Ophthalmology reasoning	30 participants: specialists, residents, and students	30 challenging published ophthalmology cases	Claude-3.5-Sonnet conversational support	Crossover human-LLM experiment	Incorrect LLM insights arose naturally	Performance deterioration and acceptance of incorrect LLM reasoning
Qazi et al. ⁴⁰	2026	General diagnostic reasoning/LLM mitigation	72 licensed physicians; 432 vignette decisions	Six clinical vignettes	ChatGPT-5.1 recommendations with trust-calibration cues	Randomized controlled mitigation trial	Three of six recommendations contained deliberate clinically significant errors	Diagnostic reasoning and top-choice accuracy with vs without behavioral nudge
Rosbach et al. ⁶¹	2026	Computational pathology	28 pathology experts	20 H&E image patches; 560 AI-assisted estimates	AI tumor-cell-percentage estimator	Counterbalanced within-subject experiment	Inaccurate AI predictions with and without time pressure	Negative consultation, anchoring, and time-pressure effects
Spitzer et al. ³⁵	2026	Radiology reasoning	101 radiologists	20 multimodal radiology cases; 2020 assessments	Multimodal GPT-4	Randomized between-group vignette experiment	Correct and incorrect diagnoses and explanations	Adherence to incorrect LLM diagnoses by explanation format
Vinas et al. ⁶⁰	2026	Treatment decisions	223 physicians in two experiments	Repeated treatment decisions for fictitious patients	AI patient-sensitivity classification	Two controlled behavioral experiments	Patients deliberately misclassified despite equal treatment response	Persistence of treatment choices despite contradictory outcome evidence

AF atrial fibrillation, AFL atrial flutter, AI artificial intelligence, BI-RADS Breast Imaging Reporting and Data System, BMI body mass index, CADx computer-aided diagnosis, CDSS clinical decision-support system, CNN convolutional neural network, ECG electrocardiography, EM emergency medicine, GLMM generalized linear mixed model, H&E hematoxylin and eosin, ICU intensive care unit, IM internal medicine, LLM large language model, MRI magnetic resonance imaging, PCP primary-care physician, XAI explainable artificial intelligence.

Clinician correctness under correct versus incorrect automated advice

Three adjusted cohorts; REML random-effects model with modified Hartung–Knapp confidence interval.



Pooled OR 10.19 (95% CI 7.43–13.98); $I^2 = 0\%$. ≥ 2020 sensitivity analysis yielded OR 9.22 (2.06–41.13). Absolute risk reduction was not pooled because baseline risks and adjusted model structures differed across cohorts.

Fig. 2 | Clinician correctness under correct versus incorrect automated advice, with absolute effects. Forest plot of adjusted odds ratios with study-specific absolute risk reductions (ARRs). The pooled estimate used a random-effects restricted maximum-likelihood model with a modified Hartung–Knapp confidence interval; the ≥ 2020 sensitivity analysis excludes the Novotny/Bond ECG cohort.

ARRs are reported in Supplementary Table 2. Table 2 summarizes the primary quantitative syntheses and direction-specific safety outcomes. Across the four real-world non-correction studies, publication year was not associated with a statistically reliable linear change in the probability that erroneous automated output remained uncorrected ($\beta = 0.075$ log-odds/year, 95% CI -0.066 to 0.216 ; $p = 0.30$). Given the small number of studies ($k = 4$), extreme between-study heterogeneity, and differences in workflow and outcome definition, this meta-regression should be interpreted as exploratory rather than as evidence that error interception has remained unchanged over time. Full sensitivity analyses and the exploratory temporal meta-regression are reported in Supplementary Table 3B.

Historical and repeated-reader evidence of the wrong-AI penalty The historical ECG studies showed that the conditional effect of automated advice was recognized long before contemporary machine learning. Primary-care physicians exposed to computerized reports became more likely to produce diagnoses consistent with those reports even when the reports were erroneous²². Emergency physicians benefited markedly from correct computer diagnoses, whereas incorrect diagnoses did not improve performance²³. In a two-period crossover experiment, internal-medicine residents agreed with incorrect computerized findings 67.7% of the time when those findings were shown versus 34.6% when they were absent; overall accuracy decreased from 56.7% to 48.3% in the incorrect-advice subset²⁴. Among emergency senior house officers, access to computer-generated reports did not significantly reduce major interpretation errors despite the computer itself having a low major-error rate²⁵.

Contemporary reader experiments reproduced the same pattern with more complex AI. Bernstein et al. studied five radiologists who interpreted 90 chest radiographs across four sessions. On false-negative AI cases, the false-negative rate increased from 2.7% without AI to 33.0% when the AI result was retained without localization, 26.7% when the AI record was

deleted, and 20.7% when a localization box was supplied. On false-positive cases, the false-positive rate increased from 51.4% without AI to 80.5–86.0% across assisted conditions²⁶. Mammography, pathology, glaucoma, and general diagnostic reasoning studies likewise showed that systematically wrong suggestions could redirect clinician judgments, often despite explanations or professional experience^{27–31}. Prescribing and alerting studies provided complementary mechanism-specific evidence: Goddard et al. documented 27 correct-to-wrong switches among 520 decisions; a cluster-randomized renal-risk alert trial found no significant adoption of false-positive alerts; and incorrect e-prescribing decision support increased omission and commission errors compared with no decision support^{32–34}. Contemporary LLM experiments in radiology and challenging clinical reasoning likewise showed that incorrect model output or explanations could adversely influence clinician decisions^{35,36}. The revised contemporary evidence base also included the linked Nagendran/Festor high-fidelity ICU simulation cohort, randomized LLM studies by Qazi et al., selective-prediction work by Jabbour et al., AI-generated reporting experiments by Acosta et al., clinician–AI interaction studies by Xu et al., and simulated primary-care messaging work by Biro et al.^{37–44}. Collectively, these studies broadened the evidence beyond legacy computerized decision-support systems and showed that erroneous recommendations can influence clinician reasoning across newer AI architectures and clinical tasks. The Qazi studies directly examined flawed LLM recommendations and mitigation through behavioral nudges, while the Nagendran/Festor simulation showed that clinicians could reject or escalate unsafe recommendations in a high-fidelity intensive-care workflow.

Real-world failure to correct erroneous output Four retrospective ECG studies reported extractable denominators for failure to intercept erroneous computerized interpretations, but their operational definitions were not identical. Anh et al. counted erroneous AF

Table 2 | Primary quantitative syntheses and direction-specific safety outcomes

Outcome/analysis	Contributing evidence	Effect measure	Key result	Heterogeneity/uncertainty	Interpretation
Correct vs incorrect automated advice	Gaube ⁴⁵ , Kücking ²¹ ; Novotny/Bond ^{17/18}	Adjusted odds ratio	OR 10.19 (95% CI 7.43–13.98)	k = 3; I ² = 0%; REML + modified Hartung–Knapp	Large conditional penalty when advice was wrong; exploratory because k = 3.
Contemporary sensitivity (≥2020)	Gaube ⁴⁵ , Kücking ²¹	Adjusted odds ratio	OR 9.22 (95% CI 2.06–41.13)	k = 2; modified Hartung–Knapp	Excluding the older ECG cohort preserved the point estimate but markedly widened uncertainty.
Absolute risk reductions (ARRs)	Same 3 cohorts	Study-specific percentage-point reduction in incorrect decisions/increase in correct decisions	Gaube 40.1 pp radiologists/37.5 pp IM/EM; Kücking 36.5 pp; Novotny/Bond 43.2 pp cardiology/59.0 pp non-cardiology	Not pooled because baseline risks and study structures differed	Absolute effects are reported alongside ORs; no pooled ARR was calculated.
Real-world erroneous-output non-correction	Ann ⁴⁵ , Bae ⁴⁶ , Lindow ⁴⁷ ; Escabi-Mendoza ⁴⁸	Random-effects proportion	26.6% (95% CI 6.4–65.6%)	I ² = 95.8%; 95% PI 0.3–98.0%	Exploratory implementation estimate; operational definitions differ.
Real-world temporal trend	Same 4 studies	Random-effects meta-regression	β = 0.075 log-odds/year (95% CI –0.066 to 0.216); p = 0.30	k = 4; extreme heterogeneity	No reliable evidence of linear temporal change; hypothesis-generating only.
Bidirectional clinician-CDSS transitions	Dreiseitl ⁴⁹	Opportunity-specific transitions	Wrong → correct 36/140 (25.7%); correct → wrong 50/217 (23.0%)	Benefit:harm ratio 0.72; net –14 correct decisions	Benefit and harm coexisted within the same workflow.
High-fidelity unsafe-AI interception	Festor ³⁸	Study-specific interception rate	83% unsafe recommendations stopped; 92% stopped or escalated	Single simulated ICU cohort	Escalation can operate as an explicit safety barrier.
Contemporary LLM error penalty	Qazi ³⁹	Adjusted mean difference	Diagnostic reasoning –14.0 pp; top-choice accuracy –18.3 pp	Randomized trial	Extends susceptibility to deliberately flawed LLM recommendations.
Generative-AI draft error interception	Biro ⁴⁴	Study-specific miss rate	Each of four errors missed by 65–75%; mean 2.67/4 missed; 35–45% drafts sent unedited	20-PCP controlled simulation	Contemporary evidence of residual error after clinician review.

ARR absolute risk reduction, CDSS clinical decision-support system, CI confidence interval, ECG electrocardiography, ICU intensive care unit, IM/EM internal medicine/emergency medicine, LLM large language model, OR odds ratio, PI prediction interval, pp percentage points, REML restricted maximum likelihood.

calls not corrected by the primary reader; Bae et al. used a broader correction process that could include subsequent computer interpretation; Lindow et al. counted accepted errors among records with documented physician interpretation; and Escabi-Mendoza et al. classified explicitly unrecognized errors while retaining ‘no mention’ as a separate category^{45–48}. These definitions were mapped to the common implementation-level construct of failure of error interception but were not treated as interchangeable.

Across the four real-world ECG studies, the pooled proportion of erroneous automated interpretations that remained uncorrected or unrecognized was 26.6% (95% CI 6.4–65.6%; I² = 95.8%; 95% prediction interval 0.3–98.0%; Fig. 3). The operational definitions differed across studies, and the extreme heterogeneity indicates that this estimate should not be interpreted as a common underlying failure rate. The beta-binomial robustness estimate was 29.3% (95% CI 14.9–46.5%). Excluding Bae et al., whose correction definition could include serial computerized interpretation, yielded a pooled estimate of 36.3% (95% CI 13.0–68.4%). Sensitivity analyses varying the treatment of undocumented physician responses in Lindow et al. and ‘no mention’ cases in Escabi-Mendoza et al. produced estimates within the same broad range and did not resolve the heterogeneity. We therefore interpret the pooled value as an exploratory summary demonstrating that failure to intercept erroneous automated output occurs in routine implementation, while its frequency is strongly context-dependent. Supplementary Table 3A reports, for each real-world cohort, the original study-specific definition, denominator, handling of missing or undocumented responses, harmonized synthesis definition, and corresponding sensitivity-analysis rule.

Specialty, expertise, and workflow

Error recognition varied substantially by clinical role. In Anh et al., only 5 of 77 erroneous atrial-fibrillation calls were uncorrected when a cardiologist was the primary reader, compared with 38 of 131 for internists, 35 of 140 for emergency physicians, and 14 of 34 for other specialists⁴⁵. In Escabi-Mendoza et al., the explicit non-recognition rate excluding records with no ECG mention was 11.6% among cardiology attendings, compared with 75.9% among emergency physicians, 44.0% among hospitalists or internal-medicine attendings, 51.4% among residents, and 56.3% among other providers⁴⁸. Cardiologist consultation corrected all nine erroneous interpretations for which consultation was documented in Lindow et al.⁴⁷. These comparisons were observational and may be confounded by case mix and workflow, but they consistently suggest that domain-specific verification capacity matters more than the mere presence of a human reviewer. Full specialty-stratified data are provided in Supplementary Table 4. Figure 4 maps cohort-level coverage across wrong-advice exposure, direction-specific transitions, real-world non-correction, clinical consequences, and mitigation.

Bidirectional decision transitions

Dreiseitl and Binder provided the only complete direction-specific transition matrix⁴⁹. When an initially wrong dermatologist decision conflicted with the decision-support system, 36 of 140 decisions changed toward the correct recommendation (25.7%). When an initially correct decision conflicted with the system, 50 of 217 changed toward the wrong recommendation (23.0%). Across 357 disagreement opportunities, the system therefore corrected 36 errors but created 50 new errors, yielding a benefit-to-harm event ratio of 0.72 and a net loss of 14 correct decisions. In other conditional analyses, physicians followed 18 of 84 entirely wrong OncoDoc2 recommendations (21.4%), clinicians changed to an incorrect AI-supported MRI diagnosis in 132 of 556 eligible disagreements (23.7%), and primary-care physicians accepted 36 of 77 erroneous ECG calls (46.8%)^{47,50,51}. These estimates were not pooled because the opportunity denominators and decision structures were not equivalent. The complete direction-specific transition and wrong-advice adoption data are provided in Supplementary Table 5. Figure 5 integrates the bidirectional transition evidence, selected wrong-advice adoption rates, and observed downstream management actions and patient events. Figure 6 provides three conceptual scenarios

Real-world failure to correct erroneous automated interpretations

Study-level Wilson intervals; primary logit random-effects model and beta-binomial robustness analysis.

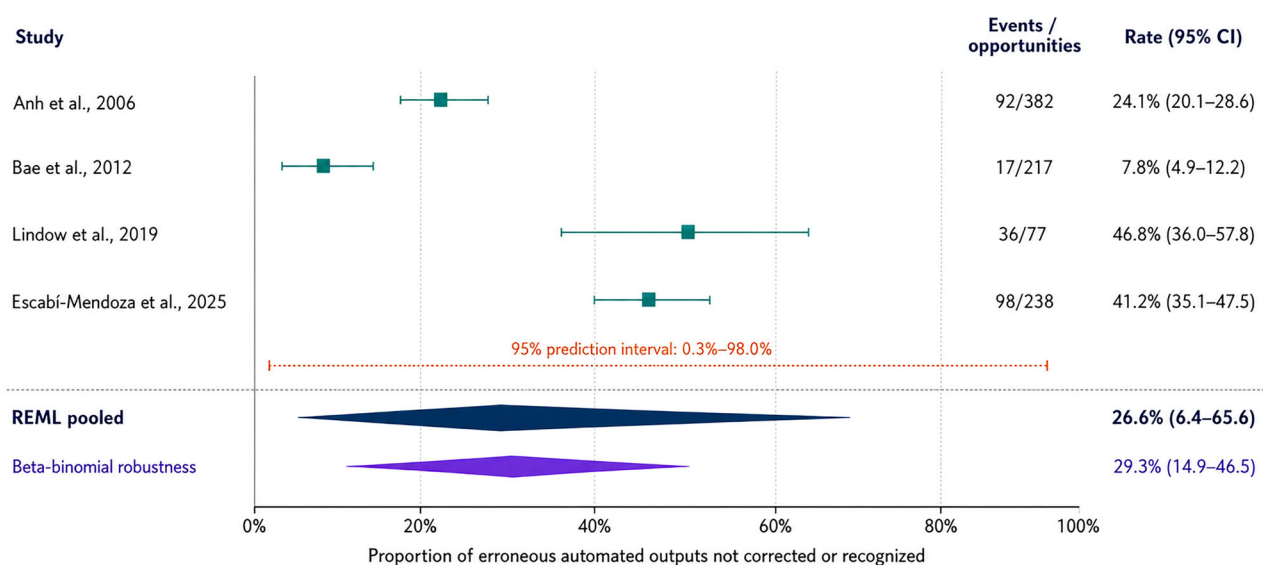


Fig. 3 | Real-world failure to correct erroneous automated interpretations. Forest plot of study-level non-correction proportions with Wilson confidence intervals, the primary logit random-effects estimate, the beta-binomial robustness estimate, and the 95% prediction interval.

illustrating beneficial correction, AI-induced harm, and resilient oversight through independent verification or escalation.

Management consequences and patient harm

Accepted or uncorrected automated errors were associated with downstream management consequences, including inappropriate treatment and delayed anticoagulation. In Southern and Arnsten, an erroneous computerized interpretation of acute ischemia increased recommendations for urgent revascularization from 10% to 30%⁵². Bae et al. and Lindow et al. documented inappropriate follow-up, treatment, and anticoagulation after accepted erroneous ECG interpretations^{46,47}. These observations demonstrate that erroneous automated output may persist into subsequent clinical management, but they do not by themselves establish patient harm.

Patient-level adverse events were much more limited: thromboembolic and bleeding events were reported in only one retrospective cohort and involved small event counts. Escabi-Mendoza et al. reported delayed anticoagulation in 14 of 76 unrecognized missed-atrial-fibrillation cases, two thromboembolic events, inappropriate anticoagulation in 10 of 86 overcalled cases, and one bleeding hospitalization⁴⁸. These observational findings demonstrate that erroneous automated output may persist into subsequent clinical management, but they do not establish that the automated error caused the adverse patient outcomes. Event-level management and patient consequences are detailed in Supplementary Table 6.

Mitigation strategies

Mitigation evidence was inconsistent. Localization reduced false-negative influence in Bernstein et al., whereas deleting the AI record produced only partial mitigation²⁶. Presenting multiple computerized diagnostic proposals in the Novotny/Bond cohort may have encouraged differential reasoning, although the comparison requires replication^{17,18}. Selective suppression of high-risk outputs reduced harmful switching in simulation, but has not yet been prospectively validated in a clinical workflow⁵¹.

By contrast, explanations and generic warnings were not dependable safeguards. Image-based explanations did not meaningfully protect clinicians from systematically biased AI, local and global explanations did not rescue performance under wrong chest-radiograph advice, and a warning that “ChatGPT can make mistakes” did not reduce diagnostic switching^{53–56}.

Trust-calibration training and attitudinal priming also lacked consistent overall benefit^{54,57}. Large observational analyses further showed substantial heterogeneity in who benefits or is harmed by AI assistance, reinforcing the need for interaction-level rather than average-effect evaluation^{58–61}. Mitigation strategies, their observed effects, and their implementation implications are summarized in Table 3.

Methodological appraisal

Among the 40 core cohorts, 26 were classified as having high overall methodological concern and 14 as having some concerns; none was classified as low concern. Common limitations included convenience sampling, small reader samples, simulated or purported AI, absence of a same-case pre-AI judgment, incomplete accounting for repeated clinician or case observations, and incomplete direction-specific outcome reporting. Final domain-level consensus judgments are presented in Supplementary Table 7 and summarized visually in Supplementary Fig. 1.

Discussion

This review provides complementary evidence on human–AI interaction safety across controlled and real-world settings. Across three adjusted cohorts, clinician correctness was substantially higher when automated advice was correct than when it was wrong, and the ≥ 2020 sensitivity analysis yielded a similar point estimate after exclusion of the older ECG cohort. Across four real-world ECG studies, erroneous output sometimes survived clinical review, but the prediction interval spanned almost the full possible range. Direction-specific evidence further showed that automated support can both correct an initially wrong decision and overturn an initially correct one. Together, these findings support evaluating human oversight as a conditional error-interception capability rather than inferring safety from average AI performance or the mere presence of a clinician reviewer.

The historical and contemporary evidence should not be treated as technologically equivalent. Historical studies should not be interpreted as direct estimates of the safety of contemporary generative AI; their relevance is primarily mechanistic. Earlier computerized ECG and conventional decision-support studies already showed susceptibility to incorrect automated recommendations and the possibility of both beneficial and harmful decision switching. Contemporary AI and LLM studies extend these

Evidence matrix across 40 independent core cohorts

Direct quantitative evidence remains concentrated in immediate performance and reliance outcomes; real-world error interception, consequences, and mitigation remain sparse.

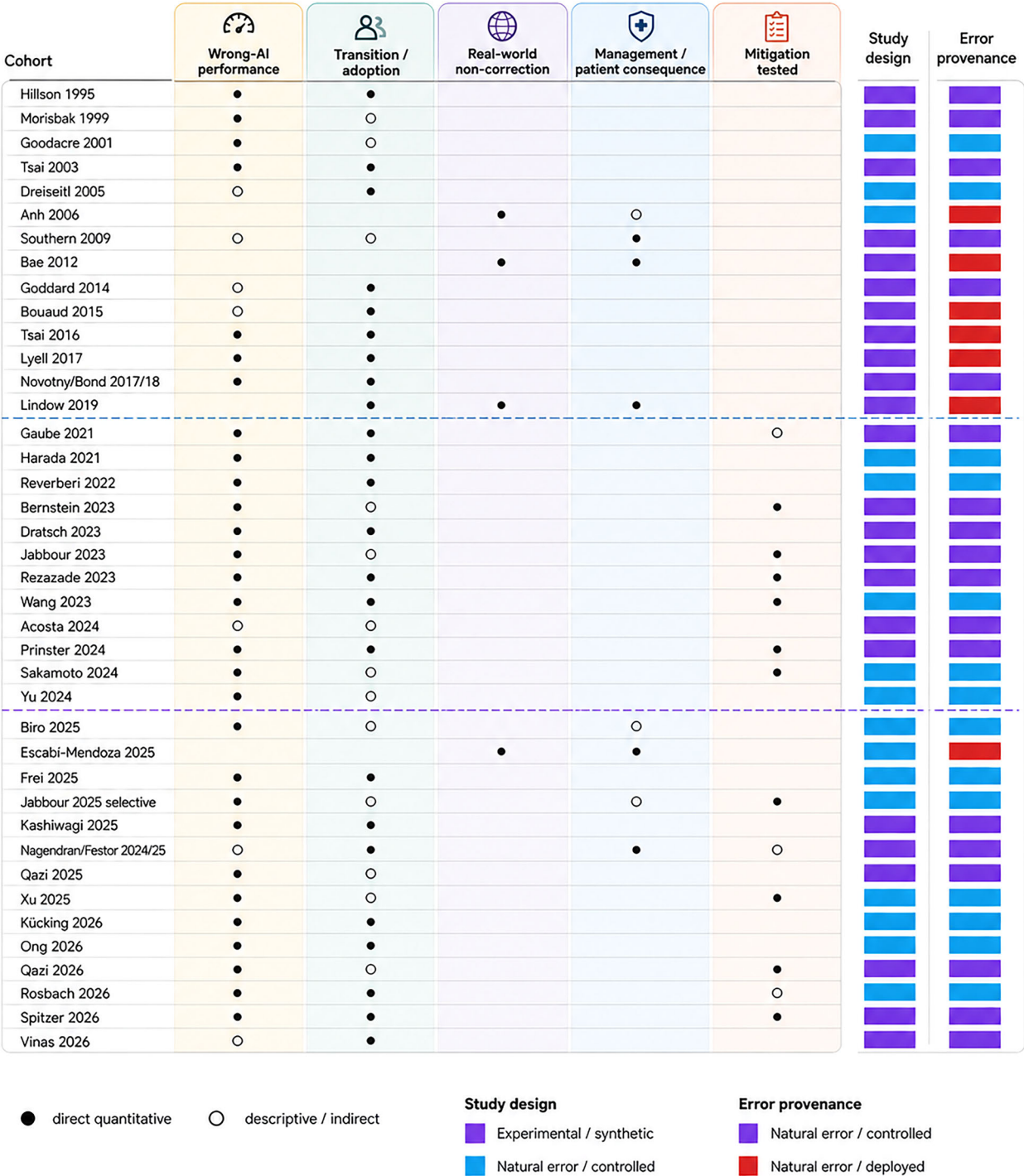


Fig. 4 | Evidence matrix across 40 independent core cohorts. Heat map showing cohort-level evidence across wrong-advice exposure, direction-specific transitions/adoption, real-world non-correction, management or patient events, mitigation, and error provenance. Error provenance is distinguished as experimentally manipulated/synthetic, naturally occurring in controlled evaluations, or naturally occurring in deployed systems.

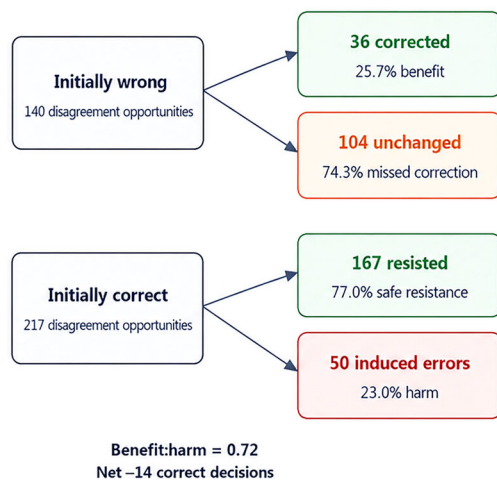
mechanisms into more capable and persuasive systems, but most modern evidence remains controlled or simulated rather than deployed. The older quantitative estimates therefore should not be interpreted as direct measures of the safety or failure rate of contemporary generative AI.

The provenance of an automated error changes what can be inferred from clinician responses to it. Deliberately manipulated errors isolate the effect of exposure to incorrect advice more cleanly because error status is controlled independently of the underlying case. Naturally occurring errors

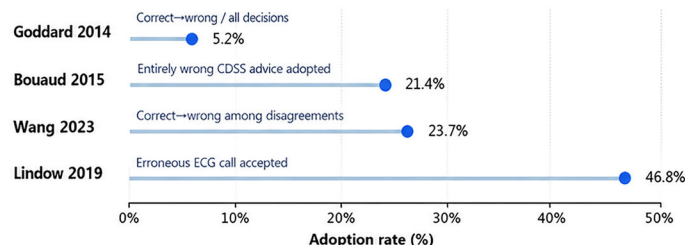
Benefit–harm transitions and observed downstream events

Direction-specific denominators separate AI-enabled correction, AI-induced errors, and observed downstream management actions and patient events.

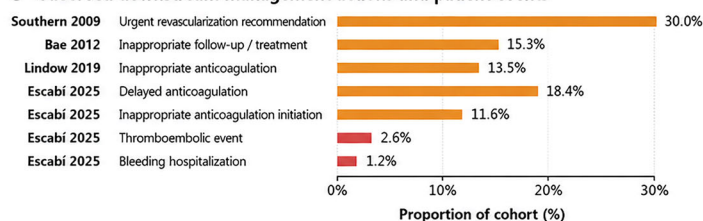
A Complete bidirectional transition matrix



B Selected wrong-advice adoption rates



C Observed downstream management actions and patient events



Orange = management actions; red = patient-level adverse events. Patient-level adverse events were reported in only one retrospective cohort.

Fig. 5 | Benefit–harm transitions and observed downstream events. Panel A shows wrong-to-correct and correct-to-wrong decision transitions. Panel B summarizes selected wrong-advice adoption rates using opportunity-specific denominators. Panel C summarizes observed downstream management actions and patient-level

events after accepted or uncorrected automated errors. Patient-level events were reported in only one retrospective cohort and should not be interpreted as establishing causality.

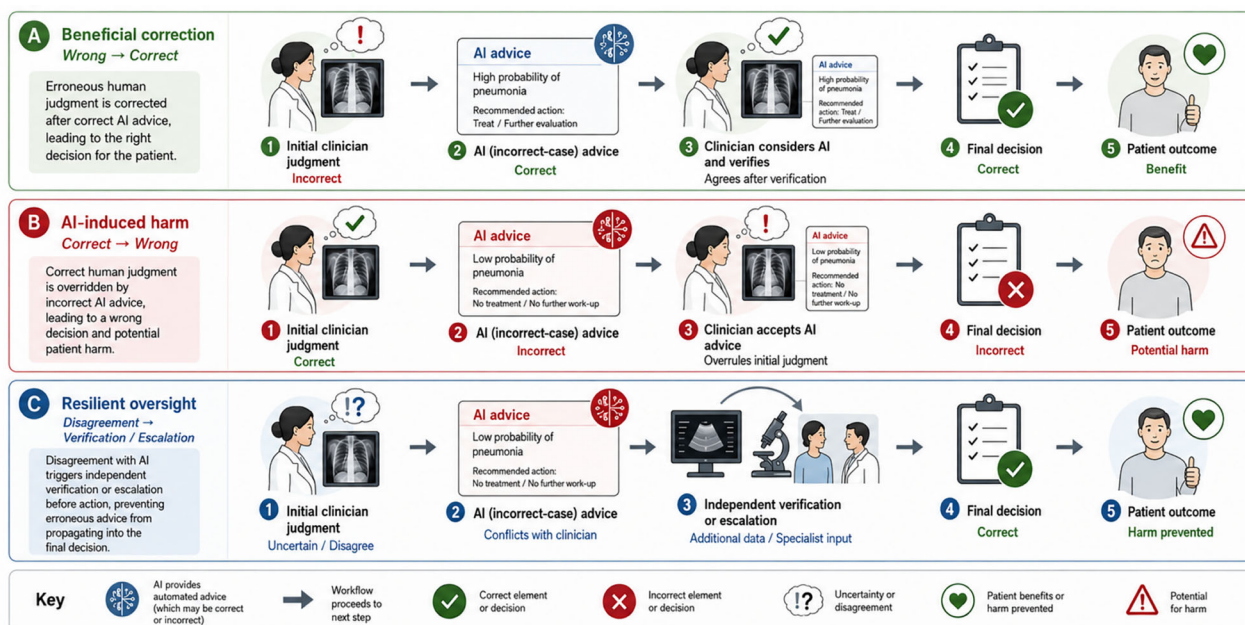


Fig. 6 | Hypothetical clinical scenarios illustrating human oversight of erroneous AI. The figure depicts three conceptual pathways: beneficial correction of an initially incorrect clinician judgment, AI-induced reversal of an initially correct

judgment, and resilient oversight in which disagreement triggers independent verification or escalation before action. These scenarios are illustrative and are not derived from individual patient-level data.

in controlled evaluations more closely represent authentic model failures but may occur preferentially in difficult or ambiguous cases, making the effects of model error and case difficulty harder to separate. Errors observed in deployed systems provide the strongest evidence about routine error interception but are additionally influenced by workflow, documentation, expertise, and downstream verification processes. These evidence sources are therefore complementary rather than interchangeable. Their inclusion

within one review reflects a common human-oversight question, whereas their provenance is explicitly retained when interpreting internal validity, generalizability, and implementation relevance.

The phrase “human in the loop” describes an organizational arrangement, not a demonstrated safety property. A clinician can formally retain final authority while lacking the expertise, time, uncertainty information, or source evidence needed to challenge a plausible

Table 3 | Strategies intended to mitigate erroneous automated influence

Study	Mitigation strategy	Comparison/design	Quantitative result	Interpretation
Bernstein et al. ²⁶	Localization box/deletion of AI record	False-negative and false-positive sham-AI outputs	Localization attenuated false-negative influence; deletion only partly mitigated anchoring	Task-specific localization may help but does not eliminate error influence.
Jabbour et al. ⁵³	Image-based explanation	Biased AI with explanation vs biased AI alone	Difference 2.3 pp (95% CI −2.7 to 7.2)	No convincing protection from systematically biased predictions.
Wang et al. ⁵¹	Selective AI suppression	Suppress predictions at high estimated risk of misleading clinician	Estimated 41.7% reduction in misleading events with 3.2% loss of correcting events	Promising trade-off; requires prospective validation.
Prinster et al. ⁵⁵	Local vs global explanations	Incorrect AI advice across explanation conditions	Accuracy remained low under wrong AI; explanation type did not reliably rescue performance	Explanations are not equivalent to independent verification.
Sakamoto et al. ⁵⁷	Trust-calibration training	Calibration intervention vs control	Diagnostic accuracy 41.5% vs 46.0%; calibration accuracy 61.5%	Training alone was not consistently protective.
Jabbour et al. ⁴¹	Selective prediction/abstention	Inaccurate AI shown vs selectively withheld vs no AI	Accuracy 56% with inaccurate AI, 64% selective prediction, 66% no AI	Recovered most accuracy but increased missed diagnoses/treatments when AI abstained.
Qazi et al. ⁴⁰	Behavioral trust-calibration nudge	Same LLM recommendations with vs without baseline-accuracy + ensemble-confidence cues	Diagnostic reasoning +7.6 pp (95% CI 1.4–13.9); top-choice accuracy +10.9 pp	Contemporary randomized evidence that case-specific uncertainty cues can attenuate LLM automation bias.
Festor et al. ³⁸	Senior/second-opinion escalation	Unsafe recommendation rejected vs rejected or escalated	83% stopped; 92% stopped or escalated	Escalation can function as an explicit workflow safety barrier.
Kiyak et al. ⁵⁶	Generic warning	“ChatGPT can make mistakes” vs no warning	Switching 15.9% vs 15.3%; OR 1.09 (95% CI 0.46–2.59)	Generic warnings were not protective.

AI artificial intelligence, CI confidence interval, LLM large language model, OR odds ratio, pp percentage points.

recommendation. The specialty gradients in real-world ECG studies illustrate this distinction: cardiologists corrected erroneous output more often than non-specialists, and specialist consultation functioned as a separate safety layer. Human oversight should therefore be evaluated as a conditional capability comprising independent judgment, access to confirmatory evidence, domain expertise, time for verification, and an escalation pathway.

This framing also clarifies why asking users simply to “check the AI” is inadequate. Verification complexity may exceed the resources available at the point of care, especially for high-dimensional images, long records, or fluent generative explanations^{8,9}. The appropriate safety question is not whether a clinician signed the final decision, but whether the workflow enabled meaningful independent verification of the failure modes that matter.

Post-assistance accuracy collapses distinct transitions. An unchanged correct decision, an error corrected by AI, a correct decision overturned by AI, and a persistent human error can produce the same aggregate accuracy under different clinical mechanisms. The Dreiseitl transition matrix shows why this matters: beneficial and harmful transitions occurred in the same workflow, and the harmful transitions slightly outnumbered the beneficial ones. A complete four-cell transition matrix—correct-to-correct, correct-to-wrong, wrong-to-correct, and wrong-to-wrong—should become a standard outcome, accompanied by disagreement-opportunity denominators and a benefit-to-harm event ratio.

This approach also prevents misleading comparisons across studies. A raw “AI-adoption rate” can appear low when calculated over all decisions even though the conditional adoption rate is high when the AI is wrong and disagrees with an initially correct clinician. Conversely, a system may have a high overall acceptance rate because most advice is correct. Safety reporting must therefore condition reliance on AI correctness and on whether a genuine disagreement opportunity existed.

The I² of 95.8% in the real-world non-correction synthesis is not merely statistical noise. It reflects genuine variation in disease prevalence, signal quality, clinician expertise, documentation, workflow, access to overreading, and how correction was defined. The prediction interval of 0.3–98.0% means that the pooled mean cannot forecast safety in a new deployment. Local validation should measure non-correction in the intended workflow, stratified by reader role, error type, urgency, and availability of confirmatory information. Reporting a single pooled percentage without this context would imply a transportability that the evidence does not support.

The prominence of ECG studies in the earlier evidence base reflected both the long history of computerized ECG interpretation as a deployed clinical decision-support technology and incomplete capture of several recent human–AI studies in the original review. After inclusion of additional contemporary evidence, ECG interpretation accounted for 10 of 40 core cohorts. Nevertheless, an important asymmetry remains: the pooled real-world non-correction estimate derives entirely from deployed computerized ECG systems, whereas most contemporary AI and LLM evidence comes from controlled experiments or simulated clinical workflows. The ECG implementation estimate should therefore not be extrapolated directly to generative AI.

This review is complementary to recent syntheses of average human–LLM collaboration and broad collaboration patterns^{4–6}. Those reviews establish that collaboration effects are heterogeneous. The present analysis identifies one mechanism underlying that heterogeneity: the value of AI assistance depends strongly on whether the advice is correct and whether the workflow can recognize the cases in which it is not.

The evidence supports a layered rather than symbolic approach to oversight. First, workflows should preserve an independent clinician judgment before revealing AI output when feasible. Second, disagreement should trigger active verification rather than passive reconsideration. Third, high-risk or poorly calibrated outputs should be suppressed, accompanied by abstention, or routed to specialist review. Fourth, interfaces should expose task-specific uncertainty and verifiable source evidence rather than relying on generic explanations. Fifth, post-deployment monitoring should count AI-induced errors, clinician-corrected AI errors, known-output non-

correction, and downstream management consequences. These principles align with lifecycle risk management and responsible clinical machine learning^{13,62}.

Mitigation trials should compare concrete interface and workflow mechanisms. “Explanation” is too broad an intervention category: localization may focus verification, while narrative rationales may increase perceived coherence without increasing truthfulness. Similarly, warnings are unlikely to change behavior when they are generic, continuously displayed, or disconnected from the specific uncertainty of the case. Prospective trials should test independent-first workflows, structured counterfactual prompts, alternative-hypothesis generation, selective abstention, specialist escalation, and audit feedback using the same direction-specific safety outcomes.

Strengths include cohort-level linkage of overlapping reports, explicit separation of experimentally manipulated and naturally occurring errors, integration of historical and contemporary technologies, separation of all-decision and disagreement-opportunity denominators, use of adjusted comparative effects where available, study-specific absolute effects, a contemporary sensitivity analysis, a prediction interval for real-world non-correction, and a beta-binomial robustness analysis.

Evidence linking erroneous automated output to patient-level harm remains sparse. Although inappropriate treatment, delayed management, thromboembolism, and bleeding were observed downstream of accepted or uncorrected errors, patient-level adverse events were reported in only a single retrospective cohort and involved few events. Residual confounding, underlying disease severity, clinician decision-making, and subsequent care processes could all have contributed to these outcomes. The available data therefore support a plausible pathway from automated error to altered management but do not establish a causal effect of automation bias on patient harm.

Several limitations remain. The primary adjusted meta-analysis contained only three cohorts, and the real-world proportion meta-analysis contained four ECG studies with extreme heterogeneity and non-identical correction definitions. The published Novotny/Bond model did not clearly document adjustment for repeated physician and case observations, while the two-study contemporary sensitivity analysis was necessarily imprecise. Most contemporary AI and LLM studies were controlled experiments or simulations rather than deployed systems. Patient-level adverse events were reported in only one retrospective cohort with small event counts, precluding causal attribution. Several historical studies provided only partial extractable data. Search completeness should also be interpreted with some caution, as citation-based checking of known eligible studies improved coverage but could not exclude the possibility that relevant studies were missed when responses to erroneous AI output were secondary analyses or were not apparent in searchable titles, abstracts, or keywords. Finally, the archived project record did not preserve study-level reasons for the 16 non-retrieved reports or verbatim platform-exported search histories; these limitations are reported transparently rather than retrospectively reconstructed.

Evidence that erroneous automated output can influence clinician decisions spans several generations of clinical decision support, but the strength and setting of that evidence differ substantially by technology era. Real-world non-correction data currently derive mainly from legacy computerized ECG systems, whereas evidence for contemporary AI and large language models remains predominantly experimental. Erroneous automated advice can overturn otherwise correct clinical judgments and persist into inappropriate or delayed management, but patient-level harm evidence remains sparse and observational. Human oversight is therefore not intrinsically protective and should be designed, tested, and monitored as a conditional error-interception capability within each technology and clinical context. Prospective studies should determine whether failures of oversight causally affect patient safety while preserving the benefits of AI assistance.

Methods

Protocol and reporting

The review was conducted and reported in accordance with the PRISMA 2020 statement and the Cochrane Handbook, with synthesis without meta-analysis reported according to SWiM when quantitative pooling was not defensible^{63–65}. The protocol and statistical analysis plan were specified before the expanded quantitative synthesis and registered in PROSPERO (CRD420261457103). The completed PRISMA checklist is provided in the Supplementary Information, and the conventional study-selection flow is presented in Supplementary Fig. 2. AI was used for language editing, grammar correction and improvement of readability. AI-assisted tools were not used to determine study eligibility, extract data, assess methodological quality, make statistical decisions independently, or generate scientific conclusions. AI-assisted tools were not used to generate underlying study data or fabricated visual content.

Information sources and search strategy

We searched MEDLINE, Embase, Web of Science Core Collection, Scopus, PsycINFO, CINAHL, IEEE Xplore, ACM Digital Library, medRxiv, and arXiv from inception to 20 July 2026. The search was deliberately focused on studies in which erroneous, misleading, discordant, or otherwise incorrect automated clinical output was followed by an observable clinician judgment, decision, correction, acceptance, or other response; it was not intended to retrieve the broader literature on clinical AI performance or human–AI collaboration without an error-exposure component. The original search identified 160 records. After removal of 10 duplicates, 150 records were screened; 79 reports were sought for retrieval, 63 were assessed for eligibility, and 38 were included. Sixteen reports could not be retrieved. The archived project record retains this aggregate count but does not contain sufficiently reliable study-level reasons for non-retrieval or a complete record of retrieval attempts; these details were therefore not retrospectively reconstructed. During revision, reviewer-suggested studies and targeted citation/recent-literature searching through 17 August 2026 identified 11 additional reports for assessment, of which eight were included and three excluded. Full source-specific search strategies, including Boolean combinations, field restrictions, and platform-specific syntax, are provided in Supplementary Table 8. Because verbatim platform-exported histories from the original search were not retained in the available project archive, the source-specific syntax was reconstructed from the prespecified search concepts, protocol, and available project records and is explicitly identified as reconstructed.

Eligibility criteria

We included studies in which licensed clinicians or clinical trainees made a diagnostic, prescribing, triage, interpretation, treatment, documentation, or management decision after viewing output from an AI, machine-learning model, automated interpretation system, clinical decision-support system, or LLM. Eligible studies had to include an incorrect, misleading, systematically biased, false-positive, false-negative, unsafe, or naturally erroneous automated recommendation, or to provide pre-AI and post-AI decisions that permitted identification of an AI-induced correct-to-wrong transition. Randomized trials, crossover studies, repeated-reader experiments, quasi-randomized studies, before–after studies, high-fidelity simulations, and real-world implementation cohorts were eligible.

We excluded standalone AI-versus-human performance comparisons, patient-only chatbot studies, studies reporting only satisfaction or readability, protocols, editorials, and studies in which the automated output was always correct. Studies with disagreeing but non-adjudicable advice were retained only as adjacent mitigation evidence. Reports sharing participants and cases were linked at the cohort level; the most complete report supplied the primary quantitative data and companion reports supplied only unique secondary outcomes.

Study selection and data extraction

Two reviewers (D.Z. and J.L.) independently screened titles and abstracts, assessed full texts, and extracted data using a piloted form. Disagreements

were resolved through discussion and, when necessary, adjudication by Z.L. We extracted clinical domain, clinician expertise, sample size, task and modality, automation type, source labeling, timing of the automated advice, whether a pre-AI independent judgment was recorded, how erroneous output arose, reference-standard construction, unit of analysis, repeated-measures structure, accuracy, reliance, confidence, decision time, mitigation, and downstream clinical consequences. Numerical data were extracted from text, tables, figures, and supplementary files. Discrepancies between reported denominators were retained transparently and explored in sensitivity analyses rather than silently reconciled.

Error provenance was classified into three categories: (1) experimentally manipulated or synthetic erroneous output, in which investigators deliberately controlled the correctness of the automated recommendation; (2) naturally occurring model errors evaluated in controlled or simulated settings; and (3) naturally occurring errors identified in deployed or routine-care systems. These categories were retained within the same review because each addressed the common review question of how clinicians respond after exposure to erroneous automated output, but they were treated as distinct sources of evidence with different implications for internal validity, ecological validity, and implementation relevance.

Outcome definitions

The primary comparative outcome was clinician correctness under correct versus incorrect automated advice. The co-primary implementation outcome was the proportion of erroneous automated outputs not corrected or recognized by the responsible clinician in real-world care. Direction-specific transitions were defined as correct-to-wrong switching (harmful negative consultation) and wrong-to-correct switching (beneficial correction). Secondary outcomes were erroneous-advice adoption, omission and commission errors, confidence and calibration, decision time, specialty differences, mitigation effects, management consequences, and patient harm. Rates based on all decisions were kept separate from rates conditioned on erroneous advice or a clinician–AI disagreement opportunity.

For the real-world implementation studies, study-specific definitions of non-correction or non-recognition were retained during extraction and then mapped to a harmonized implementation construct of failure of error interception: an erroneous automated interpretation that survived the relevant clinical review or documentation step without documented correction or recognition. Anh et al. defined non-correction as failure of the primary physician reader to correct an erroneous computerized atrial-fibrillation interpretation. Bae et al. used a broader definition in which correction could arise through physician review and/or subsequent computerized interpretation. Lindow et al. classified erroneous atrial-fibrillation or atrial-flutter interpretations according to whether they were accepted or corrected in the documented physician interpretation. Escabí-Mendoza et al. explicitly distinguished recognized from unrecognized computerized errors and reported cases with no documented recognition separately. Because these operational definitions were not identical, the pooled proportion was interpreted as an exploratory implementation-level summary rather than a universal estimate of clinician error recognition. Sensitivity analyses varied the handling of the broader Bae definition and undocumented or ‘no mention’ responses in Lindow and Escabí-Mendoza.

Methodological appraisal

Because the evidence base included randomized trials, crossover and repeated-reader studies, quasi-experimental studies, and retrospective implementation cohorts, methodological appraisal was design-specific. Two reviewers (D.Z. and J.L.) independently assessed each core cohort using prespecified domains. Disagreements were resolved through discussion and, when necessary, adjudication by Z.L. Conventional design domains were informed by RoB 2 for randomized studies and ROBINS-I principles for non-randomized studies^{66,67}. Additional human–AI interaction domains assessed whether the erroneous-output condition was clearly defined, whether the reference standard was independent of the automated system, whether a same-case clinical judgment was available before exposure to

automated advice, whether repeated observations from the same clinicians and/or clinical cases were appropriately accounted for, whether automated-output correctness could be distinguished from underlying case difficulty, and whether direction-specific outcomes were completely reported. These domains were applied consistently across all core cohorts and were not combined into a numerical quality score. TRIPOD + AI and PROBAST + AI informed assessment of model and reference-standard reporting but were not used as numerical scores^{68,69}. The final appraisal is summarized visually in Supplementary Fig. 1.

Statistical analysis

Adjusted odds ratios comparing clinician correctness under correct versus incorrect automated advice were log-transformed. Multiple strata from the same cohort were combined before meta-analysis to preserve independence. Random-effects models used restricted maximum-likelihood estimation with modified Hartung–Knapp confidence intervals^{70,71}. In addition to adjusted odds ratios, absolute risk reductions (ARRs) were calculated for each cohort contributing to the primary meta-analysis to provide an absolute measure of effect. Because baseline risks and study structures differed across cohorts, ARRs were reported at the study or specialty-stratum level and were not pooled. We audited the three cohorts contributing to the primary adjusted meta-analysis for dependence arising from repeated observations within clinicians and/or clinical cases. Gaube et al. used multilevel modeling with participant-level random effects, and Kücking et al. used a generalized linear mixed-effects model with a participant random intercept. For the Novotny/Bond cohort, the published logistic regression analysis did not clearly document adjustment for repeated observations contributed by the same physicians and cases. We therefore retained the published adjusted estimate but considered its confidence interval potentially overprecise if within-clinician or within-case correlation was not fully accounted for. A sensitivity analysis excluding this cohort was performed to assess the robustness of the pooled association.

Real-world non-correction proportions were synthesized on the logit scale with Wilson study-level confidence intervals; τ^2 , I^2 , and a 95% prediction interval were reported. A beta-binomial random-effects model was used as a robustness analysis. Temporal effects were explored using three complementary approaches: a contemporary sensitivity analysis of the primary adjusted meta-analysis restricted to studies published from 2020 onward; descriptive stratification by technology era; and an exploratory random-effects meta-regression of publication year as a continuous moderator of real-world non-correction proportions. Because these temporal analyses contained very few studies, particularly the meta-regression, they were treated as descriptive and hypothesis-generating rather than confirmatory.

Error provenance was considered a clinically and methodologically relevant source of heterogeneity. Formal provenance-specific meta-analysis was undertaken only when sufficiently comparable studies and common outcome definitions were available. Because subdivision of the primary adjusted synthesis would have resulted in strata containing only one or two independent cohorts, provenance-specific pooled estimates were not calculated. Instead, provenance was reported at the study level and incorporated into the stratified descriptive synthesis. Quantitative pooling was restricted to studies estimating sufficiently comparable outcomes; studies with incompatible outcome definitions or dependence structures were synthesized narratively. Direction-specific transitions, mitigation strategies, and clinical consequences were not pooled when opportunity denominators or dependence structures were incompatible. Publication-bias tests were not performed because no homogeneous analysis included at least 10 independent cohorts. Analyses were conducted in R using metafor and custom beta-binomial code⁷².

Data availability

All study-level data supporting the findings of this systematic review and meta-analysis are provided in the Supplementary Information and

accompanying source-data files, including extracted study characteristics, effect-size calculations, and analysis-ready datasets.

Code availability

The code generated and/or analyzed during the current study are not publicly available due to the study-specific nature of the data extraction, processing, and analytical workflows used in this systematic review and meta-analysis, but are available from the corresponding author on reasonable request.

Received: 21 July 2026; Accepted: 4 October 2026;

Published online: 23 October 2026

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 <https://doi.org/10.1038/s41591-018-0300-7> (2019).
2. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 <https://doi.org/10.1186/s12916-019-1426-2> (2019).
3. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 <https://doi.org/10.1136/bmj.m689> (2020).
4. Wang, G. et al. Human–large language model collaboration in clinical medicine: a systematic review and meta-analysis. *npj Digit. Med.* **9**, 195 <https://doi.org/10.1038/s41746-026-02382-2> (2026).
5. Strong, J. et al. Human–AI collaboration in healthcare: a scoping review. *npj Digit. Med.* <https://doi.org/10.1038/s41746-026-02918-6> (2026).
6. You, Y. & Li, X. A scoping review of human–AI collaboration patterns and task divisions in healthcare applications. *npj Digit. Med.* <https://doi.org/10.1038/s41746-026-02912-y> (2026).
7. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **19**, 121–127, <https://doi.org/10.1136/amiajnl-2011-000089> (2012).
8. Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *J. Am. Med. Inform. Assoc.* **24**, 423–431 <https://doi.org/10.1093/jamia/ocw105> (2017).
9. Kostick-Quenet, K. M. & Gerke, S. AI in the hands of imperfect users. *npj Digit. Med.* **5**, 197 <https://doi.org/10.1038/s41746-022-00737-z> (2022).
10. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health* (World Health Organization, 2021).
11. World Health Organization. *Regulatory Considerations on Artificial Intelligence for Health* (World Health Organization, 2023).
12. U.S. Food and Drug Administration. *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff* (FDA, 2022).
13. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-101 <https://doi.org/10.6028/NIST.AI.100-1> (2023).
14. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 <https://doi.org/10.1038/s41591-022-01772-9> (2022).
15. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 <https://doi.org/10.1038/s41591-020-1034-x> (2020).
16. Cruz Rivera, S. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 <https://doi.org/10.1038/s41591-020-1037-7> (2020).
17. Novotny, T. et al. The role of computerized diagnostic proposals in the interpretation of the 12-lead electrocardiogram by cardiology and non-cardiology fellows. *Int. J. Med. Inform.* **101**, 85–92 <https://doi.org/10.1016/j.ijmedinf.2017.02.007> (2017).
18. Bond, R. R. et al. Automation bias in medicine: the influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *J. Electrocardiol.* **51**, S6–S11 <https://doi.org/10.1016/j.jelectrocard.2018.08.007> (2018).
19. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit. Med.* **4**, 31 <https://doi.org/10.1038/s41746-021-00385-9> (2021).
20. Kücking, F. et al. Automation bias in AI-decision support: results from an empirical study. *Stud. Health Technol. Inform.* **317**, 298–304 <https://doi.org/10.3233/SHTI240871> (2024).
21. Kücking, F. et al. Impact of AI recommendation correctness on diagnostic accuracy in clinical decision-making. *Int. J. Med. Inform.* **207**, 106223 <https://doi.org/10.1016/j.ijmedinf.2025.106223> (2026).
22. Hillson, S. D., Connelly, D. P. & Liu, Y. The effects of computer-assisted electrocardiographic interpretation on physicians' diagnostic decisions. *Med. Decis. Mak.* **15**, 107–112 <https://doi.org/10.1177/0272989X9501500202> (1995).
23. Morisbak, B. & Gjesdal, K. Computer-based interpretation of ECG—guiding or misleading? *Tidsskr. Nor. Laegeforen.* **119**, 3441–3444 (1999).
24. Tsai, T. L., Fridsma, D. B. & Gatti, G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J. Am. Med. Inform. Assoc.* **10**, 478–483 <https://doi.org/10.1197/jamia.M1279> (2003).
25. Goodacre, S., Webster, A. & Morris, F. Do computer generated ECG reports improve interpretation by accident and emergency senior house officers? *Postgrad. Med. J.* **77**, 455–457 <https://doi.org/10.1136/pmj.77.909.455> (2001).
26. Bernstein, M. H. et al. Can incorrect artificial intelligence results impact radiologists, and if so, what can we do about it? *Eur. Radiol.* **33**, 8263–8269 <https://doi.org/10.1007/s00330-023-09747-1> (2023).
27. Dratsch, T. et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* **307**, e222176 <https://doi.org/10.1148/radiol.222176> (2023).
28. Frei, A. L. et al. Computer-aided tumor cell fraction estimation by medical students, residents, and pathologists improves inter-observer agreement while highlighting the risk of automation bias. *Virchows Arch.* <https://doi.org/10.1007/s00428-025-04163-w> (2025).
29. Kashiwagi, K. et al. Influence of artificial intelligence on ophthalmologists' judgments in glaucoma. *PLoS ONE* **20**, e0321368 <https://doi.org/10.1371/journal.pone.0321368> (2025).
30. Harada, Y., Katsukura, S., Kawamura, R. & Shimizu, T. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: an open-label randomized controlled study. *Int. J. Environ. Res. Public Health* **18**, 2086 <https://doi.org/10.3390/ijerph18042086> (2021).
31. Harada, Y., Katsukura, S., Kawamura, R. & Shimizu, T. Effects of a differential diagnosis list of artificial intelligence on differential diagnoses by physicians: an exploratory analysis of data from a randomized controlled study. *Int. J. Environ. Res. Public Health* **18**, 5562 <https://doi.org/10.3390/ijerph18115562> (2021).
32. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: empirical results assessing influencing factors. *Int. J. Med. Inform.* **83**, 368–375 <https://doi.org/10.1016/j.ijmedinf.2014.01.001> (2014).
33. Tsai, C.-Y., Wang, S.-H., Hsu, M.-H. & Li, Y.-C. J. Do false positive alerts in naïve clinical decision support system lead to false adoption by physicians? A randomized controlled trial. *Comput. Methods Prog. Biomed.* **132**, 83–91 <https://doi.org/10.1016/j.cmpb.2016.04.011> (2016).

34. Lyell, D. et al. Automation bias in electronic prescribing. *BMC Med. Inform. Decis. Mak.* **17**, 28 <https://doi.org/10.1186/s12911-017-0425-5> (2017).
35. Spitzer, P. et al. The effect of medical explanations from large language models on diagnostic accuracy in radiology. *npj Digit. Med.* **9**, 333 <https://doi.org/10.1038/s41746-026-02619-0> (2026).
36. Ong, K. T. et al. Success and failure of human–AI collaboration in clinical reasoning: an experimental study on challenging real-world cases. *Int. J. Med. Inform.* **211**, 106342 <https://doi.org/10.1016/j.ijmedinf.2026.106342> (2026).
37. Nagendran, M., Festor, P., Komorowski, M., Gordon, A. C. & Faisal, A. A. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. *npj Digit. Med.* **7**, 202 <https://doi.org/10.1038/s41746-024-01200-x> (2024).
38. Festor, P., Nagendran, M., Gordon, A. C., Faisal, A. A. & Komorowski, M. Safety of human–AI cooperative decision-making within intensive care: a physical simulation study. *PLoS Digit. Health* **4**, e0000726 <https://doi.org/10.1371/journal.pdig.0000726> (2025).
39. Qazi, I. A. et al. Automation bias in large language model assisted diagnostic reasoning among AI-trained physicians. Preprint at <https://doi.org/10.1101/2025.08.23.25334280> (2025).
40. Qazi, I. A. et al. Mitigating automation bias in physician-LLM diagnostic reasoning using behavioral nudges: a randomized controlled trial. Preprint at <https://doi.org/10.64898/2026.06.01.26354596> (2026).
41. Jabbour, S. et al. On the limits of selective AI prediction: a case study in clinical decision making. Preprint at <https://doi.org/10.48550/arXiv.2508.07617> (2025).
42. Acosta, J. N. et al. The impact of AI assistance on radiology reporting: a pilot study using simulated AI draft reports. Preprint at <https://doi.org/10.48550/arXiv.2412.12042> (2024).
43. Xu, X. et al. Explainable AI as a double-edged sword in dermatology: the impact on clinicians versus the public. Preprint at <https://doi.org/10.64898/2025.12.19.25342205> (2025).
44. Biro, J. M. et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *npj Digit. Med.* **8**, 222 <https://doi.org/10.1038/s41746-025-01586-2> (2025).
45. Anh, D. et al. Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis. *J. Electrocardiol.* **39**, 343–345 <https://doi.org/10.1016/j.jelectrocard.2006.02.002> (2006).
46. Bae, M. H. et al. Erroneous computer electrocardiogram interpretation of atrial fibrillation and its clinical consequences. *Clin. Cardiol.* **35**, 348–353 <https://doi.org/10.1002/clc.22000> (2012).
47. Lindow, T. et al. Erroneous computer-based interpretations of atrial fibrillation and atrial flutter in a Swedish primary health care setting. *Scand. J. Prim. Health Care* **37**, 426–433 <https://doi.org/10.1080/02813432.2019.1684429> (2019).
48. Escabí-Mendoza, J. et al. Diagnostic accuracy of atrial fibrillation by computerized electrocardiogram analysis versus cardiologist interpretation. *J. Electrocardiol.* **93**, 154147 <https://doi.org/10.1016/j.jelectrocard.2025.154147> (2025).
49. Dreiseitl, S. & Binder, M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artif. Intell. Med.* **33**, 25–30 <https://doi.org/10.1016/j.artmed.2004.07.007> (2005).
50. Bouaud, J. et al. Physicians' attitudes towards the advice of a guideline-based decision support system: a case study with OncoDoc2 in the management of breast cancer patients. *Stud. Health Technol. Inform.* **216**, 264–268 (2015).
51. Wang, D. Y. et al. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *J. Am. Med. Inform. Assoc.* **30**, 1684–1692 <https://doi.org/10.1093/jamia/ocad118> (2023).
52. Southern, W. N. & Arnsten, J. H. The effect of erroneous computer interpretation of ECGs on resident decision making. *Med. Decis. Mak.* **29**, 372–376 <https://doi.org/10.1177/0272989X09333125> (2009).
53. Jabbour, S. et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *JAMA* **330**, 2275–2284 <https://doi.org/10.1001/jama.2023.22295> (2023).
54. Rezazade Mehrizi, M. H. et al. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Sci. Rep.* **13**, 9230 <https://doi.org/10.1038/s41598-023-36435-3> (2023).
55. Prinster, D. et al. Care to explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology* **313**, e233261 <https://doi.org/10.1148/radiol.233261> (2024).
56. Kiyak, Y. S. et al. ChatGPT can make mistakes" warnings fail: a randomized controlled trial. *Med. Educ.* **60**, 138–142 <https://doi.org/10.1111/medu.70056> (2026).
57. Sakamoto, T., Harada, Y. & Shimizu, T. Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for determining physicians' diagnostic accuracy: quasi-experimental study. *JMIR Form. Res.* **8**, e58666 <https://doi.org/10.2196/58666> (2024).
58. Yu, F. et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 <https://doi.org/10.1038/s41591-024-02850-w> (2024).
59. Reverberi, C. et al. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.* **12**, 14952 <https://doi.org/10.1038/s41598-022-18751-2> (2022).
60. Vinas, A., Blanco, F. & Matute, H. Doctors vs algorithms: physicians, too, struggle to learn from evidence that contradicts AI suggestions. *PLoS Digit. Health* **5**, e0001490 <https://doi.org/10.1371/journal.pdig.0001490> (2026).
61. Rosbach, E. et al. Stuck on suggestions: automation bias, the anchoring effect, and the factors that shape them in computational pathology. *MELBA* **2026**, 126–142 <https://doi.org/10.59275/j.melba.2026-87b1> (2026).
62. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 <https://doi.org/10.1038/s41591-019-0548-6> (2019).
63. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 <https://doi.org/10.1136/bmj.n71> (2021).
64. Higgins, J. P. T. et al. (eds) Cochrane Handbook for Systematic Reviews of Interventions, version 6.5.1. Cochrane (2026).
65. Campbell, M. et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ* **368**, l6890 <https://doi.org/10.1136/bmj.l6890> (2020).
66. Sterne, J. A. C. et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* **366**, l4898 <https://doi.org/10.1136/bmj.l4898> (2019).
67. Sterne, J. A. C. et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **355**, i4919 <https://doi.org/10.1136/bmj.i4919> (2016).
68. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 <https://doi.org/10.1136/bmj-2023-078378> (2024).
69. Moons, K. G. M. et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 <https://doi.org/10.1136/bmj-2024-082505> (2025).
70. Veroniki, A. A. et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synth. Methods* **7**, 55–79 <https://doi.org/10.1002/jrsm.1164> (2016).

71. Röver, C., Knapp, G. & Friede, T. Hartung–Knapp–Sidik–Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med. Res. Methodol.* **15**, 99. <https://doi.org/10.1186/s12874-015-0091-1> (2015).
72. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 <https://doi.org/10.18637/jss.v036.i03> (2010).

Acknowledgements

Not applicable.

Author contributions

Di Zeng: Conceptualization; data curation; formal analysis; investigation; methodology; software; visualization; writing – original draft; writing – review & editing. Jianyong Lei: Data curation; formal analysis; investigation; methodology; validation; writing – review & editing. Zhihui Li: Conceptualization; supervision; project administration; methodology; validation; writing – review & editing.

Funding

No funding was granted for the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-026-03408-5>.

Correspondence and requests for materials should be addressed to

Jianyong Lei or Zhihui Li.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.