Supplementary information

Contents

PRISMA Checklist 2020

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction-Paragraphs 1-2
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction-Paragraph 3
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods-Search strategy and study selection-Para 2 Results -Literature search and selection- Para 1
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods -Search strategy and study selection-Para 1
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Methods -Search strategy and study selection-Para 2; Supplementary Table 4
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods -Data extraction-Para 2 Supplementary Table 5 & Supplementary Table 6
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods -Data extraction-Para 2
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Methods -Search strategy and study selection-Para 2 (PICO); Methods-Data extraction-Para 3
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Methods-Data extraction-Para 2
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Methods-Risk of bias assessment

Section and Topic	Item #	Checklist item	Location where item is reported
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Methods-Data synthesis and statistical analysis-Para 2 (Effect measures)
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods-Data synthesis and statistical analysis-Para 2; Results-Meta-analytic Results (subsections)
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Methods-Data synthesis and statistical analysis-Para 2
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Methods-Data synthesis and statistical analysis-Para 1; Throughout Results
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Methods-Data synthesis and statistical analysis-Paras 1 & 3
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Methods-Data synthesis and statistical analysis-Paras3 Results-Sensitivity and Special-Design Analyses
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Methods-Data extraction-Para 2; Methods-Data synthesis and statistical analysis- Paras 2-3
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Methods-Data synthesis and statistical analysis-Para 3; Results-Risk of Bias, Reporting Bias, and GRADE-Para 2
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Methods-Data synthesis and statistical analysis-Para 3
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Results -Literature search and selection- Para 1 Figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Results -Literature search and selection-Para 1
Study characteristics	17	Cite each included study and present its characteristics.	Results -Study characteristics; Table 1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Results -Risk of Bias, Reporting Bias, and GRADE-Para 1; Figure 3; Supplementary Table 1 & 2
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Results -Meta-analytic Results (subsections) Figure 2

Section and Topic	Item #	Checklist item	Location where item is reported
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Results -Study characteristics
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Results -Meta-analytic Results-(subsections) Figure 2
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Results -Meta-analytic Results-Time/Efficiency-Para 1
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Results -Sensitivity and Special-Design Analyses- (Entire section)
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Results -Risk of Bias, Reporting Bias, and GRADE-Para 2
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Results -Risk of Bias, Reporting Bias, and GRADE-Para 3; Supplementary Table 3
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion, Paras 1-6
	23b	Discuss any limitations of the evidence included in the review.	Discussion, Para 7
	23c	Discuss any limitations of the review processes used.	Discussion, Para 7
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion, Para 8 ("Practice implications") & Para 9 ("Future research")
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Methods -Protocol and registration-Para 1
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Methods -Protocol and registration-Para 1
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Methods -Protocol and registration-Para 1
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Fund
Competing interests	26	Declare any competing interests of review authors.	Ethics declarations -Competing interests
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability; Code availability; Supplementary Table 6

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Supplementary Table 1. Risk of bias assessment of randomized controlled trials using the RoB 2 tool.

Risk-of-bias judgments for the nine randomized trials included in the review, assessed across five RoB 2 domains: D1, randomization process; D2, deviations from intended interventions; D3, missing outcome data; D4, measurement of the outcome; D5, selection of the reported result. Overall ratings were “low risk” in 2/9 trials and “some concerns” in 7/9 trials. SC = some concerns.

Citation	Design	Tool	Overall	D 1	D2	D3	D4	D5	Notes
Goh 2024 · JAMA Network Open	Parallel RCT	ROB 2	Low Risk	Low Risk	Low Risk	Low Risk	Low Risk	Low Risk	CONSORT reporting / preregistered; mixed-effects modeling; objective outcomes; blinded assessment.
Goh 2025 · Nature Medicine	Parallel RCT	ROB 2	Some Concern	Low Risk	Some Concern	Low Risk	Low Risk	Low Risk	Randomization with a priori sample-size calculation; participants could not be blinded to AI use (D2); high inter-rater reliability.
Goh 2025 · Communications Medicine	Randomized pre/post (context-level)	ROB 2	Some Concern	Low Risk	Some Concern	Low Risk	Low Risk	Low Risk	Context-level randomization with single-blinded administrators; intervention not blindable (D2).
Kim 2025 · European Radiology	Randomized reader–case crossover	ROB 2	Some Concern	Low Risk	Some Concern	Low Risk	Low Risk	Some Concern	Reader-by-case unit-of-analysis correlation and lack of blinding to AI; crossover-sequence analyses require caution.
Tailor 2025 · JAMA Ophthalmolog	Parallel RCT (QI)	ROB 2	Some Concern	Low Risk	Some Concern	Low Risk	Some Concern	Low Risk	Block randomization; subjective rating scales with unblinded participants (D4/D2).

y									
Gorenshtein 2025 · Journal of Neurology	Three-arm RCT (H+AI vs H vs AI)	R oB 2	Some Concern	Low Risk	Some Concern	Low Risk	Some Concern	Low Risk	Patient randomization with provider rotation; intervention not blindable and outcomes partly subjective.
Yu He Ke 2025; NPJ Digital Medicine	Randomized crossover trial	R oB 2	Some Concern	Some Concern	Some Concern	Low Risk	Low Risk	Some Concern	randomized crossover in routine workflow; blinded quality review; partial adherence and reliance on “actual-use” analyses elevate D2 concerns; overall Some concerns (SC).
Baker 2023 · JAAOS	Parallel RCT, blinded assessors	R oB 2	Low Risk	Low Risk	Low Risk	Low Risk	Low Risk	Low Risk	Blinded adjudication of objective documentation quality; reporting is adequate.
Wu 2025 · Critical Care	Parallel RCT	R oB 2	Some Concern	Low Risk	Some Concern	Low Risk	Low Risk	Low Risk	Stratified randomization; participants could not be blinded to AI use (D2).

Footnote: D1. Randomization process; D2. Deviations from intended interventions; D3. Missing outcome data; D4. Measurement of the outcome;D5. Selection of the reported result. SC, Some Concern

Supplementary Table 2. Risk of bias assessment of the non-randomized study using the ROBINS-I tool.

Risk-of-bias judgments for the single non-randomized study included in the review, assessed across seven ROBINS-I domains: D1, confounding; D2, selection of participants; D3, classification of interventions; D4, deviations from intended interventions; D5, missing data; D6, measurement of outcomes; D7, selection of the reported result. The study was rated “moderate risk” overall, with concerns primarily related to confounding, participant selection, and selective reporting.

Citation	Design	Tool	Overall	D1	D2	D3	D4	D5	D6	D7	Notes_summary
McDuff 2025 · Nature (AMIE)	Clinician–case assisted diagnostic study; within-pair random assignment of assist condition	ROBINS-I	Moderate Risk	Moderate Risk	Moderate Risk	Low Risk	Moderate Risk	Low Risk	Low Risk	Moderate Risk	Within-pair randomized assist condition; clinicians unblinded; blinded outcome adjudication; challenging CPC cases; likely no preregistration.

Footnote: D1. Confounding; D2. Selection of participants; D3. Classification of interventions; D4. Deviations from intended interventions; D5. Missing data; D6. Measurement of outcomes; D7. Selection of the reported result

Supplementary Table 3. GRADE Summary of Findings: Human–AI Collaboration vs Human-Only Across Clinical Performance and Efficiency Outcomes

Outcome	Comparison	Pooled effect & model	Studies (k)	Certainty of evidence*	Key reasons for rating down (GRADE domains)
Diagnostic/interpretive correctness/accuracy (binary)	H+AI vs H	RR = 1.63 (95% CI 0.99–2.70; random-effects model with Paule–Mandel τ^2; I^2 = 63.8%)	2	Low ⊕⊕◯◯	① Risk of bias (RoB 2 mostly some concerns); ② Inconsistency (substantial heterogeneity, I^2=63.8%)*; ③ Imprecision (CI crosses the line of no effect)

Outcome	Comparison	Pooled effect & model	Studies (k)	Certainty of evidence*	Key reasons for rating down (GRADE domains)
Composite/diagnostic reasoning percentage score (continuous)	H+AI vs H	MD = +4.88 percentage points [95% CI +0.65 to +9.12; I^2 = 35.6%; fixed‐effect sensitivity +5.24 p.p. (95% CI 2.06–8.42)]	2	Moderate ⊕⊕⊕◯	① Indirectness (composite score construction not fully homogeneous across trials); ② Imprecision (k=2; prediction interval spans zero)
Time/efficiency (continuous, minutes)	H+AI vs H	MD = +0.40 min (95% CI −4.18 to +4.97; random-effects model with Paule–Mandel τ^2 and HKSJ-adjusted CIs; I^2 = 70.1%)	3	Very low ⊕◯◯◯	① Risk of bias (crossover/adherence/non-ITT issues—RoB2 D2); ② Inconsistency (I^2=70.1%, effects in opposite directions); ③ Indirectness (task/workflow heterogeneity); ④ Imprecision (CI covers both “faster” and “slower”)

RCTs start at high certainty and were rated down per GRADE domains. Abbreviations: RR, risk ratio; MD, mean difference; p.p., percentage points; τ^2, between-study variance; HKSJ, Hartung–Knapp–Sidik–Jonkman adjustment.

* In a small-k sensitivity analysis using a random-effects model with REML τ^2 and HKSJ-adjusted CIs, the pooled RR for diagnostic correctness was 1.59 (95% CI 0.08–32.74; I^2 = 63.8%; τ^2 = 0.0763; 95% PI 0.02–163.67).

Key GRADE judgments by outcome. Based on 10 peer-reviewed studies, we graded three core outcomes.

1) Diagnostic/interpretive correctness (binary): Pooled from Kim 2025 and Wu 2025; effects were directionally consistent but varied in magnitude (I^2 = 63.8%), and the 95% CI crossed the null (0.99–2.70). Principal RoB 2 concerns were lack of blinding to AI use (D2) and limited handling of reader–case/context clustering (D1/D5). The paired reader–case study by McDuff 2025 (Nature; NEJM CPC case set) showed concordant improvements in Top-1/Top-10 accuracy (McNemar P<0.01) and is cited for external consistency only, without upgrading certainty. Overall certainty: low.

2) Composite/diagnostic reasoning percentage scores (continuous): Pooled from Goh 2024 (JAMA Network Open) and Goh 2025 (Nature Medicine); effects were concordant with moderate heterogeneity (I^2 =35.6%). Downgrades were for indirectness (non-identical construction of “composite scores” across trials) and imprecision (k=2 with a prediction interval crossing zero). Overall certainty: moderate.

3) Time/efficiency (minutes): Pooled from Goh 2025 (Nat Med), Goh 2024 (JNO), and Yu He Ke 2025 (NPJ Dig Med); no overall difference with high heterogeneity (I^2 = 70.1%). Downgrades reflected risk of bias (crossover non-adherence, partial “as-treated” analyses, inability to blind, potential learning/carry-over), inconsistency (mixed directions), indirectness (substantial variation in task and workflow), and imprecision (CI spanning both faster and slower). Wu 2025 reported a median reduction from 32.0 to 16.2 minutes but lacked variance estimates and was not pooled. Overall certainty: very low.

Important but non-pooled outcomes (qualitative): Documentation/report quality: Baker 2023—higher PDQI-9 with H+AI but ~36% factual errors; Gorenshtein 2025—H+AI ≈ H on AIGERS, whereas AI-only was significantly worse than H. Cross-disciplinary understanding: Tailor 2025—improved understanding (+9.0 percentage points, 95% CI 0.3–18.2), satisfaction, and clarity, with ~26% errors (mostly low-risk). Guideline concordance/fairness stratification: Goh 2025 (Communications Medicine) suggested improvements but provided only stratified differences without denominators, so not pooled.

Summary: H+AI shows a moderate advantage over H alone in correctness and composite scores (with low and moderate certainty, respectively), whereas time efficiency is highly context-dependent (very low certainty). Major uncertainties arise from unavoidable non-blinding and behavior-related deviations, design/task heterogeneity, and small k leading to imprecision. Recommendations: Future work should prioritize preregistered, multicenter parallel RCTs; standardize and objectify primary outcomes; pre-specify and transparently report handling of pairing/clustering and crossover sequence effects; and employ blinded outcome adjudication and ITT analyses to enhance comparability and external validity.

Supplementary Table 4. Database-specific search strategies

Database	Search strategy
PubMed (n=314)	(("large language model*" OR "LLM" OR "ChatGPT" OR "GPT-4" OR "GPT-3" OR "Claude" OR "Gemini" OR "generative AI" OR "generative artificial intelligence") AND ("clinical decision*" OR "diagnosis"[MeSH] OR "diagnostic reasoning" OR "treatment recommendation*" OR "patient care"[MeSH] OR "medical decision*" OR "clinical reasoning") AND ("physician*" OR "clinician*" OR "doctor*" OR "healthcare professional*" OR "medical professional*") AND ("randomized" OR "RCT" OR "prospective" OR "cohort" OR "MRMC" OR "multi-reader" OR "reader study" OR "clinical trial"[Publication Type])) AND ("0001/01/01"[Date - Publication] : "2025/06/28"[Date - Publication])
Web of Science (n=458)	TS=(("large language model*" OR LLM OR ChatGPT OR GPT-4* OR GPT-3* OR Claude OR Gemini OR "generative AI" OR "generative artificial intelligence") AND ("clinical decision*" OR "diagnostic reasoning" OR "treatment recommendation*" OR "clinical reasoning" OR diagnosis OR "patient care" OR "medical decision*") AND (physician* OR clinician* OR doctor* OR "healthcare professional*" OR "medical professional*") AND (randomized OR RCT OR prospective OR cohort OR MRMC OR "multi-reader" OR "reader study" OR "clinical trial*"))
	Search window: from database inception to June 28, 2025 (inclusive).
Embase (n=391)	('large language' NEAR/2 model* OR llm OR chatgpt OR gpt-4* OR gpt4* OR gpt-3* OR gpt3* OR claude OR gemini OR 'generative ai' OR 'generative artificial intelligence') AND (clinical NEAR/3 decision* OR 'diagnostic reasoning' OR treatment NEAR/3 recommendation* OR 'clinical reasoning' OR diagnosis/de OR 'patient care'/de OR 'medical decision making'/de) AND (physician* OR clinician* OR doctor* OR 'physician'/de) AND (randomized OR rct OR prospective OR cohort OR mrmc OR 'multi-reader' OR 'reader study' OR 'randomized controlled trial'/de OR 'prospective study'/de OR 'cohort analysis'/de)

	Search window: from database inception to June 28, 2025 (inclusive). Results were restricted using the Embase.com date filter.
Cochrane Library (n=51)	#1 ("large language" NEXT model* OR llm* OR chatgpt OR (gpt NEXT 4*) OR gpt4* OR (gpt NEXT 3*) OR gpt3* OR claude OR gemini OR (generative NEXT ai) OR (generative NEXT artificial NEXT intelligence)):ti,ab,kw; #2 ((clinical NEXT decision*) OR "diagnostic reasoning" OR (treatment NEXT recommendation*) OR (clinical NEXT reasoning) OR diagnosis OR (patient NEXT care) OR (medical NEXT decision*)):ti,ab,kw; #3 (physician* OR clinician* OR doctor* OR (healthcare NEXT professional*) OR (medical NEXT professional*)):ti,ab,kw; #4 (randomized OR rct OR prospective OR cohort OR mrmc OR (multi NEXT reader) OR (reader NEXT study) OR (clinical NEXT trial*)):ti,ab,kw; #5 #1 AND #2 AND #3 AND #4
	Search window: from database inception to June 28, 2025 (inclusive).

Supplementary Table 5. Study-level screening and eligibility decisions for full-text candidate records (n = 95)
Caption: PRISMA-aligned screening outcomes for all 95 records with reasons for exclusion and identification of studies meeting the PICO criteria.

Supplementary Table 5a. Erratum reviewed and impact on inclusion (n = 1)

Erratum (n = 1)	Title	Note
1	Publisher Correction: GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial	In our analysis, we included the study 'GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial', taking into account the erratum that was published to acknowledge the equal contributions of Ethan Goh and Robert J. Gallo, as well as the joint supervision by Jason Hom, Jonathan H. Chen, and Adam Rodman. The erratum did not involve any correction of the data, thus the study's findings remain valid for our analysis.

Supplementary Table 5b. Preprints and corresponding journal versions with inclusion decisions (n = 12)

Preprint (n = 12)	Title of Preprint	Corresponding journal version title	Note
1	Evaluating GPT-4 as a Clinical Decision Support Tool in Ischemic Stroke Management	GPT-4 as a Clinical Decision Support Tool in Ischemic Stroke Management: Evaluation Study	Retain only the journal version. And, **exclude**: Offline/retrospective comparison of AI versus human performance

			(comparing the stroke management recommendations provided by GPT-4 with expert opinions and past actual treatments).
2	Evaluating the Accuracy and Reliability of Large Language Models in Assisting with Pediatric Differential Diagnoses: A Multicenter Diagnostic Study	Large Language Models for Pediatric Differential Diagnoses	Retain only the journal version. And, **has been excluded after screening**: Offline comparison of model performance alone versus physicians (comparing the diagnostic accuracy of GPT-3 with pediatricians).
3	The Impact of Large Language Models on Diagnostic Reasoning Among LLM-Trained Physicians: A Randomized Clinical Trial	Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial	Retain only the journal version. And, **has been included**.
4	**The Effect of Medical Explanations from Large Language Models on Diagnostic Decisions in Radiology**	No corresponding journal.	**Preprint included**.
5	ChatGPT Influence on Medical Decision-Making, Bias, and Equity: A Randomized Study of Clinicians Evaluating Clinical Vignettes	Physician Clinical Decision Modification and Bias Assessment in a Randomized Controlled Trial of AI Assistance	Retain only the journal version. And, **has been included**.
6	Assessing Retrieval-Augmented Large Language Model Performance in Emergency Department ICD-10-CM Coding Compared to Human Coders	No corresponding journal	**Exclude**: RAG-LLM–generated codes vs previously provider/coder-assigned codes
7	Evaluation of Large Language	No corresponding journal	**Exclude:**

	Models in Percutaneous Coronary Intervention Decision-Making		The study retrospectively evaluated how different LLMs, including ensemble models, assess the need for PCI compared to past clinical decisions, expert opinions, records, or guidelines. It did not include a 'human-AI collaboration' component, nor did it involve real-time use of LLMs in clinical decision-making.
8	ChatGPT Achieves Comparable Accuracy to Specialist Physicians in Predicting the Efficacy of High-Flow Oxygen Therapy	ChatGPT Achieves Comparable Accuracy to Specialist Physicians in Predicting the Efficacy of High-Flow Oxygen Therapy	Retain only the journal version. And, **has been exclude:** Offline comparison between AI and specialists
9	ChatGPT Assisting Diagnosis of Neuro-ophthalmology Diseases Based on Case Reports	ChatGPT Assisting Diagnosis of Neuro-Ophthalmology Diseases Based on Case Reports	Retain only the journal version. And, **exclude:** This study is a case report-based comparison between AI and clinicians (comparing the diagnostic performance of ChatGPT with that of doctors).
10	**Real-World Validation of MedSearch: A Conversational Agent for Real-Time, Evidence-Based Medical Question-Answering**	No corresponding journal	**Preprint Include.**
11	Analysis of Artificial Intelligence Decision-Making in Hormone Receptor-Positive / HER2-Negative Early Breast Cancer within Real-World Practice	No corresponding journal	**Excluded**: This study is a controlled comparison of AI versus humans (comparing the consistency/effectiveness of GPT-4o with clinical physicians in decision-making for adjuvant therapy of early-stage HR+/HER2- breast cancer), without a 'human+AI collaboration'

			experimental arm, and is categorized as C (H vs AI).
12	**From Tool to Teammate: A Randomized Controlled Trial of Clinician-AI Collaborative Workflows for Diagnosis**	No corresponding journal	**Preprint, Include.**

Supplementary Table 5c. Full-text screening of PICO-aligned studies and final inclusion decisions (n = 82)

NO	Title	Design/Context	Human-only arm	Human+AI arm	AI-only arm	Main comparison	Primary outcome domain / notes	Screening decision
1	International Society on Thrombosis and Haemophilia-A: Critical Appraisal	Guideline/methodological appraisal	No	No	No	—	Non-interventional; unrelated to AI/H–AI collaboration	Exclude
2	Association Between a Germline OCA2 Polymorphism and ER-Negative Breast Cancer Survival	Genetic epidemiology/prognosis	No	No	No	—	Unrelated to AI/clinical collaboration	Exclude
3	Retrieval-Augmented Therapy Suggestion for Molecular Tumor Boards: Algorithmic Development and Validation Study	Algorithm development & validation; tumor board recommendation generation	No	No	Yes	AI-only vs reference/gold standard	Algorithmic performance; lacks H+AI vs H comparator	Exclude
4	The clinician's perspective on sarcoma pathology reporting: impact on treatment	Survey/opinion study	No	No	No	—	No AI intervention	Exclude

	decisions?							
5	Estimation of midazolam concentrations using Bayesian pharmacokinetic software	Software/model evaluation	No	No	No	—	PK software; not LLM collaboration	Exclude
6	Single standardized ultrasonographic measurement…	Imaging measurement methodology (ultrasound)	No	No	No	—	No AI/collaboration	Exclude
7	ChatGPT-4.0 vs Human expertise for epilepsy diagnosis	Comparative offline case/questionnaire study	Yes	No	Yes	AI-only vs Human	Diagnostic/classification accuracy; agreement	Exclude
8	Clinical decision-making in benzodiazepine deprescribing	Education/simulation; AI as standardized patient; medical students	No	No	No	—	Non-licensed trainees; no eligible control/outcomes	Exclude
9	Patient Confidence in Urology: LLMs vs Urologists	Patient survey/perception comparison	Yes	No	Yes	AI-only vs Human	Patient confidence/trust (subjective)	Exclude
10	Impact of AI-Assisted Diagnosis on Patients' Trust and Intention to Seek Care	Randomized web-based vignette RCT	Yes	Yes	No	H+AI vs Human	Patient trust/visit intention (patient-level)	Exclude
11	Economic Outcomes With Precision Diagnostic Testing in Stable Chest Pain (PRECISE)	Diagnostic strategy RCT (non-LLM)	Yes	No	No	—	Non-LLM strategy trial	Exclude
12	Exploring ChatGPT as a supportive tool for sialendoscopy decision making and patient information	Narrative/feasibility report	No	No	Yes	AI-only vs narrative	Descriptive feasibility; no H/H+AI arm	Exclude
1	The Exceptional Responders Initiative:	Protocol/feasibility	No	No	No	—	Unrelated to	Exclude

3	Feasibility of an NCI Pilot Study						LLM-assisted collaboration	
14	Ductal carcinoma in situ diagnosis: French national guidelines (3)	Guideline	No	No	No	—	Non-interventional guideline	Exclude
15	International External Validation of ML Model for 90-Day Mortality after Gastrectomy	External validation of prediction model	No	No	Yes	AI-only vs outcomes	Discrimination/calibration; no H+AI vs H	Exclude
16	Comparative evaluation of a language model vs specialists applying IBD guidelines	Vignette/case interpretation	Yes	No	Yes	AI-only vs Human	Guideline concordance/accuracy	Exclude
17	Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study	Parallel-group RCT (clinical vignettes)	Yes	Yes	No	H+AI vs Human	Composite diagnostic reasoning score; time	Exclude (MedRxiv, Updated in PMID: 39466245)
18	Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial	Parallel-group RCT (larger sample)	Yes	Yes	Yes	H+AI vs Human (AI-only arm also reported)	Expert ratings/diagnostic reasoning	Include
19	AI-Based EMG Reporting: A Randomized Controlled Trial	Prospective single-center RCT; real EDX workflow; AI draft then clinician edits; AI-only reference	Yes	Yes	Yes	H+AI vs Human (AI-only reported)	Report quality; clinician acceptance/usability	Include
20	Decision support system and outcome prediction in necrotizing soft-tissue infections	Retrospective/prospective cohort	No	No	No	—	Prognostic tool; non-trial	Exclude

21	Enhancing Pulmonary Disease Prediction Using LLMs with Hybrid RAG	Model development/external validation	No	No	Yes	AI-only vs gold standard	Algorithmic discrimination/calibration	Exclude
22	Assessing the ChatGPT aptitude in Dermatology	Question bank/objective items	Yes	No	Yes	AI-only vs Human	Answer accuracy	Exclude
23	Understanding Anterior Shoulder Instability Through ML	Prediction model study	No	No	Yes	AI-only vs outcome	Outcome prediction	Exclude
24	AI-powered standardized patients: impact on intern physicians	Web/vignette experiment	Yes	No	Yes	AI-only vs Human	Guideline adherence/decision quality	Exclude
25	Evaluating ChatGPT, Gemini and other LLMs in orthopaedic diagnostics	Prospective comparative evaluation	Yes	No	Yes	AI-only vs Human (and AI vs AI)	Diagnostic accuracy/agreement	Exclude
26	ML-based prediction of appendicitis in ED patients	Model development/validation	No	No	Yes	AI-only vs outcome	Diagnostic prediction	Exclude
27	ChatGPT-assisted classification of postoperative bleeding after MIGS using EHR	Annotation/classification assistance (offline)	No	No	Yes	AI-only vs labels	Algorithmic performance	Exclude
28	Diagnostic capabilities of ChatGPT in ophthalmology	Offline interpretation	Yes	No	Yes	AI-only vs Human	Diagnostic accuracy; recommendation appropriateness	Exclude
29	GPT-4 as a Clinical Decision Support Tool in Ischemic Stroke: Evaluation Study	Offline retrospective/simulation	No	No	Yes	AI-only vs reference standard	Agreement/AUC; not a parallel trial	Exclude
3	Evaluation of AI summaries on	Randomized crossover	Yes	Yes	No	H+AI vs Human	Document	Include

0	interdisciplinary understanding of ophthalmology notes	simulated reading					comprehension; task completion; time	
31	Evaluating chain-of-thought prompting in BCID2 interpretation	Offline/simulated interpretation	Yes	No	Yes	AI-only vs Human	Interpretation accuracy; stewardship appropriateness	Exclude
32	Accuracy of ChatGPT-4 in blood gas analysis	Offline questions/case interpretation	Yes	No	Yes	AI-only vs Human	ABG interpretation accuracy	Exclude
33	Deep Learning-Based Knee Joint Effusion Classification	Imaging DL prospective study	No	No	Yes	AI-only vs ground truth	AUC/sensitivity/specificity	Exclude
34	A large language model improves clinicians' diagnostic performance in complex critical illness cases	Reader–case/simulation study	Yes	Yes	No	H+AI vs Human	Diagnostic/management accuracy; time	Include
35	A comparative study of GPT-4o and ophthalmologists in glaucoma diagnosis	Offline/case interpretation	Yes	No	Yes	AI-only vs Human	Diagnostic accuracy/agreement	Exclude
36	Evaluating GPT-4 (with Vision) on chest radiograph findings	Offline imaging interpretation	Yes	No	Yes	AI-only vs Human (sometimes AI-only)	Finding detection/accuracy	Exclude
37	Impact of an Evidence-Based LLM Diagnostic Decision Support System: RCT	Vignette/offline emergency cases	Yes	Yes	No	H+AI vs Human	Diagnostic/management performance; error rate; time	Exclude
3	Customized LLM to support and improve	Method/exploratory	No	No	Yes	AI-only vs	Usability/potential;	Exclude

8	cervical cancer screening					reference	no comparator arm	
39	CAESAR-P/N instruments for ICU end-of-life experience	Instrument development	No	No	No	—	Unrelated to AI	Exclude
40	Follow-up of tuberculosis patients with TB-info® software (2004)	Retrospective/legacy non-LLM tool	Yes	No	No	Tool vs usual care	Legacy software; non-LLM	Exclude
41	ProFit prosthetic fit/alignment assessment (transtibial)	Tool/instrument development	No	No	No	—	Non-AI/LLM collaboration	Exclude
42	Recommendations for diabetic macular edema management by retina specialists vs LLM platforms	Offline/vignette comparison	Yes	No	Yes	AI-only vs Human	Guideline concordance of recommendations	Exclude
43	Auditing the inference processes of medical-image classifiers via generative AI + physicians	Method/audit	No	No	Yes	AI-only	Methodological auditing; not performance vs clinicians	Exclude
44	Diagnostic efficacy of LLMs in the pediatric emergency department: pilot	Offline case set/simulated interpretation	Yes	No	Yes	AI-only vs Human	Diagnostic accuracy/agreement; no parallel clinical arm	Exclude
45	Leveraging Guideline-Based CDSS with LLMs: Case Study in Breast Cancer	System/case study	No	No	Yes	AI-only	No randomized control	Exclude
46	Evaluation of ChatGPT in predicting 6-month outcomes after traumatic brain injury	Prediction model study	No	No	Yes	AI-only vs outcome	Prognostic prediction	Exclude
47	Risk stratification of pulmonary nodules… PET/CT strategy comparison	Imaging strategy comparison	No	No	No	—	Unrelated to LLMs	Exclude
4	GPT-4 assistance for improvement of	Parallel-group RCT	Yes	Yes	No	H+AI vs Human	Diagnostic/manag	Include

8	physician performance on patient care tasks: RCT	(clinical/simulated)					ement performance; time	
49	GPT-4 analysis of cardiovascular MR reports in suspected myocarditis	Text processing/retrospective evaluation	No	No	Yes	AI-only vs gold standard	Information extraction/consistency	Exclude
50	AI in clinical decision-making: ChatGPT-4 vs Llama2 for otolaryngology cases	Prospective/vignette comparison	Yes	No	Yes	AI-only vs Human (also AI vs AI)	Diagnostic/decision accuracy	Exclude
51	AI suggestions for level of amputation in diabetic foot ulcers vs clinicians	Offline agreement/correlation comparison	Yes	No	Yes	AI-only vs Human	Recommendation agreement	Exclude
52	Currently available LLMs do not provide musculoskeletal treatment recommendations concordant with evidence	Offline guideline concordance study	No	No	Yes	AI vs Guideline	Appropriateness vs guideline	Exclude
53	Performance of GPT-4o in management of toxicological exposures vs emergency medicine residents	Vignette/case interpretation	Yes	No	Yes	AI-only vs Human	Management recommendation accuracy/safety; time	Exclude
54	Arthrosis diagnosis and treatment recommendations in clinical practice: exploratory with GPT-4	Exploratory/offline evaluation	No	No	Yes	AI-only	Recommendation quality	Exclude
55	Clinical feasibility of AI Doctors in outpatient CNS tumor settings	Prospective/simulation feasibility	Yes	No	Yes	AI-only vs Human	Diagnostic/recommendation accuracy; feasibility	Exclude
5	Identifying pathological subtypes of	Conventional imaging ML	No	No	Yes	AI-only vs	Classification AUC	Exclude

6	NSCLC using PET/CT radiomics					ground truth		
57	Performance of GPT-4 for automated prostate biopsy decision-making based on mpMRI	Multi-center evidence study	Yes	No	Yes	AI-only vs Human / guideline	Decision correctness/consistency	Exclude
58	Risk stratification of thyroid nodules: assessing ChatGPT for text-based analysis	Offline text evaluation	No	No	Yes	AI-only vs reference	Usability/consistency	Exclude
59	LLM-powered breast cancer staging from PET/CT reports	Information extraction vs gold standard	Yes	No	Yes	AI-only vs gold standard	Staging extraction accuracy	Exclude
60	Evaluating large language & large reasoning models as decision-support tools in emergency internal medicine	Clinical/simulated decision-support evaluation	Yes	No	No	—	No H+AI arm; not parallel with clinicians using AI	Exclude
61	ChatGPT-4o and OpenAI-o1: Accuracy in refractive surgery	Model/version comparison	Yes	No	Yes	AI-only vs Human / AI vs AI	Diagnostic/recommendation accuracy	Exclude
62	Evaluating multimodal ChatGPT for emergency decision-making of ocular trauma cases	Vignette/case interpretation	Yes	No	Yes	AI-only vs Human	Decision/management accuracy; time	Exclude
63	An Institutional LLM for Musculoskeletal MRI Improves Protocol Adherence and Accuracy	Real-world before–after	Yes	Yes	No	Before–after (H+AI vs historical H)	Protocol adherence; report accuracy; time	Exclude
64	Analyzing Generative AI/ML in auto-assessing schizophrenia's negative symptoms	Algorithm/automatic scale scoring	No	No	Yes	AI-only vs clinician ratings	Score agreement/validity	Exclude
65	Evaluation of a deep learning segmentation tool for spinal cord lesions	Conventional DL segmentation tool evaluation	No	No	Yes	AI-only vs ground truth	Dice/sensitivity	Exclude

	in MS							
66	AI chatbot vs pathology faculty & residents (GU treatment-planning conference)	Conference/live Q&A comparison	Yes	No	Yes	AI-only vs Human	Answer accuracy/appropriateness	Exclude
67	AI versus spinal surgeons in management of controversial spinal surgery scenarios	Questionnaire/case interpretation	Yes	No	Yes	AI-only vs Human	Management recommendation accuracy/agreement	Exclude
68	Comparing diagnostic accuracy of ChatGPT to clinical diagnosis in general surgery consults	Retrospective/simulation comparison	Yes	No	Yes	AI-only vs Human	Diagnostic accuracy	Exclude
69	Benchmark of computational methods to detect digenism in sequencing data	Computational genomics benchmark	No	No	Yes	AI-only vs gold standard	Algorithmic performance	Exclude
70	Automated detection of retinal artery occlusion via self-supervised deep learning	Imaging DL automatic detection	No	No	Yes	AI-only vs reference	Detection performance	Exclude
71	How intelligent is AI? Using ChatGPT to diagnose epilepsy and classify seizures	Case/question interpretation	Yes	No	Yes	AI-only vs Human	Diagnostic/classification accuracy	Exclude
72	Can ChatGPT outperform a neurosurgical trainee? A prospective comparative study	Prospective comparison	Yes	No	Yes	AI-only vs Human	Diagnostic/recommendation accuracy	Exclude
73	ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial	Parallel-group RCT; discharge summaries	Yes	Yes	No	H+AI vs Human	Document quality score; defect rate; time	Include
74	Large language models improve clinical decision making of medical students	Education/simulation; medical students	No	No	No	—	Non-licensed trainees; no	Exclude

	through simulation and feedback: RCT						eligible clinician population	
75	A randomized controlled trial on clinician-supervised generative AI for decision support	Three-arm simulated RCT; non-licensed participants	Yes	Yes	Yes	Guideline (paper) vs ChatGPT vs Supervised-ChatGPT	Decision accuracy/time/workload; wrong population	Exclude
76	Physician clinical decision modification & bias assessment in a randomized controlled trial of AI assistance	Parallel-group RCT (clinical vignettes)	Yes	Yes	No	H+AI vs Human	Composite diagnostic reasoning score; bias/over-reliance	Include
77	Clinical & economic impact of a large language model in perioperative medicine: randomized crossover trial	Randomized crossover RCT; real preoperative clinic	Yes	Yes	No	H+AI (PEACH) vs Human	Time efficiency; documentation quality; economics	Include
78	Assessing the utility of artificial intelligence throughout the triage outpatients: a prospective randomized controlled clinical study	Prospective; ChatGPT and staff triaged separately	Yes	No	Yes	AI-only vs Human	Agreement/ratings only	Exclude
79	Comparison of ChatGPT and Internet Research in Occupational Medicine: Randomized Controlled Trial	Information retrieval/decision RCT	Yes	No	Yes	AI-only vs alternative tools	Performance/time; mixed population; no H+AI arm	Exclude
80	Impact of AI systems for UGI bleeding on clinician trust and learning (poster)	Conference abstract/poster	Yes	Yes	No	H+AI vs Human	Trust/learning (subjective); abstract only	Exclude
81	Towards accurate differential diagnosis with large language models (AMIE)	Parallel-group RCT; clinical vignettes; baseline unassisted	Yes	Yes	Yes	H+AI vs Human (primary);	Diagnostic reasoning	Include

						AI-only vs unassisted Human (secondary)	accuracy & quality	
82	Human-AI collaboration in LLM-assisted brain MRI differential diagnosis: a usability study	Offline reader study (brain MRI DDx)	Yes	Yes	No	H+AI vs Human	Diagnostic accuracy; time; usability	Include
Total included (n = 10)：18, 19, 30, 34, 48, 73, 76, 77, 81, 82 Abbreviations. LLM, large language model; RCT, randomized controlled trial; EDX, electrodiagnostic testing; H+AI, human with AI assistance; CDSS, clinical decision support system. Notes. Numbers (NO) correspond to the master screening list supplied by the author.								

Supplementary Table 6. Raw data extraction from the Article and Preprints document

Article					
Study (First author, Year)	**Journal**	**Arms**	**Primary outcome(s) & metric**	**H+AI vs H (key result)**	**AI-only (if reported)**
McDuff 2025	Nature	H vs H+AI; AI-only	Diagnostic reasoning	Top-10 accuracy: H 33.6% (95%	AMIE model-only

Article					
		reported	quality/accuracy (Top-10, Top-1) on CPC cases	CI 29.2–38.0) → H+AI 51.7% (45.7–57.7). Top-1: H 29.2% (23.6–34.8) → H+AI 59.1% (51.1–67.1). McNemar test P<0.01 (favor H+AI).	performance reported in paper (not needed for meta main effect).
Goh 2024	JAMA Network Open	H vs H+AI	Composite diagnostic reasoning score; time	Primary: no sig difference—median 76% vs 74%; adjusted diff +2 p.p. (95% CI −4 to +8), P=0.60. Time −82 s (95% CI −195 to +31), P=0.20.	Model-only median 92%; +16 p.p. vs H (95% CI +2 to +30), P=0.03.
Goh 2025	Nature Medicine	H vs H+AI (parallel RCT); AI-only analytic	Total performance score; domain scores; time	Primary: +6.5% (95% CI +2.7 to +10.2), P<0.001 (favor H+AI). Domains: management +6.1% (P=0.001); diagnosis +12.1% (P=0.009). Time longer with H+AI: +119 s (95% CI +17 to +221), P=0.02.	Model-only benchmark analyses reported.
Wu 2025	Critical Care	H vs H+AI; AI-only	Diagnostic accuracy; differential quality; time	Top diagnosis accuracy: H 27% → H+AI 58%. DDx quality median 3.0 → 5.0. Time: H 1920 s → H+AI 972 s.	AI-only top accuracy 60%; DDx quality 5.0.

Article					
Goh 2025	Communications Medicine	H vs H+AI	Clinical decision accuracy vs guideline; attitudes; interaction patterns	Accuracy improved by subgroup: White male +18 p.p. (47%→65%); Black female +18 p.p. (63%→80%). 90% of physicians anticipate important future AI role.	—
Gorenshtein 2025	Journal of Neurology	H vs H+AI vs AI-only	EMG/EDX report quality (AIGERS 0–1); usability/trust	AIGERS: H+AI 0.94 (0.11) vs H 0.94(0.13) (P=0.32; no difference). Usability/trust moderate (e.g., trust 3.7/5; efficiency 2.0/5).	AI-only AIGERS 0.70 (0.26) (significantly worse, P<0.001).
Yu He Ke 2025	NPJ Digital Medicine	H+AI (PEACH-assisted) vs Human-only	Primary: documentation time per encounter (min/case). Secondary: documentation quality (blinded human preference; completeness of problem list; errors), user acceptance, and economic impact	Overall: mean difference +0.56 min (95% CI −3.27 to 4.38; p=0.78; ns). Moderate-complexity cases: −5.77 min (−10.93 to −0.61; p=0.03). More-experienced physicians: −4.60 min (−8.70 to −0.50; p=0.03). Blinded raters preferred H+AI notes in 57.1% of pairs (16/28); problem lists were more complete (p=0.05).	-

Article					
				Modeled annual savings SGD 197,501 (USD 146,297), sensitivity range SGD 48,979–197,499.	
Kim 2025	European Radiology	H vs H+AI (reader/usability)	Brain MRI DDx accuracy; efficiency; confidence; LLM output quality	Accuracy higher with H+AI: 61.4% (70/114) vs 46.5% (53/114), P=0.033; no time or confidence difference. LLM responses contained hallucinations in 11.5% (per response basis 5.4%).	—
Baker 2023	JAAOS	Typing vs Dictation vs H+AI (ChatGPT)	Documentation quality (PDQI-9), time, efficiency	PDQI-9: H+AI 35.6 > Dictation 31.6 > Typing 30.4; time: Dictation fastest (43.7 s) < H+AI (69.8 s) < Typing (96.8 s).	H+AI hallucinations in 36% of notes (content errors).
Tailor 2025	JAMA Ophthalmology	H vs H+AI (LLM plain-language summaries)	Cross-disciplinary understanding, satisfaction, clarity	Non-ophthalmology clinicians preferred PLS (85%); understanding +9.0 p.p. (95% CI 0.3–18.2; P=0.01); satisfaction +21.5 p.p. (P<0.001); clarity +23.0 p.p.	Safety: 26% PLS had errors—mostly low-risk; no severe-risk errors reported.

Article					
				(P<0.001); reduced jargon comprehension gap.	
Preprints					
Study (First author, Year)	**Journal**	**Arms**	**Primary outcome(s) & metric**	**H+AI vs H (key result)**	**AI-only (if reported)**
Spitzer, 2025	medRxiv (preprint)	H (control, no LLM) vs H+AI (GPT-4 with Standard / Differential Dx / Chain-of-Thought explanations)	Physician diagnostic accuracy (% correct) on text+imaging cases; follow/override behavior (secondary)	Chain-of-Thought improved accuracy by +12.2% vs control (p=0.001). In subset with consistent LLM advice: H 77.60% vs Standard 89.58% vs Differential 95.83% vs CoT 94.02%.	LLM advice baseline: top-5 accuracy ≈78.3%; top-1 ≈73.3%.
Everett, 2025	medRxiv (preprint)	AI-first vs AI-second clinician workflows (custom GPT system); secondary comparison vs	Total graded vignette score (diagnosis + evidence + next steps)	+9.8% and +6.8% higher total scores vs conventional resources (p<0.0004; p<0.00001); no significant difference between AI-first and AI-second workflows.	AI-alone score ≈90%; not significantly different from either workflow.

Article					
		conventional resources			
Castaño-Villegas, 2025	medRxiv (preprint)	H+AI (MedSearch RAG assistant) vs H (traditional search)	Expert-scored answer validity (1–3 scale), response time, number of searches, acceptability	Arrived at final answers in ~½ the time (~3 min faster) with ~66% fewer searches; validity mean ≈2.8/3, significantly higher across evaluated aspects ($p<0.01$).	Not applicable (tool used within H+AI workflow; no AI-only arm).
H: Human-only; H+AI: Human-AI collaboration; AI-only: AI model performance alone; AIGERS: AI-Generated EMG Report Score; CI: Confidence Interval; CoT: Chain-of-Thought; CPC: Clinico-pathological correlation; DDx: Differential Diagnosis; EMG/EDX: Electromyography / Electrodiagnosis; ns: Non-significant; p.p.: Percentage points; PDQI-9: Physician Documentation Quality Index-9; PLS: Plain-language summaries; RAG: Retrieval-Augmented Generation; SGD: Singapore Dollars; USD: US Dollars.					