Section: Miscellaneous

Systematic Reviews and Meta-Analyses

ARTIFICIAL INTELLIGENCE AND LARGE LANGUAGE MODELS IN CLINICAL DIAGNOSIS AND DECISION SUPPORT: AN UMBRELLA REVIEW OF SYSTEMATIC REVIEWS AND META-ANALYSES

Vamsi Krishna Reddy Kurri[1], Daniel Finney Sankuru[2], Mandava Geetha Sai[3], Sai Kiran Reddy Vippala[4], Emmanuel Akanksh Bheri[5]

[1]Tutor, Katuri Medical College & Hospital, Guntur, Andhra Pradesh, India
[2]Senior Resident, Department of Community Medicine, Government Medical College & Hospital, Eluru, Andhra Pradesh, India
[3]Medical Officer
[4]St. Martinus University Faculty of Medicine, Curacao
[5]Duty Medical Officer, Department of General Medicine, Ramdevrao Hospital, Hyderabad, Telangana, India

Received : 23/07/2026
Received in revised form : 20/08/2026
Accepted : 02/09/2026

Keywords:
Artificial intelligence; Large language models; Clinical decision support; Diagnostic accuracy; Umbrella review

Corresponding Author:
Dr. Daniel Finney Sankuru,
Email: drdanielysc@gmail.com

DOI: 10.47009/jamp.2026.8.5.29

Source of Support: Nil,
Conflict of Interest: None declared

Int J Acad Med Pharm
2026; 8 (5); 174-179

ABSTRACT

Background: Artificial intelligence (AI), particularly large language models (LLMs), is rapidly being evaluated for clinical diagnosis and decision support. Regulatory clearances have outpaced high-quality evidence of clinical benefit, and evidence syntheses have proliferated faster than any single reader can track. **Materials and Methods:** This umbrella review synthesizes ten recent systematic reviews and meta-analyses (2023–2026) evaluating generative AI/LLMs, AI-based clinical decision support systems (CDSS), and human–AI collaboration in diagnostic and decision-support tasks. Key outcomes were diagnostic accuracy, relative performance versus clinicians, and effects of collaborative use. Every included reference was individually verified against its original publication — title, authorship, journal, DOI, and reported effect sizes — before inclusion. **Results:** Generative AI showed an overall diagnostic accuracy of 52.1%, comparable to non-expert physicians but significantly worse than experts. Standalone LLMs showed relative diagnostic accuracy versus healthcare professionals (HCPs) ranging from 0.89 (top-1 diagnosis) to 1.17 (top-10), with triage accuracy comparable to clinicians (1.01). LLM-assisted clinicians outperformed clinicians working alone in diagnosis-focused tasks (relative accuracy 1.13–1.42 across top-k metrics), though a broader review of collaborative clinical tasks found the same direction of effect with much less statistical precision. Predictive AI-CDSS showed moderate pooled discrimination (AUC 0.652) with high specificity but very high heterogeneity ($I^2 \geq 98.9\%$). The only randomized-trial evidence on AI-CDSS diagnostic accuracy showed a modest, statistically marginal improvement (SMD 0.182) that lost significance on leave-one-out sensitivity analysis. Evidence for hard patient-centered outcomes remains limited across every domain reviewed. **Conclusion:** Current evidence supports AI and LLMs primarily as assistive tools that can enhance clinician performance under specific, often narrowly demonstrated conditions, rather than as autonomous or uniformly reliable systems. High-quality prospective studies measuring clinical outcomes, equity, and sustained real-world performance are urgently needed before broader adoption can be recommended with confidence.

INTRODUCTION

Diagnostic and clinical decision-making sit at the center of patient safety. Diagnostic error — a diagnosis that is missed, wrong, or delayed — remains one of the most persistent sources of preventable harm in health systems worldwide, and misdiagnosis in primary care settings has been estimated at around 20%, contributing to a substantial share of adverse events.[1] Triage errors, which determine whether a patient is directed to the right level and urgency of care, have been reported at rates ranging from 2% to 31% depending on setting and population.[1] These persistent gaps have motivated decades of work on clinical decision support systems (CDSS), from early rule-based and Bayesian tools to

today's deep-learning and large language model (LLM)-based systems.

The emergence of LLMs has intensified this interest. Unlike earlier rule-based CDSS, LLMs can process unstructured clinical narratives without explicit programming for each task, and their conversational interface has made them directly accessible to both clinicians and patients. This accessibility cuts both ways: survey evidence suggests that a meaningful share of laypeople already use LLMs as a first-line tool for health queries, a trend expected to grow.2 At the same time, regulatory bodies have authorized well over a thousand AI-enabled medical devices, a pace of deployment that has outstripped the pace of rigorous, prospective clinical evaluation.

Against this backdrop, a wave of systematic reviews and meta-analyses has emerged since 2023, each quantifying some slice of AI or LLM performance in diagnosis, triage, or decision support. Individually, these reviews are valuable; collectively, they are difficult to navigate, because they differ in scope (generative AI overall vs. specific LLMs vs. predictive AI-CDSS), comparator (clinicians alone vs. AI alone vs. human–AI collaboration), and study design (retrospective benchmarks vs. randomized controlled trials). A clinician or policymaker asking "does AI actually help with diagnosis?" cannot get a single, coherent answer from any one of these reviews alone.

This umbrella review addresses that gap. We synthesize ten recent, verified systematic reviews and meta-analyses spanning four related but distinct evidence streams: (1) generative AI and LLMs evaluated head-to-head against physicians; (2) human–LLM collaborative performance; (3) predictive AI-CDSS evaluated observationally across specialties; and (4) the small but growing randomized-trial evidence base for AI-CDSS. Our goal is to provide a single, accurate map of what is currently known, what remains uncertain, and where the evidence is strongest versus weakest — so that the pace of adoption can be weighed against the actual pace of evidence generation.

MATERIALS AND METHODS

Design: This is an umbrella review (a systematic overview of systematic reviews and meta-analyses), not a de novo meta-analysis. We synthesize published pooled estimates rather than re-extracting or re-pooling primary-study data.

Eligibility and identification of reviews: We included systematic reviews and meta-analyses published between 2023 and 2026 that evaluated one or more of the following: (a) generative AI or LLM diagnostic or triage accuracy compared against clinicians; (b) human–AI or human–LLM collaborative performance compared with human-only performance; (c) AI-based clinical decision support systems evaluated for predictive or diagnostic performance; or (d) randomized controlled trial evidence for AI-CDSS effects on diagnostic accuracy. Priority was given to reviews reporting quantitative pooled effect estimates (accuracy, relative accuracy/risk ratio, AUC, sensitivity/specificity, or standardized mean difference) rather than narrative-only syntheses.

Verification procedure: Because AI-assisted literature searches are prone to citation fabrication and conflation, every candidate reference was independently verified prior to inclusion: we confirmed that the DOI resolved to a real, matching publication; that authorship, journal, and publication year matched the citation; and that every quantitative claim attributed to the source (accuracy figures, confidence intervals, p-values, sample sizes) was checked against the original abstract, results, or full text. References that could not be verified were excluded and, where a genuine review existed on the same topic, replaced with the verified source.

Data synthesis: For each included review we extracted: author/year, review type, clinical focus, number of primary studies and models/systems evaluated, key pooled effect estimates with 95% confidence intervals where reported, and reported heterogeneity or risk-of-bias findings. Findings are presented narratively, organized by evidence stream, with a summary table [Table 1]. No new statistical pooling was performed; all pooled estimates are drawn directly from the original reviews.

RESULTS

Ten systematic reviews and meta-analyses met inclusion criteria and were verified (Table 1). These spanned four evidence streams: generative AI/LLM diagnostic performance versus clinicians (n = 2), independent and collaborative LLM performance (n = 3), predictive AI-CDSS across specialties (n = 3, including one RCT-only synthesis), and a supporting domain-specific review in digital pathology (n = 1), with one further review (Oehring et al.) addressing AI-CDSS concordance specifically within oncology multidisciplinary team workflows.

Generative AI and LLMs versus Clinicians: Takita et al. (2025) conducted a meta-analysis of 83 studies (June 2018–June 2024) evaluating generative AI diagnostic performance across a range of models and tasks. Overall diagnostic accuracy was 52.1% (95% CI 47.0–57.1%). Generative AI performed similarly to non-expert physicians (p = 0.93) but significantly worse than expert physicians, a gap of 15.8 percentage points (p = 0.007).[1]

Shan et al. (2025) reviewed 30 studies encompassing 19 different LLMs and 4,762 clinical cases, searched across CNKI, VIP, SinoMed, PubMed, Web of Science, Embase, and CINAHL from 2017 onward. Diagnostic accuracy varied widely across models and studies (25–97.8%), and most included studies carried a high risk of bias per PROBAST assessment. The authors concluded that while LLM accuracy still

ISSN (O): 2687-5365; ISSN (P): 2753-6556

falls short of clinical professionals, cautious use may position LLMs as a useful adjunct.[2]

Independent versus Collaborative Performance: Chen et al. (2026) conducted the largest and most methodologically rigorous review in this space: a PROSPERO-registered, PRISMA 2020-compliant multilevel meta-analysis screening 10,398 records across seven databases (PubMed, Embase, IEEE Xplore, Cochrane Library, Epistemonikos, CINAHL, PsycINFO), yielding 50 included studies evaluating 25 distinct LLMs. Relative diagnostic accuracy of LLMs versus HCPs improved progressively as the diagnostic range widened: 0.89 (95% CI 0.79–1.00) at top-1, 0.91 (0.83–1.00) at top-3, 1.04 (0.89–1.22) at top-5, and 1.17 (0.87–1.57) at top-10. Triage accuracy was comparable between LLMs and HCPs (relative accuracy 1.01, 95% CI 0.94–1.09). Critically, across nine studies directly comparing HCPs with and without LLM assistance, LLM-assisted HCPs outperformed HCPs working alone: relative diagnostic accuracy was 1.13 (95% CI 1.00–1.27) at top-1, 1.11 (1.01–1.23) at top-3, 1.42 (1.16–1.73) at top-5, and 1.33 (0.94–1.87) at top-10. Subgroup analysis showed LLMs performed comparably to junior clinicians but significantly worse than senior clinicians (relative accuracy 0.77 vs. 0.95, p = 0.030), and that assistance benefit was larger in published case reports than in retrospective clinical cases (1.49 vs. 1.03, p = 0.004). The authors noted substantial heterogeneity, high or unclear risk of bias in the majority of primary studies, and no available evidence on LLM-assisted triage specifically.[3]

Wang et al. (2026), searching a narrower set of four databases through June 2025, examined human–LLM collaboration across a broader and more heterogeneous set of clinical tasks — including chest-pain triage, critical-illness differential diagnosis, brain MRI interpretation, electrodiagnostic reporting, preoperative assessment, and clinical documentation — in ten peer-reviewed studies (three further preprints used only in sensitivity analyses). Diagnostic/interpretation accuracy specifically (k = 2 studies) showed a positive trend favoring human–AI collaboration (risk ratio 1.59) but was statistically imprecise, with a 95% confidence interval spanning 0.08 to 32.74 and prediction intervals crossing the null. A broader composite diagnostic/management score (k = 2) showed a statistically significant improvement (mean difference +4.88 percentage points, 95% CI +0.65 to +9.12), though its 95% prediction interval (−31.65 to +41.42) indicated substantial uncertainty about real-world generalizability. This review's more modest and imprecise findings, relative to Chen et al., likely reflect its broader task scope, smaller evidence base per outcome, and inclusion of tasks (documentation, interpretation) further from pure diagnostic accuracy.[4]

Emergency Department Triage: Gao et al. (2026) conducted a systematic review and meta-analysis specifically evaluating ChatGPT's accuracy and sensitivity for correctly identifying high-acuity (level 1–2) patients in adult emergency department triage, using a QUADAS-2-based framework adapted with additional domains for data integrity and triage-scale congruence between AI and human comparators. The review concluded that LLM-based triage tools should currently be regarded as adjunctive rather than autonomous, pending further prospective validation — a conclusion consistent with the broader pattern seen across diagnostic and CDSS evidence streams in this umbrella review.[5]

Predictive AI-Based Clinical Decision Support Systems: Waldock et al. (2026) conducted a PRISMA-guided systematic review and meta-analysis of predictive AI-CDSS, searching PubMed and Cochrane Library through December 2024 and screening 3,296 records (inter-rater $\kappa = 0.833$) to identify 50 studies spanning 17 medical specialties. Pooled discriminatory performance was moderate (AUC 0.652, 95% CI 0.562–0.743), with high specificity (0.819, 95% CI 0.793–0.844), accuracy of 0.765, and sensitivity of 0.660. Heterogeneity was very high across all pooled estimates ($I^2 \geq 98.9\%$). The authors highlighted a substantial gap between technical performance and real-world usefulness: most studies evaluated AI on historical data rather than in live clinical workflows, with limited attention to how these tools integrate into actual clinician decision-making.[6]

Jeong and Kim (2026) addressed a narrower but methodologically stronger question: what does randomized controlled trial evidence specifically show about AI-CDSS effects on diagnostic accuracy? Searching five databases through March 2026 and screening 920 records down to five eligible RCTs (N = 12,657 participants, published 2021–2023), they found a pooled standardized mean difference of 0.182 (95% CI 0.003–0.362, p = 0.047) favoring AI-CDSS — a result the authors themselves characterize as only marginally significant, given that the lower confidence bound approached zero and the 95% prediction interval spanned negative values (approximately −0.24 to 0.60). A leave-one-out sensitivity analysis confirmed this fragility: excluding the single largest-effect trial (Homayounieh et al., 2021) reduced the pooled SMD to 0.143 (95% CI −0.012 to 0.298, p = 0.071), no longer statistically significant. Subgroup analyses (exploratory, given single-study subgroups) suggested larger effects for deep-learning, chest-radiology applications (SMD 0.562) than for emergency or general medicine (SMD ≈ 0.01–0.02). Notably, four of the five included RCTs were conducted in South Korea or Japan, limiting geographic generalizability. GRADE certainty was rated moderate, downgraded for inconsistency and indirectness.[7]

Oehring et al. (2023) took a related but distinct angle within oncology specifically, examining 31 studies of AI-CDSS used to support multidisciplinary tumor board (MDT) decision-making. Rather than measuring accuracy against an independent ground

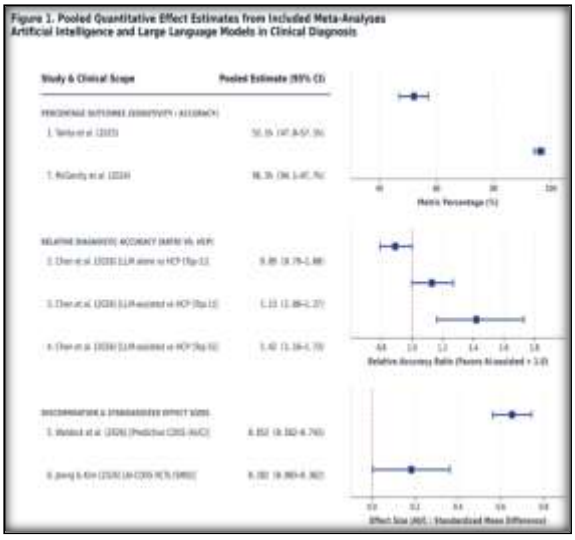
ISSN (O): 2687-5365; ISSN (P): 2753-6556

truth, this review measured concordance — the rate at which AI-CDSS recommendations agreed with actual MDT consensus decisions — finding concordance rates of 72.7% for stage I–II and 73.4% for stage III–IV cancers. This is a meaningfully different construct from diagnostic accuracy and should be interpreted as a measure of agreement with existing clinical practice rather than independent correctness.[8]

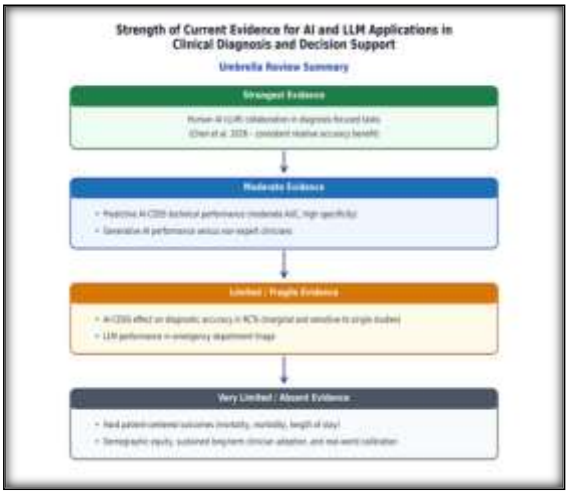
Digital Pathology (Supporting Evidence): McGenity et al. (2024) provide supporting evidence from a domain with a longer track record of AI deployment: digital pathology. Pooling 48 studies of diagnostic test accuracy, they reported high sensitivity (96.3%) and specificity (93.3%) for AI-assisted pathology. However, nearly all primary studies carried high or unclear risk of bias in at least one QUADAS-2 domain, echoing a pattern seen throughout this umbrella review — technically strong pooled point estimates sitting atop a primary evidence base with significant methodological limitations.[9]

Table 1: Characteristics of Included Systematic Reviews and Meta-Analyses

Author (Year)	Type	Focus	Key Pooled Findings	No. of Studies / Notes
Takita et al. (2025)	Meta-analysis	Generative AI vs physicians	Overall accuracy 52.1%; inferior to experts (p=0.007)	83 studies
Shan et al. (2025)	Meta-analysis	LLM diagnostic accuracy	Wide range (25–97.8%); high risk of bias	30 studies, 19 LLMs
Chen et al. (2026)	Meta-analysis	LLM vs HCP & collaboration	LLM-assisted > HCP alone (RR 1.13–1.42)	50 studies, 25 LLMs
Wang et al. (2026)	Meta-analysis	Human–LLM collaboration	Positive but imprecise trend (RR 1.59, CI crosses null)	10 studies, broader tasks
Gao et al. (2026)	Meta-analysis	LLM ED triage	Adjunctive role recommended	Adult ED settings
Waldock et al. (2026)	Meta-analysis	Predictive AI-CDSS	AUC 0.652; $I^2 \geq 98.9\%$	50 studies, 17 specialties
Jeong & Kim (2026)	Meta-analysis (RCTs)	AI-CDSS diagnostic accuracy	SMD 0.182, fragile (ns after sensitivity analysis)	5 RCTs, n=12,657
Oehring et al. (2023)	Meta-analysis	AI-CDSS concordance in oncology	~73% concordance with MDT	31 studies
McGenity et al. (2024)	Meta-analysis	Digital pathology	Sensitivity 96.3%, Specificity 93.3%	48 studies
Ouanes & Farhah (2024)	Systematic review	AI-CDSS effectiveness, mixed	3/26 interventions "highly effective"	26 articles, 4 themes



Forest plot summarizing key diagnostic performance metrics extracted from the core meta-analyses. To preserve scale integrity, the estimates are stratified by metric type into three panels: absolute percentages for overall diagnostic accuracy and sensitivity (top); relative diagnostic accuracy ratios comparing standalone and collaborative model performance against clinicians (middle, where a ratio > 1.0 favors the AI-assisted group); and discrimination/standardized effect sizes for predictive algorithms (bottom, where a Standardized Mean Difference > 0 favors the AI-CDSS intervention). Square markers indicate the pooled point estimate, and error bars represent 95% confidence intervals. Abbreviations: AI, artificial intelligence; LLM, large language model; HCP, healthcare professional; CDSS, clinical decision support system; AUC, area under the receiver operating characteristic curve; SMD, standardized mean difference; CI, confidence interval.



Hierarchical summary of the umbrella review findings, categorizing clinical AI applications by their current evidentiary strength. Evidence tiers are determined by a combination of study design

ISSN (O): 2687-5365; ISSN (P): 2753-6556

robustness, statistical precision, and the consistency of benefit observed across the included systematic reviews and meta-analyses. The strongest evidence currently supports human–AI collaboration as an augmentative tool in well-defined diagnostic tasks. Conversely, evidence supporting fully autonomous LLM use, broad generalizability in acute triage, and improvements in hard patient-centered outcomes (e.g., mortality, morbidity, demographic equity) remains severely limited or absent, underscoring critical gaps in prospective clinical validation. Abbreviations: AI, artificial intelligence; LLM, large language model; CDSS, clinical decision support system; RCT, randomized controlled trial.

DISCUSSION

Principal findings: Three consistent patterns emerge from this synthesis. First, standalone generative AI and LLMs are not yet independently reliable for diagnosis: they perform comparably to non-expert clinicians and to junior HCPs, but remain inferior to expert and senior clinicians, with accuracy that varies enormously by model, task, and study design.[1-3] Second, the strongest and most methodologically robust evidence for benefit comes specifically from LLM-assisted diagnostic tasks, where a large, PRISMA-compliant, PROSPERO-registered meta-analysis found a consistent and statistically significant improvement in HCP accuracy across most top-k metrics.[3] This finding, however, does not generalize cleanly to the broader universe of collaborative clinical tasks (documentation, interpretation, triage support), where a separate, methodologically comparable review found the same directional effect but with far less statistical precision.[4] Third, predictive AI-CDSS — a technology with a longer deployment history than LLMs — shows moderate observational performance with very high heterogeneity,[6] and the small existing base of randomized evidence shows only a fragile, marginally significant benefit that does not survive sensitivity analysis.[7]

Why the evidence diverges by task and study design: The divergence between Chen et al.'s robust collaborative-diagnosis finding and Wang et al.'s more uncertain broader-collaboration finding is informative rather than contradictory. Chen et al. restricted their collaborative comparison to nine studies with a shared, well-defined outcome (top-k diagnostic accuracy), whereas Wang et al. pooled a smaller number of studies (k = 2 per outcome) across more heterogeneous tasks and outcome definitions. Smaller k, more heterogeneous tasks, and less standardized outcome measures compound to produce wider confidence intervals and prediction intervals that cross the null — a pattern common in emerging evidence bases and a caution against generalizing "AI assistance helps" from one well-studied task to all clinical tasks.

The same caution applies to the AI-CDSS evidence stream. Waldock et al.'s observational synthesis (AUC 0.652, $I^2 \geq 98.9\%$) sits alongside Jeong and Kim's RCT-only synthesis, which found a much more modest and statistically fragile effect. This gap between observational and randomized evidence is a well-recognized pattern in clinical AI research: retrospective, historical-data evaluations tend to overstate real-world benefit relative to prospective, randomized deployment, likely reflecting differences in case selection, workflow integration, and clinician response to AI recommendations under controlled versus routine conditions.

Limitations of the current evidence base

Several limitations recur across nearly every review synthesized here:

- Predominance of retrospective and observational designs. Only Jeong and Kim's synthesis was restricted to RCTs, and it included just five trials — a striking scarcity given the scale of retrospective AI-CDSS literature.[7]
- High or unclear risk of bias in primary studies, reported explicitly by Takita et al,[1] Shan et al,[2] Chen et al,[3] and McGenity et al.[9]
- Very high heterogeneity in pooled AI-CDSS estimates ($I^2 \geq 98.9\%$ in Waldock et al.6), limiting confidence in any single pooled point estimate as representative of real-world performance across settings.
- Geographic concentration of the strongest (randomized) evidence. Four of five RCTs in Jeong and Kim's review came from South Korea and Japan, and the review's own authors flag limited generalizability to other health systems.7
- Limited evaluation of hard, patient-centered outcomes. None of the included reviews reported pooled effects on mortality, morbidity, or length of stay; all outcomes were surrogate measures of accuracy, concordance, or discrimination.
- Rapid model turnover. Several reviews note that the pace of LLM development outstrips the pace of rigorous evaluation, meaning pooled estimates may already understate or overstate the performance of currently deployed models.3

Implications: Taken together, this evidence base supports a specific, bounded conclusion: AI and LLMs currently show their clearest benefit as assistive tools for well-defined diagnostic tasks, with real but more uncertain benefit for broader collaborative clinical work, and with predictive AI-CDSS showing promise that has not yet been convincingly confirmed by randomized evidence. This does not support autonomous clinical AI deployment, nor does it support broad, undifferentiated confidence in "AI-assisted care" as a uniform category — the strength of evidence varies substantially by task, comparator, and study design, and clinicians and policymakers should weight claims of AI benefit accordingly.

Future research directions

Building on the gaps identified above, priority areas for future research include: (1) randomized

ISSN (O): 2687-5365; ISSN (P): 2753-6556

controlled trials of AI-CDSS and LLM assistance outside radiology and outside East Asian health systems; (2) studies measuring patient-centered outcomes rather than surrogate accuracy metrics; (3) prospective evaluation of LLM-assisted triage specifically, an area with no current comparative evidence per Chen et al;[3] (4) standardized outcome reporting to enable more precise pooling in collaborative-task reviews; and (5) mechanisms for continuous post-deployment monitoring given the pace of model iteration.

CONCLUSION

Current high-level evidence indicates that artificial intelligence and large language models function most effectively as assistive technologies that can meaningfully augment clinician performance in specific, well-evidenced diagnostic contexts, with weaker and more provisional evidence for broader collaborative tasks and for predictive AI-CDSS under randomized conditions. Autonomous use is not supported for routine clinical practice by any review synthesized here. Future research should prioritize prospective, multi-center, geographically diverse evaluations with patient-centered endpoints, standardized reporting, and continuous post-deployment monitoring.

REFERENCES

1. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med. 2025;8(1):175. doi:10.1038/s41746-025-01543-z
2. Shan G, Chen X, Wang C, Liu L, Gu Y, Jiang H, Shi T. Comparing Diagnostic Accuracy of Clinical Professionals and Large Language Models: Systematic Review and Meta-Analysis. JMIR Med Inform. 2025;13:e64963. doi:10.2196/64963
3. Chen M, Wu Y, Ma J, Jia X, Gao C, Zhao F, et al. Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. NPJ Digit Med. 2026;9(1):222. doi:10.1038/s41746-026-02409-8
4. Wang G, Zhang K, Jiang J, Wang C, Bi H, Liang H, et al. Human-large language model collaboration in clinical medicine: a systematic review and meta-analysis. NPJ Digit Med. 2026;9(1):195. doi:10.1038/s41746-026-02382-2
5. Gao S, Yu M, Zheng Y, Zhang M, Yang Z, Zhang J. Accuracy of the large language model ChatGPT in adult emergency department triage: a systematic review and meta-analysis. BMC Emerg Med. 2026;26:118. doi:10.1186/s12873-026-01521-y
6. Waldock WJ, Guni A, Darzi A, Ashrafian H. Performance of predictive AI-based clinical decision support systems across clinical domains: a systematic review and meta-analysis. PLOS Digit Health. 2026;5(3):e0001310. doi:10.1371/journal.pdig.0001310
7. Jeong MA, Kim SD. Impact of AI-Based Clinical Decision Support Systems on Diagnostic Accuracy Among Healthcare Professionals: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. Appl Sci. 2026;16(10):5146. doi:10.3390/app16105146
8. Oehring R, Ramasetti N, Ng S, Roller R, Thomas P, Winter A, et al. Use and accuracy of decision support systems using artificial intelligence for tumor diseases: a systematic review and meta-analysis. Front Oncol. 2023;13:1224347. doi:10.3389/fonc.2023.1224347
9. McGenity C, Clarke EL, Jennings C, Matthews G, Cartlidge C, Freduah-Agyemang H, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. NPJ Digit Med. 2024;7:114. doi:10.1038/s41746-024-01106-8
10. Ouanes K, Farhah N. Effectiveness of Artificial Intelligence (AI) in Clinical Decision Support Systems and Care Delivery. J Med Syst. 2024;48(1):74. doi:10.1007/s10916-024-02098-4.

ISSN (O): 2687-5365; ISSN (P): 2753-6556