GMR GLOBAL MEDICAL REVIEWS

Volume 1 · Issue 1 · 2026 · Article e0003

OPEN ACCESS

DOI registration pending

Evidence map and narrative synthesis

MULTIMODAL AND GENERATIVE AI-ENABLED CLINICAL DECISION SUPPORT IN MEDICINE: A PRISMA 2020-INFORMED EVIDENCE MAP AND NARRATIVE SYNTHESIS

B.N. Sadykov, T.M. Saliev, K.K. Toguzbaeva, G.S. Bismildina, A. Tolegenkyzy, D.M. Ospanbekova

How to cite: B.N. Sadykov, T.M. Saliev, K.K. Toguzbaeva, G.S. Bismildina, A. Tolegenkyzy, D.M. Ospanbekova. Multimodal and Generative AI-Enabled Clinical Decision Support in Medicine: A PRISMA 2020-Informed Evidence Map and Narrative Synthesis. Global Medical Reviews. 2026;1(1):e0003.

ABSTRACT

Background: Multimodal artificial intelligence (AI), vision-language models, multimodal large language models and tool-using generative AI agents are increasingly proposed for clinical decision support, triage, diagnostic reasoning, treatment preparation and documentation. The literature is expanding quickly, but study designs range from pragmatic trials to external validation studies, simulated consultations, expert-concordance studies and benchmark-like evaluations. This creates a risk of overstating clinical readiness if technical performance is treated as equivalent to patient benefit.

Objective: To identify and synthesize empirical clinical and near-clinical evidence on multimodal and generative AI-enabled clinical decision support, while keeping strict multimodal AI evidence separate from adjacent text-based generative AI clinical decision-support studies.

Materials and methods: A structured, PRISMA 2020-informed evidence map and narrative synthesis was prepared for peer-reviewed human or clinician-facing studies published from 2018 to 9 July 2026. Strict core eligibility was operationalized as an AI system integrating at least two modalities, or explicit multimodal tools, in a clinical decision task. A review-specific expanded contextual set included generative AI or large-language-model clinical decision-support studies using patient-specific clinical information and reporting workflow or clinical endpoints. The auditable set was assembled from PubMed/MEDLINE source-level verification, publisher pages, registry checks when relevant and backward/forward citation chasing. Embase, Scopus and Web of Science strategies are supplied for replication, but these sources were not included in the quantitative flow because archived searches and exports were unavailable. Risk of bias and applicability were assessed with design-appropriate domains, and evidence maturity was judged using explicit criteria for directness, validation, clinical relevance, safety and replication.

Results: Twenty-five reports were assessed in the auditable screening set. Seventeen empirical studies were included: 12 in the strict multimodal AI core synthesis and 5 in the expanded generative AI clinical decision-support contextual set. Strict core studies covered ophthalmology triage, ICU prognosis, pathology, radiology report generation, echocardiography, oncology decision preparation, lymphadenopathy ultrasound diagnosis, emergency-care assessment, multimodal diagnostic evaluation, simulated telehealth, hematology tumor-board support and dental imaging. Strict multimodal studies most often reported diagnostic discrimination, expert concordance, report quality, usability or simulated consultation performance. The most mature real-world clinical-outcome evidence came from adjacent generative AI clinical decision-support studies rather than from strict multimodal systems. Safety, equity, cost-effectiveness and post-deployment surveillance were inconsistently reported.

Conclusions: Current evidence supports multimodal and generative AI primarily as assistive, auditable and human-in-the-loop clinical support. It is most defensible for triage, structured documentation, decision preparation, second reading and specialist-workflow support. Evidence for autonomous deployment or direct improvement in patient-important outcomes remains limited. Future work should prioritize pragmatic trials, prospective safety monitoring, external validation, subgroup fairness analyses and transparent local calibration.

Limitations: Complete archived exports from all intended subscription databases, independent duplicate screening and prospective registration were unavailable. Study designs and endpoints were heterogeneous, and meta-analysis was not appropriate.

Registration: This evidence map was not prospectively registered, and no public protocol was prepared.

Funding: This work received no external funding.

Keywords: multimodal artificial intelligence; large multimodal models; vision-language models; generative AI; clinical decision support; PRISMA 2020; human-in-the-loop; external validation; equity; patient outcomes

Key messages

- Strict multimodal AI studies often show strong diagnostic, triage, reporting or expert-concordance signals, but patient-important outcome evidence remains sparse.
- Text-based generative AI clinical decision-support trials are clinically important, but they are analyzed separately from strict multimodal AI evidence.
- The most defensible near-term role is assistive and auditable: triage support, second reading, documentation, decision preparation and clinician-facing workflow support.
- Safety surveillance, local calibration, subgroup fairness analysis, automation-bias monitoring and model-drift governance are still underdeveloped.

Introduction

Clinical decision-making rarely depends on a single data source. Physicians combine symptoms, examination findings, laboratory values, images, waveforms, prior documentation, medication history, guidelines, patient context and health-system constraints. Multimodal AI is attractive because it attempts to approximate this information environment rather than forcing clinical reasoning into a single image, text or numeric input. The World Health Organization has highlighted the potential clinical uses of large multi-modal models, but also stresses the risks of inaccurate outputs, bias, privacy breaches, automation bias and weak governance [1].

The recent generation of vision-language models, multimodal large language models and tool-using AI agents has changed the type of clinical tasks that can be evaluated. Instead of classifying one image or answering one text question, these systems can combine photographs, radiographs, pathology slides, ultrasound videos, echocardiography videos, clinical notes, structured electronic health record variables, guidelines and retrieval tools. Published examples now include eye emergency triage [2], ICU prognosis [3], pathology assistance [4], radiology report generation [5], echocardiography interpretation [6], oncology decision preparation [7], lymphadenopathy diagnosis [8], simulated multimodal telehealth consultations [9], hematology tumor-board support [10] and dental imaging support [11].

The central problem is that technical performance does not automatically translate into clinical effectiveness. A high area under the curve, a high expert-preference score or strong tumor-board concordance may be useful, but it does not prove fewer deaths, fewer treatment failures, fewer missed diagnoses, better referral appropriateness, lower cost or improved equity. The gap between model capability and patient-important benefit is especially important in AI, because impressive demonstrations can be rapidly interpreted by institutions and vendors as implementation evidence before safety, workflow and subgroup performance have been adequately tested [12-14].

There is also a conceptual problem in the literature. Some of the strongest real-world clinical AI studies evaluate text-based large language model clinical decision support embedded in electronic medical record workflows [15-17], referral coordination [18] or referral justification [19]. These studies may be more clinically mature than many strict multimodal evaluations, but they do not necessarily meet a strict definition of multimodal AI. If such trials are mixed silently with image-text or multi-input systems, the review question becomes blurred and conclusions become vulnerable to overstatement.

For that reason, this review separates two evidence layers. The strict core synthesis includes AI systems that integrate at least two modalities, or explicit multimodal tools, in a clinical decision task. The expanded contextual synthesis includes generative AI or LLM-enabled clinical decision-support systems that use patient-specific clinical information and report clinical, process, safety or workflow endpoints. This separation is not merely semantic. It determines how the evidence should be interpreted, which risk-of-bias tools are most relevant and whether conclusions apply to multimodal reasoning, generative workflow support or both.

The aim of this evidence map is to provide a clinically conservative synthesis of what has been shown, what remains indirect and what conditions are needed before routine deployment. The emphasis is not only on whether models perform well, but on whether they are evaluated in contexts that resemble real clinical decisions, preserve clinician authority, report harms and address generalizability beyond academic development settings.

Objectives

The primary objective was to identify and synthesize empirical clinical and near-clinical evidence on multimodal AI systems used for diagnosis, triage, prognosis, documentation, treatment decision preparation, referral support or clinical workflow support.

A secondary objective was to compare the maturity of strict multimodal AI evidence with an expanded set of generative AI clinical decision-support studies that are not always multimodal in the strict sense but provide important real-world evidence on workflow, safety or patient-level outcomes.

The review questions were: (1) which clinical functions have the strongest evidence; (2) whether reported gains are patient-important or mainly process/technical; (3) what role human-in-the-loop oversight plays; and (4) where safety, equity, local calibration and post-deployment monitoring remain insufficient.

Materials and methods

Design and reporting framework

This review was prepared as a structured, PRISMA 2020-informed evidence map and narrative synthesis. PRISMA 2020 and PRISMA-S principles were used to structure eligibility criteria, information-source reporting, study selection, synthesis and limitations [20-22]. The evidence map reports only directly verifiable source-level records and does not claim database-level yields for sources without archived exports. The PubMed query and replication strategies for other sources are reported with their execution status in Supplementary Table S1, and reporting items and their manuscript locations are summarized in Supplementary Table S5.

Protocol and registration

The review was not prospectively registered, and no publicly accessible protocol was prepared. The eligibility framework and separation of strict multimodal AI from the expanded generative AI clinical decision-support layer were applied during final report-level classification and synthesis. In the absence of a protocol, these decisions should not be interpreted as prospectively specified.

Eligibility criteria

Eligibility was defined by population/setting, index technology, comparator, outcome relevance and publication type. The key methodological choice was to separate strict multimodal AI from adjacent text-based generative AI clinical

decision support. Table 1 shows the decision rules applied in the auditable screening set.

Table 1. Eligibility criteria.

Domain	Definition used in this review
Population/setting	Human patients, clinicians, clinical services or simulated clinical consultations involving patient cases. Animal-only studies and synthetic tasks without patient-level clinical context were excluded.
Strict core intervention/index technology	Multimodal AI, vision-language models, multimodal large language models or AI agents integrating at least two modalities or explicit multimodal tools in a clinical decision task.
Expanded contextual intervention	Generative AI or LLM-enabled clinical decision support using patient-specific clinical information with a clinical, process, safety or workflow endpoint, even when not strictly multimodal.
Comparator	Standard care, unassisted clinicians, expert panels, tumor boards, single-modality AI, guideline CDSS, pre/post workflow comparison or reference diagnosis.
Outcomes	Patient-important outcomes; diagnostic or triage accuracy; treatment or referral decisions; documentation/process outcomes; usability; safety/harms; equity/fairness; economic or resource outcomes.
Eligible designs	Pragmatic trials, cluster trials, prospective or retrospective validation studies, external validation studies, diagnostic accuracy studies, prediction/prognostic studies, usability/implementation studies and carefully defined simulated clinical evaluations.
Exclusions	Narrative reviews, guidance documents, editorials, preprints for the main synthesis, pure engineering benchmarks without a clinical workflow/outcome, and patient-education tools without decision-making or routing endpoints.
Time window	Peer-reviewed studies or reports published from 1 January 2018 to 9 July 2026.

Information sources and search strategy

The auditable information set comprised directly verified records from PubMed/MEDLINE, publisher pages for Nature Portfolio, The Lancet Digital Health, Cell Reports Medicine, BMJ Digital Health & AI, Frontiers, Springer Nature and other journal-specific pages, registry checks when relevant, and backward/forward citation chasing from highly relevant reports. Google Scholar was used only as a citation-discovery aid rather than as a reproducible bibliographic database. The final source-verification and citation update was completed on 9 July 2026.

Search terms combined concepts for multimodal AI, multimodal large language models, vision-language models, medical foundation models, generative AI, AI agents, clinical decision support, diagnosis, triage, tumor boards, treatment decisions, referral and clinical workflow. Supplementary Table S1 distinguishes the PubMed query and sources consulted in the current review from replication strategies prepared for Embase, Scopus and Web of Science. Because no archived RIS/CSV exports or source-specific yields were available for those subscription databases, they were not counted as searched sources in Figure 1.

Data management

Records, source details and eligibility decisions were maintained in a structured review log. Duplicate reports were checked using DOI, title, author list, journal and publication year. Multiple reports arising from the same underlying study were linked to prevent double counting, while report-level records were retained when they contributed distinct methods or outcomes. Because archived database exports were unavailable for some subscription sources, only records that could be verified directly were included in the quantitative screening flow.

Selection process

Eligibility was applied in two review-specific passes during final classification. Pass 1 applied strict multimodal AI criteria. Pass 2 applied the expanded generative AI clinical decision-support contextual criteria. The two-pass structure prevented text-only LLM clinical decision-support studies from being merged with strict multimodal AI evidence. Screening was conducted in a single, non-duplicated review stream rather than by two independent reviewers. Potentially eligible and uncertain reports were retained for full-text assessment, and final classifications were based on the complete report and available supplementary material. Reports were classified as core, expanded contextual, excluded or background-only according to Table 1. Excluded and background-only reports, with the primary reason for each decision, are listed in Supplementary Table S2. The absence of independent duplicate screening is reported as a limitation.

Data collection process

Data were extracted in a single, non-duplicated review stream using the structured template in Supplementary Table S4. Extraction relied on the main publication, online supplementary material and registry information when relevant. No unpublished outcome data were incorporated. When multiple reports described the same underlying study, records were linked at study level and the most complete report was used for each data item. Unclear or unavailable information was coded as not reported, and no missing values were imputed.

Data items and outcome prioritization

All outcomes relevant to the review questions were eligible for extraction. Patient-important outcomes were prioritized for interpretation, followed by treatment or referral decisions, clinical process and workflow outcomes,

diagnostic or triage performance, report quality or expert concordance, usability, safety, equity and resource outcomes. The study authors' prespecified primary endpoint was retained when explicitly identified; otherwise, the principal outcome aligned with the stated study objective was extracted. Other variables included publication year, country and setting, study design, sample size and unit of analysis, participant or clinician characteristics, AI system and version, input modalities or tools, comparator or reference standard, validation setting, prospective status, human-in-the-loop design, safety and harms reporting, equity or fairness analysis, implementation constraints, funding and conflicts of interest. Compatible results and time points relevant to the review questions were retained when available.

Effect measures

Effect estimates were extracted and presented in the metric reported by each study. These included area under the receiver operating characteristic curve, sensitivity, specificity, accuracy, F1 score, odds ratios, risk ratios, hazard ratios, mean or median differences, concordance or preference proportions, consultation or processing time and usability scores. Reported 95% confidence intervals and P values were retained when available. No common effect metric was imposed, estimates were not transformed, and no numerical imputation or recalculation was undertaken for synthesis.

Risk of bias and applicability assessment

Risk of bias and applicability were assessed by design. Randomized and cluster-randomized trials were evaluated with RoB 2 domains; diagnostic accuracy studies with QUADAS-2 domains; and prognostic or prediction-model studies with PROBAST domains and TRIPOD+AI reporting considerations. Early-stage AI evaluations, simulated studies, safety assessments and implementation reports were appraised using DECIDE-AI and CONSORT-AI principles where applicable [12-14,23,24]. Reporting guidance was used to judge completeness and clinical-stage appropriateness, not as a substitute for a risk-of-bias tool. All judgements were made within the same non-duplicated review stream, with risk of bias and applicability considered separately. Domain-level and study-level assessments are presented in Supplementary Tables S3A-S3D.

Assessment of reporting bias

Because meta-analysis was not performed and each clinical-function synthesis contained few heterogeneous studies, funnel plots and regression-based tests for small-study effects were not appropriate. Risk of missing results was considered qualitatively by examining whether protocols or registrations were available, whether prespecified outcomes could be identified, whether reporting emphasized favorable technical endpoints while omitting clinical or safety outcomes, and whether null or unfavorable findings were represented. These considerations informed the interpretation of evidence maturity but were not converted into a numerical score.

Assessment of certainty and evidence maturity

Formal GRADE ratings were not applied because the included designs, interventions and endpoints were highly heterogeneous and no common effect estimate was generated. Evidence maturity was instead judged at the outcome-domain level using an explicit framework that considered directness to a clinical decision, prospective or live evaluation, independence of external validation, comparator quality, presence of patient-important outcomes, sample size and precision, safety and equity reporting, and consistency or replication across settings. Ratings were reported as very low to low, low, low to moderate or moderate. Ratings were lowered for simulated or benchmark-like evaluation, retrospective single-site design, unclear reference standards, limited calibration, absent external validation, sparse harm reporting or lack of subgroup analysis. A high rating required replicated prospective evidence with patient-important outcomes and acceptable risk of bias; no outcome domain met that threshold.

Synthesis methods

Meta-analysis was not performed because the evidence differed substantially by clinical domain, AI architecture, input modality, comparator, validation design and endpoint. Studies were assigned to the strict multimodal core or the expanded generative AI clinical decision-support set during final classification. Within each evidence layer, studies were grouped by clinical function: diagnostic or triage support, prognosis or risk stratification, documentation or report generation, tumor-board or treatment decision preparation, simulated telehealth or usability, and expanded generative AI clinical decision support. Heterogeneity was explored narratively by comparing study design, prospective versus retrospective evaluation, external validation, human-in-the-loop configuration and endpoint type. No missing data were imputed. The expanded evidence layer was interpreted contextually and was not pooled with strict multimodal studies. No formal sensitivity analysis was conducted. Conclusions gave greater weight to prospective, externally validated, real-world and patient-outcome studies than to benchmark-like or simulated evaluations.

Results

Study selection

The auditable screening set contained 25 unique, directly verified reports, with no duplicate reports remaining before eligibility assessment. These reports were assessed at report level because they arose from targeted source verification rather than from a conventional set of archived database exports. Seventeen empirical studies met inclusion criteria: 12 entered the strict multimodal AI core synthesis and 5 entered the expanded generative AI clinical decision-support contextual set. Eight reports were excluded from the empirical synthesis or retained only as background: four reviews, guidance documents or conceptual reports; one report without a clinical decision endpoint; one text-only background trial; and two benchmark-like reports without sufficient clinical workflow relevance. The study-selection process is summarized in Figure 1, and report-level decisions are listed in Supplementary Table S2.

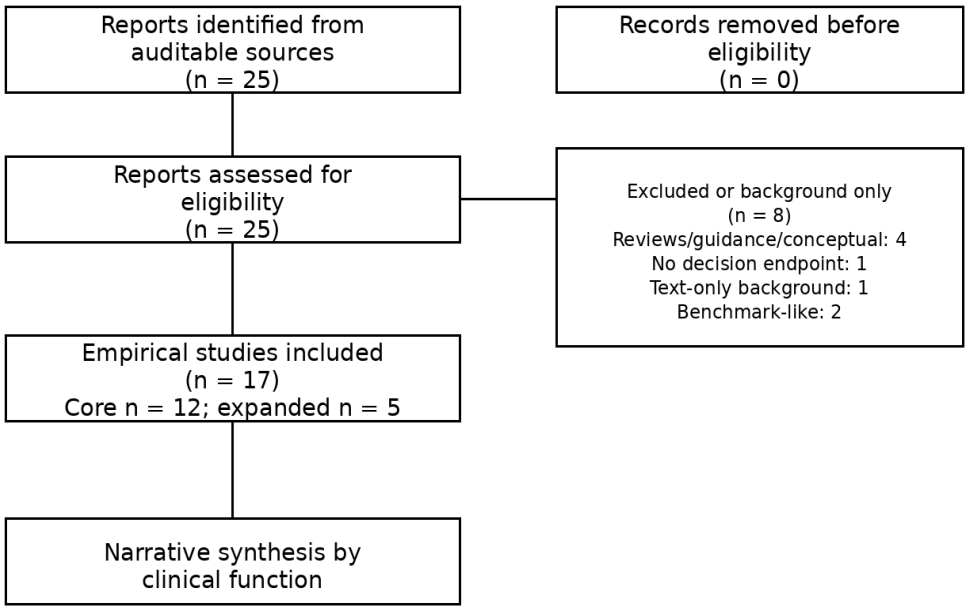

Figure 1. Study selection flow for the evidence map. The figure reports the combined accessible, source-verified set used for this review and should not be interpreted as a complete database-export PRISMA flow; source-specific retrieval counts and archived exports were unavailable.

Characteristics of included studies

The strict core set covered 12 studies across ophthalmology, intensive care, pathology, radiology, echocardiography, oncology, ultrasound diagnosis, emergency care, simulated telehealth, hematology and dentistry. The expanded contextual set added 5 real-world generative AI clinical decision-support studies that were clinically informative but not strict multimodal AI. Table 2 summarizes the included evidence.

Table 2. Included studies and evidence classification.

Study	Clinical function/design	Sample and setting	AI input or tools	Layer	Main quantitative finding and interpretation
Chen et al. [2]	Ophthalmology emergency triage; diagnostic validation with external and pilot referral testing	2,405 internal participants; 103 external participants; prospective pilot referral test	Ocular images + clinical metadata	Core	Internal triage AUC 0.982 (95% CI 0.966-0.998); external AUC 0.988; internal primary-diagnosis accuracy 80.8%. Routing relevance is high, but outcome impact remains indirect.
Lin et al. [3]	ICU prognosis; retrospective development, validation and external testing	3,798 ICU patients from MIMIC-IV and an external hospital	Clinical parameters + chest X-rays	Core	External test AUC 0.82 and F1 score 0.61 versus APACHE II AUC 0.62 and F1 score 0.50. Local calibration and actionability remain decisive.
Lu et al. [4]	Pathology diagnostic assistance; proof-of-concept multimodal evaluation	105 multiple-choice diagnostic cases; 260 open-ended cases assessed by 7 pathologists	Histopathology images + natural-language prompts/clinical context	Core	On 235 consensus open-ended cases, diagnostic accuracy was 78.7% versus 52.3% for GPT-4V. Strong near-clinical signal, but no prospective deployment.
Tanno et al. [5]	Radiology report generation; clinician-AI collaboration evaluation	606 chest radiographs (306 MIMIC-CXR; 300 Indian dataset); 27 board-certified radiologists	Chest X-rays + report text	Core	Clinician-AI reports were equivalent or preferred in 53.6% and 71.2% of cases, versus 44.4% and 51.2% for AI alone in the two datasets. Clinician review remains essential.
Christensen et al. [6]	Echocardiography interpretation; foundation-model development and external validation	1,032,975 videos from 224,685 studies in 99,870 patients; 5,000-video external test set	Echocardiography video + report text	Core	External LVEF mean absolute error was 7.1%; device-detection AUCs ranged from 0.84 to 0.97. Clinical utility still requires prospective workflow testing.
Ferber et al. [7]	Precision oncology decision preparation; realistic multimodal case evaluation	20 realistic multimodal oncology cases	LLM + pathology tools + radiology segmentation + guideline/web tools	Core	Among 245 assessable statements, 223 (91.0%) were factually correct, 16 (6.5%) incorrect and 6 (2.4%) potentially harmful. Appropriate tool use was 87.5%, and decision accuracy increased from 30.3% for GPT-4 alone to 87.2% with the integrated agent. Simulated cases limit effectiveness inference.

Study	Clinical function/design	Sample and setting	AI input or tools	Layer	Main quantitative finding and interpretation
Cao et al. [8]	Lymphadenopathy diagnosis; retrospective, prospective and multicenter validation with reader assistance	7,371 patients; 147,420 BUS/CDFI key frames; multicenter retrospective and prospective cohorts	B-mode ultrasound videos + color Doppler videos + clinical information	Core	In the prospective external cohort, radiologists' mean AUC increased from 0.767 (95% CI 0.705-0.829) to 0.899 (95% CI 0.853-0.944) with AI assistance; external false-positive rate decreased by 9.8%.
Russ et al. [25]	Emergency-care assessment; exploratory feasibility/usability pilot	20 participants in an emergency-care feasibility study	Conversational interface + sensors/vital signs + report generation	Core	Mean System Usability Scale score was 90.6 ± 7.9; at least 80% gave the most positive ratings on key usability and safety items. Sample size was very small and no clinical-outcome comparator was used.
Mahajan et al. [26]	Multimodal diagnostic performance; visual-input contribution study	160 Clinical Picture Quiz cases	Clinical visual data + text cases	Core	Across 160 cases, adding images did not produce a statistically significant aggregate improvement over text-only prompting. The evaluation remains benchmark-like and should not be overinterpreted.
Saab et al. [9]	Simulated multimodal telehealth; randomized blinded OSCE-style evaluation	105 simulated consultations; 18 specialist physicians	Text chat + skin images + ECGs + clinical documents	Core	AMIE was superior on 29 of 32 overall evaluation axes and 7 of 9 multimodal axes. The signal is strong but simulated, with no real-world patient outcomes.
Zoller et al. [10]	Hematology tumor-board support; external validation and prospective silent validation	45 high-complexity benchmark cases; 555 external cases; 64 prospective silent-validation cases	Case-grounded LLM agent + guidelines + case memory/tools	Core	Concordance with tumor-board decisions was 81.8% externally and 82.8% prospectively; hallucinations occurred in 2 of 664 outputs (0.3%). Survival and toxicity outcomes were not tested.
Liu et al. [11]	Dental imaging support; multimodal LLM development/evaluation	450 internal test cases with 4,950 VQA pairs; external test of 19 cases with 3 junior dentists	Orthopantomography images + text/decision-support tasks	Core	ToothXpert achieved an F1 score of 73.73% and processed an external case in a mean 7.09 s; its F1 score exceeded two junior dentists by 1.96 and 2.99 percentage points.
Agweyu et al. [15]	Primary-care CDSS; pragmatic cluster-randomized trial	9,691 patients; 103 clinical officers; 16 primary-care facilities	Text-based LLM/EMR-guideline CDSS	Context	Fourteen-day treatment failure was 2.2% with CDSS versus 2.0% with usual care (adjusted OR 0.77, 95% CI 0.55-1.08; P=0.13). Process gains did not translate into a significant patient-outcome benefit.
Agweyu et al. [16]	Primary-care CDSS safety evaluation	1,469 clinical records from 16 primary-care facilities	LLM-based CDSS record review	Context	Hallucinations were identified in 50 encounters (3.4%, 95% CI 2.5-4.5), and clinical management guidance was aligned with local guidelines in 1,455 encounters (99.0%, 95% CI 98.4-99.5). Potentially harmful recommendations were generated in 115 encounters (7.8%, 95% CI 6.5-9.3), of which 67 were incorporated into the final clinical documentation.
Obong'o et al. [17]	Implementation/usability of EHR-integrated generative AI-CDSS	System logs from all consultations over 8 months; qualitative data from 42 staff members	Usage logs + clinician interviews	Context	Clinicians rated 31% of AI responses; 99.5% of rated responses received a positive rating. Uptake varied by clinician, case complexity and workflow, limiting transferability.
Tao et al. [18]	Primary-to-specialist care transition; randomized controlled trial	2,069 patients and 111 specialists across three randomized groups	Text-based LLM chatbot for referral/coordination	Context	Specialist consultation time fell from 4.41 to 3.14 min (28.7% reduction; P<0.001), while physician-rated care coordination increased from 1.73 to 3.69 (P<0.001).
Saban et al. [19]	CT referral justification; retrospective real-world cohort	6,356 patients in a large European cohort	Text-based LLMs compared with guideline CDSS/expert reference	Context	CT justification accuracy was 92.4% for experts, 88.8% for GPT-4 and 85.2% for Claude. The evidence is clinically relevant but retrospective and not strictly multimodal.

Synthesis by clinical function

Diagnostic and triage support

Diagnostic and triage studies produced the strongest strict multimodal performance signals. EE-Explorer combined ocular images and clinical metadata and reported high triage discrimination in internal and external evaluation, making it directly relevant to eye emergency routing [2]. The lymphadenopathy study was especially important because it used B-mode and color Doppler ultrasound videos with clinical information, included retrospective and prospective multicenter validation, and assessed reader assistance; junior radiologists appeared to benefit from AI-supported interpretation [8]. These studies support multimodal AI as a second-reader or triage layer, but they still do not prove downstream reductions in missed diagnosis, unnecessary referral or patient harm.

Prognosis and risk stratification

PrismICU and EchoCLIP illustrate a different evidence pattern. They show that imaging plus structured or textual clinical information can improve prognostic or interpretation tasks [3,6]. The clinical question is not only whether discrimination improves, but whether predictions are calibrated locally and linked to an actionable management pathway. A model that identifies higher risk without changing treatment, monitoring intensity or resource allocation has limited clinical value.

Documentation and report generation

The chest-radiograph vision-language model study is important because it evaluates clinician-AI collaboration rather than autonomous report generation alone [5]. This is a realistic implementation pathway: AI-generated preliminary reports may reduce workload or improve completeness only if clinicians can review, correct and override them. The main risks are plausible but not fully quantified: false reassurance, subtle omission, anchoring bias and increased downstream work when outputs require extensive correction.

Treatment decision preparation and tumor-board support

The oncology AI agent and HemaGuide show how generative AI may support higher-level decision preparation when constrained by tools, guidelines, retrieval and case memory [7,10]. This architecture is more defensible than unconstrained free-form generation because recommendations can be grounded, audited and compared with expert decision processes. The limitation is equally clear: expert concordance and prospective silent validation are not the same as improved survival, reduced toxicity, fewer inappropriate treatments or better quality of life.

Simulated telehealth, dentistry and usability evidence

Multimodal AMIE performed strongly in randomized blinded simulated telehealth consultations using skin photographs, ECGs and clinical documents [9]. This is a meaningful translational step beyond text-only chatbot evaluation, but it remains simulated and should not be interpreted as deployment evidence. The emergency-care platform had high usability in a small feasibility study, while the dental imaging MLLM expands the field into oral health decision support [11,25]. These studies justify further prospective evaluation, not autonomous clinical use.

Expanded generative AI-CDSS contextual evidence

The expanded generative AI-CDSS set provides the strongest real-world workflow evidence, even though several studies are not strict multimodal AI. The Kenyan cluster-randomized primary-care trial is the key example: it evaluated a generative AI-enabled CDSS embedded in clinical workflow across 16 primary-care facilities and found documentation/process gains without a statistically significant reduction in 14-day treatment failure [15]. This distinction matters. A system can improve the apparent structure of care while failing to demonstrate a short-term patient-outcome benefit. The associated safety and implementation studies add useful information about adverse-event review, clinician experience and adoption patterns [16,17]. The referral-transition and CT-referral studies show that text-based generative AI may influence routing and justification decisions, but they should remain separate from strict multimodal AI conclusions [18,19].

Risk of bias, applicability and evidence maturity

Risk of bias and applicability varied substantially by design. Ten studies were judged to have some concerns, four had high risk for clinical-effect inference, and three were dominated by high applicability concerns; no study provided low-risk evidence across all domains relevant to clinical effectiveness. Recurrent limitations included retrospective or curated sampling, small external cohorts, preference-based or simulated endpoints, incomplete reporting of calibration and missing data, and absence of live workflow or patient-important outcomes. The most clinically oriented strict multimodal evidence came from externally validated or prospective diagnostic studies, especially lymphadenopathy diagnosis and eye triage [2,8]. The weakest applicability came from simulated consultations, small feasibility studies and benchmark-like evaluations [9,25,26]. The expanded generative AI-CDSS evidence included pragmatic trials and implementation studies, but these were not strict multimodal AI and were not used to inflate claims about multimodal systems [15-19]. Design-specific assessments are provided in Supplementary Tables S3A-S3D, while Table 3 summarizes evidence maturity across outcome domains.

Table 3. Maturity and strength of evidence by outcome domain.

Outcome domain	Evidence base	Overall evidence maturity	Main reason for judgement
Patient-important outcomes	Mainly expanded generative AI-CDSS trial evidence	Low to moderate	One pragmatic cluster RCT assessed treatment failure, but strict multimodal core evidence rarely tested hard outcomes [15].
Diagnostic/triage accuracy	EE-Explorer, lymphadenopathy model and related strict multimodal studies	Moderate	External/prospective validation exists for some studies, but calibration, workflow effect and downstream harm remain uncertain [2,8].
Prognosis/risk stratification	PrismICU and EchoCLIP-type evidence	Low to moderate	Performance gains are promising, but actionability and local calibration are insufficiently tested [3,6].
Documentation/report generation	Radiology VLM and CDSS documentation outcomes	Low to moderate	Clinician-AI collaboration is plausible, but error propagation and time burden need prospective testing [5,15].
Treatment decision preparation	Oncology AI agent and HemaGuide	Low to moderate	Concordance and auditability are strengths; patient benefit, toxicity and cost effects are not proven [7,10].
Usability and implementation	Emergency-care pilot and Kenya adoption study	Low	Useful implementation signals, but small samples and setting-specific workflows limit generalizability [17,25].

Outcome domain	Evidence base	Overall evidence maturity	Main reason for judgement
Safety and harms	Sparse adverse-event review and limited error reporting	Low	Rare harms, automation bias, delayed harm and subgroup harms are not adequately measured [16].
Equity/fairness	Mostly indirect or narrative reporting	Very low to low	Few studies report disaggregated performance by sex, age, language, ethnicity, geography, facility type or data completeness.
Economic/resource outcomes	Limited process-cost and referral/resource evidence	Very low to low	Cost-effectiveness, workload substitution and downstream resource use remain underdeveloped [15,19].

Reporting biases and missing evidence

Formal quantitative assessment of reporting bias was not possible. Across outcome domains, the available literature was dominated by positive technical, diagnostic or workflow findings, whereas null clinical-effect results, detailed harm analyses and negative implementation experiences were uncommon. Protocols, registrations and clearly prespecified clinical endpoints were inconsistently available, particularly for simulated and benchmark-like evaluations. Publication bias and selective outcome reporting therefore could not be excluded and were treated as reasons to avoid high evidence-maturity ratings.

Discussion

This evidence map shows a field with strong technical progress but uneven clinical maturity. Strict multimodal AI systems increasingly perform well in diagnostic, triage, report-generation and expert-concordance tasks [2-11,25,26]. The stronger clinical test, however, is whether an AI system changes a decision that matters to patients, clinicians or health systems. On that standard, the evidence remains limited. Most strict multimodal studies do not yet demonstrate fewer treatment failures, fewer missed diagnoses, reduced mortality, reduced complications or cost-effectiveness.

The most important interpretive point is that strict multimodal AI and text-based generative AI-CDSS should not be conflated. Text-based clinical decision-support studies may be more pragmatic and closer to patient outcomes than many multimodal evaluations [15-19]. Yet they answer a different question. If they are used to support claims about multimodal AI without subgroup separation, the review becomes methodologically weak. A defensible synthesis should keep strict multimodal systems and expanded generative AI-CDSS evidence in separate layers.

The most plausible implementation model is assistive rather than autonomous. Across domains, the defensible functions are triage prioritization, second reading, structured documentation, preliminary report generation, referral preparation, guideline retrieval and tumor-board preparation. These uses are clinically relevant because they can reduce cognitive load or improve consistency while preserving clinician authority. They are also reversible and auditable. By contrast, autonomous high-stakes diagnosis or treatment selection remains insufficiently supported by current evidence.

Architecture matters. Systems that are retrieval-grounded, guideline-linked, tool-constrained and auditable appear more credible than unconstrained conversational models. HemaGuide and the oncology AI agent illustrate why this matters: they restrict unsupported generation and allow recommendations to be linked to case features, guidelines or tools [7,10]. That does not remove risk, but it creates a clearer audit trail and makes error analysis more feasible.

Safety reporting is not yet mature. Absence of a reported harm signal in short studies should not be interpreted as proof of safety. AI-related harm can occur through missed diagnoses, over-referral, under-referral, unnecessary treatment, false reassurance, alert fatigue, clinician deskilling, automation bias or delayed downstream consequences. These outcomes require prospective monitoring and predefined adjudication rather than informal post hoc discussion [12,13,16].

Equity is a major unresolved issue. Multimodal AI could narrow gaps by giving non-specialist facilities access to structured decision support, triage logic and specialist-like preparation. It could also widen gaps if development datasets come mainly from tertiary academic centers, high-income settings, English-language documentation or clean electronic health records. Future studies should report performance by sex, age, geography, language, facility level, device type, data completeness and other clinically relevant subgroups whenever feasible.

The practical message for journals and health systems is modest but important. Multimodal and generative AI can already support clinical work, but the current evidence does not justify broad autonomous deployment. Implementation should require local validation, clinician training, human override, audit logs, predefined safety review, subgroup monitoring and a mechanism to pause or update the system after data drift. These conditions should be treated as minimum safeguards, not optional refinements.

Strengths and limitations

The main strength of this review is the explicit separation of strict multimodal AI evidence from expanded generative AI-CDSS evidence. This avoids a common interpretive error in medical AI reviews. The synthesis also includes recent 2025-2026 studies, applies design-specific risk-of-bias logic and avoids claiming clinical effectiveness from benchmark-like or simulated studies.

The limitations are substantial. This is a PRISMA 2020-informed evidence map rather than a prospectively registered systematic review with archived exports from all intended databases. Direct RIS/CSV exports from Embase, Scopus, Web of Science and Google Scholar were not available for this version; database-level yields, deduplication across those sources and quantitative retrieval completeness could therefore not be independently audited. Screening, data extraction and study-level appraisal were

conducted in a structured but non-duplicated review workflow. The evidence base is heterogeneous, and several included studies use simulated, benchmark-like or feasibility designs. Formal meta-analysis, quantitative assessment of reporting bias and GRADE certainty ratings were not appropriate. These limitations reduce confidence in broad claims of effectiveness and make a conservative narrative synthesis more appropriate.

Implications for practice

- Implement multimodal and generative AI first in assistive, auditable and reversible tasks: triage, second reading, documentation support, referral preparation and tumor-board preparation.
- Avoid autonomous deployment in high-stakes diagnosis or treatment decisions without prospective local validation, predefined safety monitoring and clear clinician override.
- Require local calibration, subgroup performance checks, audit logs, adverse-event review and a withdrawal or update mechanism after model drift.
- Treat patient-facing and clinician-facing systems differently; workflow, liability, informed use and safety controls are not the same.

Implications for research

- Conduct pragmatic and cluster-randomized trials with patient-important endpoints rather than only AUC, expert preference or simulation metrics.
- Compare human-in-the-loop models directly: alert-only, second-reader, preliminary-report, recommendation-with-rationale and mandatory sign-off designs.
- Report calibration, missingness, site-level data shift, language support and subgroup performance.
- Measure harms prospectively, including missed diagnoses, over-referral, under-referral, unnecessary treatment, automation bias, clinician deskilling and delayed harm.
- Include community clinics, rural facilities, low-resource settings and non-English workflows.

Conclusions

The available evidence supports assistive use of multimodal and generative AI-enabled clinical decision support in selected clinical functions, but remains insufficient for autonomous decision-making or broad claims of patient-important outcome benefit. The most consistent strict multimodal signals were observed for diagnostic discrimination, triage classification, report generation and expert-concordance outcomes. Real-world patient-outcome evidence is currently more mature in adjacent generative AI-CDSS studies, which are therefore analyzed separately from strict multimodal claims. A more defensible implementation pathway is assistive, human-in-the-loop, locally validated and auditable. The next phase of research should move from model demonstrations toward pragmatic trials, prospective safety surveillance, fairness audits and transparent post-deployment monitoring.

Other information

Registration and protocol: This evidence map was not prospectively registered, and no public protocol was prepared. The two-layer classification was applied during final report-level assessment and synthesis and is therefore described as review-specific rather than prospectively specified. The absence of registration and a protocol is a limitation.

Funding: This work received no external funding.

Competing interests: The authors declare no competing interests.

Data availability: The extracted screening decisions, documented search and replication strategies, accessible-set flow, design-specific risk-of-bias assessments, data extraction template and PRISMA 2020 checklist are provided in the manuscript and Supplementary Tables S1, S2, S3A-S3D, S4 and S5. Archived database exports and source-specific retrieval logs were not available. No individual participant data were used.

Author contributions: A detailed CRediT statement was not provided in the source manuscript.

References

1. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2025.
2. Chen J, Wang R, Luo Y, et al. EE-Explorer: a multimodal artificial intelligence system for eye emergency triage and primary diagnosis. Am J Ophthalmol. 2023;252:253-264. doi:10.1016/j.ajo.2023.04.007.
3. Lin J, Yang J, Yin M, et al. Development and validation of multimodal models to predict the 30-day mortality of ICU patients based on clinical parameters and chest X-rays. J Imaging Inform Med. 2024;37:1312-1322. doi:10.1007/s10278-024-01066-1.
4. Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. Nature. 2024;634:466-473. doi:10.1038/s41586-024-07618-3.
5. Tanno R, Barrett DGT, Sellergren A, et al. Collaboration between clinicians and vision-language models in radiology report generation. Nat Med. 2025;31:599-608. doi:10.1038/s41591-024-03302-1.
6. Christensen M, Vukadinovic M, Yuan N, et al. Vision-language foundation model for echocardiogram interpretation. Nat Med. 2024;30:1481-1488. doi:10.1038/s41591-024-02959-y.
7. Ferber D, El Nahhas OSM, Woelflein G, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat Cancer. 2025;6:1337-1349. doi:10.1038/s43018-025-00991-6.
8. Cao R, Zhu Y, Zhao H, et al. A multimodal feature disentanglement model for lymphadenopathy diagnosis based on BUS and CDFI ultrasound videos: a retrospective, prospective, multicenter study. Eur Radiol. 2026;36(7):6119-6132. doi:10.1007/s00330-026-12409-7.
9. Saab K, Park C, Strother T, et al. Advancing conversational diagnostic AI with multimodal reasoning. Nat Med. 2026;32:1726-1736. doi:10.1038/s41591-026-04371-0.
10. Zoller J, Kalz M, Wu X, et al. Clinical decision support in hematological malignancies using a case-grounded AI agent. Nat Med. 2026. doi:10.1038/s41591-026-04494-4.
11. Liu X, Hung KF, Yu W, et al. Developing and evaluating multimodal large language model for orthopantomography analysis to support clinical dentistry. Cell Rep Med. 2026;7:102652. doi:10.1016/j.xcrm.2026.102652.
12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364-1374. doi:10.1038/s41591-020-1034-x.
13. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. DECIDE-AI: reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence. BMJ. 2022;377:e070904. doi:10.1136/bmj-2022-070904.

14. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
15. Agweyu A, Mwaniki P, et al. Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial. Nat Med. 2026. doi:10.1038/s41591-026-04503-6.
16. Agweyu A, et al. Safety of a large language model-based clinical decision support system in African primary healthcare. Nat Health. 2026;1:607-618. doi:10.1038/s44360-026-00082-5.
17. Obong'o C, Njenga G, Otiangala D, et al. Mixed-methods evaluation of clinician experiences and adoption patterns of an EHR-integrated generative AI-based clinical decision support uptake by clinicians in Kenya. BMJ Digit Health AI. 2026;2:e000207. doi:10.1136/bmjdhai-2025-000207.
18. Tao X, et al. An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nat Med. 2026;32:934-942. doi:10.1038/s41591-025-04176-7.
19. Saban M, Alon Y, Luxenburg O, Singer C, Hierath M, Karoussou Schreiner A, et al. Comparison of CT referral justification using clinical decision support and large language models in a large European cohort. Eur Radiol. 2025;35:6150-6159. doi:10.1007/s00330-025-11608-y.
20. 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.
21. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160. doi:10.1136/bmj.n160.
22. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev. 2021;10:39. doi:10.1186/s13643-020-01542-z.
23. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529-536. doi:10.7326/0003-4819-155-8-201110180-00009.
24. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170:51-58. doi:10.7326/M18-1376.
25. Russ P, et al. Feasibility of a multimodal AI-based clinical assessment platform in emergency care: an exploratory pilot study. Front Digit Health. 2025;7:1657583. doi:10.3389/fdgth.2025.1657583.
26. Mahajan A, Fry C, Zhou L, Bates DW. Evaluating the effect of visual data on multimodal artificial intelligence diagnostic performance. Lancet Digit Health. 2025;7:100938. doi:10.1016/j.landig.2025.100938.
27. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7:e2440969. doi:10.1001/jamanetworkopen.2024.40969.
28. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616:259-265. doi:10.1038/s41586-023-05881-4.
29. Schouten D, Nicoletti G, Dille B, Chia C, Vendittelli P, Schuurmans M, et al. Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications. Med Image Anal. 2025;105:103621. doi:10.1016/j.media.2025.103621.
30. Thirunavukarasu AJ, Li S, Qin P, Nie D, Sanghera R, Lim E, et al. Clinical artificial intelligence applications of vision-language foundation models. PLOS Digit Health. 2026;5(6):e0001453. doi:10.1371/journal.pdig.0001453.

Supplementary Table S1. Search documentation status and reproducibility strategies.

Source	Status in the current review	Query or documentation note
PubMed/MEDLINE	Consulted for source-level retrieval and verification; final update 9 July 2026. A source-specific yield and export filename were not separately archived; PubMed records contributed to the combined accessible set (n=25).	(("multimodal artificial intelligence"[tiab] OR "multimodal AI"[tiab] OR "multimodal large language model*"[tiab] OR "large multimodal model*"[tiab] OR "vision-language model*"[tiab] OR "vision language model*"[tiab] OR "medical foundation model*"[tiab] OR "generative AI"[tiab] OR "large language model*"[tiab] OR "AI agent*"[tiab]) AND ("clinical decision support"[tiab] OR "Decision Support Systems, Clinical"[MeSH] OR diagnosis[tiab] OR triage[tiab] OR "tumor board"[tiab] OR "treatment decision"[tiab] OR referral[tiab] OR "clinical workflow"[tiab]) AND (clinical[tiab] OR patient*[tiab] OR clinician*[tiab] OR physician*[tiab] OR human*[tiab])) AND ("2018/01/01"[dp] : "2026/07/09"[dp])
Publisher pages, registry checks and citation chasing	Consulted for direct record/full-text verification; final update 9 July 2026. No database-style export was available; records contributed to the combined accessible set (n=25).	Nature Portfolio, The Lancet Digital Health, Cell Reports Medicine, BMJ Digital Health & AI, Frontiers, Springer Nature and other journal pages; registry checks when relevant; backward and forward citation chasing from highly relevant reports.
Embase	Replication strategy only. Not counted as a searched database in the current quantitative flow because no archived result set, platform-specific yield or export filename was available.	('multimodal artificial intelligence':ti,ab OR 'multimodal ai':ti,ab OR 'multimodal large language model*':ti,ab OR 'large multimodal model*':ti,ab OR 'vision-language model*':ti,ab OR 'vision language model*':ti,ab OR 'medical foundation model*':ti,ab OR 'generative ai':ti,ab OR 'large language model*':ti,ab OR 'ai agent*':ti,ab) AND ('clinical decision support':ti,ab OR 'decision support system'/exp OR diagnosis:ti,ab OR triage:ti,ab OR 'tumor board':ti,ab OR 'treatment decision':ti,ab OR referral:ti,ab OR 'clinical workflow':ti,ab) AND [humans]/lim AND [2018-2026]/py
Scopus	Replication strategy only. Not counted as a searched database in the current quantitative flow because no archived result set, source-specific yield or export filename was available.	TITLE-ABS-KEY(("multimodal artificial intelligence" OR "multimodal AI" OR "multimodal large language model*" OR "large multimodal model*" OR "vision-language model*" OR "medical foundation model*" OR "generative AI" OR "large language model*" OR "AI agent*") AND ("clinical decision support" OR diagnosis OR triage OR "tumor board" OR "treatment decision" OR referral OR "clinical workflow") AND (clinical OR patient* OR clinician* OR physician* OR human*)) AND PUBYEAR > 2017
Web of Science Core Collection	Replication strategy only. Not counted as a searched database in the current quantitative flow because no archived result set, source-specific yield or export filename was available.	TS=(("multimodal artificial intelligence" OR "multimodal AI" OR "multimodal large language model*" OR "large multimodal model*" OR "vision-language model*" OR "medical foundation model*" OR "generative AI" OR "large language model*" OR "AI agent*") AND ("clinical decision support" OR diagnosis OR triage OR "tumor board" OR "treatment decision" OR referral OR "clinical workflow") AND (clinical OR patient* OR clinician* OR physician* OR human*)) Timespan: 2018-2026

Source	Status in the current review	Query or documentation note
Google Scholar/citation searching	Used only as a citation-discovery aid. Query-specific yields and a stable export were not archived; the proposed queries below are replication aids rather than completed database searches.	Run as separate narrow searches and record the first 100 results per query: "multimodal AI" "clinical decision support" medicine; "vision-language model" "clinical" "decision support"; "large multimodal model" diagnosis clinical; "generative AI" "clinical decision support" randomized trial. Conduct forward/backward citation chasing from Chen 2023, Lu 2024, Tanno 2025, Agweyu 2026, Saab 2026 and Zoller 2026.

Supplementary Table S2. Excluded and background-only reports.

Report	Topic	Decision	Primary reason
Li et al. 2026	Community-codesigned LLM-powered chatbot for primary care	Excluded	Patient education/e-learning outcome; no clinician decision-making or routing endpoint.
Goh et al. 2024 [27]	Large language model influence on diagnostic reasoning	Background only	Important RCT but text-only; used as context, not strict multimodal evidence.
Kim et al. 2025	Multimodal LLMs for chest X-rays with incomplete context	Excluded	Retrospective benchmark/report-generation evaluation without clinical workflow or patient/process endpoint.
Ekingen and Ucdal 2026	Multimodal and unimodal LLMs versus experts in aortic dissection management	Excluded	Multiple-choice/vignette benchmark, not a real clinical workflow or outcome study.
Moor et al. 2023 [28]	Foundation models for generalist medical artificial intelligence	Background only	Conceptual/perspective article, not primary clinical evaluation.
Schouten et al. 2025 [29]	Navigating the landscape of multimodal AI in medicine	Background only	Review/scoping article, useful for context and citation chasing.
Thirunavukarasu et al. 2026 [30]	Clinical AI applications of vision-language foundation models	Background only	Review/overview, not primary empirical evaluation.
WHO 2025 [1]	Ethics and governance of AI for health: guidance on large multi-modal models	Background only	Guidance document, used for governance context, not empirical evidence.

Supplementary Table S3A. QUADAS-2 domain assessment of diagnostic accuracy studies.

Domain judgements use Low, High or Unclear. They were made within the single review stream and were not independently duplicated.

Study	Patient selection	Index test	Reference standard	Flow/timing	Applicability and overall judgement
Chen et al. [2]	Unclear - cohort assembly and spectrum selection were incompletely reported.	Unclear - threshold handling and model lock status were not fully transparent.	Unclear - adjudication and blinding details were incomplete.	Low/unclear - external and pilot testing were reported, but exclusions and missing data were incompletely described.	Some concerns; small external cohort and no downstream outcome evaluation.
Cao et al. [8]	Unclear - multicenter retrospective and prospective cohorts were used, but consecutive enrollment was not fully documented.	Low - independent internal, retrospective external and prospective external evaluations were reported.	Unclear - reference-standard blinding and adjudication across centers require fuller reporting.	Low/unclear - large cohorts were retained, but exclusions and missing video/frame handling were not fully transparent.	Some concerns; strong diagnostic applicability but no downstream biopsy or harm outcome.
Liu et al. [11]	High - curated test material and only 19 external cases.	Unclear - model/version locking and threshold procedures were incompletely reported for clinical use.	Unclear - reference-standard construction and blinding were not fully described.	High - very small external evaluation and limited reporting of case-level exclusions.	High applicability concern; routine dental workflow and patient outcomes were not tested.
Saban et al. [19]	Low/unclear - real-world, consecutive CT referrals across centers, but short sampling windows may affect representativeness.	Low/unclear - identical prompts/settings were used, but model versions are time-sensitive.	Unclear - ESR iGuide with radiologist input was used as the reference, not an outcome-based standard.	Low - the retrospective cohort was evaluated consistently across comparison systems.	Some concerns; clinically relevant but text-only, retrospective and without prospective ordering outcomes.

Supplementary Table S3B. PROBAST domain assessment of prediction-model studies.

Study	Participants	Predictors	Outcome	Analysis	Applicability and overall judgement
Lin et al. [3]	Unclear - retrospective ICU cohorts and selection procedures were incompletely described.	Low/unclear - clinical parameters and chest radiographs were available at prediction, but preprocessing transparency was limited.	Low - 30-day mortality is objective.	High - potential overfitting, incomplete missing-data reporting and limited calibration assessment.	High risk for clinical-effect inference; transportability and actionability across ICUs remain uncertain.
Christensen et al. [6]	Low/unclear - very large retrospective single-system cohort with patient-level splitting and an external dataset.	Low - echocardiography video and paired report text were clearly defined.	Low/unclear - quantitative labels and device outcomes were clinically interpretable, but some were report-derived.	Some concerns - large-scale validation is a strength; calibration and multiple-video clustering remain relevant.	Some concerns; technical benchmark without prospective workflow or outcome impact.

Supplementary Table S3C. RoB 2 domain assessment of randomized trials.

Study	Randomization	Deviations from intervention	Missing outcome data	Outcome measurement	Selection of reported result / overall
Agweyu et al. [15]	Low - clinical officers were randomized as clusters and allocation was reported.	Some concerns - shared facilities, variable uptake and protocol deviations could reduce between-group contrast.	Some concerns - withdrawals, loss to follow-up and exposure misclassification affected the primary analysis set.	Low/some concerns - an expert-adjudicated 14-day composite was used; rare safety outcomes remained imprecise.	Some concerns overall; registration and a prespecified primary outcome were reported.
Tao et al. [18]	Low/unclear - a three-arm randomized design with balanced baseline groups was reported.	Some concerns - participants and staff were not blinded to chatbot use.	Low/unclear - the final analysis included 2,069 participants; attrition details require the trial flow.	Some concerns - consultation time was objective, but coordination and communication outcomes were perception-based.	Some concerns overall; a frozen model and defined trial endpoints were reported.

Supplementary Table S3D. DECIDE-AI-oriented appraisal of early-stage, simulated, safety and implementation studies.

Study	Appraisal approach	Sampling and setting	Comparator/outcome and applicability	Overall
Lu et al. [4]	DECIDE-AI	Curated diagnostic questions and expert ratings; no prospective workflow or patient-level comparator.	Benchmark-like pathology tasks and no prospectively specified clinical endpoint.	High applicability concern.
Tanno et al. [5]	DECIDE-AI / CONSORT-AI principles	Retrospective datasets; blinded clinician comparison is a strength.	Preference-based endpoints may reflect style; no live clinical deployment.	Some concerns.
Ferber et al. [7]	DECIDE-AI	Twenty simulated oncology cases; no randomized clinical decisions.	Statement accuracy, tool use and simulated decision accuracy do not establish patient benefit.	High risk for clinical-effect inference.
Russ et al. [25]	DECIDE-AI	Uncontrolled feasibility study with n=20 and subjective usability outcomes.	Single exploratory setting; no diagnostic accuracy or clinical-outcome comparator.	High risk for clinical-effect inference.
Mahajan et al. [26]	DECIDE-AI	Public quiz cases and repeated prompt-based benchmarking.	No real-time workflow, clinician interaction or patient outcome.	High applicability concern.
Saab et al. [9]	DECIDE-AI	Randomized, blinded exploratory simulation with specialist ratings; not a preregistered clinical trial.	OSCE-style telehealth scenarios are not equivalent to routine patient care.	Some concerns.
Zoller et al. [10]	DECIDE-AI	External and prospective silent validation are strengths; concordance was the main endpoint.	No treatment implementation, survival, toxicity or quality-of-life outcomes.	Some concerns.
Agweyu et al. [16]	Observational safety appraisal / DECIDE-AI	Retrospective record review; subtle or delayed harms may be missed.	Direct primary-care relevance, but rare downstream harms and causal attribution remain uncertain.	Some concerns.
Obong'o et al. [17]	Mixed-methods / DECIDE-AI	Self-selection, self-report and descriptive usage analysis; no counterfactual comparison.	One provider network and implementation context.	High risk for effectiveness inference.

Supplementary Table S4. Data extraction template.

Domain	Variable	Coding or format	Notes
Identification	Study ID and full citation	Author-year; reference number	Use one record per report and link multiple reports from the same study.
Bibliographic	Publication year and status	Year; peer reviewed/preprint	Preprints excluded from the main synthesis.
Context	Country, health system and clinical setting	Free text; single/multicenter	Record facility type and resource setting when available.
Clinical scope	Clinical domain and decision function	Diagnosis, triage, prognosis, documentation, treatment preparation, referral or workflow	Allow more than one function when explicitly defined in the report.
Design	Study design	RCT, cluster RCT, prospective/retrospective validation, diagnostic accuracy, implementation, feasibility or simulation	Record prospective status separately.
Population	Sample size and unit of analysis	Patients, encounters, images, cases, clinicians or facilities	Record development, validation and external cohorts separately.
Population	Participant characteristics	Age, sex, disease spectrum, clinician experience	Code not reported rather than assume absence.
Technology	AI system and version	Model name, version/date, vendor or open-source status	Record model updates and retrieval/tool components.
Technology	Input modalities and tools	Images, video, text, structured EHR, waveforms, sensors, guidelines, web or case memory	Classify strict multimodal versus expanded generative AI-CDSS.
Comparator	Comparator or reference standard	Standard care, unassisted clinicians, expert panel, tumor board, single-modality model or guideline CDSS	Describe blinding and adjudication where reported.
Task	Clinical task and intended user	Free text	Specify patient-facing, clinician-facing or silent evaluation.
Outcomes	Primary endpoint	Definition, time point and measurement method	Retain the authors' prespecified primary endpoint.
Outcomes	Secondary endpoints	Clinical, diagnostic, process, usability, safety, equity or resource outcomes	Record all outcome time points.

Domain	Variable	Coding or format	Notes
Results	Effect estimates	AUC, sensitivity, specificity, accuracy, OR/RR/HR, mean difference, concordance, time or usability score	Include 95% CI and P value when available.
Validation	Validation setting	Internal, temporal, geographic, external, prospective silent or live deployment	Record number of centers and independence from development data.
Implementation	Human-in-the-loop design	No/yes; second reader, preliminary report, recommendation with rationale or mandatory sign-off	Record override and escalation mechanisms.
Safety	Safety and harms reporting	Prespecified, post hoc or not reported	Include hallucination, missed diagnosis, over/under-referral and delayed harm.
Equity	Equity/fairness assessment	Subgroups and performance metrics	Record sex, age, language, ethnicity, geography, facility type, device and missingness.
Implementation	Workflow and implementation constraints	Training, latency, interoperability, usability, workload and local calibration	Record model-drift and audit-log provisions.
Other	Funding and conflicts of interest	Source, role and author/vendor relationships	Record not reported where applicable.
Review process	Reviewer notes and verification status	Free text; verified/not verified	Document unclear items, correspondence and reasons for classification.

Supplementary Table S5. PRISMA 2020 checklist

Section	Topic	Item	PRISMA 2020 requirement	Location/status in this review
TITLE	Title	1	Identify the report as a systematic review.	Title page: identified as a PRISMA 2020-informed evidence map and narrative synthesis; a complete database-export systematic review is not claimed.
ABSTRACT	Abstract	2	Provide a structured summary of the review, including background, objectives, methods, results, limitations, conclusions, registration and funding.	Structured Abstract, including a separate limitations statement; registration/protocol and funding reported.
INTRODUCTION	Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction.
INTRODUCTION	Objectives	4	Provide an explicit statement of the objective(s) or question(s) addressed by the review.	Objectives.
METHODS	Eligibility criteria	5	Specify inclusion and exclusion criteria and how studies were grouped for synthesis.	Methods - Eligibility criteria; Table 1.
METHODS	Information sources	6	Specify all databases, registers, websites, organizations, reference lists and other sources searched or consulted, and the date each source was last searched.	Partially met. Consulted sources and the final verification date are reported; source-specific yields and archived exports were unavailable for intended subscription databases. See Methods and Supplementary Table S1.
METHODS	Search strategy	7	Present the full search strategies for all databases, registers and websites, including filters and limits.	Partially met. The PubMed query and replication strategies are shown in Supplementary Table S1, with explicit status labels; Embase, Scopus and Web of Science strategies were not supported by archived searches/exports.
METHODS	Selection process	8	Specify methods used to decide whether a study met inclusion criteria, including number of reviewers, independence and any automation tools.	Methods - Data management and Selection process; single non-duplicated workflow reported.
METHODS	Data collection process	9	Specify methods used to collect data, including number of reviewers, independence, author contact and automation tools.	Methods - Data collection process; Supplementary Table S4.
METHODS	Data items - outcomes	10a	List and define all outcomes for which data were sought and specify whether all compatible results were collected.	Methods - Data items and outcome prioritization; Supplementary Table S4.
METHODS	Data items - other variables	10b	List and define all other variables for which data were sought and describe assumptions about missing or unclear information.	Methods - Data items and outcome prioritization; Supplementary Table S4.
METHODS	Risk of bias assessment	11	Specify methods used to assess risk of bias, including tools, reviewers, independence and automation tools.	Methods - Risk of bias and applicability assessment; single non-duplicated appraisal reported; Supplementary Tables S3A-S3D.
METHODS	Effect measures	12	Specify the effect measure(s) used for each outcome.	Methods - Effect measures; Table 2.
METHODS	Synthesis methods	13a	Describe how studies were judged eligible for each synthesis.	Methods - Synthesis methods; strict core and expanded contextual evidence layers.

Section	Topic	Item	PRISMA 2020 requirement	Location/status in this review
METHODS	Synthesis methods	13b	Describe methods required to prepare data for presentation or synthesis, including handling of missing data and conversions.	Methods - Data collection process, Data items and Effect measures.
METHODS	Synthesis methods	13c	Describe methods used to tabulate or visually display results.	Table 2; Figure 1; Supplementary Tables S2-S4.
METHODS	Synthesis methods	13d	Describe methods used to synthesize results and justify the choice of methods.	Methods - Synthesis methods; narrative synthesis and rationale for no meta-analysis.
METHODS	Synthesis methods	13e	Describe methods used to explore possible causes of heterogeneity.	Methods - Synthesis methods; stratification by evidence layer, clinical function, design and validation context.
METHODS	Synthesis methods	13f	Describe sensitivity analyses conducted to assess robustness.	No formal sensitivity analysis was conducted. The expanded generative AI-CDSS layer is a separate contextual synthesis, not a sensitivity analysis.
METHODS	Reporting bias assessment	14	Describe methods used to assess risk of bias due to missing results in a synthesis.	Methods - Assessment of reporting bias.
METHODS	Certainty assessment	15	Describe methods used to assess certainty or confidence in the body of evidence.	Methods - Assessment of certainty and evidence maturity; Table 3; formal GRADE not used.
RESULTS	Study selection	16a	Describe search and selection results from records identified to studies included, ideally using a flow diagram.	Results - Study selection; Figure 1; Supplementary Table S2. The figure represents the combined accessible verification set, not a complete database-export flow.
RESULTS	Study selection	16b	Cite studies that might appear eligible but were excluded and explain why.	Supplementary Table S2.
RESULTS	Study characteristics	17	Cite each included study and present its characteristics.	Table 2.
RESULTS	Risk of bias in studies	18	Present risk-of-bias assessments for each included study.	Supplementary Tables S3A-S3D; Results - Risk of bias, applicability and evidence maturity.
RESULTS	Results of individual studies	19	For all outcomes, present summary statistics and effect estimates with precision for each study.	Table 2; narrative synthesis.
RESULTS	Results of syntheses	20a	For each synthesis, briefly summarize characteristics and risk of bias among contributing studies.	Results - Synthesis by clinical function; Supplementary Tables S3A-S3D.
RESULTS	Results of syntheses	20b	Present results of all statistical syntheses, including precision and heterogeneity.	Not applicable: meta-analysis was not performed; reason reported in Methods.
RESULTS	Results of syntheses	20c	Present results of investigations of possible causes of heterogeneity.	Narrative comparison by evidence layer, clinical function, prospective/external validation and endpoint type.
RESULTS	Results of syntheses	20d	Present results of sensitivity analyses.	Not applicable: no formal sensitivity analysis was conducted; the expanded contextual evidence layer is reported separately.
RESULTS	Reporting biases	21	Present assessments of risk of bias due to missing results for each synthesis assessed.	Results - Reporting biases and missing evidence.
RESULTS	Certainty of evidence	22	Present assessments of certainty or confidence in the body of evidence.	Table 3 and Results - Risk of bias, applicability and evidence maturity; categories are evidence-maturity ratings, not GRADE ratings.

Section	Topic	Item	PRISMA 2020 requirement	Location/status in this review
DISCUSSION	Interpretation	23a	Provide a general interpretation of results in the context of other evidence.	Discussion.
DISCUSSION	Limitations of evidence	23b	Discuss limitations of the evidence included in the review.	Discussion - Strengths and limitations.
DISCUSSION	Limitations of review processes	23c	Discuss limitations of review methods and processes.	Discussion - Strengths and limitations.
DISCUSSION	Implications	23d	Discuss implications for practice, policy and future research.	Implications for practice; Implications for research.
OTHER INFORMATION	Registration	24a	Provide registration information, including register name and registration number, or state that the review was not registered.	Abstract; Methods - Protocol and registration; Other information: not registered.
OTHER INFORMATION	Protocol	24b	Indicate where the review protocol can be accessed, or state that no protocol was prepared.	Methods - Protocol and registration; Other information: no public protocol.
OTHER INFORMATION	Amendments	24c	Describe and explain amendments to information provided at registration or in the protocol.	Not applicable because the review was not registered and no public protocol was published.
OTHER INFORMATION	Support	25	Describe sources of financial or non-financial support and the role of funders or sponsors.	Abstract and Other information: no external funding.
OTHER INFORMATION	Competing interests	26	Declare competing interests of review authors.	Other information.
OTHER INFORMATION	Availability of data, code and materials	27	Report which review materials are publicly available and where they can be found.	Other information; Supplementary Tables S1, S2, S3A-S3D, S4 and S5. Archived database exports and source-specific retrieval logs were unavailable.