Computational Diagnostics
AN ELDENHALL JOURNAL

ELDENHALL
OPEN ACCESS PUBLISHING

SYSTEMATIC REVIEW OPEN ACCESS

CITATION
Padhi, A. (2025). Clinical evaluation readiness of multimodal foundation models in healthcare: A systematic review and evidence-gap map of 125 studies from benchmark to bedside. *Computational Diagnostics*, *1*(1), Article 4.

DOI
To be assigned on publication

SECTION
Clinical AI Evaluation & Governance

ARTICLE HISTORY
**Received:** 6 January 2025
**Published:** 7 March 2025

ACADEMIC EDITOR
Assigned on acceptance

PEER REVIEW
Single-blind, two reviewers



# Clinical Evaluation Readiness of Multimodal Foundation Models in Healthcare: A Systematic Review and Evidence-Gap Map of 125 Studies from Benchmark to Bedside

**Ansuman Padhi**[1,*]

[1] Mangalayatan University, Aligarh, Uttar Pradesh 202145, India
[*] Correspondence: ansuman_padhi@outlook.com

## HIGHLIGHTS

- 125 verified evaluations of multimodal foundation models mapped onto a six-level clinical evaluation readiness ladder.
- Half the literature (49.6%) is benchmark-only and 87.2% never ran prospectively on live clinical data.
- Only four studies (3.2%) were comparative trials, and none reported post-deployment surveillance.
- Fairness reporting fell to zero in all 16 studies that reached prospective evaluation.
- Prospective evidence is ~18 months old, rising from 0% before 2025 to 25.7% in 2026.

**Background:** Multimodal foundation models, pre-trained at scale across images, text, waveforms and structured clinical data, now rival specialist clinicians on selected benchmarks. Whether that performance reflects benefit or safety in real clinical environments is unestablished.

**Methods:** Systematic review and evidence-gap map. Eight pre-specified domains were searched using 100+ structured queries across journals, preprints, proceedings and trial registries (January 2021 to August 2026). Eligible studies evaluated a model combining two or more modalities, pre-trained at scale, on a health task. All records were primary-source verified. Studies were assigned to a pre-specified six-level Clinical Evaluation Readiness Ladder (CERL) anchored to TRIPOD+AI, DECIDE-AI and CONSORT-AI, and coded on seven reporting domains, with Wilson 95% confidence intervals.

**Results:** Of 182 records, 125 were included. Sixty-two (49.6%, 95% CI 41.0-58.2) were benchmark-only (CERL-0) and 47 (37.6%) retrospective; 109 (87.2%) never ran prospectively. Sixteen (12.8%) reached prospective evaluation, 11 (8.8%) live clinician-facing use, 4 (3.2%) comparative trials. None reported post-deployment surveillance. External validation reached 39.2%, clinician comparator 44.0%, fairness 12.0%, safety 8.0% and robustness 3.2%. Fairness reporting was zero in all 16 prospective studies; prospective evaluation rose from 0% pre-2025 to 25.7% in 2026.

**Conclusions:** The evidence base is broad at the benchmark end and near empty at the bedside. Prospective evidence is roughly eighteen months old, equity reporting is thinnest closest to patients, and no system has been followed post-deployment.

**Keywords:** multimodal foundation models; medical vision-language models; clinical artificial intelligence; prospective clinical validation; evidence-gap map; systematic review; algorithmic fairness in healthcare; post-deployment surveillance

**Plain-language summary.** Artificial intelligence systems that read scans, notes and clinical data together are often described as ready for the clinic. This review checked all 125 published evaluations of such systems in medicine. Most were tested only on public datasets; seven in eight were never run on live patient care; four were tested in a trial; and none was monitored after being deployed. Reporting on whether these systems work equally well for different patient groups disappeared entirely in the studies conducted closest to patients.

## 1. Introduction

The dominant paradigm in clinical artificial intelligence has changed twice in five years. The first shift replaced hand-engineered features with task-specific deep learning. The second, beginning around 2021, replaced task-specific models with foundation models, that is, large networks pre-trained on broad data at scale and adapted to many downstream tasks by fine-tuning, prompting or transfer (Moor et al., 2023). When such a model ingests more than one data type, whether images and text, or images, waveforms and structured records, it is a multimodal foundation model. Medicine is intrinsically multimodal, and the appeal is immediate: a single backbone that reads the radiograph, the report, the notes and the vital signs resembles how clinicians actually reason.

Capability has advanced faster than most forecasts. Vision-language models generate radiology reports at a standard that radiologists sometimes prefer to those written by their colleagues (Tanno et al., 2025; Zambrano Chaves et al., 2025), and computed tomography foundation models trained on paired volumes and reports now transfer across hundreds of downstream tasks (Blankemeier et al., 2026). Pathology foundation models transfer across dozens of oncological tasks from a single backbone (Ding et al., 2025; Lu et al., 2024a; Xiang et al., 2025; Xu et al., 2024). General-purpose multimodal assistants answer licensing-style questions containing figures (Jin et al., 2024; Kaczmarczyk et al., 2024; Tu et al., 2024). Promptable segmentation backbones now span nine or more imaging modalities from a single set of weights (Ma et al., 2024; Zhao et al., 2025). Ophthalmic, dermatological, echocardiographic and pathology-dialogue systems are evaluated directly against specialists (Christensen et al., 2024; Lu et al., 2024b; Yan et al., 2025). A dense literature of benchmark results has accumulated, and with it a strong rhetorical current suggesting that clinical deployment is imminent or already underway.

There is reason for caution about that inference. Benchmark performance is a weak surrogate for clinical utility. Curated datasets are enriched for unambiguous cases, are frequently drawn from the same public repositories used in pre-training, and rarely represent the case mix, image quality, missingness or workflow pressure of routine care. For task-specific clinical artificial intelligence the pattern is well documented: most published models never undergo external validation, far fewer are evaluated prospectively, and only a small fraction are tested in a comparative trial reporting patient-relevant outcomes. Whether multimodal foundation models are repeating that trajectory, escaping it, or compressing it is an empirical question that has not been asked systematically.

Two properties of foundation models make the question more urgent rather than less. First, their generality dissolves the notion of a fixed intended use, which is the anchor of both regulatory evaluation and clinical validation; a model that can do anything has not thereby been shown to do anything safely (Dong et al., 2026). Second, many are proprietary, are versioned opaquely and are updated without notice, so an evaluation may not describe the artefact a clinician later encounters. Both properties weaken the inferential link from any published evaluation to any deployed system.

### *1.1. What existing reviews do and do not cover*

Several relevant syntheses exist. Reviews of multimodal foundation models in medical imaging (Huang et al., 2025), of vision and vision-language models in ophthalmology (Jin et al., 2026), and of vision-language models in clinical artificial intelligence generally (Thirunavukarasu et al., 2026) address model capability within a modality or a specialty. A scoping review of silent trials examines one rung of the evaluation pathway across model types (Tikhomirov et al., 2026). Consensus and perspective pieces argue that benchmark-derived readiness claims require governance (Dong et al., 2026), and that foundation models in biomedical imaging must be converted from hype into demonstrable reality (Muneer et al., 2026). A translational-readiness review has been conducted for imaging artificial intelligence in dentistry (Ardila et al., 2026).

These are complementary to, but distinct from, the question posed here. Specialty-siloed reviews cannot describe the overall translational position of the field, because specialties differ substantially in data availability, regulatory maturity and deployment infrastructure; a synthesis restricted to one specialty cannot

establish whether the pattern it observes is local or general. Capability-oriented reviews index what models can do rather than what has been shown about their use. The appearance of commentary calling for precisely this analysis, in the absence of the analysis itself, is the characteristic signature of an unfilled evidence gap.

To the author's knowledge, no systematic review has characterised, across specialties and modalities, the position of the multimodal foundation model literature on the pathway from benchmark evaluation to clinical deployment, appraised against the reporting standards developed for each stage of that pathway.

### 1.2. *A framework for the question: the Clinical Evaluation Readiness Ladder*

Answering the question requires a defensible way to classify what kind of evaluation a study performed. Existing reporting guidelines already partition this space, but each addresses a single stage. TRIPOD+AI governs prediction model development and validation (Collins et al., 2024). DECIDE-AI governs early-stage live clinical evaluation (Vasey et al., 2022). CONSORT-AI and SPIRIT-AI govern randomised trials and their protocols (Cruz Rivera et al., 2020; Liu et al., 2020). TRIPOD-LLM addresses studies using large language models (Gallifant et al., 2025). What is missing is an ordinal scheme that spans them and allows a whole literature to be positioned at once.

A six-level ordinal classification anchored to those standards was therefore pre-specified and is termed the Clinical Evaluation Readiness Ladder (Table 1). Each study is assigned the highest level for which it reports primary empirical data. The ladder is ordinal but not strictly hierarchical in value: a well-conducted external validation may inform practice more than a poorly conducted live evaluation, so level and quality are treated as separate axes throughout.

**Table 1.** The Clinical Evaluation Readiness Ladder (CERL). Each study is assigned the highest level for which it reports primary empirical data.

| Level | Label | Definition | Anchor standard |
|---|---|---|---|
| CERL-0 | Benchmark | Evaluation confined to public, curated, synthetic or examination-style data. No institutional patient data. | Not applicable |
| CERL-1a | Retrospective, single site | Real patient data from one institution, analysed retrospectively and offline. | TRIPOD+AI |
| CERL-1b | Retrospective, external | As CERL-1a, with external, multi-site, temporal or geographic validation. | TRIPOD+AI |
| CERL-2 | Silent or shadow | Model runs prospectively on live clinical data; outputs are withheld from clinicians and cannot influence care. | TRIPOD+AI |
| CERL-3 | Live clinician-facing | Outputs are available to clinicians in real workflow; human-AI interaction, usability, workload or safety assessed. | DECIDE-AI |
| CERL-4 | Comparative trial | Randomised or rigorous quasi-experimental comparison reporting patient, clinician, workflow or system outcomes. | CONSORT-AI, SPIRIT-AI |
| CERL-5 | Post-deployment surveillance | Routine monitoring of a system in sustained clinical use: drift detection, algorithmovigilance, incident reporting. | Not applicable |

### 1.3. *Objectives*

The primary objective was to determine the distribution of published evaluations of multimodal foundation models in healthcare across CERL levels. Secondary objectives were to quantify the reporting of external validation, clinician comparison, fairness, safety, robustness and prospective design; to compare these across clinical domain, model provenance, geographic setting and publication venue; to describe temporal trends; and to produce an openly available dataset and evidence-gap map identifying where evidence is concentrated and where it is absent.

## 2. Methods

This review is reported according to the PRISMA 2020 statement (Page et al., 2021). The search is reported according to PRISMA-S (Rethlefsen et al., 2021) and the synthesis according to the SWiM reporting guideline (Campbell et al., 2020).

### 2.1. *Protocol and registration*

A protocol specifying the research questions, eligibility criteria, search domains, extraction fields, the Clinical Evaluation Readiness Ladder and the analysis plan was written before screening began and is provided unaltered as Supplementary File 1. No changes were made to the primary outcome or to the eligibility criteria after screening commenced.

The review was not registered with PROSPERO. PROSPERO's stated scope is restricted to reviews addressing at least one human health-related outcome, and it explicitly excludes mapping reviews. The present study evaluates the characteristics and reporting completeness of a research literature rather than a human health-related outcome, and it incorporates an evidence-gap map, placing it outside that scope. The a priori protocol in Supplementary File 1 serves the transparency function that registration provides, and the complete extraction dataset is deposited openly so that every judgement reported here can be independently re-derived.

### 2.2. *Eligibility criteria*

Studies had to evaluate a model satisfying both of two criteria. The first was multimodality: the model accepted, or jointly represented, two or more distinct data modalities, for example image and text, image and structured data, waveform and text, video and text, or audio and text. Unimodal systems were excluded, as were pipelines combining independent unimodal models only at a final decision rule without any shared or jointly trained representation. The second was foundation-model character: the model was pre-trained at scale on broad data and adapted to downstream tasks by fine-tuning, prompting, in-context learning, adapters, or zero-shot or few-shot transfer, rather than being trained end to end for a single task from random initialisation.

Any health or clinical task was eligible, in any specialty, setting or country, whether performed on patient data or directed at clinicians or patients. Preclinical, molecular, veterinary and laboratory-science applications without an articulated clinical task were excluded.

Studies had to report at least one empirical result: discriminative or generative performance, agreement with a reference standard, safety or error data, usability or workload data, process outcomes, or patient outcomes. All empirical designs were eligible. Narrative and systematic reviews, editorials, commentaries, correspondence without original data, protocols, registry records without results, and papers reporting architecture or datasets without clinical evaluation were excluded. Reviews retrieved by the search were retained for reference-list screening.

The publication window was 1 January 2021 to 22 August 2026, the start date reflecting the emergence of the foundation-model paradigm, restricted to English. Preprints on arXiv, medRxiv and bioRxiv were eligible and were handled as a pre-specified analytic stratum. This decision was deliberate: a substantial share of this literature is disseminated only as preprints, so restriction to peer-reviewed work would misrepresent the field, whereas pooling without distinction would obscure differences in evidentiary standard.

### 2.3. *Information sources and search strategy*

Searches covered eight pre-specified domains: radiology and medical imaging; computational pathology; ophthalmology and dermatology; cardiology and neurology; electronic health records and critical care; cross-specialty multimodal assistants and promptable segmentation models; prospective, deployment, silent-trial and randomised evaluations; and fairness, safety, robustness and evaluations conducted in low-income and middle-income countries. More than 100 distinct structured queries were executed across the journal literature, arXiv, medRxiv, bioRxiv, conference proceedings (CVPR, ECCV, NeurIPS, ICML, MICCAI, ML4H and CIKM) and trial registries.

The seventh domain was searched exhaustively and independently, because the primary outcome of the review depends on complete ascertainment at the top of the ladder. Three secondary sources were additionally consulted to corroborate the scarcity observed there (Thirunavukarasu et al., 2026; Tikhomirov et al., 2026). Full query lists, dates and yields are provided in Supplementary File 2.

### *2.4. Selection process and verification*

Records were screened on title, venue and abstract, after which every retained record was sought and verified against a primary source, that is, the publisher's page, the preprint server, or a trial registry. Verification was treated as an inclusion requirement rather than a courtesy. A record whose primary source could not be retrieved, or whose eligibility-determining fields could not be read directly, was excluded and reported as such, rather than coded from a search-engine summary or from prior knowledge. This is a deliberately conservative rule and its cost is quantified in Figure 1.

Borderline eligibility decisions, chiefly systems that are multimodal only in the sense of combining structured with unstructured records, and systems whose foundation-model character is asserted but not demonstrated, were resolved against the two criteria and are listed with reasons in Supplementary File 3.

### *2.5. Data items and readiness classification*

Extracted items comprised bibliographic details (venue, venue type and year), clinical domain, model identity and provenance (proprietary application programming interface, open-weights, or locally developed), modality combination, dataset provenance, sample size, comparator, geographic setting classified by World Bank income group, and seven cross-cutting domains: external validation; prospective data collection; clinician comparator; fairness or subgroup analysis; safety, harms or error analysis; robustness testing, including out-of-distribution, missing-modality, image-degradation, adversarial and prompt-sensitivity testing; and transparency of model version. Readiness level was assigned as the highest level for which primary empirical data were reported, using the definitions in Table 1, with verbatim supporting text captured for auditability.

### *2.6. Synthesis methods*

Synthesis is narrative and quantitative-descriptive, structured around the readiness framework. Proportions are reported with Wilson 95% confidence intervals. Differences across venue type were examined with the chi-square test with Yates's continuity correction. All comparative analyses are exploratory and hypothesis-generating; p values are reported without adjustment for multiplicity and are interpreted accordingly.

Meta-analysis of performance estimates was not undertaken. This was a considered decision rather than an omission. Included studies differ irreducibly in task, reference standard, model, model version, prompt and case mix. Pooling accuracy across them would produce a precise number with no interpretable referent, and would enact precisely the benchmark-to-clinic overreach that this review exists to characterise.

Risk of bias was appraised descriptively against the stage-appropriate standard, using PROBAST+AI for studies reporting diagnostic or prognostic performance (Moons et al., 2025), and TRIPOD+AI, TRIPOD-LLM, DECIDE-AI or CONSORT-AI for reporting completeness. No composite quality score was computed, in line with methodological guidance against numerical quality scoring.

## 3. Results

### *3.1. Study selection*

Of 182 records identified, 9 were duplicates across domains. Of 173 records screened and sought for verification, 12 could not be retrieved or verified against a primary source. Of 161 reports assessed for eligibility, 36 were excluded, most commonly for having fewer than two modalities (n = 12) or for not being foundation-model based (n = 7). In total, 125 studies were included. The selection process is shown in Figure 1.

The group excluded for unimodality is informative in itself. It contains several of the most prominent systems in the field, including image-only pathology and retinal backbones that are frequently described in secondary literature as multimodal foundation models. This indicates that the term is applied considerably more loosely in commentary than in primary reports.

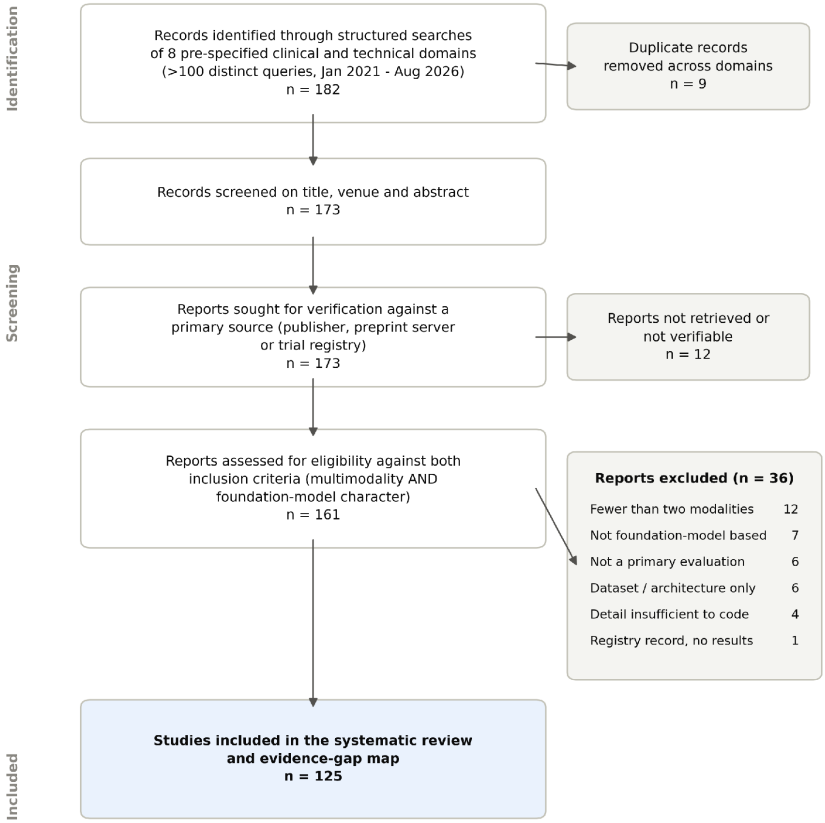


**Figure 1.** PRISMA 2020 flow diagram. Records were identified through structured searches of eight pre-specified domains. Verification against a primary source was applied as an inclusion requirement.

### 3.2. *Characteristics of included studies*

Studies were published between 2022 and 2026, with 71.2% (89 of 125) appearing in 2025 or 2026. Sixty-nine (55.2%) appeared in peer-reviewed journals, 46 (36.8%) as preprints and 10 (8.0%) in conference proceedings.

Clinical domains were dominated by imaging-adjacent specialties: cross-specialty or multispecialty evaluations (n = 24), radiology (n = 22), pathology (n = 19), ophthalmology (n = 13), electronic health records and critical care (n = 11), cardiology (n = 9), dermatology (n = 7) and neurology (n = 6). Seven further specialties, namely primary care, gastroenterology, emergency medicine, critical care, oncology, surgery and obstetrics, contributed two studies each. Psychiatry, paediatrics, orthopaedics and rehabilitation contributed none.

Models were open-weights in 55 studies (44.0%), locally developed in 37 (29.6%) and proprietary application programming interfaces in 33 (26.4%). The geographic setting was high-income in 65 studies (52.0%) and mixed or multinational in 38 (30.4%). Eighteen studies (14.4%) were conducted in upper-middle-income settings, and only 4 (3.2%, 95% CI 1.3 to 7.9) exclusively in low-income or lower-middle-income settings.

### 3.3. *Primary outcome: distribution across the readiness ladder*

The distribution is severely bottom-heavy, as shown in Figure 2 and Table 2.

**Table 2.** Distribution of 125 included studies across the Clinical Evaluation Readiness Ladder. Confidence intervals are Wilson intervals.

| Readiness level | n | % | 95% CI |
|---|---|---|---|
| CERL-0, benchmark only | 62 | 49.6 | 41.0 to 58.2 |
| CERL-1a, retrospective, single site | 17 | 13.6 | 8.7 to 20.7 |
| CERL-1b, retrospective, external | 30 | 24.0 | 17.4 to 32.2 |
| CERL-2, silent or shadow | 5 | 4.0 | 1.7 to 9.0 |
| CERL-3, live clinician-facing | 7 | 5.6 | 2.7 to 11.1 |
| CERL-4, comparative trial | 4 | 3.2 | 1.3 to 7.9 |
| CERL-5, post-deployment surveillance | 0 | 0.0 | 0.0 to 3.0 |
| *Any prospective evaluation (CERL 2 or above)* | *16* | *12.8* | *8.0 to 19.8* |
| *Live use or trial (CERL 3 or above)* | *11* | *8.8* | *5.0 to 15.1* |

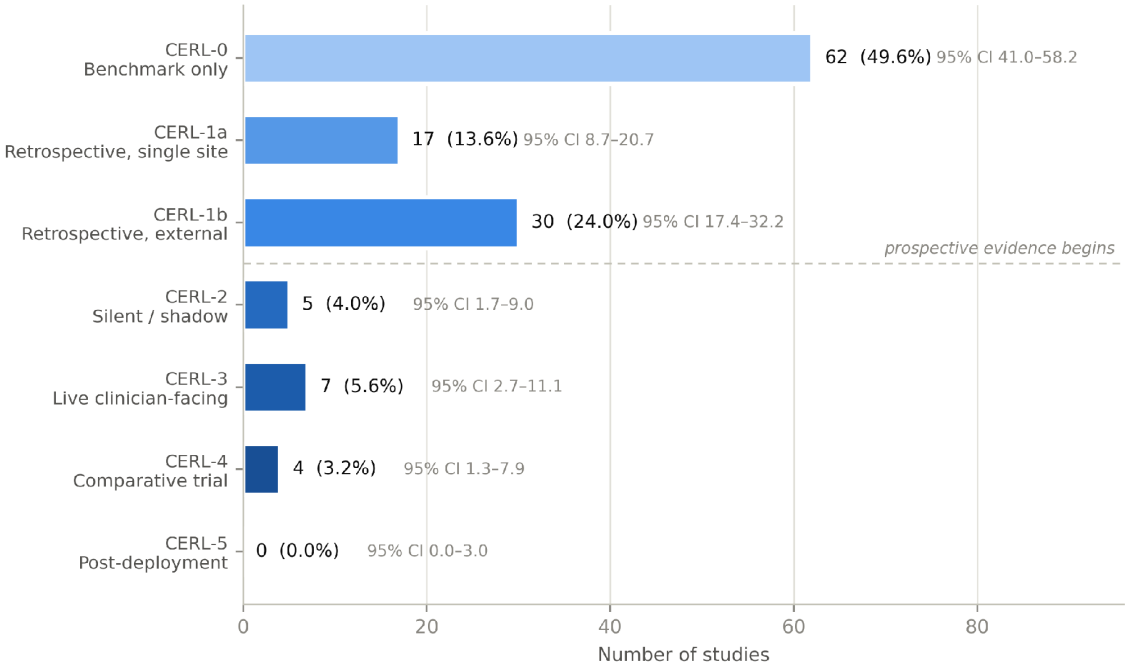


**Figure 2.** Distribution of 125 multimodal foundation-model evaluations across the Clinical Evaluation Readiness Ladder, with Wilson 95% confidence intervals. The dashed rule marks the point at which prospective evidence begins; 87.2% of the literature lies below it.

Sixty-two studies (49.6%, 95% CI 41.0 to 58.2) sat at CERL-0, with evaluation confined to public benchmarks, curated published case series, or examination items, and no institutional patient data at any point. A further 17 (13.6%) were single-site retrospective (CERL-1a) and 30 (24.0%) were retrospective with external or multi-site validation (CERL-1b). Together, 109 of 125 studies (87.2%) never observed the model operating prospectively on live clinical data.

Above that line the literature thins abruptly. Five studies (4.0%, 95% CI 1.7 to 9.0) reported silent or shadow prospective evaluation (CERL-2): a neuroimaging foundation model deployed silently for one week on 1,155 consecutive studies (Kondepudi et al., 2026); a sepsis-prediction system running in silent mode across two emergency departments (Shashikumar et al., 2025); a fine-tuned pathology model in a four-month real-time silent trial for lung cancer biomarker detection (Campanella et al., 2025); a multimodal respiratory-failure model assessed prospectively against physician predictions; and a silent trial of large language models supporting community health workers in Rwanda (Shimelash et al., 2026).

Seven studies (5.6%, 95% CI 2.7 to 11.1) reached CERL-3, that is, live clinician-facing use. These comprised prospective artificial-intelligence-assisted radiologist reporting for chest radiography across three hospitals (Bai et al., 2026); multicentre prospective endoscopy reporting (Jiang et al., 2026); a multidisease retinal screening pilot; an ophthalmic multimodal diagnostic system across three centres; an emergency department decision-support system explicitly framed as a DECIDE-AI stage-one evaluation (Leibovitch et al., 2026); a 20-patient emergency department feasibility pilot; and a prospective multicentre guideline-based tumour board study.

Four studies (3.2%, 95% CI 1.3 to 7.9) were comparative trials (CERL-4). Two were randomised trials of clinician-facing diagnostic support in ophthalmology (Jia et al., 2026; Wu et al., 2025) and two were randomised trials of ambient documentation systems (Afshar et al., 2025; Lukac et al., 2025). Only two of the four therefore tested a multimodal foundation model as a diagnostic or decision-support intervention.

No study reported post-deployment surveillance. The CERL-5 count was 0 of 125 (95% CI 0.0 to 3.0). Not one of the 125 studies followed a system after it entered sustained clinical use, and none reported drift monitoring, algorithmovigilance or incident reporting.

Only one study, the ophthalmic randomised trial of an eyecare foundation model, reported any patient-level outcome, in the form of referral compliance and self-management adherence at follow-up (Wu et al., 2025). No study reported a clinical health outcome such as morbidity, mortality, or change in diagnostic yield.

**Table 3.** Reporting of cross-cutting evaluation domains across 125 included studies. Confidence intervals are Wilson intervals.

| Evaluation domain | n | % | 95% CI |
|---|---|---|---|
| Clinician comparator | 55 | 44.0 | 35.6 to 52.8 |
| External validation | 49 | 39.2 | 31.1 to 48.0 |
| Prospective data collection | 16 | 12.8 | 8.0 to 19.8 |
| Fairness or subgroup analysis | 15 | 12.0 | 7.4 to 18.9 |
| Safety, harms or error analysis | 10 | 8.0 | 4.4 to 14.1 |
| Robustness testing | 4 | 3.2 | 1.3 to 7.9 |

### *3.4. Cross-cutting reporting domains*

Reporting was incomplete across every domain examined, as shown in Figure 3A and Table 3. External validation was reported by 49 studies (39.2%, 95% CI 31.1 to 48.0) and a clinician comparator by 55 (44.0%, 95% CI 35.6 to 52.8). Fairness or subgroup analysis appeared in 15 (12.0%, 95% CI 7.4 to 18.9), safety, harms or error analysis in 10 (8.0%, 95% CI 4.4 to 14.1), and robustness testing in 4 (3.2%, 95% CI 1.3 to 7.9).

The most consequential finding in this section is an interaction rather than a marginal proportion. Fairness reporting rose modestly with retrospective rigour, from 9.7% at CERL-0 (6 of 62) to 11.8% at CERL-1a (2 of 17) and 23.3% at CERL-1b (7 of 30), and then fell to zero in every study that reached prospective evaluation: 0 of 5 at CERL-2, 0 of 7 at CERL-3 and 0 of 4 at CERL-4, giving 0 of 16 overall. The studies conducted closest to patients reported least about whom the model works for.

**Table 4.** Fairness or subgroup reporting by readiness level. Reporting falls to zero in every study that reached prospective evaluation.

| Readiness level | Fairness reported (n/N) | % |
|---|---|---|
| CERL-0 | 6/62 | 9.7 |
| CERL-1a | 2/17 | 11.8 |
| CERL-1b | 7/30 | 23.3 |
| CERL-2 | 0/5 | 0.0 |
| CERL-3 | 0/7 | 0.0 |
| CERL-4 | 0/4 | 0.0 |
| *All prospective studies (CERL 2 or above)* | *0/16* | *0.0* |

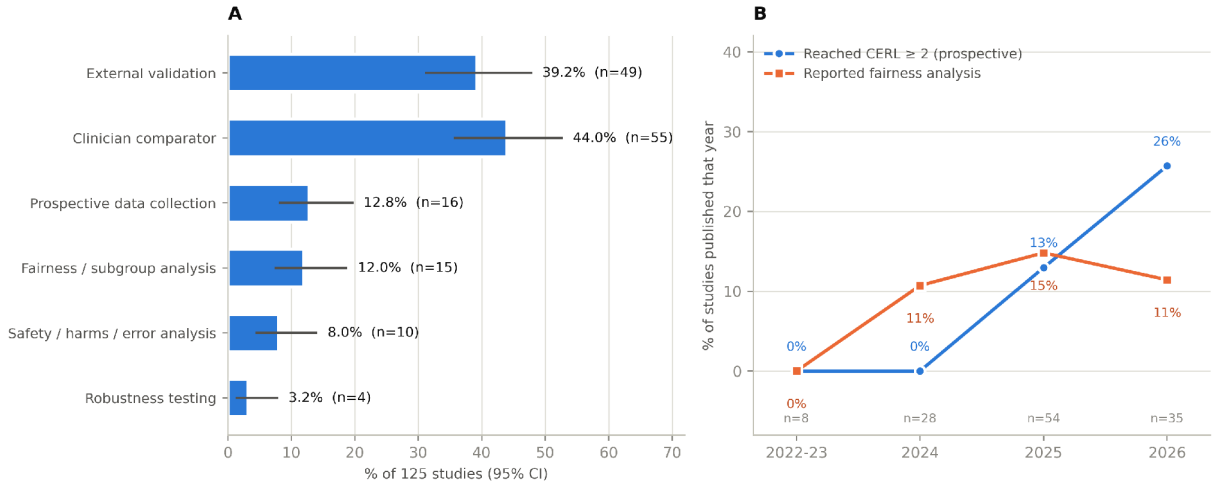


**Figure 3.** Panel A shows reporting of cross-cutting evaluation domains with Wilson 95% confidence intervals. Panel B shows change over time in the proportion of studies reaching prospective evaluation (CERL 2 or above) and reporting fairness analysis. The years 2022 and 2023 are pooled because 2022 contributed a single study.

Substantive fairness evidence, where it existed, was concentrated in a small number of dedicated audits rather than distributed through the literature. The largest of these, an analysis of 858,884 chest radiographs across five international datasets, found that vision-language foundation models consistently underdiagnosed female, younger and Black patients, with amplified disparities in intersectional subgroups and larger fairness gaps than those of board-certified radiologists (Yang et al., 2025). Dermatology contributed a cluster of skin-tone analyses, several of which reported fairness as a secondary outcome of model development rather than as a primary question (Nijjer et al., 2025; Zhou et al., 2024).

Safety evidence showed a similar concentration. Dedicated adversarial and hallucination studies demonstrated that sub-visual prompts embedded in medical images induced harmful outputs across every tested model (Clusmann et al., 2025), that prompt injection degraded surgical decision support by corrupting intermediate perceptual processing rather than merely overriding final answers (Zhang et al., 2026), and that a majority of generated chest-radiograph outputs contained hallucinations, with normal radiographs paradoxically attracting the most severe errors (Wang & Li, 2026). These findings were not, however, reflected in the safety reporting of the broader corpus.

### *3.5. Subgroup and temporal analyses*

Prospective evaluation was substantially more frequent in peer-reviewed journals (14 of 69, 20.3%, 95% CI 12.5 to 31.2) than in preprints (2 of 46, 4.3%, 95% CI 1.2 to 14.5) or conference proceedings (0 of 10), with a chi-square statistic of 6.32 on 1 degree of freedom ($p = 0.012$). The preprint and proceedings literature is therefore not a leading indicator of the journal literature at the top of the ladder; it is a differently shaped literature, overwhelmingly concentrated at CERL-0.

Prospective evaluation was distributed across provenance categories without a clear gradient: proprietary 6 of 33 (18.2%), locally developed 5 of 37 (13.5%) and open-weights 5 of 55 (9.1%). Notably, the two randomised trials of ambient documentation evaluated commercial systems whose underlying base models were not identified in the published reports, so the evaluated artefact cannot be reconstructed from the record (Afshar et al., 2025; Lukac et al., 2025).

The evidence-gap map in Figure 4 shows that concentration and readiness are inversely related. The three largest domains, namely cross-specialty, radiology and pathology, contributed 65 studies but only 4 at CERL-2 or above. Conversely, every study in emergency medicine, critical care and primary care sat at CERL-2 or above, reflecting not maturity but the fact that these domains have been entered late and directly at the deployment end, largely by systems combining text with structured data rather than by imaging-based systems. Ophthalmology is the only domain with a full column, having benchmark, retrospective, live and randomised evidence all present.

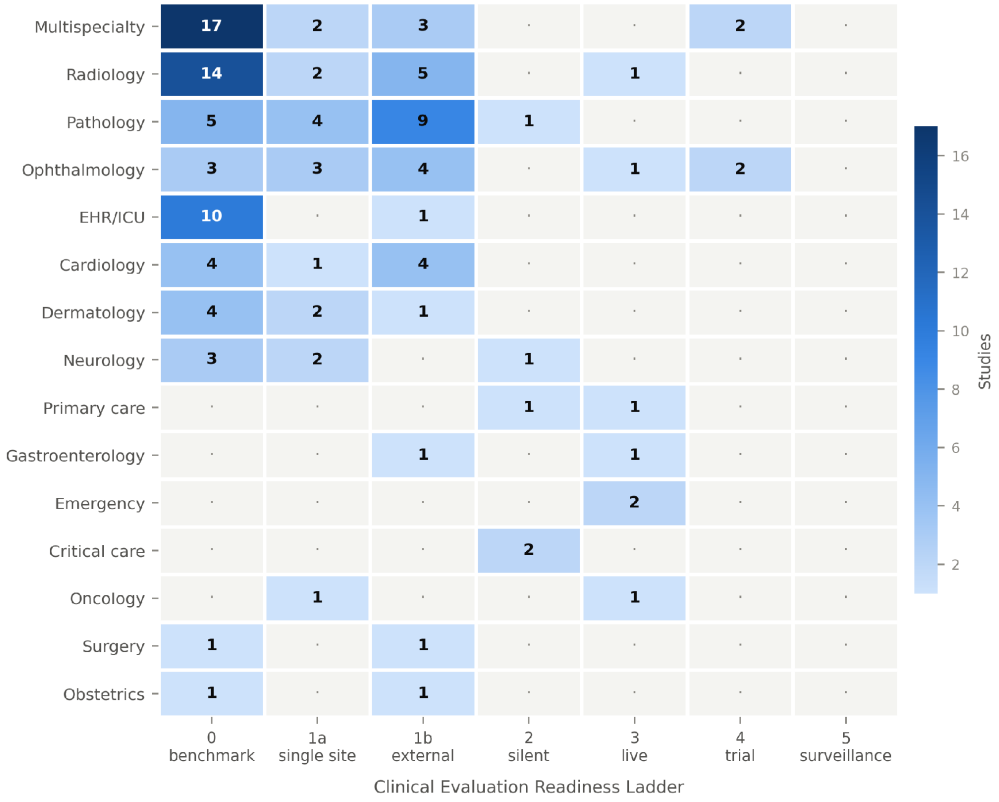


**Figure 4.** Evidence-gap map showing clinical domain by readiness level. Cell values are study counts and shading encodes count. Blank cells are empty. The CERL-5 column is empty across every domain.

No study published before 2025 reached CERL-2 or above, with 0 of 8 in 2022 to 2023 and 0 of 28 in 2024. The proportion rose to 13.0% in 2025 (7 of 54) and 25.7% in 2026 (9 of 35), as shown in Figure 3B. The entire prospective evidence base for multimodal foundation models in medicine is therefore approximately eighteen months old. Fairness reporting showed no comparable trajectory, remaining between 11% and 15% from 2024 onward.

### *3.6. Risk of bias and reporting completeness*

Formal appraisal was constrained by the reporting deficits that the review documents. Among the 109 retrospective and benchmark studies, applicability concerns under PROBAST+AI were high in the majority. Benchmark-only evaluations by construction address a population that is not the intended-use population, and many used datasets plausibly contained within the pre-training corpora of the models being evaluated, a contamination risk that only a minority of studies acknowledged and none quantified.

Among the 11 studies at CERL-3 or above, adherence to DECIDE-AI was partial. One study explicitly framed itself as a DECIDE-AI stage-one evaluation and reported adoption over time, disengagement, safety events and workflow outcomes (Leibovitch et al., 2026). The remainder reported accuracy and, variably, reading time or clinician preference, without the human-factors, error-analysis or implementation-context items that the guideline specifies. Reporting of model version and access date, which is the minimum needed to identify which artefact was evaluated, was absent or partial in a substantial share of studies evaluating proprietary application programming interfaces.

## 4. Discussion

### *4.1. Principal findings*

Across 125 verified studies, the evidence base for multimodal foundation models in healthcare is broad at the benchmark end and close to empty at the bedside. Half the literature has never encountered institutional patient data. Seven in eight studies never observed the model operating prospectively. Comparative trials number four, and half of those evaluate documentation rather than clinical decision-making. One study reports a patient-level outcome and none reports a clinical health outcome. No system anywhere in this literature has been followed after deployment.

Three features of the distribution deserve emphasis beyond the headline scarcity. The first is an equity inversion. Fairness reporting improves with retrospective rigour and then vanishes entirely at the point of prospective evaluation. This is the opposite of the pattern the field needs. Subgroup performance is a property of a model in a population, and a prospective study is the first opportunity to measure it in the population that will actually be exposed. It is precisely there that no study measured it. The chest-radiograph audit that did look found that underdiagnosis fell disproportionately on female, younger and Black patients, and that the models were less equitable than the clinicians against whom they were compared (Yang et al., 2025). That finding is not reassuring about what the sixteen unmeasured prospective studies would have shown.

The second is the absent tail. The complete absence of post-deployment surveillance is arguably the most consequential negative finding of this review, because foundation models are the class of system for which it matters most. Proprietary models are updated silently under stable names, so performance measured in March may not describe the artefact a clinician queries in September. A literature with no surveillance rung has no mechanism to detect that, and the regulatory frameworks now being applied to high-risk medical artificial intelligence assume evidence of exactly the kind that nobody is generating.

The third is geography. Four studies, amounting to 3.2% of the literature, were conducted exclusively in low-income or lower-middle-income settings, despite these being the settings in which diagnostic workforce shortages make generalist artificial intelligence most attractive, and in which case mix, imaging hardware and disease prevalence differ most from the datasets on which these models were pre-trained. The single silent trial conducted with community health workers in Rwanda is instructive, since it found near-parity for one frontier model and markedly lower referral accuracy for another, a divergence invisible in benchmark rankings (Shimelash et al., 2026).

### *4.2. Comparison with existing literature*

These findings extend, and are consistent with, the siloed reviews that preceded them. A review of vision-language foundation models in clinical artificial intelligence reported that results are generally drawn from retrospective studies, with no prospective clinical trials (Thirunavukarasu et al., 2026). A scoping review of medical artificial intelligence silent trials identified none involving foundation models (Tikhomirov et al., 2026). Reviews confined to medical imaging and to ophthalmology reached similar conclusions within their own boundaries (Huang et al., 2025; Jin et al., 2026). The contribution of the present review is to show that this is not a property of any one specialty or modality but of the field as a whole, and to quantify it on a common scale that permits comparison across domains and over time.

The temporal analysis reconciles an apparent tension between these findings and the optimism of the field. The prospective literature is not absent; it is new. Zero studies before 2025, 13.0% in 2025 and 25.7% in 2026 describes a field at an inflection rather than a field in stasis. The appropriate reading is therefore neither that multimodal foundation models are unready in principle, nor that deployment is imminent, but that the evidence needed to distinguish those possibilities has only just begun to be generated.

### *4.3. Implications for practice, policy and research*

For clinicians and health systems, benchmark performance should not be treated as evidence of clinical utility for this class of system, and procurement should ask which readiness level a claimed capability rests on. At present, for the overwhelming majority of multimodal foundation models, the accurate answer is CERL-0 or CERL-1.

For researchers, the most valuable studies now are not further benchmarks. Silent and shadow evaluations are inexpensive, carry no patient risk, and are the fastest way to discover whether benchmark performance survives contact with local data, yet there are five in the entire literature. Prospective studies should report subgroup performance as a primary, pre-specified outcome rather than as an appendix.

For journals and editors, requiring authors to state the evaluation level of their contribution, and to justify claims of clinical readiness against it, is a low-cost editorial intervention. The readiness ladder presented here is offered as a candidate instrument for that purpose.

For regulators, the absence of any post-deployment surveillance literature, combined with silent versioning of proprietary models, is a specific and addressable gap. Evidentiary expectations for high-risk medical artificial intelligence presuppose a surveillance literature that does not currently exist.

### *4.4. Strengths and limitations*

The principal strength of this review is verification discipline. Every included study was confirmed against a primary source, and 12 records were excluded rather than coded from secondary description. A second strength is the pre-specified ordinal framework, which allows a heterogeneous literature to be positioned on one scale without pooling incommensurable performance estimates. A third is that exhaustive independent searching of the deployment domain, corroborated against secondary sources, makes under-ascertainment at the top of the ladder, where the primary outcome is most sensitive, the least likely failure mode.

Several limitations qualify the findings. First, the search used structured topic-based querying across eight domains rather than a single exhaustive Boolean export from bibliographic databases. Recall at CERL-0, where the literature is largest and most repetitive, is therefore likely to be incomplete, and the figure of 49.6% for benchmark-only studies should be read as a lower bound on benchmark dominance rather than as a census. Because under-ascertainment falls disproportionately on the most numerous and least advanced stratum, this bias is conservative with respect to the principal conclusion.

Second, extraction relied on what studies reported, so absence of reporting is coded as absence and may in some cases understate what was done; for fairness and safety, however, that distinction is itself the finding. Third, the English-language restriction probably under-represents Chinese-language clinical evaluation, which matters because four of the six unambiguous prospective studies were conducted in China. Fourth, readiness assignment required judgement at two boundaries, namely prospectively collected cohorts analysed offline, and systems that are multimodal only in combining structured with unstructured records; reasonable reviewers may classify a minority of studies differently, and all such decisions are published for re-derivation. Fifth, trial registries could not be searched systematically, so registered but unreported trials are under-counted and the true CERL-4 pipeline is larger than the published record shows. Sixth, no formal quantitative risk-of-bias scoring was applied. Seventh, this review was conducted by a single author, so screening and extraction were not performed in duplicate by independent reviewers; the complete dataset is deposited openly to allow independent re-coding.

### *4.5. Conclusions*

Multimodal foundation models have accumulated an extensive benchmark literature and almost no bedside literature. Half of published evaluations have never seen institutional patient data, seven in eight have never run prospectively, four have been tested in a trial, none has been followed after deployment, and equity reporting disappears precisely in the studies conducted closest to patients. The prospective evidence base that does exist is approximately eighteen months old and growing, which makes this an opportune moment to set expectations for what the next generation of studies should report. The readiness ladder, the evidence-gap map and the openly deposited dataset are offered as instruments for doing so.

## Ethics approval and consent to participate

Ethical approval and informed consent were not required. This study is a systematic review of previously published research. It involved no human participants, no animal subjects, and no access to individual-level patient data. All data analysed were derived from publicly available published reports, preprints and trial registry records. No identifiable personal information was accessed, generated or reported at any stage.

## Consent for publication

Not applicable. This study contains no individual person's data in any form.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The author has no affiliation with, and has received no support from, any developer, vendor or commercial sponsor of any artificial intelligence system evaluated in the studies included in this review.

## Data availability

All data supporting the findings of this study are contained within the article and its supplementary files. The complete study-level extraction dataset (125 records, 15 fields), the full search strategy with dates and yields, the analysis code and the figure-generation code are provided as Supplementary Files 1 to 4 and are additionally deposited in a public repository under a Creative Commons Attribution licence. No proprietary or restricted-access data were used.

## Declaration of generative AI and AI-assisted technologies in the writing process

No AI tools were used to generate or develop the research content of this manuscript. AI was used only for minor grammar, language, and readability improvements, with all intellectual content and final revisions reviewed and approved by the authors.

## CRediT authorship contribution statement

Ansuman Padhi: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Software, Visualization, Writing – original draft, Writing – review and editing, Project administration.

## Acknowledgements

The author thanks the developers and maintainers of the open bibliographic infrastructure on which the verification procedure in this review depended, in particular Crossref, arXiv and medRxiv, whose openly accessible metadata made independent confirmation of every included record possible.

# References

Afshar, M., Baumann, M. R., Resnik, F., Hintzke, J., Sullivan, A. G., Wills, G., Lemmon, K., Dambach, J., Mrotek, L. A., Quinn, M., Abramson, K., Kleinschmidt, P., Brazelton, T. B., Leaf, M. A., Twedt, H., Kunstman, D., Patterson, B., Liao, F., Rasmussen, S., ... Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI*, *2*(12), Article AIoa2500945. https://doi.org/10.1056/AIoa2500945

Ardila, C. M., Vivares-Builes, A. M., & Pineda-Vélez, E. (2026). From algorithmic performance to clinical translation: Translational readiness of imaging-based artificial intelligence in dentistry. A systematic review. *Healthcare*, *14*(13), Article 1952. https://doi.org/10.3390/healthcare14131952

Bai, Y., Zhang, R., Lei, Y., Duan, X., Yao, J., Ju, S., Wang, C., Yao, W., Guo, Y., Zhang, G., Wan, C., Yuan, Q., Chen, L., Tang, W., Zhu, B., Wang, X., Sun, T., Zhou, W., Tao, D., ... Du, B. (2026). A DeepSeek-powered AI system for automated chest radiograph interpretation in clinical practice. *Nature Communications*, *17*(1), Article 6141. https://doi.org/10.1038/s41467-026-72680-6

Blankemeier, L., Kumar, A., Cohen, J. P., Liu, J., Liu, L., Van Veen, D., Gardezi, S. J. S., Yu, H., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Holland, R., Truyts, C., Bluethgen, C., Wu, Y., Lian, L., Jensen, M. E. K., Ostmeier, S., ... Chaudhari, A. S. (2026). Merlin: A computed tomography vision-language foundation model and dataset. *Nature*, *652*(8112), 1318–1328. https://doi.org/10.1038/s41586-026-10181-8

Campanella, G., Kumar, N., Nanda, S., Singi, S., Fluder, E., Kwan, R., Muehlstedt, S., Pfarr, N., Schüffler, P. J., Häggström, I., Neittaanmäki, N., Akyürek, L. M., Basnet, A., Jamaspishvili, T., Nasr, M. R., Croken, M. M., Hirsch, F. R., Elkrief, A., Yu, H., ... Vanderbilt, C. (2025). Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nature Medicine*, *31*(9), 3002–3010. https://doi.org/10.1038/s41591-025-03780-x

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, *368*, l6890. https://doi.org/10.1136/bmj.l6890

Christensen, M., Vukadinovic, M., Yuan, N., & Ouyang, D. (2024). Vision-language foundation model for echocardiogram interpretation. *Nature Medicine*, *30*(5), 1481–1488. https://doi.org/10.1038/s41591-024-02959-y

Clusmann, J., Ferber, D., Wiest, I. C., Schneider, C. V., Brinker, T. J., Foersch, S., Truhn, D., & Kather, J. N. (2025). Prompt injection attacks on vision language models in oncology. *Nature Communications*, *16*(1), Article 1239. https://doi.org/10.1038/s41467-024-55631-x

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & the SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, *26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., Vaidya, A. J., Jaume, G., Shaban, M., Kim, A., Williamson, D. F. K., Robertson, H., Chen, B., Almagro-Pérez, C., Doucet, P., Sahai, S., Chen, C., Chen, C. S., Komura, D., ... Mahmood, F. (2025). A multimodal whole-slide foundation model for pathology. *Nature Medicine*, *31*(11), 3749–3761. https://doi.org/10.1038/s41591-025-03982-3

Dong, Y., Cheng, J., Ding, C., & Lu, R. (2026). Governing clinical readiness claims derived from medical AI benchmark results. *Journal of Medical Systems*, *50*(1), Article 119. https://doi.org/10.1007/s10916-026-02445-7

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L. H., Landman, A., Lehmann, L., McCoy, L. G., Miller, T., Moreno, A., Munch, N., Restrepo, D., Savova, G., ... Bitterman, D. S. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*, *31*(1), 60–69. https://doi.org/10.1038/s41591-024-03425-5

Huang, S.-C., Jensen, M., Yeung-Levy, S., Lungren, M. P., Poon, H., & Chaudhari, A. S. (2025). A systematic review and implementation guidelines of multimodal foundation models in medical imaging [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-5537908/v1

Jia, H., Qian, B., Qu, Y., Wang, J., Chen, J., Li, T., Zheng, C., Han, J., Zhang, G., Jin, Y., Lee, S., Chen, X., Chen, H., Xu, J., Xu, K., Rong, Y., Jiang, Y., Yang, Y., Li, W., ... Sun, X. (2026). AI-based clinician decision support system for diagnosis of inherited retinal diseases: A multicenter, randomized trial. *Nature Medicine*. Advance online publication. https://doi.org/10.1038/s41591-026-04545-w

Jiang, R., Chen, B., Dong, Z., Zeng, X., You, H., Li, Y., Deng, Y., Mu, G., Wang, J., Huang, L., Li, J., Cheng, D., Zhou, W., & Yu, H. (2026). Domain specific multimodal large language model for automated endoscopy reporting with multicenter prospective validation. *npj Digital Medicine*, *9*(1), Article 394. https://doi.org/10.1038/s41746-026-02569-7

Jin, Q., Chen, F., Zhou, Y., Xu, Z., Cheung, J. M., Chen, R., Summers, R. M., Rousseau, J. F., Ni, P., Landsman, M. J., Baxter, S. L., Al'Aref, S. J., Li, Y., Chen, A., Brejt, J. A., Chiang, M. F., Peng, Y., & Lu, Z. (2024). Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digital Medicine*, *7*(1), Article 190. https://doi.org/10.1038/s41746-024-01185-7

Jin, K., Yu, T., Ying, G.-S., Ge, Z., Li, K. Z., Zhou, Y., Shi, D., Wang, M., Goktas, P., & Grzybowski, A. (2026). A systematic review of vision and vision-language foundation models in ophthalmology. *Advances in Ophthalmology Practice and Research*, *6*(1), 8–19. https://doi.org/10.1016/j.aopr.2025.10.004

Kaczmarczyk, R., Wilhelm, T. I., Martin, R., & Roos, J. (2024). Evaluating multimodal AI in medical diagnostics. *npj Digital Medicine*, *7*(1), Article 205. https://doi.org/10.1038/s41746-024-01208-3

Kondepudi, A., Rao, A., Zhao, C., Lyu, Y., Harake, S., Banerjee, S., Ogle, J., Joshi, R., Meissner, A.-K., Hou, X., Jiang, C., Chowdury, A., Srinivasan, A., Athey, B., Gulani, V., Pandey, A., Lee, H., & Hollon, T. (2026). Health system learning enables generalist neuroimaging models. *Nature Medicine*, *32*(8), 2831–2837. https://doi.org/10.1038/s41591-026-04497-1

Leibovitch, L., Ahituv, A., Gorenshtein, A., Aran, D., Sorka, M., Miron, K., & Shelly, S. (2026). Prospective evaluation of a large language model clinical decision support system in the emergency department. *Nature Medicine*. Advance online publication. https://doi.org/10.1038/s41591-026-04601-5

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & the SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, *26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., Parwani, A. V., Zhang, A., & Mahmood, F. (2024a). A visual-language foundation model for computational pathology. *Nature Medicine*, *30*(3), 863–874. https://doi.org/10.1038/s41591-024-02856-4

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Zhao, M., Chow, A. K., Ikemura, K., Kim, A., Pouli, D., Patel, A., Soliman, A., Chen, C., Ding, T., Wang, J. J., Gerber, G., Liang, I., Le, L. P., Parwani, A. V., Weishaupt, L. L., & Mahmood, F. (2024b). A multimodal generative AI copilot for human pathology. *Nature*, *634*(8033), 466–473. https://doi.org/10.1038/s41586-024-07618-3

Lukac, P. J., Turner, W., Vangala, S., Chin, A. T., Khalili, J., Shih, Y.-C. T., Sarkisian, C., Cheng, E. M., & Mafi, J. N. (2025). Ambient AI scribes in clinical practice: A randomized trial. *NEJM AI*, *2*(12), Article AIoa2501000. https://doi.org/10.1056/AIoa2501000

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, *15*(1), Article 654. https://doi.org/10.1038/s41467-024-44824-z

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Liu, N., ... van Smeden, M. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, *388*, e082505. https://doi.org/10.1136/bmj-2024-082505

Moor, M., Banerjee, O., Shakeri Hossein Abad, Z., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Muneer, A., Zhang, K., Hamdi, I., Qureshi, R., Waqas, M., Fouad, S., Ali, H., Anwar, S. M., & Wu, J. (2026). Foundation models in biomedical imaging: Turning hype into reality. *Nature Biomedical Engineering*, *10*(8), 1557–1575. https://doi.org/10.1038/s41551-026-01762-z

Nijjer, K., Bui, R., Jiu, D., Ahmed, A., Wang, P., Liu, B., Zhu, K., & Zhu, L. (2025). Adapting large language models to mitigate skin tone biases in clinical dermatology tasks: A mixed-methods study [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.00055

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, *10*(1), Article 39. https://doi.org/10.1186/s13643-020-01542-z

Shashikumar, S. P., Mohammadi, S., Krishnamoorthy, R., Patel, A., Wardi, G., Ahn, J. C., Singh, K., Aronoff-Spencer, E., & Nemati, S. (2025). Development and prospective implementation of a large language model based system for early sepsis prediction. *npj Digital Medicine*, *8*(1), Article 290. https://doi.org/10.1038/s41746-025-01689-w

Shimelash, N., Rutunda, S., Menon, V., Emmanuel-Fabula, M., Uwimbabazi, A., Rugege, C., Nshimiyimana, C., Rwema, I., Kandekwe, M., Berhe, D. F., Wong, R., Remera, E., Hezagira, E., Gill, J., Archer, L., Riley, R. D., Denniston, A. K., Liu, X., & Mateen, B. A. (2026). A "silent trial" assessing the accuracy of large language models for assisting community health workers in low-resource settings [Preprint]. medRxiv. https://doi.org/10.64898/2026.02.16.26346409

Tanno, R., Barrett, D. G. T., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., Singhal, K., Schaekermann, M., May, R., Lee, R., Man, S., Mahdavi, S., Ahmed, Z., Matias, Y., Barral, J., ... Ktena, I. (2025). Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine*, *31*(2), 599–608. https://doi.org/10.1038/s41591-024-03302-1

Thirunavukarasu, A. J., Li, S., Qin, P., Nie, D., Sanghera, R., Lim, E., Yu, J., & Zhang, L. (2026). Clinical artificial intelligence applications of vision-language foundation models. *PLOS Digital Health*, *5*(6), Article e0001453. https://doi.org/10.1371/journal.pdig.0001453

Tikhomirov, L., Semmler, C., Prizant, N., Bhasin, S., Kenyon, G., van der Vegt, A., Erdman, L., Kurian, N. C., Thompson, H., Palmer, L. J., Mohamud, A., Gichoya, J. W., Soremekun, S., Sendak, M. P., Anderson, J. A., Pfohl, S. R., Stedman, I., Ehrmann, D., Verspoor, K., ... McCradden, M. D. (2026). A scoping review of silent trials for medical artificial intelligence. *Nature Health*, *1*(5), 532–554. https://doi.org/10.1038/s44360-025-00048-z

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Palepu, A., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., ... Natarajan, V. (2024). Towards generalist biomedical AI. *NEJM AI*, *1*(3), Article AIoa2300138. https://doi.org/10.1056/AIoa2300138

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., ... McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, *28*(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9

Wang, H., & Li, Z. (2026). HalluCXR: Benchmarking and mitigating hallucinations in medical vision-language models for chest radiograph interpretation [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.20469

Wu, Y., Qian, B., Li, T., Qin, Y., Guan, Z., Chen, T., Jia, Y., Zhang, P., Zeng, D., Moroi, S., Raman, R., Thinggaard, B. S., Pedersen, F., Ñehe, J. A. O., Kamalden, T. A., Zhou, Y., Jin, Y., Li, H., Ran, A. R., ... Sheng, B. (2025). An eyecare foundation model for clinical assistance: A randomized controlled trial. *Nature Medicine*, *31*(10), 3404–3413. https://doi.org/10.1038/s41591-025-03900-7

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.-H., Willens, S., Olguin, F. M., Nirschl, J. J., Neal, J., Diehn, M., Yang, S., & Li, R. (2025). A vision-language foundation model for precision oncology. *Nature*, *638*(8051), 769–778. https://doi.org/10.1038/s41586-024-08378-w

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., ... Poon, H. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*, *630*(8015), 181–188. https://doi.org/10.1038/s41586-024-07441-w

Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., Tschandl, P., Hu, M., Ju, L., Tan, G., Tang, V., Ng, A. B., Powell, D., Bonnington, P., See, S., Magnaterra, E., Ferguson, P., Nguyen, J., Guitera, P., Banuls, J., Janda, M., Mar, V., Kittler, H., Soyer, H. P., & Ge, Z. (2025). A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, *31*(8), 2691–2702. https://doi.org/10.1038/s41591-025-03747-y

Yang, Y., Liu, Y., Liu, X., Gulhane, A., Mastrodicasa, D., Wu, W., Wang, E. J., Sahani, D., & Patel, S. (2025). Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, *11*(13), Article eadq0305. https://doi.org/10.1126/sciadv.adq0305

Zambrano Chaves, J. M., Huang, S.-C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., Awadalla, H., Gong, J., Hu, H., Yang, J., Li, C., Gao, J., Gu, Y., Wong, C., Wei, M., ... Poon, H. (2025). A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications*, *16*(1), Article 3108. https://doi.org/10.1038/s41467-025-58344-x

Zhang, Z., Qadir, M. I., Carstens, M., Zhang, E. H., Loiselle, M. S., Martinus, F. M., Mroczkowski, M. K., Clusmann, J., Kather, J. N., & Kolbinger, F. R. (2026). Prompt injection attacks on vision-language models for surgical decision support. *npj Digital Surgery*, *1*(1), Article 15. https://doi.org/10.1038/s44484-026-00014-6

Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H. H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., Piening, B., Bifulco, C., Wei, M., Poon, H., & Wang, S. (2025). A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, *22*(1), 166–176. https://doi.org/10.1038/s41592-024-02499-w

Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., & Gao, X. (2024). Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nature Communications*, *15*(1), Article 5649. https://doi.org/10.1038/s41467-024-50043-3

## Supplementary Material

Supplementary File 1. A priori review protocol, comprising research questions, eligibility criteria, search domains, extraction fields, the Clinical Evaluation Readiness Ladder and the analysis plan.

Supplementary File 2. Full search strategy, listing all queries by domain, dates executed and yields, reported in accordance with PRISMA-S.

Supplementary File 3. Complete study-level dataset comprising 125 records and 15 fields, including readiness assignment with supporting evidence, together with the log of borderline eligibility decisions and the reasons for each.

Supplementary File 4. Analysis code, figure-generation code and the automated verification script that re-derives every quantitative statement in this article from the dataset.

**How to cite this article.** Padhi, A. (2025). Clinical evaluation readiness of multimodal foundation models in healthcare: A systematic review and evidence-gap map of 125 studies from benchmark to bedside. *Computational Diagnostics*, *1*(1), Article 4.

 Received 6 January 2025; published 7 March 2025.