

EVALUATION

Evaluating clinical language models: benchmarks, blind spots, and what the evidence supports

ABSTRACT

Vendor accuracy claims for clinical language models almost always resolve to a score on an examination-style benchmark. This paper reviews what the published evaluation literature has measured, corrects three widely repeated claims that the underlying papers do not support, and sets out the properties an evaluation needs before an accuracy figure carries information. It concludes that examination performance and performance on real records are separated by a measured gap, and that the tasks most often automated have barely been evaluated at all.

Type Critical review of evaluation practice **References** 28 **Reading time** 15 min

Last reviewed September 2026

1 What “accurate” has referred to

When a clinical software vendor describes its language model as accurate, the claim usually resolves to a score on a multiple-choice examination. That is not a rhetorical complaint. It is the measured shape of the literature.

A systematic review published in *JAMA* in 2025 assessed 519 studies of large language models in health care. Only 5% used real patient care data; 44.5% used medical examination questions. Accuracy was the primary metric in 95.4% of studies, and only 15.8% assessed fairness, bias, or toxicity. The task distribution is more telling: question answering 84.2%, summarization 8.9%, dialogue 3.3%, billing codes and prescriptions 0.2% each.¹ Those last two describe much of what an ambulatory practice delegates to software.

5%

of 519 health-LLM studies used real patient care data

44.5%

evaluated on medical examination questions

0.2%

task coverage for billing codes, and the same for prescriptions

88%

of 56 medical benchmarks with no contamination-risk mitigation

This paper makes three claims. The headline figures in circulation descend from a narrow lineage of examination-derived benchmarks whose provenance is routinely misstated. The properties that would make a benchmark score interpretable, contamination control chief among them, are missing from most medical benchmarks. And the few studies that measured clinical work rather than examination performance yield a less flattering, more useful picture. It is a review of others' published research and contains no new data.

2 Where the famous numbers come from

Three papers supply most of the accuracy figures repeated in sales material. Each is misquoted in a specific way.

The Kung paper is not "ChatGPT passed the USMLE"

Kung et al. tested ChatGPT on 350 USMLE items drawn from a starting set of 376.² These were practice materials rather than the administered examination, and the model was GPT-3.5. Censoring indeterminate responses gave Step 1 75.0%, Step 2CK 61.5%, and Step 3 68.8%; without censoring the same figures were 45.4%, 54.1%, and 61.5%. The paper described this as performance at or near the passing threshold, which is around 60% and varies by Step and by year.

The popular claim therefore fails three ways: the items were practice items, the uncensored figures for two of three Steps fall below the threshold, and the study's own language stops short of the claim.

The twenty-point margin belongs to a vendor preprint

The claim usually paired with Kung, that the model exceeded the passing score by more than 20 points, comes from Nori et al., who evaluated GPT-4 on USMLE practice materials and MultiMedQA.³ That work is an arXiv preprint from March 2023, not peer reviewed, authored by the research laboratory of the company that produced the model. The figure carries the weight of a vendor self-report, and attributing it to Kung merges two papers testing two different models.

Med-PaLM 2 did not outperform physicians

Singhal et al. introduced MultiMedQA and reported Flan-PaLM at 67.6% on MedQA, more than 17 points above the prior state of the art; the paper's own framing was that this exposed the limits of the benchmarks rather than clinical readiness.⁴ The follow-up reported Med-PaLM 2 at 86.5% on MedQA, 72.3% on MedMCQA, and 75.0% on PubMedQA, rising to 81.8% with self-consistency.⁵

The "outperforms physicians" claim attaches to a different measurement in the same paper. Across 1,066 MultiMedQA questions, physician raters compared long-form answers pairwise and preferred Med-PaLM 2 on 8 of 9 axes; on better reflecting medical consensus it was preferred 72.9% of the time. A separate bedside-consultation set of 20 questions had specialists preferring it to generalist physician answers 65% of the time.⁵ That is a rater preference over text: no patient, no encounter, no outcome.

The 86.5% measures multiple-choice accuracy, the preference result measures prose, and the two are routinely compressed into one sentence that neither supports. Ayers et al. is the precedent: raters preferred chatbot answers in 78.6% of 585 evaluations, but the physician comparators were unpaid volunteers on a public forum averaging 52 words against the chatbot's 211.⁶

3 What a benchmark score does not control for

MedCheck audited 56 medical LLM benchmarks against a lifecycle rubric and found an evaluation infrastructure largely without controls. 88% (49 of 56) had no contamination-risk mitigation of any kind, 89% could not test robustness, 91% did not evaluate uncertainty awareness, half did not align with ICD or SNOMED CT terminology, only 38% used authentic real-world scenarios, and 80% had no maintenance plan.⁷ That audit is itself a non-peer-reviewed preprint.

Why contamination makes a score an upper bound

A benchmark score is meant to estimate performance on unseen items. If the items were in the training data, the score measures retrieval rather than generalization, and the measured value is at least as high as the true one. The direction of the error is known; the magnitude is not, because nobody measured it. A contaminated benchmark yields an upper bound of unknown tightness, which is a weaker object than an estimate.

Examination items are the worst case: widely reproduced online, stable across years, likely to be in a large web crawl. Set beside the finding that 44.5% of the literature runs on examination questions while 5% uses real patient data,¹ the field's central quantity is an upper bound of unknown tightness on a task nobody deploys.

Benchmarks aimed at the actual work

MedHELM is the most developed attempt to cover clinical work: a clinician-validated taxonomy of 5 categories, 22 subcategories, and 121 tasks, assembled from 37 benchmarks and applied to 9 frontier models. Its argument is that licensing-examination benchmarks do not represent real clinical work.⁸ The 15.8% fairness figure carries a cost accuracy metrics cannot detect. Omiye et al. put nine questions to four models across five runs each and found all four reproduced debunked race-based claims, including race-adjusted eGFR justified by false assertions about muscle mass in Black patients; the authors concluded the models were not ready for clinical use.⁹ No MedQA item asks any of that.

4 What happens when the record is real

Hager et al. evaluated models on 2,400 real patient cases from MIMIC-IV across four abdominal pathologies.¹⁰ Hospitalists achieved 87.5% to 92.5% diagnostic accuracy; the open models achieved 58.8% (Llama 2 Chat), 67.8% (OASST), and 65.1% (WizardLM), a gap of 16 to 25 percentage points.

Accuracy fell further when a model had to gather information autonomously rather than receive it, from 58.8% to 45.5% in one case. Laboratory-value interpretation ranged from 24.1% to 77.2% depending on model and direction, guideline adherence was inconsistent, and the models hallucinated nonexistent tools every two to five patients.

The autonomous-collection result transfers directly to purchasing decisions. A demonstration presents a curated case: relevant history assembled, pertinent labs surfaced, question well posed. Live use requires the system to decide what to look at, which is the step at which accuracy fell. A demonstration that hands the model its context measures a task the deployed system will not face.

Tool hallucination is a separate failure mode: not a wrong answer but a confident invocation of a capability that does not exist. Inside an EMR that is an order or an interface the model believes is available, so the error surfaces as an action rather than as a sentence.

LIMITATIONS

This study evaluated open-weight models available at the time, on four abdominal pathologies, retrospectively, against structured MIMIC-IV records. It establishes that the gap is measurable and that it widens when the model must gather its own information. It does not establish the size of that gap for current proprietary models, for other presentations, or in prospective use. No study reviewed here repeats the design on current frontier models, so the present gap is unknown rather than small.

5 Hallucination, and the denominator problem

“Large language models hallucinate X% of the time” is not a claim that can be evaluated, because the published rates use at least four incompatible denominators. Table 1 sets them side by side for the sole purpose of showing that they cannot be placed side by side.

TABLE 1

Measured hallucination rates in clinical text, with the unit each is measured in. The column that governs interpretation is the second, not the third.

Study	Denominator	Rate	What limits the reading
Asgari 2025 ¹¹	Sentences in generated summaries (12,999 annotated)	1.47%, of which 44% graded major	Scales with note length; major errors clustered in the Plan section (21%)
Palm 2025 ¹²	Notes containing at least one hallucination	31% of ambient notes vs 20% of physician notes (p=0.01)	Note-level presence, not error density; 97 encounters, 194 notes
Chelli 2024 ¹³	Generated bibliographic references	GPT-3.5 39.6%, GPT-4 28.6%, Bard 91.4%	Reference generation is a distinct task from note generation
Bhattacharyya 2023 ¹⁴	Generated references (115 across 30 short papers)	47% entirely fabricated; 46% authentic but inaccurate	Two error classes with different consequences, collapsed by casual citation
Omar 2025 ¹⁵	Adversarial vignettes seeded with one fabricated element	50–53.3% to 80–82.7% by model; 66% overall	Deliberately adversarial input; an upper bound on elaboration, not a base rate

The per-sentence figure comes from 12,999 clinician-annotated sentences across 450 note and transcript pairs, with the major hallucinations concentrated in the Plan section. Omission was more than twice as common, at 3.45% of 49,590 annotated items.¹¹ Omission appears in no vendor hallucination statistic and is harder to catch on review, since nothing incorrect is present to notice.

The note-level figure comes from a blinded comparison of 194 notes across 388 paired specialist reviews. Reviewers preferred the ambient note 47% of the time against 39% for the physician draft, and rated overall quality at near parity, while those same ambient notes carried more hallucinations.¹² Preference and error rate moved in opposite directions within one study, a warning about evaluating documentation tools by reviewer preference.

The adversarial result matters most for design. Omar et al. seeded each of 300 physician-validated vignettes with a single fabricated element, a fake laboratory value, sign, or condition, and measured whether the model elaborated on it. Elaboration ran from 50–53.3% for GPT-4o to 80–82.7% for a distilled DeepSeek-R1, 66% overall, and a mitigation prompt cut the overall rate to 44%. Setting temperature to zero produced no significant improvement.¹⁵ Determinism is not accuracy. A temperature-zero model returns the same output every time, including the same fabrication.

6 The input layer: speech recognition

Every claim about a voice-driven tool rests on a transcript, which has its own error rate, usually omitted from the tool’s accuracy figure. Zhou et al. remains the reference measurement: 217 clinical documents comprising 83 office notes, 75 discharge summaries, and 59 operative notes. Errors per 100 words were

7.4 in raw speech-recognition output, 0.4 after professional transcriptionist editing, and 0.3 in the physician-signed note. Quoted alone, that sequence reads as a review pipeline working.

The second series in the same study is the one that matters. The proportion of remaining errors that were clinically significant was 5.7% at recognizer output, 8.9% after the transcriptionist, and 6.4% in the signed note.¹⁶ Human review cut error volume roughly twentyfold and did not reduce the share of surviving errors that carried clinical consequence. Review removes the errors that are easy to see; a plausible wrong laterality survives a read-through that catches a garbled word. A pipeline reporting a post-review error rate reports a quantity that improves while the risk does not.

Goss et al. found 1.3 errors per note in 100 emergency department dictated notes, up to 15% of errors judged critical, and clinicians perceiving error rates as lower than measured; those figures reach this paper at second hand and should be checked against the original.¹⁷

Whisper is often cited here, and usually incorrectly. Koencke et al. found it produced entirely hallucinated phrases or sentences in roughly 1% of transcriptions, text with no counterpart in the audio, and that 38% of those hallucinations contained explicit harms, including perpetuating violence and implying false authority; fabrication was disproportionately triggered by longer non-vocal pauses.¹⁸ That study used aphasia speech corpora and control speech, not medical dictation. It does not establish an error rate for clinical transcription, and the circulated accounts of Whisper fabricating in hospital transcription derive from journalism rather than from this paper. What it establishes is a mechanism and a disparate impact: any speaker who pauses is at elevated risk, patients with aphasia among them. In ambient documentation the disfluent speaker is frequently the patient.

7 Trials that measured outcomes rather than scores

A small number of randomized trials measured what happened when clinicians used these systems. They do not point in one direction.

TABLE 2

Randomized trials of LLM-assisted clinical work. Three measure rated reasoning scores on vignettes; the fourth measures a clinical event in real patients.

Trial	Design and participants	Primary comparison	Result
Goh 2024, diagnostic reasoning ¹⁹	Single-blind RCT; 50 physicians, 6 vignettes, 244 cases	LLM plus conventional resources vs conventional resources	76% vs 74% median; adjusted difference 2 points (95% CI -4 to 8; P = .60). LLM alone 92%
Goh 2024, management reasoning ²⁰	RCT, preprint; 92 attendings and residents	GPT-4 access vs no access	+6.5 percentage points (95% CI 2.7–10.2, P<0.001); +119.3 s per case
Everett 2026 ²¹	RCT; 70 clinicians, 254 cases	Three workflow positions vs conventional resources	Baseline 75%; second opinion 82%; first opinion 85%; AI alone 87%
Agweyu 2026 ²²	Pragmatic cluster RCT; 16 facilities, 103 clinical officers, 9,347 patients analyzed	Generative CDS vs usual care, real patients	Treatment failure at 14 days 2.2% vs 2.0%; adjusted OR 0.77 (95% CI 0.55–1.08, P = 0.13)

The diagnostic-reasoning trial is routinely reported as evidence that AI beats doctors. The accurate reading is narrower and more useful: physicians using the model did no better than physicians without it, and no faster (519 seconds against 565, $P = .20$), while the model alone did better than both arms.¹⁹ That is an integration failure, not a capability claim, and it is invisible to every benchmark described above. A benchmark measures a model; this trial measured a model plus a clinician plus an interface, and the combination underperformed its best component.

The opposite generalization is equally unsupported. Goh's management-reasoning trial found management decisions up 6.1% ($P=0.001$) and diagnostic decisions up 12.1% ($P=0.009$), with augmented physicians not differing from GPT-4 alone (-0.9% , $P=0.8$). It is a medRxiv preprint, and its publication status should be confirmed before it is cited as peer-reviewed.²⁰

Everett et al. suggests the operative variable is workflow position rather than model quality. Against a 75% baseline, AI as a second opinion reached 82% (+6.8%, 95% CI 4.0–9.6%) and AI as a first opinion 85% (+9.9%, 95% CI 4.7–15%), while AI alone reached 87%.²¹ The same class of model produced different collaborative results depending on when the clinician saw its output: a property of the product, not of the model.

The one trial with real patients and a clinical primary outcome is the least flattering and the most relevant. Its secondary outcomes showed improved documented diagnostic quality and treatment planning, with no differences in antibiotic or antimalarial prescribing, and its 33 serious adverse events were adjudicated as unrelated to the intervention.²² Across all four trials, documentation quality and rated reasoning move more readily than patient outcomes.

LIMITATIONS

Three of these four trials used vignettes rather than patients, and their outcome is a reasoning score rated by humans rather than a clinical event. The fourth used real patients, but its primary outcome was treatment failure at 14 days with a low event rate, 2.0% in control, so a null result there stays compatible with small true effects in both directions. None of the four measured billing accuracy, prescription safety, or documentation burden, leaving the highest-volume administrative applications without randomized evidence.

8 Reporting standards, and what they would and would not fix

Four reporting guidelines now cover this territory, and they are frameworks rather than evidence. TRIPOD-LLM extends TRIPOD+AI to language models with a modular checklist covering development, tuning, prompting, and evaluation.²³ DECIDE-AI covers live clinical evaluation between offline validation and large comparative trials, the stage most relevant to software already shipping AI features, with 27 reporting items.²⁴ CONSORT-AI added 14 items to CONSORT 2010 through a two-round Delphi and a 34-participant pilot.²⁵ SPIRIT-AI added 15 items to SPIRIT 2013 through the same process, covering trial protocols rather than trial reports.²⁶

Applied to the studies in section 2, these checklists would have forced disclosure of the items used, the model version and evaluation date, whether contamination was assessed, how indeterminate responses were handled, and who funded and authored the evaluation. That is most of the correction work this paper has done by hand.

What they do not do is measure anything. A reporting guideline improves the description of a study, not the study, and adherence is voluntary. A perfectly reported evaluation of a contaminated benchmark is still an evaluation of a contaminated benchmark.

The assurance proposals aim at that gap. Shah et al. proposed a nationwide network of health AI assurance laboratories issuing performance reports usable across populations and settings.²⁷ The Coalition for Health AI published an Assurance Standards Guide setting out a six-stage lifecycle from problem definition through ethical design, engineering, deployment and monitoring; it is an industry consensus standard, not peer-reviewed research.²⁸ Both remain proposals. Neither has produced published evidence that a model reviewed under one performs better in deployment.

9 What follows

The answerable question is not whether a clinical language model is accurate. It is: on what task, measured how, against which comparator, and when. The literature reviewed here supports five properties that separate an evaluation worth reading from a number worth ignoring.

1. **Task match.** The evaluation covers the work the software will perform rather than a proxy for it. Coding and prescribing carry the thinnest published evidence and are among the most commonly automated tasks.¹
2. **Data provenance.** Real records rather than curated vignettes, and an explicit statement of whether the model was handed its context or had to gather it.¹⁰
3. **A stated denominator.** Every error rate names its unit: per sentence, per note, per reference, per case. A figure quoted without its unit conveys nothing.^{11,12,13,14,15}
4. **Contamination handling.** Either evidence that the evaluation items could not have been in the training data, or a plain statement that this was not assessed, in which case the score is an upper bound rather than an estimate.⁷
5. **An outcome, or an honest absence of one.** Reasoning scores and preference ratings measure real things, none of which is a patient outcome. A null outcome is information about the intervention, not a defect in the study.^{19,22}

None of this requires evidence that does not yet exist. The reporting frameworks are published, benchmarks aimed at clinical content exist, and randomized trials with clinical endpoints have been run. What is missing is coverage: the field has measured examination performance extensively and the clinical and administrative work these systems are sold to do barely at all. Until that changes, an accuracy claim should be read as a statement about a benchmark, and a demonstration as a statement about a curated case. Neither is a statement about a practice.

REFERENCES

Entries 1, 7 and 8 are the meta-evaluations on which the argument rests. Entries 2 to 22 are primary studies; of those, 10, 16, 17 and 19 to 22 are the ones that measured real records, real transcripts, or real clinicians, and they carry the most weight here. Entries 3, 7 and 20 are non-peer-reviewed preprints and entry 28 is an industry consensus document, each identified as such where the text relies on it. Entries 23 to 28 are reporting frameworks and proposals rather than evidence. Every figure is quoted at the precision its source reports.

1. Bedi S, Liu Y, Orr-Ewing L, et al. *Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review*. JAMA. 2025;333(4):319–328.
[doi:10.1001/jama.2024.21700](https://doi.org/10.1001/jama.2024.21700) SYSTEMATIC REVIEW

2. Kung TH, Cheatham M, Medenilla A, et al. *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models*. PLOS Digital Health. 2023;2(2):e0000198. [doi:10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198) BENCHMARK
3. Nori H, King N, McKinney SM, et al. *Capabilities of GPT-4 on Medical Challenge Problems*. arXiv preprint arXiv:2303.13375. March 2023. arxiv.org/abs/2303.13375 PREPRINT
4. Singhal K, Azizi S, Tu T, et al. *Large language models encode clinical knowledge*. Nature. 2023;620(7972):172–180. [doi:10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2) BENCHMARK
5. Singhal K, Tu T, Gottweis J, et al. *Toward expert-level medical question answering with large language models*. Nature Medicine. 2025;31:943–950. [doi:10.1038/s41591-024-03423-7](https://doi.org/10.1038/s41591-024-03423-7) BENCHMARK
6. Ayers JW, Poliak A, Dredze M, et al. *Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum*. JAMA Internal Medicine. 2023;183(6):589–596. [doi:10.1001/jamainternmed.2023.1838](https://doi.org/10.1001/jamainternmed.2023.1838) CROSS-SECTIONAL
7. Chen W, Yu G, Cheung Y-F, et al. *Beyond the Leaderboard: Rethinking Medical Benchmarks for Large Language Models*. arXiv preprint arXiv:2508.04325v2. Version 2 dated 29 April 2026. arxiv.org/abs/2508.04325 PREPRINT
8. Bedi S, Cui H, Fuentes M, et al. *Holistic evaluation of large language models for medical tasks with MedHELM*. Nature Medicine. 2026;32(3):943–951. [doi:10.1038/s41591-025-04151-2](https://doi.org/10.1038/s41591-025-04151-2) BENCHMARK
9. Omiye JA, Lester JC, Spichak S, et al. *Large language models propagate race-based medicine*. npj Digital Medicine. 2023;6:195. [doi:10.1038/s41746-023-00939-z](https://doi.org/10.1038/s41746-023-00939-z) BENCHMARK
10. Hager P, Jungmann F, Holland R, et al. *Evaluation and mitigation of the limitations of large language models in clinical decision-making*. Nature Medicine. 2024;30(9):2613–2622. [doi:10.1038/s41591-024-03097-1](https://doi.org/10.1038/s41591-024-03097-1) BENCHMARK
11. Asgari E, Montaña-Brown N, Dubois M, et al. *A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation*. npj Digital Medicine. 2025;8:274. [doi:10.1038/s41746-025-01670-7](https://doi.org/10.1038/s41746-025-01670-7) BENCHMARK
12. Palm E, Manikantan A, Mahal H, et al. *Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe*. Frontiers in Artificial Intelligence. 2025;8:1691499. [doi:10.3389/frai.2025.1691499](https://doi.org/10.3389/frai.2025.1691499) BENCHMARK
13. Chelli M, Descamps J, Lavoué V, et al. *Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis*. Journal of Medical Internet Research. 2024;26:e53164. [doi:10.2196/53164](https://doi.org/10.2196/53164) BENCHMARK

14. Bhattacharyya M, Miller VM, Bhattacharyya D, et al. *High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content*. *Cureus*. 2023;15(5):e39238. [doi:10.7759/cureus.39238](https://doi.org/10.7759/cureus.39238) BENCHMARK
15. Omar M, Sorin V, Collins JD, et al. *Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support*. *Communications Medicine*. 2025;5:330. [doi:10.1038/s43856-025-01021-3](https://doi.org/10.1038/s43856-025-01021-3) BENCHMARK
16. Zhou L, Blackley SV, Kowalski L, et al. *Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists*. *JAMA Network Open*. 2018;1(3):e180530. [doi:10.1001/jamanetworkopen.2018.0530](https://doi.org/10.1001/jamanetworkopen.2018.0530) CROSS-SECTIONAL
17. Goss FR, Zhou L, Weiner SG. *Incidence of speech recognition errors in the emergency department*. *International Journal of Medical Informatics*. 2016;93:70–73. PMID: 27435949. pubmed.ncbi.nlm.nih.gov/27435949 CROSS-SECTIONAL
18. Koenecke A, Choi ASG, Mei KX, et al. *Careless Whisper: Speech-to-Text Hallucination Harms*. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FACCT '24)*. arXiv:2402.08021. arxiv.org/abs/2402.08021 BENCHMARK
19. Goh E, Gallo R, Hom J, et al. *Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial*. *JAMA Network Open*. 2024;7(10):e2440969. [doi:10.1001/jamanetworkopen.2024.40969](https://doi.org/10.1001/jamanetworkopen.2024.40969) RCT
20. Goh E, Gallo R, Strong E, et al. *Large Language Model Influence on Management Reasoning: A Randomized Controlled Trial*. *medRxiv preprint*. 2024. [doi:10.1101/2024.08.05.24311485](https://doi.org/10.1101/2024.08.05.24311485) PREPRINT
21. Everett SS, Bunning BJ, Jain P, et al. *From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis*. *npj Digital Medicine*. 2026;9:409. [doi:10.1038/s41746-026-02545-1](https://doi.org/10.1038/s41746-026-02545-1) RCT
22. Agweyu A, Mwaniki P, Menon V, et al. *Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial*. *Nature Medicine*. 2026;32:3032–3039. [doi:10.1038/s41591-026-04503-6](https://doi.org/10.1038/s41591-026-04503-6) RCT
23. Gallifant J, Afshar M, Ameen S, et al. *The TRIPOD-LLM reporting guideline for studies using large language models*. *Nature Medicine*. 2025;31(1):60–69. [doi:10.1038/s41591-024-03425-5](https://doi.org/10.1038/s41591-024-03425-5) GUIDANCE
24. Vasey B, Nagendran M, Campbell B, et al. *Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI*. *Nature Medicine*. 2022;28(5):924–933. [doi:10.1038/s41591-022-01772-9](https://doi.org/10.1038/s41591-022-01772-9) GUIDANCE

25. Liu X, Cruz Rivera S, Moher D, et al. *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension*. Nature Medicine. 2020;26(9):1364–1374. [doi:10.1038/s41591-020-1034-x](https://doi.org/10.1038/s41591-020-1034-x) **GUIDANCE**
26. Cruz Rivera S, Liu X, Chan A-W, et al. *Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension*. Nature Medicine. 2020;26(9):1351–1363. [doi:10.1038/s41591-020-1037-7](https://doi.org/10.1038/s41591-020-1037-7) **GUIDANCE**
27. Shah NH, Halamka JD, Saria S, et al. *A Nationwide Network of Health AI Assurance Laboratories*. JAMA. 2024;331(3):245–249. [doi:10.1001/jama.2023.26930](https://doi.org/10.1001/jama.2023.26930) **POSITION PAPER**
28. Coalition for Health AI. *Assurance Standards Guide*. Released for public comment 26 June 2024; no stable version number is shown on the document. chai.org **STANDARD**

RELATED PAPERS

Ambient documentation and the clinician's day

What the randomized trials of ambient scribes actually measured.

Human oversight of clinical AI

The reviewer is the control, and the reviewer is measurably unreliable.

Retrieval over the chart

Assembling clinical context a model can be trusted to read.