

SAFETY

Human oversight of clinical AI: what the automation-bias literature actually shows

ABSTRACT

Clinical AI is deployed on the premise that a clinician reviews the output before it reaches a patient. This paper appraises the experimental evidence for that premise. Review reliably reduces the volume of errors and does not reliably reduce the share that are clinically significant; where the machine is wrong, reviewers perform worse than unaided clinicians, and training and experience attenuate but do not remove the effect. Design consequences and three commonly misquoted findings are set out.

Type Narrative review with design implications **References** 27 **Reading time** 15 min

Last reviewed September 2026

1 The control that everything rests on

Almost every argument for deploying artificial intelligence in clinical care ends at the same sentence. A clinician reviews the output before it reaches a patient. The suggestion is a draft; the note is not filed until a physician attests to it. That sentence carries the whole deployment: it distinguishes assistance from autonomy, keeps a diagnostic feature outside the regulatory definition of a device, and converts a model error into a near miss.

It is also an empirical claim, and a narrow one: that a human reviewer, under real conditions, detects machine error at a rate high enough to make the combined system safer than either component alone. That claim has been tested for four decades, in prescribing, radiology, dictation and now language-model use. This paper collects what the tests found.

The short version is not that review is worthless. It is more specific and less comfortable. Human review reliably reduces the number of errors that survive into the record. It does not reliably reduce the share of survivors that are dangerous. And when the machine is wrong, the reviewer performs worse than with no machine at all. Those findings mean that “a clinician reviews it” is not yet a control. It becomes one when somebody specifies which errors it catches, at what cost, and how that is re-measured.

2 Automation bias, in its measured form

Goddard, Roudsari and Wyatt's systematic review retained 74 studies from 13,821 records and supplies the working definition.¹ Automation bias is over-reliance on an automated aid, producing two error classes: omission errors, which the reviewer misses because the system did not flag them, and commission errors, which the reviewer commits because the system flagged them wrongly. Mediators group into user factors such as cognitive style and experience, attitudinal factors such as trust and confidence, and environmental factors such as workload and time pressure. No pooled frequency estimate is reported: the phenomenon is well evidenced, its base rate is not.

It is not a multitasking problem

Lyell and Coiera screened nine databases covering 1983 to 2015 and analyzed 40 studies, only 6 of them healthcare-focused.² Their result contradicts the dominant human-factors assumption. Automation bias was not principally a divided-attention artifact. It appeared in single tasks, typically diagnosis rather than monitoring, and it tracked verification complexity: how hard it was to check the machine's suggestion.

That names a variable the designer controls. If the bias were driven by multitasking, the remedy would be workload policy; because it is driven by the cost of verification, the remedy is interface design. It also disposes of the standard mitigation: telling clinicians to check more carefully asks for effort where effort is most expensive. Reducing the cost of checking changes the price.

Verification cost inside an electronic record has been rising. Rule and colleagues analyzed outpatient progress notes at one academic medical center across the decade to 2018.³ Median note length rose 60.1%, from 401 words to 642; median redundancy rose 10.9 percentage points, from 47.9% to 58.8%; and 2018 notes averaged 29.4% directly typed text. A reviewer checking a generated claim against that record is checking one artifact of uncertain provenance against another.

3 What happens when the machine is wrong

+86.6%

prescribing errors with
incorrect decision support

36.8–59.9 pts

accuracy lost by radiologists
given wrong AI suggestions

–14.0 pts

diagnostic reasoning, AI-
literacy-trained physicians

Three experiments isolate the quantity that matters. Each gave clinicians decision support that was wrong by design, measured against a control with no support at all. That comparison, not the one against correct support, tests whether review protects.

Lyell and colleagues ran a randomized laboratory experiment with 120 final-year medical students on prescribing tasks.⁴ Correct support reduced omission errors by 38.3% to 46.6%. Incorrect support increased omission errors by 24.5% to 33.3% and commission errors by 51.7% to 65.8%, relative to no support. Total prescribing errors rose 86.6% with incorrect support and fell 58.8% with correct support: a wrong suggestion was substantially worse than silence.

Dratsch and colleagues had 27 radiologists read 50 mammograms, 38 with correct AI BI-RADS suggestions and 12 incorrect.⁵ On the incorrect cases accuracy fell from 79.7% to 19.8% among inexperienced readers, 81.3% to 24.8% among the moderately experienced and 82.3% to 45.5% among the very experienced: losses of 59.9, 56.5 and 36.8 percentage points, all at $p \leq .003$. Experience mattered and rescued nobody. The most experienced group was still wrong more often than right.

Qazi and colleagues ran a single-blind randomized trial with 44 physicians, whose treatment arm received deliberately flawed language-model output.⁶ Diagnostic reasoning scored 73.3% against 84.9%, an adjusted -14.0 percentage points (95% CI -19.7 to -8.3 ; $P < .0001$); top-choice accuracy 76.1% against 90.5%, -18.3 points. Every participant had received AI-literacy training. It did not protect them.

TABLE 1

The three experiments in which decision support was wrong by design, each against a no-support control.

Study	Design	What the reviewer saw	Effect when the machine was wrong
Lyell 2017 ⁴	Randomized, 120 students	Prescribing support, right or wrong	Prescribing errors +86.6% (correct: -58.8%)
Dratsch 2023 ⁵	Reader study, 27 radiologists	AI BI-RADS category, wrong on 12 of 50	Accuracy -59.9, -56.5, -36.8 points
Qazi 2025 ⁶	Single-blind RCT, 44 physicians	Flawed language-model output	Reasoning -14.0 points; top diagnosis -18.3

WHAT SEEDED-ERROR EXPERIMENTS DO AND DO NOT ESTABLISH

The strongest evidence here comes from studies in which errors were introduced deliberately, at rates no deployed system would match, and on vignettes rather than in live care. They measure precisely what a reviewer does given that the machine is wrong, but not how often a deployed system is wrong, and cannot be combined into an unconditional error rate for any product. The conditional failure of review is well established; its base rate is not.

4 Review reduces the number of errors, not their danger

The most important result for anyone designing a review step was published in 2018 and concerns dictation, not AI. Zhou and colleagues examined 217 clinical documents at three stages of a dictation pipeline.⁷ Errors per 100 words fell from 7.4% in raw speech-recognition output to 0.4% after a professional transcriptionist edited it, and 0.3% in the physician-signed note: roughly a twentyfold cut in error volume. That is the finding usually quoted.

The second is not. The proportion of remaining errors judged clinically significant was 5.7% in raw output, 8.9% after the transcriptionist and 6.4% in the signed note. It did not fall. A pipeline that removed most errors left the danger profile of the survivors unchanged.

Why this generalizes

The mechanism is not specific to speech recognition. Reviewers catch errors that look wrong, meaning locally implausible: a garbled word, a value outside a familiar range. Clinical significance is a different property. A dropped negation leaves a grammatical sentence; a reversed laterality reads normally; a plausible dose on the wrong indication reads as a dose. Catchability and danger are close to independent, so a low residual error count is not evidence of safety. Nor is a reviewer's confidence: Goss and colleagues found 1.3 errors per note across 100 emergency department attending notes, up to 15% of them critical — figures that reach this paper at second hand — with clinicians consistently perceiving error rates as lower than measured.⁸

The same shape in generated notes

The review step for ambient documentation has the same structure: a clinician reads generated text against a memory of an encounter. Asgari and colleagues annotated 12,999 sentences for hallucination and 49,590 for omission across 450 note-transcript pairs.⁹ Hallucination ran at 1.47%, 44% of it graded major; omission at 3.45%, 16.7% major. Major hallucinations clustered in the Plan section, at 21%, and negations were 30% of hallucinations; negations and omissions leave fluent text and survive a read-through.

Blinded evaluations of note quality disagree. Palm and colleagues compared 194 notes in 388 paired specialist reviews: quality was near parity, but 31% of ambient notes contained hallucinations against 20% of physician-drafted notes ($p = 0.01$).¹⁰ Reddy and colleagues, comparing 11 AI scribe tools against 18 human note takers on five standardized cases, found human notes higher on the modified PDQI-9 in all five and AI lower on all ten domains, the largest gap 43.8 against 20.3.¹¹ Note quality is a property of a product, not a category, and self-report captures neither: in a randomized trial across 238 physicians, clinically significant inaccuracies were rated as occurring only “occasionally”.¹²

Two findings bound what a reviewer can detect at all. Koenecke and colleagues found a widely used speech-to-text model produced entirely hallucinated phrases in roughly 1% of transcriptions, 38% of them containing explicit harms, triggered disproportionately by longer non-vocal pauses; that work used

aphasia and control speech corpora, not medical dictation, so it evidences a mechanism and a disparate impact rather than a clinical error rate.¹³ A scoping review of emergency ambient documentation, used here for framing only, reports higher word error rates for Black speakers than white speakers.¹⁴ A reviewer cannot detect an error whose trigger they cannot perceive.

HOW FAR THE ZHOU RESULT CARRIES

No study here reproduces the 5.7% / 8.9% / 6.4% pattern for language-model-generated notes. What carries over is the structural claim: reviewers remove errors that look wrong, and clinical significance is not strongly correlated with looking wrong. Whether the proportions hold for ambient documentation is what a deployment should measure.

5 Collaboration when nothing is deliberately broken

Trials of ordinary use answer a different question, and their results are quieter. Goh and colleagues randomized 50 US-licensed physicians across 6 vignettes and 244 cases.¹⁵ Physicians with a language model scored a median 76% on diagnostic reasoning; those with conventional resources 74%. The adjusted difference was 2 points (95% CI -4 to 8; $P = .60$), and time per case did not differ. The model alone scored 92%, 16 points above the control arm (95% CI 2 to 30; $P = .03$).

The trial is widely cited as evidence that AI beats doctors. That takes the least useful number in it. The finding with content is that physicians given the model did no better than those without it, and the combined system underperformed one of its own components: an integration failure, not a capability claim.

The mirror-image over-reading, that language models do not help clinicians, is equally unsupported. Goh's management-reasoning trial, a preprint at the time of writing, randomized 92 clinicians and found GPT-4 access improved management reasoning by 6.5 percentage points (95% CI 2.7 to 10.2; $P < 0.001$) at a cost of 119.3 extra seconds per case.¹⁶ The augmented physicians again did not differ from the model alone.

Where the loss happens

Everett and colleagues randomized 70 clinicians across 254 cases and varied the position of the model in the workflow rather than the model itself.¹⁷ Conventional resources scored 75%; the model as a second opinion 82% (+6.8%); as a first opinion 85% (+9.9%); the model alone 87%, not significantly different

from either collaborative arm. What moved the result was where the model sat in the sequence of work, not how good it was. Draft-then-review, the arrangement usually described as the safety control, was not the best-performing one tested.

A 2026 systematic review and meta-analysis pooled 10 peer-reviewed studies.¹⁸ Diagnostic accuracy for human plus model against human alone gave a risk ratio of 1.59 (95% CI 0.08 to 32.74), neither significant nor informative. Composite diagnostic and management scores improved by 4.88 percentage points (95% CI 0.65 to 9.12), but the 95% prediction interval ran from −31.65 to 41.42 and crossed the null. Factual error rates stayed around 26% to 36%, and in three-arm comparisons collaboration did not universally outperform the model alone. The authors name the supervision effort a “vigilance tax”. The prediction interval is the honest summary: the average effect is small, and the next study could land almost anywhere.

One trial used real patients and a clinical primary outcome. Agweyu and colleagues ran a pragmatic cluster-randomized trial across 16 primary care facilities, analyzing 9,347 patients.¹⁹ Treatment failure at 14 days was 2.2% against 2.0%, adjusted OR 0.77 (95% CI 0.55 to 1.08; P = 0.13). Documented diagnostic quality and treatment planning improved, with no difference in prescribing and no safety signals. The intervention moved the documentation, not the patient outcome.

TABLE 2
Clinician–model collaboration in ordinary use, showing both comparisons that matter.

Study	vs human alone	vs model alone
Goh 2024, diagnosis ¹⁵	+2 points (P = .60)	Model alone +16 points over control
Goh 2024, management ¹⁶	+6.5 points (P < 0.001)	−0.9% (P = 0.8)
Everett 2026 ¹⁷	+6.8% second opinion; +9.9% first	Model alone 87% (P = 0.20)

6 Alert fatigue is the same problem, thirty years older

Interruptive alerting is automation bias with a longer record: a machine judges, a human adjudicates, and adjudication degrades with exposure.

The claim that clinicians override 90% or more of alerts is not what its usual source says. van der Sijs and colleagues reviewed 17 studies and found override rates ranging from 49% to 96% across different alert types, systems and denominators.²⁰ The 96% is the top of a heterogeneous range, not a central estimate. Ancker and colleagues' data on 1,592,528 alerts from 112 clinicians gives baseline acceptance of roughly 25% for drug alerts and 33% for reminders: override between two-thirds and three-quarters.²¹ The interpretation is misquoted more seriously than the number. van der Sijs states that

overriding is often clinically justified and places the error-producing conditions in the system rather than the clinician: low specificity, low sensitivity, unclear alert information, workflow disruption. An override rate measures alert quality as much as clinician attention.

Ancker's second finding inverts the folk model. Override rates showed no association with general workload; what predicted override was repetition. Acceptance fell 30% for each additional reminder in the same encounter (IRR 0.70), and where a repeated alert had been overridden once, 87.9% of subsequent instances were overridden too. The remedy is not more slack for clinicians but fewer repeat alerts.

What low specificity costs

Wong and colleagues externally validated a widely deployed proprietary sepsis model across 38,455 hospitalizations, with sepsis in 2,552.²² At the recommended threshold, sensitivity was 33%, specificity 83% and positive predictive value 12%: the model missed 1,709 of the 2,552 sepsis patients while alerting on 6,971 hospitalizations, 18% of all patients. Those are 2021 figures for a model the vendor later revised, not any product's current performance; the point is what a positive predictive value of 12% does to the person receiving the alerts. A later validation across 145,885 emergency department encounters found, within six hours, sensitivity 14.7% and PPV 7.6%, a median lead time of 0 minutes, and half of all alerts firing after sepsis onset.²³

The design paper is not called what people call it

This literature is often summarized by reference to “Bates' laws of clinical decision support”. No paper by that title exists. The canonical source is Bates and colleagues, *Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality*, drawn from eight years of implementation experience.²⁴ Its principles are unglamorous: speed matters most, anticipate needs and deliver in real time, fit the workflow, small interface changes have large effects, ask for more information only when truly needed, monitor and maintain. Twenty-three years on, most AI features do not meet them.

One argument for over-alerting is legal rather than clinical: suppressing a warning creates liability. Kesselheim and colleagues concluded that more parsimonious, better-tailored warnings would not expose vendors, purchasers or users to greater litigation risk, while excessive warnings cause clinicians to pay less attention to vital alerts.²⁵ It is a legal-policy analysis, not an empirical study, but it removes the usual excuse.

7 Design properties that follow

These findings translate into properties, not features.

1. **Verification has to be cheap.** Because automation bias tracks verification complexity,² the design target is the seconds and actions needed to confirm one claim: every assertion traceable to the span of source it came from, inside the review surface, without navigation. A confirmation step whose verification is expensive produces confirmation without verification, the failure it was added to prevent.
2. **Provenance has to be visible, and its absence has to be visible.** Omar and colleagues seeded one fabricated element into each of 300 physician-validated vignettes; six models elaborated on the fabrication rather than rejecting it, at rates from 50% to 82.7%, 66% overall.²⁶ A mitigation prompt cut that to 44%; temperature zero did not help. A model asked to justify a false premise supplies a fluent justification, so an unsupported claim has to look different from a supported one.
3. **Low-specificity output must not interrupt.** Repetition drives override,²¹ and positive predictive values of 12% and 7.6% mean most interruptions are wrong.^{22,23} Such an alert spends attention that is not replenished, including the attention available for the next alert.
4. **A confident answer with an inspectable basis is a different object from a confident answer.** Dratsch's radiologists were shown a BI-RADS category, a conclusion with no visible reasoning, which moved the most experienced readers by 36.8 points.⁵ A ranked set with supporting and contradicting evidence exposed is reviewable; a bare label is not.
5. **Workflow position is a design parameter and should be evaluated as one.** Everett's arms differed only in where the model sat in the sequence of work and produced different accuracy.¹⁷ No evidence establishes draft-then-review as the safest option.
6. **Review burden is a cost and belongs in the accounting.** The vigilance tax is measurable,¹⁸ the management-reasoning trial priced it at 119.3 seconds per case,¹⁶ and a scoping review concludes ambient systems may shift rather than eliminate documentation effort.¹⁴ Moving work from authoring to reviewing has not necessarily reduced it. DECIDE-AI covers that stage, live clinical use between offline validation and large comparative trials, with 27 reporting items from a modified Delphi of 123 then 138 experts.²⁷

8 What "human in the loop" would have to mean

As commonly used, the phrase names a position in an organizational chart. A control is a different thing: a stated failure mode, a measured detection rate against it, a known cost, and a schedule for re-measurement.

Six statements are supported here. Human review reduces error volume substantially, roughly twentyfold in the one pipeline measured end to end.⁷ It does not reliably reduce the share of survivors that are clinically significant.⁷ Where the machine is wrong, the reviewer performs worse than an unaided clinician: 86.6% more prescribing errors, 36.8 to 59.9 accuracy points in mammography, 14.0

points on diagnostic reasoning.^{4,5,6} AI-literacy training did not prevent that.⁶ Experience attenuates without removing it.⁵ And where nothing was deliberately broken, the collaborative composite has not reliably beaten the model alone.^{15,17,18}

A defensible version of the claim would be specific to an interface, evaluated on the errors that matter rather than those easy to count, priced at the review time a clinician actually spends, and re-measured as the model, the population and the workflow change. Nothing here says that is impossible. It says the claim is usually asserted rather than tested, and that the untested version describes who will be blamed rather than what will be caught.

REFERENCES

Entries 1 and 2 are the systematic reviews that define the phenomenon; entries 4, 5, 6, 15, 16, 17 and 19 are the controlled experiments and trials carrying the argument, and entry 7 is the observational study on which section 4 rests. Entry 16 is a preprint and entry 6 is a journal article whose volume and issue could not be confirmed; both are flagged in the list. Entries 24, 25 and 27 are design, policy and reporting-guideline documents rather than empirical studies, and entry 14 appears in a rapid-publication venue and is used for framing only.

1. Goddard K, Roudsari A, Wyatt JC. *Automation bias: a systematic review of frequency, effect mediators, and mitigators*. Journal of the American Medical Informatics Association. 2012;19(1):121–127. [doi:10.1136/amiajnl-2011-000089](https://doi.org/10.1136/amiajnl-2011-000089) SYSTEMATIC REVIEW
2. Lyell D, Coiera E. *Automation bias and verification complexity: a systematic review*. Journal of the American Medical Informatics Association. 2017;24(2):423–431. [doi:10.1093/jamia/ocw105](https://doi.org/10.1093/jamia/ocw105) SYSTEMATIC REVIEW
3. Rule A, Bedrick S, Chiang MF, et al. *Length and Redundancy of Outpatient Progress Notes Across a Decade at an Academic Medical Center*. JAMA Network Open. 2021;4(7):e2115334. [doi:10.1001/jamanetworkopen.2021.15334](https://doi.org/10.1001/jamanetworkopen.2021.15334) RETROSPECTIVE ANALYSIS
4. Lyell D, Magrabi F, Raban MZ, et al. *Automation bias in electronic prescribing*. BMC Medical Informatics and Decision Making. 2017;17:28. [doi:10.1186/s12911-017-0425-5](https://doi.org/10.1186/s12911-017-0425-5) RANDOMIZED EXPERIMENT
5. Dratsch T, Chen X, Rezazade Mehrizi M, et al. *Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance*. Radiology. 2023;307(4). [doi:10.1148/radiol.222176](https://doi.org/10.1148/radiol.222176) READER STUDY
6. Qazi IA, Ali A, Khawaja AU, et al. *Automation Bias in Large Language Model–Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy — A Randomized Clinical Trial*. NEJM AI. [doi:10.1056/Aloa2501001](https://doi.org/10.1056/Aloa2501001). Posted as a preprint 26 August 2025, [doi:10.1101/2025.08.23.25334280](https://doi.org/10.1101/2025.08.23.25334280); volume, issue and pages of the journal version not confirmed. RCT

7. 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
8. 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](https://pubmed.ncbi.nlm.nih.gov/27435949/) CROSS-SECTIONAL
9. 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
10. 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) COMPARATIVE STUDY
11. Reddy A, Gunnink E, Wheat CL, et al. *Rapid Evaluation of Artificial Intelligence Technology Used for Ambient Dictation in Primary Care: Comparing the Quality of Documentation of Artificial Intelligence-Generated and Human-Produced Clinical Notes*. Annals of Internal Medicine. 2026;179(6):765–772. [doi:10.7326/ANNALS-25-02772](https://doi.org/10.7326/ANNALS-25-02772) CROSS-SECTIONAL
12. Lukac PJ, Turner W, Vangala S, et al. *Ambient AI Scribes in Clinical Practice: A Randomized Trial*. NEJM AI. 2025;2(12). [doi:10.1056/Aloa2501000](https://doi.org/10.1056/Aloa2501000) RCT
13. 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](https://arxiv.org/abs/2402.08021) BENCHMARK
14. Wolfe J, Welsh SM. *Ambient Documentation Systems in Emergency Medicine: A Scoping Review of Clinical Precision, Patient Experience, Throughput, and Quality*. Cureus. 2026;18(4):e106643. [doi:10.7759/cureus.106643](https://doi.org/10.7759/cureus.106643) SCOPING REVIEW
15. 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
16. Goh E, Gallo R, Strong E, et al. *Large Language Model Influence on Management Reasoning: A Randomized Controlled Trial*. medRxiv. 2024. [doi:10.1101/2024.08.05.24311485](https://doi.org/10.1101/2024.08.05.24311485) PREPRINT
17. 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
18. Wang G, Zhang K, Jiang J, et al. *Human–large language model collaboration in clinical medicine: a systematic review and meta-analysis*. npj Digital Medicine. 2026;9:195. [doi:10.1038/s41746-026-02382-2](https://doi.org/10.1038/s41746-026-02382-2) META-ANALYSIS

19. 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 **CLUSTER RCT**
20. van der Sijs H, Aarts J, Vulto A, et al. *Overriding of Drug Safety Alerts in Computerized Physician Order Entry*. *Journal of the American Medical Informatics Association*. 2006;13(2):138–147. doi:10.1197/jamia.M1809 **SYSTEMATIC REVIEW**
21. Ancker JS, Edwards A, Nosal S, et al. *Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system*. *BMC Medical Informatics and Decision Making*. 2017;17:36. doi:10.1186/s12911-017-0430-8 **COHORT**
22. Wong A, Otles E, Donnelly JP, et al. *External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients*. *JAMA Internal Medicine*. 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626 **EXTERNAL VALIDATION**
23. Ostermayer DG, Braunheim B, Mehta AM, et al. *External validation of the Epic sepsis predictive model in 2 county emergency departments*. *JAMIA Open*. 2024;7(4):ooae133. doi:10.1093/jamiaopen/ooae133 **EXTERNAL VALIDATION**
24. Bates DW, Kuperman GJ, Wang S, et al. *Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality*. *Journal of the American Medical Informatics Association*. 2003;10(6):523–530. doi:10.1197/jamia.M1370 **POSITION PAPER**
25. Kesselheim AS, Cresswell K, Phansalkar S, et al. *Clinical Decision Support Systems Could Be Modified To Reduce ‘Alert Fatigue’ While Still Minimizing The Risk Of Litigation*. *Health Affairs*. 2011;30(12). doi:10.1377/hlthaff.2010.1111 **POSITION PAPER**
26. 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 **BENCHMARK**
27. 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 **REPORTING GUIDELINE**

RELATED PAPERS

Approval gates

A specification for where human decision authority has to sit.

Evaluating clinical language models

What a vendor has measured when it says its model is accurate.

Ambient documentation and the clinician's day

The measured effect of ambient scribes, at the size the trials support.