Editorial

DOI: 10.5582/bst.2026.01255

Matching clinical claims to evidence for the use of artificial intelligence (AI) in medicine

Kenji Karako*

Department of Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan.

SUMMARY: Technical performance alone does not establish the clinical value of artificial intelligence (AI). Clinical studies can reveal favorable expert ratings, limited clinician engagement, and uncertain effects on patient outcomes within the same intervention. We present an editorial synthesis with six complementary evidence domains: technical performance, external validity, human-AI interaction, workflow integration, clinical utility and outcomes, and lifecycle performance. These domains identify the questions addressed by a study. Each claim should be assessed with respect to study design, risk of bias, performance or effect estimates, uncertainty, and the evaluated setting. Intended use, autonomy, and potential harm determine the evidence required across domains. Journals should distinguish improvements in intermediate measures from demonstrated patient or health-system benefits and require reassessment when changes potentially invalidate earlier findings. The framework supports evaluation of AI-enabled care as delivered in a clinical system.

Keywords: clinical validation, human factors, workflow integration, patient outcomes, lifecycle evaluation

1. Introduction

The growing volume of medical artificial intelligence (AI) research has not been matched by evidence of benefit in routine care. A 2025 systematic review of 1,263 intensive-care AI studies classified 74% as involving early development or initial validation, 24% as involving external validation, and 25 studies (2%) as involving clinical integration; none met the review's definition of full implementation (*1*). A scoping review drawing on 218 systematic reviews published between September 2023 and September 2024 identified 4,667 primary studies, of which 88.2% were preclinical and 2.4% were randomized trials (*2*). These proportions describe the literature sampled by the reviews, rather than a census of current clinical use. Even among randomized trials, primary endpoints frequently concern diagnostic yield or performance (*3*). In perioperative anesthesia, a review of eight reports of randomized trials found small study populations and substantial heterogeneity for some pooled outcomes (*4*).

Earlier contributions to BioScience Trends described AI-enabled medical support systems and data platforms (*5*), applications from diagnosis to treatment (*6*), and explainability (*7*). These developments provide a basis for asking how AI changes care. Interpretability may help clinicians examine a recommendation, but clinical value also depends on whether they can act on it, the consequences of doing so, and the resources required. Evaluation must therefore consider the model together with its interface, users, workflow, and patients.

The editorial problem is an inferential one. A benchmark characterizes performance with data under specific conditions, whereas a claim about improved care concerns an intervention in a clinical setting. Incomplete information, changing prevalence, local protocols, and patient preferences can affect that intervention. A Nature Medicine editorial has already argued that stronger clinical claims require stronger evidence, from computational performance through clinical usefulness and workflow integration to patient outcomes (*8*). We have developed this principle into a six-domain synthesis that treats external validity and lifecycle performance as distinct domains, that extends assessment to agentic systems, and that specifies claim limits that editors can apply to individual studies, with examples that distinguish measured findings from unestablished benefits.

2. Clinical findings and their limits

In a four-week prospective pilot of the SHAKED decision support system in a tertiary emergency department, expert reviewers rated 99 of 100 sampled

outputs as clinically appropriate. The proportion of intervention-wing encounters in which clinicians opened the SHAKED interface declined from 68% in week 1 to 30% in week 4. No difference in emergency-department length of stay was detected between the wings studied (*9*). The study illustrates that output quality, sustained use, and operational effects require separate assessment.

A trial randomizing 103 clinical officers across 16 Kenyan primary-care facilities included 9,347 patient encounters in its primary analysis. The trial did not establish a reduction in 14-day treatment failure (adjusted odds ratio: 0.77, 95% confidence interval: 0.55-1.08). Secondary assessments favored the quality of clinical documentation. No intervention-related serious adverse events or safety signals were identified during the trial (*10*).

In a single-center quality-improvement study of discharge-date estimates, case managers had lower mean absolute errors than an AI tool at 48 and 24 hours before actual discharge (*11*). These time points were anchored to the observed discharge date. The difference does not establish why human estimates were more accurate: they were visible to care teams and could themselves have influenced the timing of discharge. More broadly, a systematic review found that balanced assessments of patient-relevant benefits and harms remain uncommon (*12*). Together, these examples show why favorable performance, uncertain benefit, and possible harms must retain their distinct meanings.

3. Evidence domains and the claims they support

We organized editorial assessment into six complementary evidence domains (Figure 1 and Table 1). The framework is a conceptual proposal. It builds on existing work on implementation outcomes (*13*) and the lifecycle approach of FUTURE-AI (*14*). Its practical contribution is a set of domain-specific questions and claim limits for editors, illustrated by clinical studies, rather than a new principle of proportional evidence. A study may inform several domains, with different results and different degrees of uncertainty in each.

"Technical performance" concerns discrimination, calibration, task accuracy, robustness, and failure modes under specified conditions. "External validity" concerns whether model performance is sustained across relevant times, sites, populations, and devices. It also concerns whether findings on human-AI interaction, workflow, and clinical outcomes are applicable to other users and care settings; external validation of model performance alone does not establish this. Claims should remain limited to the settings and shifts evaluated. If investigators recalibrate or otherwise adapt a model, the updated version needs to be evaluated with data independent of those used for adaptation. Repeated validation of model performance can enhance knowledge of prediction but does not by itself estimate the effect of using the system on patient care.

"Human-AI interaction" concerns changes in judgment and behavior, including changes in decisions, appropriate overrides, usability, and trust calibration. Simulations and live clinical studies answer different questions, and their settings should be explicit.

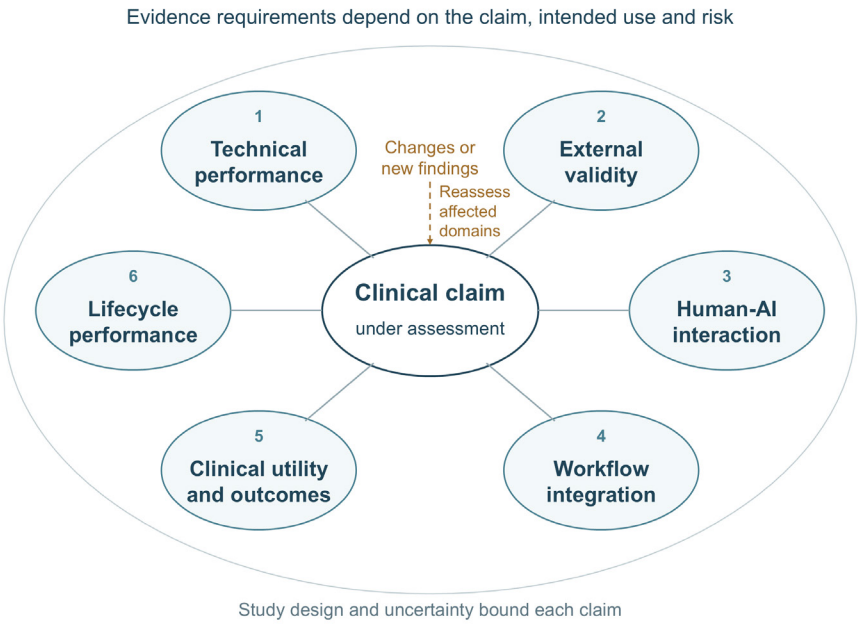


Figure 1. Complementary evidence domains for clinical artificial intelligence (AI) claims. The six domains inform claim-specific assessment; they may overlap and need not all apply to every claim. Solid lines connect domains to the claim being assessed, and the dashed arrow indicates reassessment after changes or new findings. The outer boundary represents shared considerations of the claim, intended use, and risk. Numbers identify domains without implying rank or sequence. Interpretation depends on methodological quality, comparator relevance, estimates, uncertainty, and applicability. This conceptual synthesis is not a validated grading instrument and is informed by references 8, 13, and 14.

Table 1. Evidence domains and editorial assessment of clinical AI claims

Evidence domain	Editorial question	Examples of evidence	Claim scope and limits
Technical performance	How does the model perform under specific conditions?	Discrimination, calibration, task accuracy, robustness, failure analysis.	Performance in terms of tested data and tasks; no inference of patient benefit.
External validity	Which findings are applicable to other settings and users?	Independent model validation; replication of interaction, workflow, or outcome findings across settings and users.	Findings in evaluated settings; model validation alone does not prove that clinical effectiveness transfers.
Human-AI interaction	How does AI affect users and decisions?	Changes in decisions, appropriate overrides, usability, and trust calibration; simulations or live evaluation.	Effects apply to the users and settings studied; safety must be directly assessed.
Workflow integration	How does AI affect the provision of care?	Timeliness, workload, communication, access, alert burden, implementation hazards.	Operational effects in the workflow evaluated; adoption or saved time alone do not establish patient benefit or net clinical utility.
Clinical utility and outcomes	What benefit or harm follows compared to relevant care?	Comparative patient outcomes, safety, quality of life, equity, resource use.	Effects supported by design, estimates and uncertainty; noninferiority requires a justified margin.
Lifecycle performance	What happens during continued use and after a change?	Version-linked outcome monitoring, incident review, postdeployment studies, reassessment.	Performance and harms are observed within the evaluated periods and versions; monitoring arrangements alone do not establish sustained benefit.

The domains are complementary and may overlap within a study. Each claim should be appraised against study design, risk of bias, comparator relevance, performance or effect estimates, uncertainty, and applicability. Intended use, autonomy, potential harm, and reversibility guide evidence requirements. Findings during continued use may prompt reassessment of affected domains. AI, artificial intelligence.

"Workflow integration" concerns timeliness, workload, communication, access, and hazards introduced elsewhere in care. A tool may increase workload while still providing net benefit or save time without improving care. Neither adoption nor reduced task time alone establishes clinical utility.

"Clinical utility and outcomes" concerns the consequences of using AI compared to an appropriate alternative. Relevant outcomes depend on intended use and may include morbidity, symptoms, quality of life, safety, equity, and healthcare system resource use. A claim of patient benefit requires patient-relevant evidence; a claim of efficiency requires a meaningful comparison of resource use while assessing possible diminishment of quality or safety. Noninferiority on prespecified clinical outcomes, assessed against a justified margin, may support efficiency-based adoption when resource benefits are also demonstrated. Study design should follow the causal question. Randomized trials are valuable for estimating intervention effects when feasible. Well-designed quasi-experimental or observational studies may also contribute, provided their assumptions, confounding, and other limitations are addressed. A retrospective model-validation study alone cannot establish improved care. The comparator should reflect feasible contemporary practice, including existing decision-making support and the consequences of replacing an established service.

"Lifecycle performance" concerns performance and harms observed during continued use and after relevant changes. Version control, monitoring, incident investigation, and change management provide the means to collect and respond to evidence; their existence is not evidence of sustained benefit. Claims should specify the populations, period of observation, versions, and outcomes assessed. Plans for monitoring and reassessment should be specified during development, including what would trigger an investigation, renewed evaluation, modification, or discontinuation. An update, new use, or detected deterioration should prompt reassessment in the affected domains.

Across domains, the required evidence depends on intended use, autonomy, potential harm, and reversibility. Methodological appraisal should examine risk of bias, comparator relevance, performance or effect estimates, precision, and applicability. A low-risk administrative application does not require the same outcome trial as an autonomous treatment system. Equally, monitoring a high-risk system does not compensate for inadequate evidence before deployment.

4. Evaluation of agents and changing systems

Agentic systems make the boundaries of the evaluated intervention particularly important. The Mother-Child AI Agent combines a prediction model using longitudinal electronic health records with a tool that retrieves recommendations from medical literature and guidelines

(*15*). Predictive validation and recommendation retrieval should be distinguished from evidence that autonomous clinical action improves outcomes. Systems described as agents differ in their authority to recommend, initiate, or perform actions, and studies should state that authority explicitly.

When a system selects tools or follows multistep plans, its evaluation should address errors across the full sequence, repeated-run variability, tool failures, human escalation, and recovery from a partial failure. The relevant configuration can include the model version, prompts, sources retrieved, available tools, and requirements for human approval. Testing components in isolation may miss unsafe interactions between them. These questions belong within the same evidence domains, with evaluations proportionate to the actions the system can take. World Health Organization guidance on large multimodal models similarly emphasizes stakeholder participation and post-release auditing (*16*).

5. Regulation and evidence over time

Regulatory initiatives reinforce the need to consider evidence over the life of a system. The US Food and Drug Administration's January 2025 draft guidance describes lifecycle management and marketing-submission recommendations for AI-enabled device software functions (*17*). Its August 2025 final guidance on predetermined change control plans addresses planned modifications, the methods for developing and validating them, and impact assessment (*18*). An August 2026 discussion paper seeks feedback on the regulation of generative-AI-enabled medical devices; it does not establish draft or final regulatory expectations (*19*). International Medical Device Regulators Forum principles also address representative data, independent testing, human-AI team performance, and monitoring (*20*).

Japan's Second Basic Plan on Artificial Intelligence provides a national policy context for AI development, deployment, and governance (*21*). The Pharmaceuticals and Medical Devices Agency has separately solicited public comment on aspects for the review of specific categories of software as a medical device (*22*). A national strategy, drafts submitted for comment, and device-specific guidance have different purposes and authority. For journals, their relevance is that system changes and continuing use raise scientific questions that a single predeployment evaluation cannot answer.

6. Conclusion

The clinical claims made for medical AI should match what studies directly measure and the settings in which their findings apply. The six complementary evidence domains indicated here help distinguish technical performance, use in practice, and patient or health-system benefit, while retaining methodological appraisal within each domain. This conceptual synthesis complements established reporting guidelines for AI trials, early clinical evaluations, prediction models, and diagnostic-accuracy studies (*23–27*).

Authors should make the scope and uncertainty of their findings explicit in titles, summaries, and conclusions. Journals should distinguish intermediate improvements from demonstrated clinical benefit, report harms alongside benefits, and welcome informative neutral or negative results. Changes in systems, intended uses, or care settings may require reassessment of affected domains.

The framework requires evaluation before use as a formal editorial instrument, and particularly for consistency between reviewers and its ability to improve the alignment of claims with evidence. For now, it raises a practical question: what did the study directly measure, and how far do its findings extend?

Funding: This work was supported by JSPS KAKENHI Grant Number JP26K21655.

Conflict of Interest: The author has no conflicts of interest to disclose.

References

1. Berkhout WEM, van Wijngaarden JJ, Workum JD, van de Sande D, Hilling DE, Jung C, Meyfroidt G, Gommers D, Buijsman SNR, van Genderen ME. Operationalization of artificial intelligence applications in the intensive care unit: A systematic review. JAMA Netw Open. 2025; 8:e2522866.
2. Wang Q, Yang Y, Qiu L, Xu Y, Sun R, Xue P, Qiao Y. A quantitative analysis of global AI medical studies: Gaps in randomized controlled trials. NPJ Digit Med. 2026; 9:403.
3. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: A scoping review. Lancet Digit Health. 2024; 6:e367-e373.
4. Shimada K, Inokuchi R, Ohigashi T, Iwagami M, Tanaka M, Gosho M, Tamiya N. Artificial intelligence-assisted interventions for perioperative anesthetic management: A systematic review and meta-analysis. BMC Anesthesiol. 2024; 24:306.
5. Karako K, Song P, Chen Y, Tang W. New possibilities for medical support systems utilizing artificial intelligence (AI) and data platforms. Biosci Trends. 2023; 17:186-189.
6. Huang T, Ma L, Zhang B, Liao H. Advances in deep learning: From diagnosis to treatment. Biosci Trends. 2023; 17:190-192.
7. Karako K, Tang W. Applications of and issues with machine learning in medicine: Bridging the gap with explainable AI. Biosci Trends. 2024; 18:497-504.
8. Show us the evidence for the value of medical AI. Nat Med. 2026; 32:1163.
9. Leibovitch L, Ahituv A, Gorenshtein A, Aran D, Sorka M, Miron K, Shelly S. Prospective evaluation of a large language model clinical decision support system in the emergency department. Nat Med. 2026; doi: 10.1038/s41591-026-04601-5.

10. Agweyu A, Mwaniki P, Menon V, Korom R, Isaaka L, Wanyama C, Gill J, Kiptinness S, Adan N, Emmanuel-Fabula M, Riley RD, Archer L, Denniston AK, Liu X, Mateen BA. Generative AI-enabled clinical decision support system in primary care: A pragmatic, cluster-randomized trial. Nat Med. 2026; 32:3032-3039.
11. Kantheti HS, Dolan C, Dale J. Artificial intelligence-generated discharge dates and estimation accuracy in hospitalized patients. JAMA Netw Open. 2026; 9:e2632033.
12. Wilhelm C, Steckelberg A, Rebitschek FG. Benefits and harms associated with the use of AI-related algorithmic decision-making systems by healthcare professionals: A systematic review. Lancet Reg Health Eur. 2025; 48:101145.
13. van de Sande D, Chung EFF, Oosterhoff J, van Bommel J, Gommers D, van Genderen ME. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. NPJ Digit Med. 2024; 7:58.
14. Lekadir K, Frangi AF, Porras AR, *et al*. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025; 388:e081554.
15. Liu S, Zheng W, Kang J, *et al*. Prediction of maternal and infant outcomes from longitudinal electronic health records with a Mother-Child AI agent. Nat Med. 2026; doi: 10.1038/s41591-026-04694-y.
16. World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Geneva: World Health Organization; 2024. *https://iris.who.int/handle/10665/375579* (accessed September 7, 2026).
17. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft Guidance for Industry and Food and Drug Administration Staff. January 2025. *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing* (accessed September 7, 2026).
18. U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. Guidance for Industry and Food and Drug Administration Staff. August 2025. *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence* (accessed September 7, 2026).
19. U.S. Food and Drug Administration. Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback. August 2026. *https://www.fda.gov/medical-devices/digital-health-center-excellence/considerations-regulation-generative-ai-enabled-medical-devices-discussion-paper-and-request* (accessed September 7, 2026).
20. International Medical Device Regulators Forum. Good machine learning practice for medical device development: Guiding principles (IMDRF/AIML WG/N88 FINAL:2025). January 2025. *https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles* (accessed September 7, 2026).
21. Cabinet Office, Government of Japan. Japan's Second Artificial Intelligence Basic Plan. July 14, 2026. Provisional English translation. *https://www8.cao.go.jp/cstp/ai/ai_plan/aiplan_eng_20260714.pdf* (accessed September 8, 2026).
22. Pharmaceuticals and Medical Devices Agency. Solicitation of comments on aspects for the review of software as a medical device. July 1, 2026. *https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/devices/0058.html* (accessed September 8, 2026). (in Japanese).
23. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med. 2020; 26:1364-1374.
24. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nat Med. 2020; 26:1351-1363.
25. 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. Nat Med. 2022; 28:924-933.
26. Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024; 385:e078378.
27. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, Golub RM, Denniston AK, Shetty S, Moher D, Bossuyt PM, Darzi A, Ashrafian H, STARD-AI Steering Committee. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025; 31:3283-3289.

Received September 10, 2026; Revised October 12, 2026; Accepted October 15, 2026.

**Address correspondence to:*
Kenji Karako, Department of Surgery, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8655, Japan.
E-mail: tri.leafs@gmail.com

Released online in J-STAGE as advance publication October 20, 2026.