PREMIER JOURNAL OF SCIENCE REVIEW

Reframing Clinical AI Evaluation in the Era of Generative Models: Toward Multidimensional, Stakeholder-Informed, and Safety-Centric Frameworks for Real-World Health Care Deployment

Matthew Abikenari[1], M. Hassan Awad[2], Sammy Korouri[3], Kimia Mohseni[4], Derek Abikenari[5], René Freichel[6], Yaseen Mukadam[7], Ubaid Tanzim[8], Amin Habib[9] and Ahmed Kerwan[10]

[1]Department of Neurosurgery, Stanford University School of Medicine, Stanford, CA, USA
[2]Department of Management, The College of Business and Economics, California State University, Los Angeles, CA, USA
[3]Department of Medicine, University of California, Irvine, CA, USA
[4]Department of Psychobiology, University of California, Los Angeles, CA, USA
[5]Department of Philosophy, California State Polytechnic University, Pomona, CA, USA
[6]Department of Psychology, University of Amsterdam, Amsterdam, Netherlands
[7]Department of Cardiology, Royal Brompton Hospital, Guys and St Thomas' Foundation Trust, London, UK
[8]Department of Oncology, University College London Hospitals NHS Foundation Trust, London, UK
[9]Department of Radiology, Southend University Hospital NHS Foundation Trust, Southend-on-Sea, UK
[10]Harvard T.H. Chan School of Public Health, Harvard University, MA, USA

Correspondence to: Matthew Abikenari, mattabi@stanford.edu

Additional material is published online only. To view please visit the journal online.

Cite this as: Abikenari M, Awad MH, Korouri S, Mohseni K, Abikenari D, Freichel R, Mukadam Y, Tanzim U, Habib A and Kerwan A. Reframing Clinical AI Evaluation in the Era of Generative Models: Toward

ABSTRACT

The integration of artificial intelligence (AI) in the form of large language models (LLMs) and generative models into clinical practice has progressed ahead of metrics available to measure their performance in real-world settings. Traditional benchmarks such as area under the receiver operating characteristic curve or bilingual evaluation understudy (BLEU) scores are inadequate to meet clinical nuance, patient safety, explainability, and workflow integration. This scoping review maps the evolving landscape of clinical AI evaluation, combining academic and industry architectures, including clinical risk evaluation of LLMs for hallucination and omission (CREOLA), hospital deployments, and radiological tool reviews. We explore stakeholder tensions between academia, business viability, regulation, and frontline usability, and reveal how these perceptions build competing evaluation imperatives. In particular, we highlight the novel challenges created by generative models: hallucination, omission, narrative incoherence, and epistemic misalignment. The current paper elucidates that a strategy of layered, stakeholder-engaged design needs to integrate risk stratification, contextual awareness, and continuous postdeployment surveillance. Equity, interpretability, and clinician trust are not thought of as footnotes, but as central columns upon which evaluation is built. This review offers a synthesizing overview of how health systems, developers, and regulators can coconstruct adaptive and ethically grounded evaluation frameworks, ensuring that AI tools enhance, rather than erode, clinical judgment, patient safety, and health equity in real-world care.

Keywords: Clinical AI evaluation, Generative models, Stakeholder-informed frameworks, Safety-centric deployment, Healthcare equity

Introduction

The Evaluation Gap in Clinical Artificial Intelligence (AI) Deployment

As AI becomes increasingly integrated into health care delivery through diagnostics, administrative sorting, radiology, and documentation, recognition grows that contemporary evaluation methods lag behind clinical usefulness. Conventionally, AI systems have been measured using machine learning metrics, including accuracy, area under the receiver operating characteristic curve (AUROC), F1 score, and precision-recall curves. In health care, these statistical surrogates do not necessarily translate to safety, clinical usefulness, or workflow alignment. Most importantly, in the case of generative AI, whose responses are necessarily context-dependent, linguistic, and probabilistic, these scores present an incomplete picture of performance and even mask failure modes relevant to real-world usage.[1,2]

Multiple reports, such as a 2022 review article by Liu et al. of clinical AI resilience and a 2023 World Health Organization technical series on the ethics and governance of LLMs in health settings, highlight the dangers of publishing inadequately tested models. Risks cited are hallucination, tone deafness for the context, omission of critical patient information, and clinician independence loss through incomprehensible decision support. These dangers necessitate a more comprehensive conceptualization of evaluation—one that takes into account not only predictive accuracy but also interpretability, usability, safety, fairness, and impact on real-world operation. More powerful models demand equally sophisticated measures—ones that capture not only predictive accuracy but also the multiple dimensions of clinical AI performance. Furthermore, one-dimensional, static measures are no longer sufficient to capture the multilayered dynamic nature of this performance.[3,4]

This review integrates insights from contemporary clinical models with an analysis of how emerging models like clinical risk evaluation of LLMs for hallucination and omission (CREOLA) redefine evaluation in terms of harm potential, narrative consistency, and stakeholder-centered design. It extends the vision of past scoping reviews through the integration of newly emerging challenges of LLMs and elucidates a vision-oriented evaluation paradigm that respects the changing character of clinical practice in the digital world.[5,6]

As these insights deepen, it becomes increasingly clear that evaluation must not only account for generative model behavior but also expand to accommodate real-world translational contexts where AI interfaces with biological heterogeneity, systems-level constraints, and community-driven care. For instance, AI is increasingly reshaping biomarker discovery and clinical stratification across diverse sectors—from neurosurgical pain syndromes to neurodegenerative and affective disorders—while influencing how health systems structure stakeholder-aligned delivery. In neuropathic pain, for instance, recent work in trigeminal neuralgia (TN) has demonstrated how deep learning

Multidimensional, Stakeholder-Informed, and Safety-Centric Frameworks for Real-World Health care Deployment. Premier Journal of Science 2025;11:100089

DOI: https://doi.org/10.70389/PJS.100089

Received: 12 June 2025
Revised: 10 July 2025
Accepted: 10 July 2025
Published: 15 July 2025

Ethical approval: N/a
Consent: N/a
Funding: No industry funding
Conflicts of interest: N/a
Author contribution: Matthew Abikenari, M. Hassan Awad, Sammy Korouri, Kimia Mohseni, Derek Abikenari, René Freichel, Yaseen Mukadam, Ubaid Tanzim, Amin Habib and Ahmed Kerwan – Conceptualization, Writing – original draft, review and editing
Guarantor: Matthew Abikenari
Provenance and peer-review: Unsolicited and externally peer-reviewed
Data availability statement: N/a

architectures such as U-Net and convolutional neural networks can extract imaging-based radiomic features from the trigeminal nerve, quantifying neurovascular compression with new granularity.[7] When combined with proteomic data, such as TN-specific cerebrospinal fluid and plasma biomarkers that reveal inflammatory and axonal injury signatures, AI-driven pipelines create precision diagnostic models that move beyond symptom description toward biological subtyping and targeted treatment stratification. Similarly, in geriatric depression, multiplex cytokine analysis has highlighted how changes in soluble factors like epidermal growth factor and growth-regulated oncogene alpha correlate with antidepressant remission, raising the prospect of using machine learning classifiers to forecast treatment response based on immunological fingerprints, especially in the context of multimodal behavioral and pharmacologic therapies.[8,9]

Beyond diagnostics, AI's role in health systems architecture is deepening through its interaction with stakeholder-driven service design. Studies on restorative entrepreneuring, such as Occupy Medical, illustrate how value-based and community-grounded care models can be mapped onto system-level innovation through participatory data collection and inclusive service algorithms.[10] Embodied embeddedness—where marginalized populations cocreate service logic—parallels the ethos of coproduction in AI evaluation, where interpretability, contextual fidelity, and social utility are jointly optimized.[10] When AI systems are built to reflect not only molecular or imaging signatures but also sociocultural context and infrastructural constraints, especially for at-risk populations such as the unhoused or elderly, they become not just diagnostic tools but also instruments of equity. These case studies point toward a future where AI-driven biomarker platforms are embedded within ethically attuned, stakeholder-sensitive care ecosystems that enable adaptive precision medicine across neurological, psychiatric, and social domains.

Stakeholders and Tensions in Clinical AI Evaluation

Clinical AI tool development, deployment, and assessment occur within a complex stakeholder environment whose goals, interests, and constraints often significantly diverge. Academic researchers are most focused on methodological novelty and rigor, typically validating models on past data in optimal conditions. Their evaluation priorities are governed by publication pressures, reproducibility standards, and peer-review culture, all potentially prioritizing algorithmic precision over translational impact.[11]

In contrast, industry developers are motivated by user acquisition, regulatory clearance, and speed-to-market. As much as many firms want to build clinically significant tools, business performance emphasis can lead to the rush to deploy and test superficially. Regulators, like the U.S. Food and Drug Administration (FDA), European Medicines Agency (EMA), and the UK's Medicines and Healthcare products Regulatory Agency (MHRA), are increasingly under pressure to design oversight procedures that strike a balance between innovation and patient safety. Even though recent efforts, such as the FDA's Predetermined Change Control Plan, acknowledge postmarket surveillance as critical, most regulatory models remain ill-equipped to deal with continuously evolving systems and unstable, generative models.[12,13]

Clinicians, who engage with these tools in settings where decisions carry significant implications for patient survival, long-term health, or quality of life, prioritize usability, reliability, safety, and interpretability. Such evaluation metrics as AUROC or F1 score are meaningless and too theoretical for on-the-spot decision-making in clinical practice. If an AI tool is difficult to use, involves considerable verification, or produces results that are not consistent with clinical reasonability, it will not be deployed even if its statistical performance is good. End-users and risk-takers are patients who are often excluded from AI development debates. Their representation at the time of evaluation is required to uphold justice, transparency, and credibility.

The tension caused by assessing frameworks that are biased toward the perception of one population versus another can lead to balkanized tools that are lab-effective but clinic-ineffective. A model can receive regulatory approval and commercial application but fail to establish clinical trust and meaningful outcomes. Such tensions must be resolved using coproduction approaches that involve the clinicians, patients, regulators, and developers at every stage of development. Frameworks such as Matheny et al.'s cogovernance model and the AI4People framework are positive initiatives to navigate this pluralist space, but there is much to be done yet to prepare inclusive, lasting, and ethically sound assessment methods.[14–16] The development, evaluation, and deployment of clinical AI occur within a dynamic ecosystem of stakeholders, each with distinct incentives and constraints. Misalignment between these groups often leads to gaps in how AI systems are validated and trusted. Table 1 provides a comparative overview of stakeholder-specific evaluation priorities, illuminating how these tensions shape the emerging landscape of clinical AI regulation and implementation.

The Challenge of Generative AI in Clinical Settings

The development of generative AI, particularly transformer-based LLMs such as GPT-4, Med-PaLM, and LLaMA-Med, has transformed the possibilities for automation in clinical documentation, education, and decision support. Such models are capable of producing entire consultation summaries, incorporating multimodal data, producing discharge instructions, and even developing patient-facing educational content. Their capacity to produce flexible, naturalistic language has created new opportunities to humanize care, but also new challenges in evaluating them.

Unlike classical classification or regression models, LLM outputs are nondeterministic, context-dependent, and linguistically complex. As such, conventional

Table 1 | Stakeholder evaluation priorities and tensions in clinical AI

Stakeholder	Primary Evaluation Priority	Tension Points	Examples
Academic Researchers	Accuracy, novelty, reproducibility	Static metrics (e.g., AUROC) vs. real-world impact	Benchmarking vs. contextual use limitations[11–13,15]
Industry Developers	Time-to-market, scalability, UX design	Commercial pressure vs. clinical safety	Minimal fine-tuning of LLMs before release[11,16]
Regulators (FDA, EMA, MHRA)	Public safety, accountability, compliance	Oversight models not suited for generative models	Static validation vs. dynamic model behaviors[11–13]
Clinicians	Trust, interpretability, workflow integration	Lack of usability and misaligned feedback burden	CREOLA implementation for clinical traceability[13,14,16]
Patients	Transparency, fairness, representation	Lack of inclusion, digital divide, and power asymmetries	LLM outputs miss linguistic/cultural nuance[4,11,15]

FDA = U.S. Food and Drug Administration; EMA = European Medicines Agency; MHRA = Medicines and Healthcare products Regulatory Agency; UX = User Experience.
This table outlines how divergent priorities among clinical AI stakeholders shape evaluation challenges, drawing attention to systemic tensions between innovation, usability, safety, and equity.

Table 2 | Failure modes in generative clinical AI and mitigation strategies

Failure Mode	Clinical Risk	Example from Literature	Mitigation Strategy
Hallucination	Confident but false information introduced	CREOLA: Hallucination taxonomy (Asgari et al.)	Risk-stratified grading, structured human oversight
Omission	Key clinical facts absent in generated text	CREOLA omission subtypes	Prompt engineering, explicit content recall scoring
Contextual Incoherence	Misplaced or irrelevant content in clinical setting	Kumarapeli et al. on narrative coherence	Context-aware LLM fine-tuning, human review cycles
Narrative Disruption	Logical inconsistency across clinical timelines	Kumarapeli et al. on cognitive fidelity	Alignment to decision flow, temporal annotation layers
Epistemic Misalignment	Inconsistency with clinician reasoning or logic	CREOLA and clinician feedback loops	Pretraining on clinical notes with embedded reasoning
Opacity in Output	Inability to trace model output to input stimuli	Zusterzeel R et al. hospital feedback issues	Post hoc interpretability overlays, clinician feedback portals

Each mitigation approach is aligned to real-world deployment studies or published taxonomies and presumes multidisciplinary codesign of LLM-based tools.
An overview of common failure types encountered in generative medical AI applications, associated clinical risks, literature examples, and mitigation strategies aligned with real-world deployment.

metrics of word overlap (BLEU and Recall-Oriented Understudy for Gisting Evaluation) are inadequate for identifying whether an output is clinically correct, contextually appropriate, or safe. Hallucination risk—the prediction with confidence of incorrect but plausible content—is particularly concerning in clinical use cases where incorrect content may not be caught by time-strapped providers.[17,18] Evaluating these types of models is done using a multidimensional approach, including semantic accuracy, conformity to clinical logic, appropriateness of tone, and traceability of content. Given the complex linguistic and probabilistic nature of LLM outputs, generative models introduce new categories of failure, from hallucinations and negation errors to narrative drift and epistemic misalignment. Recognizing and categorizing these failure types is crucial for designing effective evaluation and remediation pathways. Table 2 outlines the major failure modes observed in generative clinical AI, along with representative examples and corresponding mitigation strategies discussed in recent literature.

Additionally, generative AI has a tendency to aggravate the intrinsic opacity of machine learning systems. Outputs may be persuasive and fluent without being factually grounded or internally consistent. This increases the stakes for accountability and interpretability. Clinicians must be able to determine how a model produced its output, especially in medicolegal or ethically sensitive cases. Risk-stratified scoring systems, such as those proposed by CREOLA, represent a key step toward operationalization of safety-focused assessment. They allow developers and clinical evaluators to order errors by potential for harm, and to prioritize review and remediation efforts appropriately. This degree of granular, harm-directed evaluation is what is needed to ensure that generative AI augments, rather than subtracts from, clinical safety.[5,6]

Emerging Frameworks and the Move Toward Multidimensional Evaluation

The limitations of traditional AI evaluation in medicine have brought forth more nuanced and clinically aware frameworks. Among such seminal advances is the clinical risk evaluation for omission and LLM artifacts (CREOLA) framework proposed by Asgari et al. Rather than merely measuring LLM output accuracy using surface-level similarity metrics, CREOLA introduces a clinically aware taxonomy that splits errors into hallucinations and omissions, with subtypes of fabrication, contextual mismatch, negation failures, and causal inconsistencies. The taxonomy is then integrated with a harm-based grading schema borrowed from regulatory device classifications, allowing for systematic risk stratification of model outputs.[5,6]

CREOLA's innovation lies not just in error detection but also in placing them within a clinical risk landscape, making the framework actionable by both frontline clinicians and regulatory reviewers. By prioritizing fast iteration and formal feedback from clinicians, it promotes dynamic coevolution of model design and evaluation. In addition, the utilization of a graphical user interface enables scalable annotation and consensus building and represents an uncommon example of evaluation that is rigorous and workflow-integrated.

In a complementary direction, the work of Kumarapeli et al. opens the evaluative horizon to include the subtle, narrative functions of free-text within electronic health records (EHRs).[19–21]

Their review brings to the fore the reality that clinical documentation is not merely a catalog of facts but also a site of cognitive synthesis, decision explication, and communicative reasoning. As it is, the assessment of generative AI in this domain needs to pay heed not just to factual accuracy but also to epistemic transparency, stylistic aptness, and allegiance to the clinician's line of reasoning. This redirects the evaluative objective away from simulating human-like language toward augmenting clinical reasoning and documentation quality in context-responsive manners.

Frameworks developed out of hospital system implementations provide a real-world perspective as well. In addition, A recent multiapplication study evaluating AI-enabled clinical decision support across four hospital settings—triage, radiology prioritization, pneumonia detection, and deterioration prediction—demonstrates a striking disconnect between conventional performance metrics and actual clinical utility.[20]

Although AUROC scores were acceptable, uptake and trust depended largely on usability, interpretability, and perceived clinical impact. The authors propose a six-dimensional evaluation rubric of technical performance, interpretability, usability, clinical integration, fairness, and long-term safety, highlighting that successful deployment must trade off across all of these dimensions. Notably, clinician involvement throughout the model life cycle was a robust predictor of system effectiveness and acceptability.

In radiology, which is among the specialties most affected by AI, a recent scoping review of over 100 studies discovered gigantic methodological heterogeneity. The majority of the studies only looked at technical performance with no workflow analysis, user-centered validation, or postdeployment surveillance. None adhered to new standards like CONSORT-AI or DECIDE-AI, and few reported patient-centered or health economic outcomes. Authors suggest a staged evaluation process like drug trials, with model development, technical validation, simulation-based usability testing, and measurement of real-world clinical effect.[22–24]

Together, these frameworks represent an ongoing shift away from static, one-size-fits-all cutoffs and toward complex, context-sensitive, and clinically integrated evaluation strategies. Figure 1 recapitulates the multistakeholder, multidimensional framework for evaluating clinical AI tools. Their more extensive use is nevertheless impeded by resource limitations, insufficient standardization, and inconsistent regulatory oversight. Such frameworks have recently emerged to better capture the multidimensional challenges of evaluating AI in clinical settings. While CREOLA offers a harm-stratified lens on generative outputs, Kumarapeli et al. emphasize narrative coherence in documentation, and hospital-based studies contribute pragmatic, workflow-informed rubrics. To facilitate comparison, Table 3 summarizes key features, evaluative focus areas, and stakeholder involvement across these frameworks.

Cross-Cutting Themes: Safety, Equity, Interpretability, and Trust

Across diverse evaluation settings and paradigms, several issues that cut across have emerged and need

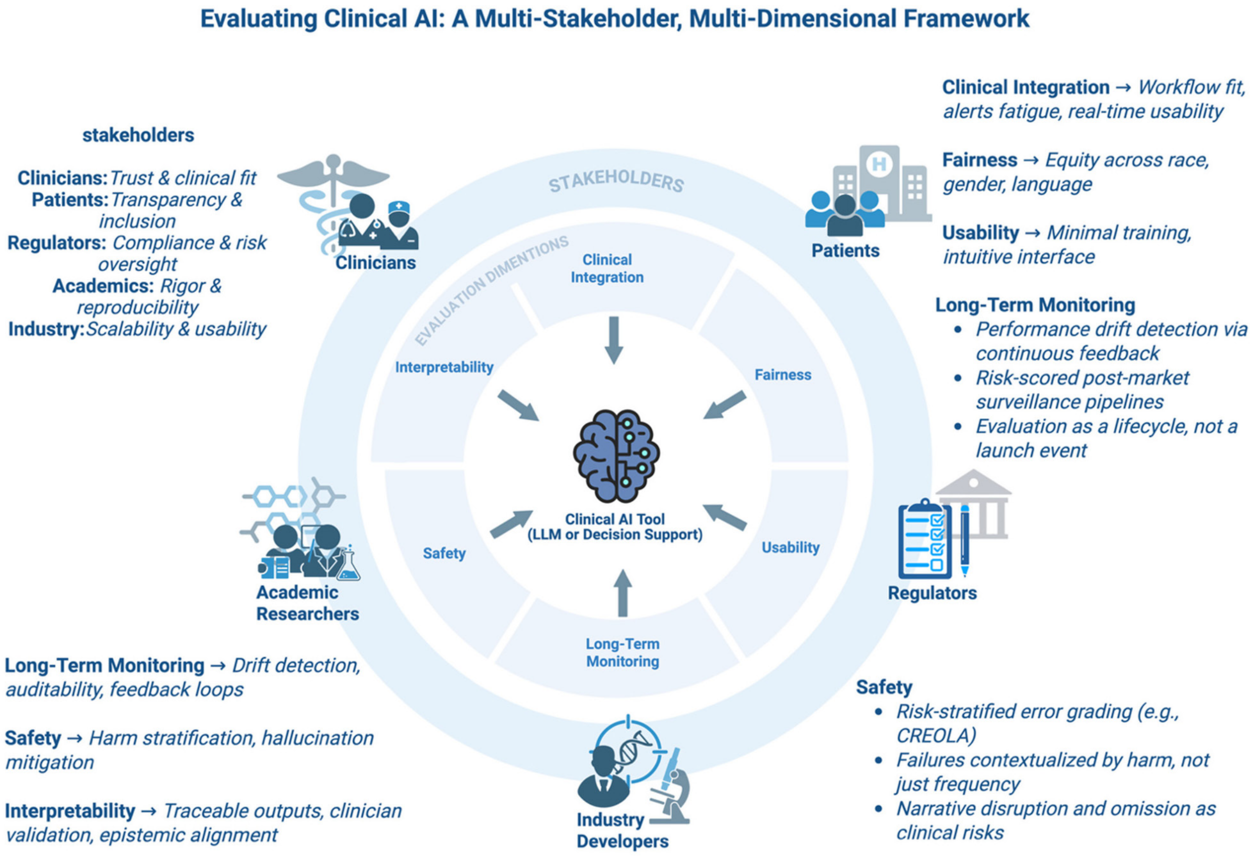
Evaluating Clinical AI: A Multi-Stakeholder, Multi-Dimensional Framework
stakeholders
Clinicians: Trust & clinical fit
Patients: Transparency & inclusion
Regulators: Compliance & risk oversight
Academics: Rigor & reproducibility
Industry: Scalability & usability
Clinicians
STAKEHOLDERS
EVALUATION DIMENTIONS
Clinical Integration
Interpretability
Fairness
Safety
Usability
Long-Term Monitoring
Clinical AI Tool (LLM or Decision Support)
Patients
Regulators
Academic Researchers
Industry Developers
Clinical Integration → Workflow fit, alerts fatigue, real-time usability
Fairness → Equity across race, gender, language
Usability → Minimal training, intuitive interface
Long-Term Monitoring
• Performance drift detection via continuous feedback
• Risk-scored post-market surveillance pipelines
• Evaluation as a lifecycle, not a launch event
Long-Term Monitoring → Drift detection, auditability, feedback loops
Safety → Harm stratification, hallucination mitigation
Interpretability → Traceable outputs, clinician validation, epistemic alignment
Safety
• Risk-stratified error grading (e.g., CREOLA)
• Failures contextualized by harm, not just frequency
• Narrative disruption and omission as clinical risks

Fig 1 | A multistakeholder, multidimensional framework for evaluating clinical AI tools. This conceptual model situates the clinical AI tool (LLMs or decision support systems) at the center of a multilayered evaluation ecosystem. Six critical evaluation dimensions—clinical integration, fairness, usability, long-term monitoring, safety, and interpretability—are depicted as core evaluation targets. Surrounding these dimensions are five stakeholder groups (clinicians, patients, regulators, academic researchers, and industry developers), each with unique priorities and tensions that shape AI validation and adoption. This framework underscores the need for harmonized, iterative, and context-sensitive assessments in the deployment of clinical AI

Table 3 | Comparative AI evaluation frameworks in clinical practice

Framework	Primary Focus	Key Evaluation Domains	Strengths	Limitations
CREOLA (Asgari et al.)	Risk-stratified clinical evaluation of LLM-generated errors	Hallucination, omission, harm grading, narrative incoherence	Risk-based, actionable, clinician-centered, scalable GUI	Requires expert annotation; focused on medical text outputs
Narrative-Centered Evaluation (Kumarapeli et al.)	Narrative quality and epistemic alignment in clinical documentation	Cognitive synthesis, reasoning transparency, stylistic fidelity	Highlights nuanced clinician reasoning in EHRs	Subjective metrics; less suited to large-scale automation
Hospital Implementation Studies (Zusterzeel R et al.)	Real-world usability and adoption of AI tools across workflows	Technical performance, interpretability, usability, safety	Grounded in deployment outcomes; six-domain framework	Generalizability limited; setting-dependent variables
Radiology AI Review (Neri et al.)	Evaluation quality of radiological AI studies	Methodology, workflow, standard adherence (CONSORT-AI)	Meta-perspective on field-wide quality	Lacks deployment or real-world surveillance data
DECIDE-AI (Vasey et al.)	Clinical trial staging for AI clinical decision support	Early-phase validation, usability, feedback, integration	Trial-based, human-centric design process	Predeployment focus; lacks longitudinal real-world data

CREOLA = clinical risk evaluation of large language models for hallucination and omission; LLM = large language model; EHR = electronic health record.
This table summarizes key characteristics, strengths, and limitations of major clinical AI evaluation frameworks, highlighting differences in scope, methodological orientation, and clinical applicability.

more nuanced investigation. Foremost among them is safety, not only in the form of avoiding explicit harm but also in considering implicit, cumulative, and systemic risks introduced by AI. The adaptive nature of AI systems, especially those that are updated in real-time via online learning or API-based LLMs, presents challenges to traditional static validation approaches. Asgari et al.'s harm stratification is one approach, but not many models are subjected to continual real-world monitoring for performance drift, unintended feedback loops, or vulnerability to adversarial inputs.[1,5,19,20,25]

Equity is a persistent blind spot in evaluation as well. Fairness is often dealt with in development pipelines—e.g., through bias mitigation algorithms or representative training data—yet not many evaluation systems check whether models work equitably across patient subgroups in actual use. Digital literacy differences, linguistic voice, and socioeconomic access can potentially widen disparities.[5,6] Furthermore, Kumarapeli et al. note that free-text generative models will systematically leave out subtleties within underrepresented groups, especially when fine-tuning data are limited to tertiary learning institutions.[19]

Interpretability remains essential to clinician trust. While attention maps, saliency tools, and SHapley Additive exPlanations values are standard for computer vision and tabular models, they have limited applicability to natural language output. CREOLA seeks to bridge the gap by annotating hallucination types with traceable references and scoring their risk impact, providing a semitransparent interface for review by clinicians. By analogy, Kumarapeli et al. propose that clinician-provided feedback loops integrated at training and validation can reveal epistemic mismatches early and propel iterative alignment.

Trust, reassuringly understood as an abstraction without physical form, is a measurable and modifiable product of these other themes. Criteria for assessment must acknowledge trust not as a product solely of correctness or brand recognition but of responsiveness to user needs, accountability for mistakes, and consistency in the face of suspicion. Clinician-to-consumer resources must have the capacity to demonstrate not only what they do but also how and why, with opportunity for override, feedback, and revision.

Lastly, innovative applications of these assessment frameworks demonstrate their real-world applicability and feasibility in real-world health care systems. For example, the CREOLA framework was just applied in a major academic medical center to assess LLM-written discharge summaries, where its harm-stratified grading facilitated clinicians to prioritize review of high-risk omissions and hallucinations and achieved measurable reductions in documentation errors and improved provider trust in the system.

Similarly, the DECIDE-AI staging approach has been piloted in European hospital consortia for guiding step-by-step validation of AI-aided sepsis prediction models, with usability and interpretability testing given high priority before deployment at scale. In radiology, multicenter CONSORT-AI-based audits have identified gaps in workflow integration and have led to revised reporting guidelines for diagnostic AI tools. These initial applications demonstrate the ways in which such frameworks can be incorporated into institutional quality improvement work, bridging technical controls to stakeholder objectives and prompting continuing oversight following deployment. The inclusion of these applications emphasizes the pragmatic character of multidimensional evaluation methods and their potential to drive responsible clinical AI adoption at scale.

Conclusion

The transition of clinical AI from predictive models to generative systems requires a concomitant change in the evaluation of these technologies. Classical performance metrics, as important as they remain, are no longer adequate to reflect the complex implications of AI in medicine. As this review shows, new frameworks such as CREOLA, stakeholder-informed approaches, and system-level assessments reflect significant steps toward more comprehensive, reliable, and context-aware evaluation methodologies.

But the field is at a juncture. Without robust, inclusive, and iteratively responsive evaluation approaches, clinical AI technologies risk replicating prior harms, eroding trust, and establishing new routes to harm. The uses of LLMs in high-stakes clinical documentation, decision-making, and communication multiply these risks even as they offer unprecedented opportunities for enhancing care.

By centering assessment on practical usefulness, patient safety, epistemic transparency, and alignment with stakeholders, future generations of frameworks can ensure that AI delivers not only efficiency but also equity, resilience, and clinical excellence. As health systems grapple with the challenges of implementation, assessment needs to become not something post facto, but an iterative, collective, and contextual endeavor.

Lastly, to ensure meaningful adoption, especially in resource-constrained settings, evaluation frameworks must be designed with scalability, modularity, and accessibility in mind. This entails simplifying technical burdens by creating lightweight, interoperable tools that can integrate into existing clinical workflows without requiring sophisticated infrastructure. For example, frameworks like CREOLA could be adapted into open-source platforms with pretrained modules and minimal annotation demands, enabling broader uptake by health systems with limited technical expertise. Moreover, stakeholder training modules and stepwise implementation guides could bridge knowledge gaps, allowing clinicians, administrators, and policymakers in low-resource environments to engage meaningfully with the evaluation process. Crucially, cross-institutional collaboration, through consortia, public-private partnerships, and global health alliances, can distribute the resource load of evaluation and generate pooled datasets for benchmarking across diverse clinical environments.

Embedding evaluation tools into electronic health record systems and leveraging cloud-based deployment can further democratize access, allowing health systems to monitor AI tool safety, usability, and bias in real-time. Adopting a tiered approach to evaluation, starting with essential safety and fairness checks and scaling up to full contextual fidelity assessment, can help ensure that even underresourced systems retain agency in shaping how AI is judged and trusted in practice.

References

1 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. BMJ. 2020;370:m3164. https://doi.org/10.1136/bmj.m3164

2 World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: World Health Organization; 2021. ISBN 9789240029200.

3 Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med. 2017;376(26):2507–9. https://doi.org/10.1056/NEJMp1702071

4 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7

5 Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274. https://doi.org/10.1038/s41746-025-01670-7

6 Goodman KE, Yi PH, Morgan DJ. AI-generated clinical summaries require more than accuracy. JAMA. 2024;331(8):637–8. https://doi.org/10.1001/jama.2024.0555

7 Abikenari M, Jain B, Xu R, Jackson C, Huang J, Bettegowda C, et al. Bridging imaging and molecular biomarkers in trigeminal neuralgia: toward precision diagnosis and prognostication in neuropathic pain. Med Res Arch. 2025;13(5):6605. https://doi.org/10.18103/mra.v13i5.6605

8 Siddarth P, Abikenari M, Grzenda A, Cappelletti M, Oughli H, Liu C, et al. Inflammatory markers of geriatric depression response to Tai Chi or health education adjunct interventions. Am J Geriatr Psychiatry. 2023;31(1):22–32. https://doi.org/10.1016/j.jagp.2022.08.004

9 Ajam Oughli H, Siddarth P, Lum M, Tang L, Ito B, Abikenari M, et al. Peripheral Alzheimer's disease biomarkers are related to change in subjective memory in older women with cardiovascular risk factors in a trial of yoga vs. memory training: Lien établi entre les biomarqueurs périphériques de la maladie d'Alzheimer et l'amélioration de la mémoire subjective chez les femmes âgées présentant des facteurs de risque cardiovasculaire dans le cadre d'un essai comparant le yoga à l'entraînement de la mémoire. Can J Psychiatry. 2025: 7067437251343291. https://doi.org/10.1177/07067437251343291

10 Awad MH, Sanchez M, Abikenari MA. The values work of restorative ventures: the role of founders' embodied embeddedness with at-risk social groups. J Bus Ventur Insights. 2022;18:e00337. https://doi.org/10.1016/j.jbvi.2022.e00337

11 Sendak MP, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey K, et al. A path for translation of machine learning products into healthcare delivery. EMJ Innov. 2020;1:114. https://doi.org/10.33590/emjinnov/19-00172

12 U.S. Food and Drug Administration (FDA). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) action plan. FDA; 2021. Available from: https://www.fda.gov/media/145022/download

13 European Medicines Agency (EMA). Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023; 2024. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf

14 Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the National Academy of Medicine. JAMA. 2020;323(6):509510. https://doi.org/10.1001/jama.2019.21579

15 Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach (Dordr). 2018;28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

16 Goodman KW. Ethics, medicine, and information technology: intelligent machines and the transformation of health care. Cambridge: Cambridge University Press; 2016. https://doi.org/10.1017/CBO9781139600330

17 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. https://doi.org/10.1038/s41591-023-02448-8

18 Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107–e108. https://doi.org/10.1016/S2589-7500(23)00021-3

19 BMC Medical Informatics and Decision Making. Evaluating the performance of artificial intelligence–based speech recognition for clinical documentation: a systematic review. BMC Med Inform Decis Mak. 2025;25:236. https://doi.org/10.1186/s12911-025-03061-0

20 Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res. 2017;19(11):e367. https://doi.org/10.2196/jmir.8775

21 Zusterzeel R, Goldstein BA, Evans BJ, Roades T, Mercon K, Silcox C. Evaluating AI-enabled clinical decision and diagnostic support tools using real-world data. Duke-Margolis Center for Health Policy; 2022.

22 European Society of Radiology (ESR). What the radiologist should know about artificial intelligence – an ESR white paper. Insights Imaging. 2019;10:44. https://doi.org/10.1186/s13244-019-0738-2

23 Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan AW, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORTAI extension. Nat Med.

2020;26(9):13641374. https://doi.org/10.1038/s41591-020-1034-x

24 Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the earlystage clinical evaluation of decision support systems driven by artificial intelligence: DECIDEAI. BMJ. 2022;377:e070904. https://doi.org/10.1136/bmj-2022-070904

25 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342