
JMIR Metascience and Research Integrity

Journal Impact Factor (JIF) (2023): 5.8
Volume 1 (2026) ISSN XXXX-XXXX Editor in Chief: Gunther Eysenbach, MD, MPH

Contents

Viewpoints

Algorithmic Research Waste in Health Care AI: The VALUE-AI (Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring) Lens for Decision-Useful Evidence (e104222) Dabeluchi Ngwu.	2
From AI Disclosure to Judgment Traceability: Addressing Research Waste in AI-Assisted Health Sciences Publishing (e106068) Tino Hsieh, Ya-Ting Chen.	16

Original Paper

Human-AI Agreement in Citation Validation of an AI-Generated Subspecialty Anesthesiology Knowledge Base: Cross-Sectional Paired Validation Study (e105947) Juan Li, Shun Huang, Yue Qiu, Wei Chen, Yang Zhang, Xuechun Zhao, Ge Qu, Yifan Bu, Xiaomei Feng, Jiawei Xu, Haobo Ma, Jingping Wang.	27
---	----

Viewpoint

Algorithmic Research Waste in Health Care AI: The VALUE-AI (Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring) Lens for Decision-Useful Evidence

Dabeluchi Chiedozi Ngwu¹, MD

Division of Thoracic Surgery, Department of Surgery, King Abdullah Hospital, Bisha, Asir, Saudi Arabia

Corresponding Author:

Dabeluchi Chiedozi Ngwu, MD

Division of Thoracic Surgery

Department of Surgery

King Abdullah Hospital

Al Nakhil

Bisha, Asir, 67714

Saudi Arabia

Phone: 966 543423453

Email: doctordabs@gmail.com

Abstract

Health care AI research is expanding rapidly, but publication volume alone does not generate evidence that reliably informs research, clinical care, policy, governance, or responsible technological development. Studies may report promising predictive performance while leaving uncertain whether the question is decision-relevant, whether the data fit the intended population, whether performance transfers across settings, whether the output can be integrated into a feasible workflow, or whether a deployed system can be monitored over time. These gaps can create avoidable losses of scientific, clinical, and societal value. This viewpoint argues that algorithmic research waste occurs when health care AI research is insufficient, relative to its stated purpose and maturity stage, to generate reliable and decision-useful evidence. The concept does not require every model to reach deployment. Exploratory studies, methodological research, replication, external validation, and negative findings can all be valuable when their aims, methods, claims, and intended decisions are aligned. Concern arises when research is presented as supporting a broader conclusion than its design and evidence can justify. This viewpoint aims to define algorithmic research waste, distinguish its principal pathways, and introduce VALUE-AI (Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring) as a conceptual metascience lens for identifying value-loss signals and routing them toward appropriate action. Four overlapping pathways are described: primary research-value loss, in which the question, design, conduct, reporting, or claims cannot support the stated purpose; contextual value loss, in which evidence is transferred without adequate attention to population, data, workflow, resource, or regulatory differences; deployment failure, in which an organization implements a technology without necessary local integration and safeguards; and postdeployment failure, in which changes in performance, use, equity, or safety are not adequately monitored or addressed. VALUE-AI is qualitative and stage-specific. The lens asks where potential research value is being lost, which actor or actors have primary control over the source of that loss, and what maturity-appropriate evidence or action is needed next. An illustrative example shows how the same model may represent useful discovery research, primary research-value loss, contextual value loss, deployment failure, or postdeployment failure, depending on its claims, evidence, setting, and intended use. VALUE-AI is not a validated reporting guideline, risk-of-bias tool, scoring system, implementation instrument, or regulatory standard. It is intended to complement established reporting, appraisal, implementation, and governance frameworks by organizing their implications around research value, attribution, and the next maturity-appropriate evidence or action. Its domains, feasibility, interrater interpretability, and effects on decisions require empirical refinement and evaluation across specialties, health systems, resource settings, and AI modalities.

(*J Meta and Res Int* 2026;1:e104222) doi:[10.2196/104222](https://doi.org/10.2196/104222)

KEYWORDS

artificial intelligence; health care AI; research waste; metascience; decision-useful evidence; external validation; implementation; responsible AI; VALUE-AI

Introduction

Research waste occurs when the resources and participant contributions invested in research fail to yield sufficiently reliable or useful knowledge. It may result from poorly prioritized questions, avoidable weaknesses in study design or conduct, incomplete reporting, nonpublication, unnecessary duplication, or failure to translate findings into practice [1-3]. A recent scoping review classified these manifestations as methodological, invisible, negligible, underreported, and structural waste, demonstrating that value can be lost throughout the research system rather than at a single methodological step [4]. Health care AI magnifies this concern. Although the number of clinical prediction models is rapidly increasing [5], systematic reviews continue to find incomplete reporting [6], high risk of bias [7], limited external validation [8], and overstated clinical claims [9]. A 2025 systematic review of 1263 intensive care AI studies found that 74% remained at early stages of development and only 25 (2%) had progressed to clinical integration [10]. Although confined to intensive care and not directly generalizable to health care AI as a whole, these findings caution against interpreting growth in publication volume as a comparable increase in decision-useful evidence.

Reviews underscore the need to distinguish technical promise from an evidence pathway that can inform care. A systematic review of 152 supervised machine-learning prediction-model studies found that 148 of 171 development or validation assessments were at high risk of bias, commonly due to inadequate sample size, poor handling of missing data, or overfitting [7]. VALUE-AI (Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring) treats these deficits as potential sources of avoidable value loss when they prevent a study from supporting the decision implied by its claims. This problem overlaps with, but is not identical to, low-value care or health care waste. Low-value care concerns services that provide no benefit or may cause harm [11], whereas health care waste is broader and encompasses inefficient or unnecessary resource use in care delivery [12]. Algorithmic research waste instead concerns the value of the knowledge produced by AI research. Economic value remains relevant because AI may alter outcomes, workload, workflow, costs, access, and equity in ways that conventional technology assessments do not fully capture [13].

This viewpoint defines algorithmic research waste, distinguishes research-level value loss from contextual and deployment failures, and presents VALUE-AI as a stage-specific metascience attribution-and-routing lens. The lens links each value-loss signal to the actor or actors who can modify it and to the next maturity-appropriate evidence or action. Evidence is decision-useful when it can credibly reduce uncertainty for a specified research, clinical, policy, governance, or implementation decision at the study's stated maturity stage.

The intended audience includes researchers, peer reviewers, editors, funders, health systems, regulators, developers, and patient and public representatives. VALUE-AI complements translational guidance [14] by examining whether health care AI research generates reliable and decision-useful evidence across the pathway from question formulation to postdeployment learning.

Definition and Conceptual Positioning

Algorithmic research waste is the avoidable loss of potential scientific, clinical, or societal value when the question, design, conduct, reporting, validation, or translation of health care AI research is insufficient, relative to its stated purpose and maturity stage, to generate reliable, decision-useful evidence for research, care, policy, or responsible technological development. This definition does not require deploying every model. Exploratory analyses, methodological studies, replications, external validations, and negative findings may be valuable when their claims align with their purpose and reduce uncertainty or enable later work [1,4]. Research value does not require universal generalizability or immediate deployment [15]. A model developed for a narrowly defined population may answer an important local question, and a carefully conducted negative validation may prevent an unsafe or unproductive implementation. The value-loss signal arises when the intended scope is unclear, local performance is presented as broadly transportable, or a decision is invited without evidence suited to that decision. This distinction protects focused, early-stage work from being mislabeled as waste while still exposing inflated claims.

The unit of appraisal is therefore the relationship among purpose, evidence, claim, and prospective use. The 4 VALUE-AI pathways are analytic locations where that relationship may fail, not a required sequence every project must follow. Discovery work can remain valuable without advancing to deployment, provided its claims are grounded in discovery-level evidence. Conversely, a strong study may lose value if its design, analysis, or reporting cannot support the clinical or policy inference readers are encouraged to draw [2,16]. Judgment requires an explicit counterfactual. The relevant comparison may be no study, an existing model, a simpler clinical rule, better data collection, independent validation, or investment in implementation capacity. Repetition is not automatically wasteful when it tests reproducibility or transportability, and novelty is not automatically valuable [3,17] when it bypasses a more consequential evidence gap. Assessors should therefore state the implied alternative use of effort and whether the information produced could reasonably change a scientific, clinical, operational, or policy decision. This keeps prioritization distinct from hindsight: an informative study can remain valuable even when its result is null or unfavorable.

The appraisal is qualitative and stage-specific, not a binary threshold or numerical score. Under this lens, actors with

primary control are those with the authority or practical capacity to make or influence the relevant decision or to change the associated process when a value-loss signal is identified. Researchers, reviewers, funders, and implementers should identify the relevant unit of analysis; the evidence supporting the judgment; whether the limitation was reasonably foreseeable and modifiable; which actor or actors had primary control; and what further evidence or action is needed to reduce the loss of value. A transparent discovery study can therefore be decision-useful without a deployment plan, whereas it becomes waste-prone if it makes claims that exceed its maturity—for example, by claiming readiness for routine clinical use without appropriate validation [18], evidence on workflow and safety [19], or governance, monitoring, and the intended response pathway [20,21]. Minor methodological choices can alter the meaning of reported performance. A review of 62 oncology machine-learning prognostic-model publications identified weaknesses in sample-size justification, calibration assessment, and model availability for external validation [22]. A companion analysis of the same publications found that comparative claims were overinterpreted or insufficiently supported [16]. VALUE-AI asks whether such weaknesses materially compromise the study’s purpose, interpretability, reproducibility, or plausible route to benefit. These examples bridge methodological appraisal and value appraisal: a deficit matters because it narrows the claims or decisions the evidence can support. The next step is to locate that loss along the evidence-to-impact pathway.

Table 1 summarizes 4 pathways that localize value loss. These categories identify where value loss occurs across the evidence-to-impact pathway; they are not 4 mutually equivalent attributes of an individual study. Primary research-value loss arises when the research question, design, conduct, reporting, or claims cannot support the study’s stated purpose. Contextual value loss arises when evidence useful under one set of conditions is transferred or generalized without adequate examination of population, infrastructure, workflow, resource, or regulatory differences [15,23]. Deployment failure occurs when a receiving organization implements a technology outside its supported conditions or without necessary workflow, training, and governance safeguards [19,24]. Postdeployment failure occurs when a deployed system is not monitored, updated, or withdrawn when performance, use, equity, or harms change [20,25]. As a practical boundary, contextual value loss applies when transfer-critical conditions or limits were insufficiently investigated, reported, or considered during evidence selection; deployment failure applies when a local organization disregards known limits or fails to establish required safeguards. Nontransportability alone is not contextual value loss. Contextual value loss arises when transfer-critical uncertainty that was reasonably foreseeable for the stated intended use is inadequately investigated or communicated by those producing the evidence, or inadequately considered by those selecting, generalizing, adapting, or applying it in a new setting. These pathways are overlapping heuristic labels rather than mutually exclusive classes: a case may fit one, several, or none. Equity is cross-cutting because proxy choices can distribute errors, benefits, and burdens differently across groups [26].

Table 1. Four pathways through which value may be lost in health care AI.^a

Pathway	Operational boundary	Typical indicators	Actors with primary control
Primary research-value loss	The study cannot support its stated scientific or clinical purpose because of avoidable weaknesses in the question, design, conduct, reporting, or claims.	No stated scientific question or intended decision appropriate to the study’s maturity stage, misaligned outcome or proxy, unsupported claims of readiness, unusable reporting.	Researchers and sponsors; funders, reviewers, and editors through priority-setting, appraisal, and publication decisions.
Contextual value loss	Evidence may be useful in one setting, but transfer assumptions or limits are inadequately examined or communicated.	Population mismatch, unavailable inputs, incompatible workflow, unexamined resources, or regulatory differences.	Original researchers and those selecting, generalizing, or adapting evidence for a new setting.
Deployment failure	A receiving organization implements a technology outside its supported conditions or without required integration and safeguards.	No response pathway, inadequate training, excessive alert burden, unclear authorization, or escalation.	Deploying health systems, vendors, and local governance bodies responsible for configuration, workflow integration, and safeguards.
Postdeployment failure	A deployed sociotechnical system is not monitored or acted on as performance, use, equity, or harms evolve.	No drift or subgroup monitoring, unreviewed overrides, absent recalibration, or deimplementation criteria.	Deployers and developers, with oversight bodies and regulators when they control monitoring requirements, corrective action, or conditions of continued use.

^aThe pathways are overlapping heuristic labels. A study or deployment may fit one, several, or none; attribution should focus on the actor or actors who typically have primary control over the relevant decision or process.

Maturity-Stage Calibration

AI studies should be judged relative to their maturity stage and intended function. Evidence sufficient to establish a signal in discovery is not sufficient to justify integration into care [14,19]. Conversely, requirements designed for deployment should not

be imposed retrospectively on exploratory work that makes no deployment claim. VALUE-AI therefore calibrates expectations to the decision being contemplated: whether to continue investigation, validate in a new setting, prepare a clinical workflow, adopt a system, or continue its use after deployment. At discovery, the central questions are whether the problem is

worth addressing and whether the data and labels correspond to that problem. The analysis should avoid preventable bias, and reporting and claims should remain proportionate to the study design and intended use [18,27]. Internal validation may be appropriate at this stage, but it should not be presented as evidence of transportability or clinical benefit [15,18]. A useful discovery study makes its intended population, prediction target, time horizon, and anticipated use explicit enough that later investigators can test the same proposition rather than reconstructing it from incomplete reporting.

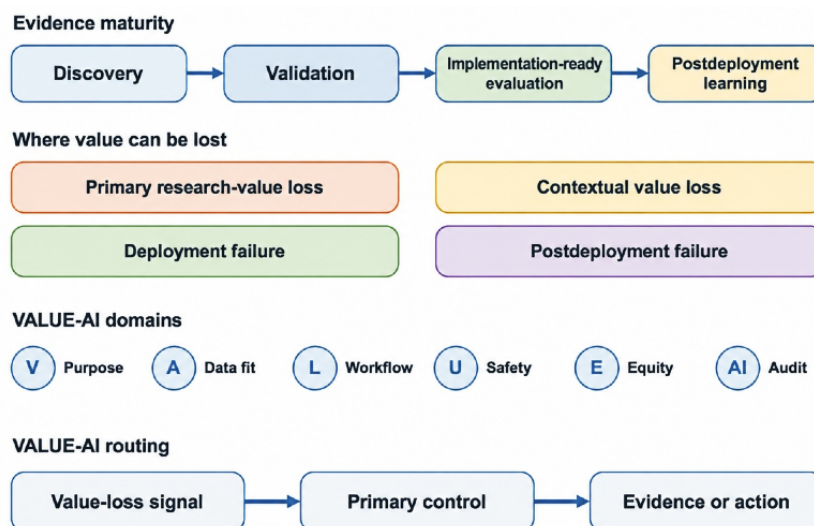
At validation, the evidentiary burden shifts toward reproducibility and performance under meaningful variation [15,18]. Independent temporal, geographic, or institutional evaluation should be selected based on the proposed use rather than treated as a ceremonial exercise. Calibration (agreement between predicted risk and observed outcomes), clinically relevant comparators, subgroup performance, uncertainty, and failure modes become increasingly important [18]. The objective is not simply to show that discrimination (the model's ability to distinguish those with and without the outcome) remains acceptable, but to learn where the model works, where it does not, and whether observed differences would change the next decision. Implementation-ready evaluation requires evidence about the combined sociotechnical intervention, not the algorithm in isolation. Investigators should specify who receives the output, when it appears, what action it is expected to trigger, what alternatives exist, and how override or escalation will work. Human factors, workflow fit, interoperability, and prospective evaluation become central at this stage [19,23]. Governance and accountability arrangements must also be established before technical performance can be responsibly translated into routine use [20,28]. A model can therefore be mature analytically yet remain immature as a clinical service.

After deployment, value depends on sustained observability and the capacity to respond to change. Performance, calibration, use, overrides, subgroup effects, workload, safety signals, and unintended consequences require monitoring against prespecified thresholds [20,25]. Data drift, workflow redesign, software updates, changes in prevalence, and new clinical practices may alter the model's meaning even when its code remains unchanged [29,30]. Documenting versions, incidents, corrective actions, and retirement decisions makes postdeployment evidence cumulative rather than anecdotal [21,28]. Recent evaluations support this staged interpretation.

Among 86 randomized trials of clinical AI, 54% used diagnostic yield or performance as the primary end point, 63% were single center, and 92% were conducted in 1 country [31]. In a systematic review of 572 machine-learning studies of intensive-care scoring systems, only 84 (14.7%) underwent external validation, and discrimination commonly declined in new data [8]. Across 91 real-time sepsis prediction studies, conventional discrimination and decision-oriented utility could diverge under external validation [32]. Study-design labels alone therefore do not establish that the end point, setting, or intervention supports the intended decision.

Stage calibration also makes stopping a legitimate and potentially valuable outcome. A failed external validation [18], an infeasible workflow test [19], or an adverse distributional finding [26] may justify redesign, restriction, or discontinuation before more resources are invested or more patients are exposed. Investigators should specify progression criteria in advance where possible. Research-assessment approaches that value complete reporting and registered reports can help preserve the evidence generated when criteria are not met [33,34]. The relevant question is then whether the project produced reliable learning and an accountable decision, not whether every model reached implementation. This reduces pressure to convert exploratory results into premature claims of readiness. Progression criteria should be tied to the claim and intended decision rather than to a generic hierarchy of study designs. A validation study may justify investigation in another setting without supporting procurement, while a prospective workflow study may establish feasibility without demonstrating patient benefit. Before moving forward, investigators and decision makers should state what findings would support continuation, what findings would require adaptation or additional evidence, and what findings would justify stopping. They should also identify who has authority to make that decision and which uncertainties can reasonably remain unresolved at the relevant stage. Figure 1 summarizes these maturity stages and the VALUE-AI routing logic. A maturity judgment should distinguish evidence that changes the next decision from evidence that merely adds another performance estimate. This prevents automatic progression when the most decision-relevant uncertainty concerns workflow, equity, safety, governance, or monitoring rather than discrimination alone. The same criteria should be communicated consistently in protocols, reports, review decisions, and local governance records [21,27,35].

Figure 1. Algorithmic research waste across evidence maturity stages and the cross-cutting VALUE-AI (Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring) lens. Health care AI research may generate decision-useful evidence at the discovery, validation, implementation-ready evaluation, and postdeployment learning stages. Value may be lost through primary research-value loss, contextual value loss, deployment failure, or postdeployment failure. VALUE-AI applies 6 cross-cutting domains: Valid Question and Purpose; Appropriate Data and Population Fit; Local Workflow and Intended Use; Usability, Safety, and Accountability; Equity and External Validity; and Auditability and Impact Monitoring. When a value-loss signal is identified, the lens directs attention to the actor or actors with primary control and the next maturity-appropriate evidence or action. These pathways are overlapping heuristic categories rather than mutually exclusive classifications.



The VALUE-AI Lens

VALUE-AI was derived from a conceptual crosswalk that aligns recurring points of value loss in the research-waste literature with the questions addressed by prediction-model reporting and risk-of-bias tools, clinical evaluation guidance, implementation frameworks, and responsible AI governance standards. This was not a formal consensus, Delphi, or systematic framework-development process. The domains were retained because each addresses a distinct decision point where evidence can lose value, and together they span question formulation through postdeployment learning. TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence) strengthens transparent reporting [27]; PROBAST (Prediction Model Risk of Bias Assessment Tool) and PROBAST+AI assess quality, risk of bias, and applicability [18,36]; CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) extend trial reporting and protocol guidance [35,37]; and DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence) addresses early live clinical evaluation [19]. NASSS (Nonadoption, Abandonment, Scale-up, Spread, and Sustainability), CFIR (Consolidated Framework for Implementation Research), RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance), and implementation-outcome frameworks examine adoption, context, implementation, and sustainability [23,24,38,39]. FUTURE-AI (Fairness, Universality, Traceability, Usability, Robustness, and Explainability for Artificial Intelligence) addresses trustworthy and deployable AI across its life cycle

[20]; the NIST (National Institute of Standards and Technology) AI Risk Management Framework addresses risk management [21]; WHO (World Health Organization) guidance addresses ethics and governance [28]; and FDA (Food and Drug Administration) guidance addresses predetermined change control plans for planned modifications to AI-enabled device software functions [40]. The 6 VALUE-AI domains do not claim that these underlying concerns are new. Their distinct contribution is to organize findings from existing tools around a metascience question: where is potential research value being lost, which actor or actors can modify the relevant source of loss, and what maturity-appropriate evidence or action is needed next? If users jointly apply the relevant existing tools and deliberately synthesize their findings around those questions, they are performing the function VALUE-AI is intended to support.

Table 2 operationalizes the 6 VALUE-AI domains across the research life cycle and links each domain to its primary object, appraisal questions, and observable signals. The domains are considered together, but each asks a distinct question about whether the work can create, preserve, and transmit value. A weakness becomes consequential when it obstructs the study's purpose or renders the associated claim unreliable, unusable, or inequitable. The Valid Question and Purpose Domain asks whether the study addresses a meaningful uncertainty and whether its intended user, decision, maturity stage, and claim are coherent. It distinguishes a technically feasible task from a consequential research question. The issue is not whether every project promises immediate clinical benefit, but whether the stated purpose justifies the data, design, and inference. A high headline metric cannot compensate for a target that is poorly aligned with the proposed scientific or clinical decision.

Table 2. Operational VALUE-AI^a domains.^b

Domain	Primary object	Core appraisal questions	Potential value-loss signals
V—Valid question and purpose	Research question, intended decision, and stated purpose	Who will use the output, when, for what decision, and at what maturity stage? Are the claims proportionate to the design?	Question lacks a meaningful uncertainty, intended user, decision consequence, or maturity-appropriate claim.
A—Appropriate data and population fit	Dataset, measurements, proxies, and target population	Who is represented? How were variables, outcomes, missingness, and proxies produced? Do they fit the intended population and purpose?	Data, measurements, missingness, proxies, or target-population fit are insufficiently justified, characterized, or reported.
L—Local workflow and intended use	Use context, workflow, and infrastructure	Where would the output appear, to whom, at what time, and with what feasible response? Which local adaptations are required?	Workflow location, response pathway, required inputs, or transfer assumptions are unspecified.
U—Usability, safety, and accountability	Human-AI interaction and governance	Can users interpret, contest, override, and escalate safely? Who authorizes use and who is responsible for action?	Arrangements for authorization, user interpretation, override, escalation, training, or failure management are unspecified or untested.
E—Equity and external validity	Evidence across groups, settings, and time	Are calibration, errors, benefits, and harms examined across relevant groups and external conditions?	Evaluation is limited to aggregate internal performance without relevant subgroup, temporal, or external assessment.
AI—Auditability and impact monitoring	Deployed sociotechnical system and decision records	Can model provenance and the decision sequence be examined? What will be monitored, by whom, and what triggers recalibration or withdrawal?	Provenance records, versioning, monitoring ownership, decision thresholds or output-interpretation rules, response actions, or deimplementation criteria are unspecified.

^aVALUE-AI: Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring.

^bDomain judgments form a life cycle profile and should not be combined into a numerical score.

The Appropriate Data and Population Fit Domain asks whether the dataset, measurements, proxies, outcomes, and sampling process support the intended purpose and population. It includes data provenance, leakage, missingness, reference standards, prevalence, representativeness, and uncertainty [18,27]. Narrow data are not inherently deficient, but their boundaries should be explicit. When broader use is proposed, evaluation should expose the model to the temporal, geographic, institutional, and demographic variation relevant to that use [15,18]. The Local Workflow and Intended Use Domain asks where the output would appear, to whom, at what time, and with what feasible response. Required inputs, interoperability, staffing, resources, escalation options, and competing processes may determine whether an output can influence care. This domain separates model performance from the operational pathway through which benefit is expected [19,23]. It also makes transfer assumptions visible when evidence developed in one service is considered for another.

The Usability, Safety, and Accountability Domain asks whether people can interpret, contest, override, and escalate model outputs safely and whether authority for action is clear. Human factors, training, automation bias, alert burden, privacy, consent or an appropriate alternative basis for data use, governance, and foreseeable harms are relevant [19,20]. Formal approval alone may be insufficient when affected patients and professionals lack a credible route to question assumptions, report problems, or obtain redress [28]. The Equity and External Validity Domain asks how evidence, benefits, burdens, errors, and access vary across groups, settings, and time [20,26]. Aggregate performance can conceal clinically important differences, while subgroup estimates can themselves be unstable or misleading [18]. Appraisal therefore considers representation, measurement

quality, calibration, uncertainty, intersectional effects, deployment context, and whether mitigation transfers burden elsewhere. The aim is reasoned distributional and transportability analysis, not a ritual list of subgroup metrics.

The Auditability and Impact Monitoring Domain asks whether the deployed sociotechnical system and its decision sequence can be examined over time. Model and data versions, outputs, user exposure, overrides, actions, outcomes, incidents, and corrective decisions may all be relevant. Feedback loops, silent updates, drift, and dependency on proprietary infrastructure can otherwise obscure change [29,30]. Monitoring should have named owners, interpretable thresholds, response authority, and criteria for recalibration, restriction, or withdrawal [20,21,25]. Taken together, the domains provide a structured explanation of where value is secure, uncertain, or being lost. They also prevent a single favorable characteristic from dominating the assessment: a useful model may still lack legitimacy, an equitable development dataset may not ensure equitable deployment, and rigorous validation may not establish workflow utility. The resulting judgment should state which domain matters, at what maturity stage, and for which claim or decision. Interactions among domains should be recorded rather than collapsed. Poor representation may weaken appropriate data and population fit, undermine external validity, and widen inequity, but the corrective action for each consequence can differ. Likewise, an opaque proprietary update may create an auditability problem if version provenance is unavailable and an accountability problem if oversight is bypassed [21]. Naming the causal pathway helps assign responsibility and prevents broad labels such as bias or implementation failure from obscuring who can act and what maturity-appropriate evidence or action is needed next.

Relationship to Existing Frameworks

Table 3 makes the relationship explicit. VALUE-AI is a metascience attribution-and-routing layer, not a replacement checklist. After users apply the instruments appropriate to the study's purpose and stage, either individually or in combination, VALUE-AI links residual deficiencies to the actor or actors able to modify them and to the next maturity-appropriate

evidence or action. This synthesis matters when instruments identify different concerns because strong reporting or governance does not resolve a weak purpose, inappropriate transfer, or absent response pathway. It prevents strength in one dimension from being treated as sufficient when another consequential gap remains. The frameworks shown are illustrative rather than exhaustive and demonstrate how VALUE-AI can synthesize complementary forms of appraisal across the AI evidence life cycle.

Table 3. How VALUE-AI^a complements established health care AI and implementation frameworks.

Framework	Primary contribution	VALUE-AI ^b domains most directly informed	Residual contribution of VALUE-AI after instrument use
TRIPOD+AI ^c [27]	Transparent reporting of prediction-model studies	V, A, and E	Interprets whether the transparently reported evidence can support the intended decision and identifies the next unresolved evidence need.
PROBAST ^d and PROBAST+AI [18,36]	Quality, risk of bias, and applicability of prediction models	A and E	Links identified bias or applicability concerns to value loss, responsibility, and the next evidence need.
CONSORT-AI ^e and SPIRIT-AI ^f [35,37]	Reporting of AI intervention trials and protocols	V, L, U, and E	Judges whether the trial question and evidence pathway justify the resources and claims at the stated maturity stage.
DECIDE-AI ^g [19]	Early live clinical evaluation of AI decision support	L, U, and AI	Interprets how early clinical findings change the value profile and readiness for the next stage.
NASSS ^h [23]	Nonadoption, abandonment, scale-up, spread, and sustainability	L, U, and AI	Distinguishes complexity foreseeable during research from complexity arising from downstream deployment choices.
CFIR ⁱ , RE-AIM ^j , and implementation outcomes [24,38,39]	Context, implementation determinants, reach, outcomes, and sustainability	L, U, E, and AI	Uses implementation findings to redirect research priorities and distinguish research-level loss from deployment failure.
FUTURE-AI ^k [20]	Trustworthy and deployable health AI across 6 principles	A, L, U, E, and AI	Judges whether trustworthy-AI evidence is proportionate to the stated purpose and is decision-useful.
NIST ^l , WHO ^m , and FDA ⁿ guidance [21,28,40]	NIST risk management; WHO ethics and health-AI governance; and FDA guidance on predetermined change control for planned modifications to AI-enabled device software functions	U, E, and AI	Links missing governance evidence to the responsible actor and the next action needed to protect research value.

^aVALUE-AI: Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring.

^bVALUE-AI does not duplicate item-level appraisal. After applicable instruments are used individually or in combination, it synthesizes findings across instruments and maturity stages, identifies the value-loss pathway and actor or actors with primary control, and routes the finding to the next maturity-appropriate evidence or action.

^cTRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence.

^dPROBAST: Prediction Model Risk of Bias Assessment Tool.

^eCONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence.

^fSPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence.

^gDECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence.

^hNASSS: Nonadoption, Abandonment, Scale-up, Spread, and Sustainability.

ⁱCFIR: Consolidated Framework for Implementation Research.

^jRE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance.

^kFUTURE-AI: Fairness, Universality, Traceability, Usability, Robustness, and Explainability for Artificial Intelligence.

^lNIST: National Institute of Standards and Technology.

^mWHO: World Health Organization.

ⁿFDA: Food and Drug Administration.

Reporting standards and risk-of-bias tools identify important deficits but do not establish whether a research question was

consequential, whether a claim aligns with the maturity stage, or whether a credible route to use exists. In 65 randomized trials

published after CONSORT-AI, median concordance was 90%, yet algorithm version and intervention or code accessibility were reported in only 20% and 42%, respectively [41]. VALUE-AI adds a judgment about purpose and consequence while directing assessors to the relevant specialist instrument. An incomplete model report routes to reporting guidance, a biased validation study to a risk-of-bias tool, and a workflow barrier to implementation science; VALUE-AI records how each finding affects the capacity to inform a decision.

Borderline Worked Example

Consider a hypothetical model developed at a tertiary cardiac center to predict postoperative deterioration. At the discovery stage, a transparent exploratory study using local data, clearly presented as hypothesis-generating and intended to test whether a signal exists, represents neither primary research-value loss nor contextual value loss merely because it is not ready for deployment. It becomes primary research-value loss if the authors claim clinical readiness without temporal or external validation, calibration assessment, clinically appropriate decision thresholds, or adequate output-interpretation rules [18,27], or without a defined clinical response pathway and corresponding workflow evidence [19]. A retrospective validation using hypothetical alerts can be informative at this point if it is presented as a test of feasibility, timing, or potential workload rather than as evidence of clinical effect. In an independent validation of a proprietary sepsis model, discrimination was lower than the developer-reported performance; the threshold flagged 18% of hospitalizations while missing 67% of sepsis cases [42]. As alerts were hypothetical, the study supports concern about transportability and potential alert burden but does not demonstrate changed management, benefit, or harm.

Suppose the model is then proposed for a resource-variable hospital where required measurements are delayed or unavailable and specialist escalation pathways differ. If the original study did not describe these dependencies or overstated transportability, contextual value loss is implicated. If the receiving hospital deploys the model despite clearly reported limits and without local validation, training, workflow design, or escalation rules, that is primarily a deployment failure rather than a defect in the original research. A recent review of cardiac-surgery AI similarly identified external validation, workflow integration, governance, and prospective evaluation as important considerations in translating technical performance into clinical value [43]. After deployment, model version, input data, output, timing, user exposure, overrides, actions, and outcomes should be preserved in accordance with applicable law and governance requirements [20,21]. These records support provenance and may permit reconstruction of the decision sequence, but they do not establish that an AI output caused or materially influenced a clinical action. In VALUE-AI, provenance records identify the model, inputs, and outputs; reconstruction traces the observable sequence; attribution evaluates whether the output influenced action; and accountability assigns authorization and responsibility. Reconstruction fidelity is therefore an empirical property to be tested against specified records and procedures, not assumed from the existence of logs. The next step is staged evidence

generation: validation and subgroup analysis [18], workflow and human-factors testing [19], and postdeployment monitoring with explicit response and withdrawal criteria [20,21,25].

VALUE-AI routes each finding by locating the relevant pathway and domain, identifying the actor with primary control, and specifying the next maturity-appropriate action. Poor external performance routes to validation; an output that arrives too late routes to workflow redesign; findings of inequitable error patterns or access barriers route to equity mitigation; and an unannounced update routes to governance review. The response may be redesign, restricted use, monitoring, or retirement, but the claim should remain within the evidence obtained. Positive controls show what stronger evidence looks like without implying that every study must reach the same end point. A prospective interventional cohort study integrated real-time diabetic retinopathy screening and safety over-read by specialists across 9 sites [44]. The population-based randomized Mammography Screening with Artificial Intelligence (MASAI) trial evaluated AI-supported mammography through interval-cancer follow-up [45]. Both connect model performance to workflow, comparators, workload, and outcomes. Neither should be generalized beyond its setting or treated as proof that deployment guarantees value. For the hypothetical deterioration model, a decision record would make this reasoning explicit. The tertiary center might continue with prospective silent-mode evaluation while withholding clinical claims; the resource-variable hospital might pause adoption until locally obtainable inputs and escalation options are tested; and a deploying institution might continue use only under defined monitoring and suspension thresholds. These are different, evidence-linked decisions about the same model, illustrating why value cannot be inferred from the model artifact alone.

Implications for Research and Governance

For researchers, VALUE-AI can be used prospectively to align claims and evidence with maturity. Discovery studies should specify their scientific purpose and limitations [27]; validation studies should prioritize transportability, calibration, subgroup performance, and reproducibility [18]; implementation-ready evaluations should examine the full intervention within the workflow [19]; and postdeployment studies should examine monitoring methods and changing performance over time [25]. Reviewers and editors can ask which VALUE-AI domain contains the most consequential unresolved value-loss signal and direct authors to the relevant detailed standard rather than expecting a single checklist to answer every question.

In practice, this means writing the intended decision before selecting the evidence used to justify it. Protocols can draw on established guidance to specify intended use, comparators, and decision pathways [35], while VALUE-AI additionally identifies the maturity stage, the principal domains at risk, and the stopping or progression rule. Reports can then separate what was measured from what is inferred and what remains hypothetical [27,37]. This separation matters because promising results may otherwise remain difficult to reproduce, compare, or translate.

Funders may reduce avoidable duplication and premature model development by rewarding replication, external validation, comparative evaluation, and publication of informative negative findings [1,4]. VALUE-AI additionally encourages investment in monitoring infrastructure and partnerships with intended users. Where several retrospective models already exist, independent validation or prospective workflow evaluation may add more value than another development study. Funding milestones can specify the evidence bottleneck, progression or stopping rules, and requirements for reporting informative negative findings. Research-assessment principles and registered reports likewise shift incentives from novelty and positive results toward rigor, transparency, and cumulative knowledge [17,33,34].

Health systems should treat publication as the start of local appraisal, not as a deployment guarantee. Before adoption, they should assess population and data fit, infrastructure, workflow, and staffing [23,24]. They should also examine reach, implementation outcomes, and sustainability [38,39]. Equity, legal and governance requirements, monitoring capacity, and response ownership require separate appraisal [20,21,28]. Before deployment, they should designate monitoring owners; define data-quality checks, clinically meaningful thresholds, subgroup review, and incident pathways; and assign authority to modify, suspend, or retire the system [20,21,25]. A scoping review identified 39 sources on clinical-AI monitoring, but only 9 described clinically tested or implemented methods [25]. Conformal Multidimensional Prediction of Sepsis Risk (COMPOSER) provides a bounded example that links outcome evaluation to site-specific analysis, data-quality surveillance, biweekly performance review, and predefined change control [46].

For regulators and policy makers, VALUE-AI separates product evidence from evidence on its use within a service. Public summaries should identify the intended population, user, workflow, model version, principal uncertainties, and postmarket obligations [20,28,40]. Of 794 AI-enabled medical devices authorized from 2023 to 2025, 43 (5.4%) had FDA-authorized predetermined change control plans; among those 43 devices, 31 (72%) publicly reported all FDA-recommended elements [47]. Of 692 publicly available FDA summaries of safety and effectiveness data for AI-enabled devices, 3.6% reported race or ethnicity, 46.1% provided comprehensive performance results, and 9.0% described prospective postmarket studies [48]. Of 1357 FDA-authorized devices listed through December 5, 2025, 34 were linked to registered prospective trials, and only 3 evaluated patient-centered outcomes such as mortality, morbidity, or readmissions [49]. These findings concern publicly traceable evidence and do not establish that unlinked devices are ineffective or wholly unevaluated.

Patients, professionals, and developers have complementary roles in this process. Patients and communities can identify outcomes, burdens, and distributional concerns that performance summaries overlook. They require clear pathways for explanation and redress [28]. How patient and public involvement shaped the research should be reported transparently, including contributors' roles and influence [50]. Professionals can test whether outputs align with actual

decisions and report workarounds or new hazards [19,20]. Developers should preserve data and model provenance [20,21], disclose relevant limitations, and support independent evaluation. For regulated AI-enabled devices, planned modifications should be disclosed and managed through applicable change-control requirements [40]. Participation does not transfer accountability to those consulted; organizations that retain authority over design, procurement, or deployment remain responsible for acting on the evidence produced [28].

Scope and Limitations

VALUE-AI is a conceptual metascience lens, not a validated reporting guideline, risk-of-bias tool, scoring system, or implementation instrument. The domains, signals, and routing questions are conceptual heuristics; they have not been empirically validated, weighted, or tested for interrater reliability. Their usefulness and interpretability require evaluation across specialties, resource settings, and AI modalities. The 4 pathways may overlap, and classifications may change as evidence and context evolve. The lens focuses primarily on prediction and decision-support systems. Applying it to generative AI, reinforcement learning, or administrative automation would require modality-specific adaptation and separate empirical evaluation addressing open-ended outputs, feedback loops, autonomy, provenance, and task-specific safety. Finally, this viewpoint is a selective conceptual synthesis rather than a systematic review of all health care AI evidence or guidance.

The next empirical task is therefore to test the lens as an appraisal process. Studies should examine interrater agreement, the reasons assessors disagree, the time and expertise required, and whether judgments change when maturity stage or intended use is specified more clearly. Construct validity can be explored by comparing VALUE-AI findings with established reporting, bias, implementation, ethics, and economic assessments while preserving the distinct question of value loss. Prospective use in funding panels, protocol review, journal assessment, and procurement could test whether the lens changes decisions or merely redescribes known concerns. Case sampling should include early exploratory work, negative validation, successful implementation, withdrawn systems, and settings with constrained infrastructure. Although VALUE-AI is not a reporting guideline, stakeholder consultation, piloting, consensus development, and evaluation of uptake offer a relevant methodological analogy for its refinement [51]. Patient and public contributors should help define acceptable evidence and consequential harm. Longitudinal evaluation is also needed because an appraisal may change as the model, workflow, population, or available alternative changes; agreement at one time point would not establish durable value. These evaluations should avoid converting VALUE-AI into an unweighted score: domain importance is context-dependent, and apparently similar deficits may have different consequences at different maturity stages. Until such testing is available, judgments should be transparent, reasoned, and accompanied by the evidence and assumptions on which they depend.

Empirical evaluation should assess not only whether assessors agree, but also whether disagreements reveal different assumptions about purpose, evidence, context, or responsibility. A useful appraisal lens may expose these underlying assumptions even when it does not yield identical classifications. Studies should therefore document the reasoning behind each judgment, the evidence that changed it, the actor assigned primary control, and the action ultimately recommended. Comparisons with usual review processes should assess whether VALUE-AI identifies neglected uncertainties, redirects resources toward more informative work, or supports earlier restriction or discontinuation when progression criteria are not met. Evaluation should also look for unintended effects, including excessive documentation, false confidence from apparently complete domain coverage, or inappropriate use of the lens to reject legitimate exploratory research. These outcomes would test whether VALUE-AI improves the quality and transparency of decisions rather than merely increasing appraisal burden.

Conclusions

Health care AI research can generate value at multiple stages, from exploratory methodological work to external validation, prospective evaluation, and postdeployment learning. Algorithmic research waste does not imply that every model must be deployed or that early-stage, locally focused, negative, or noncommercial research is unhelpful. It describes an

avoidable loss of potential value when a study's question, design, reporting, validation, implementation assumptions, or claims are insufficient to support the scientific, clinical, policy, or operational decision they purport to inform. VALUE-AI is a stage-specific metascience attribution-and-routing lens. It highlights 6 cross-cutting conditions for decision-useful evidence: a valid question and purpose; appropriate data and population fit; local workflow and intended use; usability, safety, and accountability; equity and external validity; and auditability and impact monitoring. The lens does not replace detailed reporting standards, risk-of-bias tools, clinical-evaluation guidance, implementation frameworks, or governance standards. Its contribution is to connect their findings by asking where research value is at risk, which actor or actors have practical capacity to address the source of loss, and what maturity-appropriate evidence or action should follow. The practical aim is to reduce the production of evidence that is technically impressive yet unable to guide a proportionate next decision, without creating another score or unnecessary administrative burden. Researchers, editors, funders, health systems, regulators, developers, and patient and public representatives can use this perspective to prioritize transparent questions, credible validation, locally feasible pathways, accountable oversight, and ongoing learning. VALUE-AI now requires empirical evaluation to determine whether it improves the clarity, consistency, and usefulness of decisions across research and deployment settings.

Acknowledgments

The author thanks colleagues and independent academic readers who provided informal feedback during manuscript development and gratefully acknowledges the faculty of the AI in Health Care: From Strategies to Implementation program at Harvard Medical School for scholarly insights into clinical AI and prediction-model evaluation that informed the development of this viewpoint. The author conceived the VALUE-AI lens, developed its domains, pathways, worked example, arguments, and conclusions, and wrote the manuscript. In line with the Generative Artificial Intelligence Delegation Taxonomy (GAIDeT) framework, generative AI and AI-assisted tools were used under full human supervision for limited supporting tasks: Perplexity (model version not recorded; Perplexity AI) to identify additional candidate publications to supplement database searches and to support manuscript-quality checks; ChatGPT and Codex (GPT-5.6 Sol; OpenAI) for proofreading, editing, and reformatting; and Paperpal (Cactus Communications Services Pte) and Grammarly (Grammarly Inc) for language and style review. The author screened and read all cited original sources, selected the references, verified all citations and factual claims against the original sources, and reviewed and revised all AI-assisted output. These tools did not generate the manuscript's concepts, arguments, interpretations, or conclusions. The author made all final decisions and takes full responsibility for the manuscript.

Funding

The author declares that no financial support was received for this work.

Data Availability

No new data were generated or analyzed in this article. Data sharing is not applicable because this viewpoint is based on interpretation of published literature and development of a conceptual lens.

Authors' Contributions

Conceptualization: DCN

Writing – original draft: DCN

Writing – review and editing: DCN

Conflicts of Interest

None declared.

References

1. Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PMM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? *Lancet* 2016 Apr;387(10027):1573-1586. [doi: [10.1016/s0140-6736\(15\)00307-4](https://doi.org/10.1016/s0140-6736(15)00307-4)]
2. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014 Jan;383(9913):267-276. [doi: [10.1016/s0140-6736\(13\)62228-x](https://doi.org/10.1016/s0140-6736(13)62228-x)]
3. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. *Lancet* 2014 Jan 11;383(9912):156-165. [doi: [10.1016/S0140-6736\(13\)62229-1](https://doi.org/10.1016/S0140-6736(13)62229-1)] [Medline: [24411644](https://pubmed.ncbi.nlm.nih.gov/24411644/)]
4. Rosengard LO, Andersen MZ, Rosenberg J, Fonnes S. Five aspects of research waste in biomedicine: a scoping review. *J Evid Based Med* 2024 Jun 26;17(2):351-359. [doi: [10.1111/jebm.12616](https://doi.org/10.1111/jebm.12616)] [Medline: [38798014](https://pubmed.ncbi.nlm.nih.gov/38798014/)]
5. Arshi B, Wynants L, Rijnhart E, Reeve K, Cowley LE, Smits LJ. Number of publications on new clinical prediction models: a bibliometric review. *JMIR Med Inform* 2025 Jul 04;13:e62710-e62710 [FREE Full text] [doi: [10.2196/62710](https://doi.org/10.2196/62710)] [Medline: [40614260](https://pubmed.ncbi.nlm.nih.gov/40614260/)]
6. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol* 2022 Jan 13;22(1):12 [FREE Full text] [doi: [10.1186/s12874-021-01469-6](https://doi.org/10.1186/s12874-021-01469-6)] [Medline: [35026997](https://pubmed.ncbi.nlm.nih.gov/35026997/)]
7. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021 Oct 20;375:n2281 [FREE Full text] [doi: [10.1136/bmj.n2281](https://doi.org/10.1136/bmj.n2281)] [Medline: [34670780](https://pubmed.ncbi.nlm.nih.gov/34670780/)]
8. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. *BMC Med Inform Decis Mak* 2025 Jan 06;25(1):5 [FREE Full text] [doi: [10.1186/s12911-024-02830-7](https://doi.org/10.1186/s12911-024-02830-7)] [Medline: [39762808](https://pubmed.ncbi.nlm.nih.gov/39762808/)]
9. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020 Mar 25;368:m689 [FREE Full text] [doi: [10.1136/bmj.m689](https://doi.org/10.1136/bmj.m689)] [Medline: [32213531](https://pubmed.ncbi.nlm.nih.gov/32213531/)]
10. Berkhout WEM, van Wijngaarden JJ, Workum JD, van de Sande D, Hilling DE, Jung C, et al. Operationalization of artificial intelligence applications in the intensive care unit: a systematic review. *JAMA Netw Open* 2025 Jul 01;8(7):e2522866 [FREE Full text] [doi: [10.1001/jamanetworkopen.2025.22866](https://doi.org/10.1001/jamanetworkopen.2025.22866)] [Medline: [40699572](https://pubmed.ncbi.nlm.nih.gov/40699572/)]
11. de Vries EF, Struijs JN, Heijink R, Hendriks RJP, Baan CA. Are low-value care measures up to the task? A systematic review of the literature. *BMC Health Serv Res* 2016 Aug 18;16(1):405 [FREE Full text] [doi: [10.1186/s12913-016-1656-3](https://doi.org/10.1186/s12913-016-1656-3)] [Medline: [27539054](https://pubmed.ncbi.nlm.nih.gov/27539054/)]
12. Shrank WH, Gondi S, Brailer DJ. Artificial intelligence and health care waste-promise or peril? *JAMA Health Forum* 2025 Jun 07;6(6):e251637 [FREE Full text] [doi: [10.1001/jamahealthforum.2025.1637](https://doi.org/10.1001/jamahealthforum.2025.1637)] [Medline: [40512508](https://pubmed.ncbi.nlm.nih.gov/40512508/)]
13. Hendrix N, Veenstra DL, Cheng M, Anderson NC, Verguet S. Assessing the economic value of clinical artificial intelligence: challenges and opportunities. *Value Health* 2022 Mar;25(3):331-339 [FREE Full text] [doi: [10.1016/j.jval.2021.08.015](https://doi.org/10.1016/j.jval.2021.08.015)] [Medline: [35227443](https://pubmed.ncbi.nlm.nih.gov/35227443/)]
14. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019 Oct 29;17(1):195 [FREE Full text] [doi: [10.1186/s12916-019-1426-2](https://doi.org/10.1186/s12916-019-1426-2)] [Medline: [31665002](https://pubmed.ncbi.nlm.nih.gov/31665002/)]
15. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020 Sep;2(9):e489-e492 [FREE Full text] [doi: [10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2)] [Medline: [32864600](https://pubmed.ncbi.nlm.nih.gov/32864600/)]
16. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *J Clin Epidemiol* 2023 May;157:120-133 [FREE Full text] [doi: [10.1016/j.jclinepi.2023.03.012](https://doi.org/10.1016/j.jclinepi.2023.03.012)] [Medline: [36935090](https://pubmed.ncbi.nlm.nih.gov/36935090/)]
17. Smaldino PE, McElreath R. The natural selection of bad science. *R Soc Open Sci* 2016 Sep;3(9):160384 [FREE Full text] [doi: [10.1098/rsos.160384](https://doi.org/10.1098/rsos.160384)] [Medline: [27703703](https://pubmed.ncbi.nlm.nih.gov/27703703/)]
18. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 2025 Mar 24;388:e082505. [doi: [10.1136/bmj-2024-082505](https://doi.org/10.1136/bmj-2024-082505)] [Medline: [40127903](https://pubmed.ncbi.nlm.nih.gov/40127903/)]
19. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022 May 18;377:e070904 [FREE Full text] [doi: [10.1136/bmj-2022-070904](https://doi.org/10.1136/bmj-2022-070904)] [Medline: [35584845](https://pubmed.ncbi.nlm.nih.gov/35584845/)]
20. Lekadir K, Frangi A, Porras A, Glocker B, Cintas C, Langlotz C, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 2025 Feb 5;388:e081554. [doi: [10.1136/bmj-2024-081554](https://doi.org/10.1136/bmj-2024-081554)]

21. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology (NIST). Gaithersburg, MD: National Institute of Standards and Technology; 2023 Jan 26. URL: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf> [accessed 2026-10-09]
22. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022 Apr 08;22(1):101 [FREE Full text] [doi: [10.1186/s12874-022-01577-x](https://doi.org/10.1186/s12874-022-01577-x)] [Medline: [35395724](https://pubmed.ncbi.nlm.nih.gov/35395724/)]
23. Greenhalgh T, Wherton J, Papoutsis 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 Nov 01;19(11):e367 [FREE Full text] [doi: [10.2196/jmir.8775](https://doi.org/10.2196/jmir.8775)] [Medline: [29092808](https://pubmed.ncbi.nlm.nih.gov/29092808/)]
24. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. *Implement Sci* 2022 Oct 29;17(1):75 [FREE Full text] [doi: [10.1186/s13012-022-01245-0](https://doi.org/10.1186/s13012-022-01245-0)] [Medline: [36309746](https://pubmed.ncbi.nlm.nih.gov/36309746/)]
25. Andersen ES, Birk-Korch JB, Hansen RS, Fly LH, Röttger R, Arcani DMC, et al. Monitoring performance of clinical artificial intelligence in health care: a scoping review. *JBIM Evid Synth* 2024 Dec 01;22(12):2423-2446. [doi: [10.11124/JBIES-24-00042](https://doi.org/10.11124/JBIES-24-00042)] [Medline: [39658865](https://pubmed.ncbi.nlm.nih.gov/39658865/)]
26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019 Oct 25;366(6464):447-453 [FREE Full text] [doi: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)] [Medline: [31649194](https://pubmed.ncbi.nlm.nih.gov/31649194/)]
27. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024 Apr 18;385:q902 [FREE Full text] [doi: [10.1136/bmj.q902](https://doi.org/10.1136/bmj.q902)] [Medline: [38636956](https://pubmed.ncbi.nlm.nih.gov/38636956/)]
28. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. Geneva, Switzerland: World Health Organization; 2021 Jun 28. URL: <https://www.who.int/publications/item/9789240029200> [accessed 2026-10-09]
29. Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. *NPJ Digit Med* 2024 May 09;7(1):120 [FREE Full text] [doi: [10.1038/s41746-024-01085-w](https://doi.org/10.1038/s41746-024-01085-w)] [Medline: [38724581](https://pubmed.ncbi.nlm.nih.gov/38724581/)]
30. Silva GFDS, Barcellos Filho FN, Wichmann RM, da Silva Junior FC, Chiavegatto Filho ADP. Strategies for detecting and mitigating dataset shift in machine learning for health predictions: a systematic review. *J Biomed Inform* 2025 Oct;170:104902. [doi: [10.1016/j.jbi.2025.104902](https://doi.org/10.1016/j.jbi.2025.104902)] [Medline: [40876698](https://pubmed.ncbi.nlm.nih.gov/40876698/)]
31. 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 May;6(5):e367-e373. [doi: [10.1016/s2589-7500\(24\)00047-5](https://doi.org/10.1016/s2589-7500(24)00047-5)]
32. Wang Z, Wang W, Sun C, Li J, Xie S, Xu J, et al. A methodological systematic review of validation and performance of sepsis real-time prediction models. *NPJ Digit Med* 2025 Apr 07;8(1):190 [FREE Full text] [doi: [10.1038/s41746-025-01587-1](https://doi.org/10.1038/s41746-025-01587-1)] [Medline: [40189694](https://pubmed.ncbi.nlm.nih.gov/40189694/)]
33. Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. The Hong Kong Principles for assessing researchers: fostering research integrity. *PLoS Biol* 2020 Jul;18(7):e3000737 [FREE Full text] [doi: [10.1371/journal.pbio.3000737](https://doi.org/10.1371/journal.pbio.3000737)] [Medline: [32673304](https://pubmed.ncbi.nlm.nih.gov/32673304/)]
34. Chambers CD, Tzavella L. The past, present and future of registered reports. *Nat Hum Behav* 2022 Jan;6(1):29-42. [doi: [10.1038/s41562-021-01193-7](https://doi.org/10.1038/s41562-021-01193-7)] [Medline: [34782730](https://pubmed.ncbi.nlm.nih.gov/34782730/)]
35. Rivera SC, Liu X, Chan A, 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. *BMJ* 2020 Sep 09;370:m3210 [FREE Full text] [doi: [10.1136/bmj.m3210](https://doi.org/10.1136/bmj.m3210)] [Medline: [32907797](https://pubmed.ncbi.nlm.nih.gov/32907797/)]
36. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019 Jan 01;170(1):51-58. [doi: [10.7326/m18-1376](https://doi.org/10.7326/m18-1376)]
37. 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 Sep 09;26(9):1364-1374 [FREE Full text] [doi: [10.1038/s41591-020-1034-x](https://doi.org/10.1038/s41591-020-1034-x)] [Medline: [32908283](https://pubmed.ncbi.nlm.nih.gov/32908283/)]
38. Glasgow RE, Harden SM, Gaglio B, Rabin B, Smith ML, Porter GC, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. *Front Public Health* 2019;7:64 [FREE Full text] [doi: [10.3389/fpubh.2019.00064](https://doi.org/10.3389/fpubh.2019.00064)] [Medline: [30984733](https://pubmed.ncbi.nlm.nih.gov/30984733/)]
39. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm Policy Ment Health* 2011 Mar 19;38(2):65-76 [FREE Full text] [doi: [10.1007/s10488-010-0319-7](https://doi.org/10.1007/s10488-010-0319-7)] [Medline: [20957426](https://pubmed.ncbi.nlm.nih.gov/20957426/)]
40. US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. US Food and Drug Administration. Silver Spring, MD: US Food and Drug Administration; 2025 Aug 18. URL: <https://www.fda.gov/media/166704/download> [accessed 2026-09-16]

41. Martindale APL, Llewellyn CD, de Visser RO, Ng B, Ngai V, Kale AU, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. *Nat Commun* 2024 Feb 22;15(1):1619 [FREE Full text] [doi: [10.1038/s41467-024-45355-3](https://doi.org/10.1038/s41467-024-45355-3)] [Medline: [38388497](https://pubmed.ncbi.nlm.nih.gov/38388497/)]
42. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021 Aug 01;181(8):1065-1070 [FREE Full text] [doi: [10.1001/jamainternmed.2021.2626](https://doi.org/10.1001/jamainternmed.2021.2626)] [Medline: [34152373](https://pubmed.ncbi.nlm.nih.gov/34152373/)]
43. Ngwu D, Hamadi F, Akinyemi O, Wokocha C, Arora N. From innovation to implementation: addressing the AI adoption gap in cardiac surgery. *Am Heart J Plus* 2026 Apr;64:100746 [FREE Full text] [doi: [10.1016/j.ahjo.2026.100746](https://doi.org/10.1016/j.ahjo.2026.100746)] [Medline: [41868433](https://pubmed.ncbi.nlm.nih.gov/41868433/)]
44. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *Lancet Digit Health* 2022 Apr;4(4):e235-e244. [doi: [10.1016/s2589-7500\(22\)00017-6](https://doi.org/10.1016/s2589-7500(22)00017-6)]
45. Gommers J, Hernström V, Josefsson V, Sartor H, Schmidt D, Hjelmgren A, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *Lancet* 2026 Jan;407(10527):505-514. [doi: [10.1016/s0140-6736\(25\)02464-x](https://doi.org/10.1016/s0140-6736(25)02464-x)]
46. Boussina A, Shashikumar SP, Malhotra A, Owens RL, El-Kareh R, Longhurst CA, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. *NPJ Digit Med* 2024 Jan 23;7(1):14 [FREE Full text] [doi: [10.1038/s41746-023-00986-6](https://doi.org/10.1038/s41746-023-00986-6)] [Medline: [38263386](https://pubmed.ncbi.nlm.nih.gov/38263386/)]
47. Dabek PA, Bourgeois FT, Stern AD, Zhu Y, Everhart AO. Use and public reporting of predetermined change control plans for AI-enabled medical devices. *JAMA Health Forum* 2026 May 01;7(5):e261204 [FREE Full text] [doi: [10.1001/jamahealthforum.2026.1204](https://doi.org/10.1001/jamahealthforum.2026.1204)] [Medline: [42172007](https://pubmed.ncbi.nlm.nih.gov/42172007/)]
48. Muralidharan V, Adewale BA, Huang CJ, Nta MT, Ademiju PO, Pathmarajah P, et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *NPJ Digit Med* 2024 Oct 03;7(1):273 [FREE Full text] [doi: [10.1038/s41746-024-01270-x](https://doi.org/10.1038/s41746-024-01270-x)] [Medline: [39362934](https://pubmed.ncbi.nlm.nih.gov/39362934/)]
49. Abulibdeh R, Cajas Ordóñez SA, Celi LA, Gorijavolu R, Izath N, Markussen Lunde T. 1,357 AI medical devices cleared, 3 actually tested on patient outcomes. *PLOS Digit Health* 2026 Aug 19;5(8):e0001597 [FREE Full text] [doi: [10.1371/journal.pdig.0001597](https://doi.org/10.1371/journal.pdig.0001597)] [Medline: [42616721](https://pubmed.ncbi.nlm.nih.gov/42616721/)]
50. Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. *BMJ* 2017 Aug 02;358:j3453 [FREE Full text] [doi: [10.1136/bmj.j3453](https://doi.org/10.1136/bmj.j3453)] [Medline: [28768629](https://pubmed.ncbi.nlm.nih.gov/28768629/)]
51. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010 Feb 16;7(2):e1000217 [FREE Full text] [doi: [10.1371/journal.pmed.1000217](https://doi.org/10.1371/journal.pmed.1000217)] [Medline: [20169112](https://pubmed.ncbi.nlm.nih.gov/20169112/)]

Abbreviations

CFIR: Consolidated Framework for Implementation Research

COMPOSER: Conformal Multidimensional Prediction of Sepsis Risk

CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence

DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence

FDA: Food and Drug Administration

FUTURE-AI: Fairness, Universality, Traceability, Usability, Robustness, and Explainability for Artificial Intelligence

MASAI: Mammography Screening with Artificial Intelligence

NASSS: Nonadoption, Abandonment, Scale-up, Spread, and Sustainability

NIST: National Institute of Standards and Technology

PROBAST: Prediction Model Risk of Bias Assessment Tool

RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance

SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence

TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence

VALUE-AI: Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring

WHO: World Health Organization

Edited by S Brini; submitted 09.Jun.2026; peer-reviewed by M Hoffman, M Honarmand, MB Patel, K Luz; comments to author 28.Aug.2026; revised version received 05.Oct.2026; accepted 06.Oct.2026; published 16.Oct.2026.

Please cite as:

Ngwu DC

Algorithmic Research Waste in Health Care AI: The VALUE-AI (Valid Question and Purpose, Appropriate Data and Population Fit, Local Workflow and Intended Use, Usability, Safety, and Accountability, Equity and External Validity, and Auditability and Impact Monitoring) Lens for Decision-Useful Evidence

J Meta and Res Int 2026;1:e104222

URL: <https://metascience.jmir.org/2026/1/e104222>

doi:[10.2196/104222](https://doi.org/10.2196/104222)

PMID:

©Dabeluchi Chiedozi Ngwu. Originally published in the JMIR Metascience and Research Integrity (<https://metascience.jmir.org>), 16.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Metascience and Research Integrity, is properly cited. The complete bibliographic information, a link to the original publication on <https://metascience.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

From AI Disclosure to Judgment Traceability: Addressing Research Waste in AI-Assisted Health Sciences Publishing

Tino Hsieh^{1,2}, MSN; Ya-Ting Chen², BSN

¹School of Nursing, National Yang Ming Chiao Tung University, Taipei, Taiwan

²Department of Emergency Medicine, National Taiwan University Hospital, Taipei, Taiwan

Corresponding Author:

Tino Hsieh, MSN

School of Nursing

National Yang Ming Chiao Tung University

No 155, Section 2, Linong Street

Taipei, 112

Taiwan

Phone: 886 966 737 862

Email: t22671884@gmail.com

Abstract

Generative artificial intelligence (GAI) is changing how health sciences manuscripts are produced, revised, and evaluated. Publication guidance emphasizes necessary safeguards, such as disclosure, authorship, confidentiality, transparency, verification, and human responsibility, for AI-assisted content. However, disclosure alone does not establish whether consequential claims in a manuscript are grounded in accountable authorial and domain judgment. This viewpoint addresses this governance gap. We define expression-rich but judgment-thin scholarship as AI-amplified scholarly output in which linguistic fluency, structural coherence, and surface plausibility exceed the traceability of the reasoning, evidence interpretation, uncertainty, contextual limits, and responsibility behind major claims. We frame AI-induced research waste as a conceptual extension of the established research waste literature rather than as a measured prevalence claim. The concern is not that AI-assisted writing is inherently low quality; rather, GAI can make it easier to produce large volumes of polished text whose underlying judgments may be weakly supported, insufficiently limited, or difficult to defend. Building on emerging work on claim accountability, provenance, responsible AI-assisted writing, and publication ethics, we propose domain judgment traceability as a claim-centered governance approach for health sciences publishing. The framework asks whether major clinical, methodological, statistical, ethical, policy, and implementation claims can be traced to an appropriate evidence base, an identifiable interpretive step, explicit uncertainty, population and setting limits, equity and harm considerations, proportionate recommendation strength, and accountable human authorship. This approach shifts the unit of governance from the AI tool or prompt to the major scholarly claim. It also distinguishes disclosure from defensibility: disclosure describes whether and how AI contributed to manuscript development, whereas judgment traceability examines whether the resulting claim can be reconstructed, challenged, and defended by responsible authors. We propose practical implications for authors, reviewers, editors, and journals, including claim-level verification, proportionate requests for clarification, and selective piloting of traceability prompts rather than routine AI detection or exhaustive prompt surveillance. We also identify safeguards against gatekeeping, including attention to language equity, reviewer subjectivity, and the risk of conflating linguistic polish with scholarly competence. Domain judgment traceability is presented as a provisional conceptual governance aid, not a validated scoring instrument or reporting guideline. Its purpose is to complement—not replace—existing disclosure policies and reporting standards by making accountable judgment more visible where claims may influence clinical care, research interpretation, policy, implementation, or public trust. The framework creates an empirical agenda: future studies can test whether claim-centered prompts improve reviewer agreement, citation accuracy, proportionality between evidence and recommendations, author experience, and editorial usefulness without imposing an excessive documentation burden. The goal is not to reject AI-assisted scholarship or require disclosure of every prompt but to ensure that increasingly scalable scholarly expression is proportionate, defensible, context sensitive, and answerable to human judgment in health sciences publishing.

(*J Meta and Res Int* 2026;1:e106068) doi:[10.2196/106068](https://doi.org/10.2196/106068)

KEYWORDS

generative artificial intelligence; artificial intelligence; AI; health sciences publishing; research integrity; metascience; scientific publishing; research waste; domain judgment; publication ethics; AI disclosure; scholarly communication

Background

Generative AI (GAI) has rapidly entered scholarly writing, with large-scale lexical analysis suggesting detectable traces of large language model (LLM)–assisted writing in biomedical abstracts [1]. In health sciences publishing, AI-assisted tools may be used to organize manuscripts, improve clarity, support literature navigation, identify possible omissions, and reduce language-related barriers for authors who work in an English-dominant publication environment [2-4]. These uses are not inherently problematic; peer-reviewed commentary has emphasized both potential productivity and accessibility benefits of GAI and the need for transparent, verifiable human oversight [3-5]. The research integrity concern is therefore more specific than AI use itself: fluent and well-structured AI-assisted prose may make a manuscript appear complete without showing whether its major claims are accurate, proportionate to the evidence, contextually limited, and defensible by responsible authors [6-9].

Current publishing guidance has therefore focused on disclosure, authorship, confidentiality, transparency, verification, and human responsibility. Peer-reviewed guidance and policy analyses consistently state that AI systems cannot assume authorship responsibility and that human authors remain accountable for the accuracy, integrity, and transparency of the final work [3-5,9-14]. Policy analyses also show substantial heterogeneity across journals and publishers in allowable uses, disclosure expectations, and required descriptions of human oversight [3,13,14]. This heterogeneity matters because a disclosure statement may document the presence of AI without revealing whether an author has verified the evidence, reconstructed the reasoning behind a consequential claim, or calibrated the strength of a recommendation. The emerging literature therefore increasingly distinguishes procedural transparency from substantive accountability [4,7-9,15].

Building on emerging arguments that responsible AI governance should move beyond tool disclosure toward accountability for scholarly claims [7], this viewpoint asks what claim accountability requires in health sciences publishing. It defines expression-rich but judgment-thin scholarship, distinguishes it from adjacent research integrity problems, and proposes domain judgment traceability as a claim-centered governance approach for AI-assisted health sciences manuscripts. The intended audience includes journal editors and reviewers, health sciences authors, publication ethics committees, research integrity officers, and institutions developing AI use policies. Although the argument is especially relevant to clinically oriented manuscripts, it applies more broadly to public health, health informatics, health services research, implementation science, epidemiology, policy research, qualitative inquiry, and biomedical scholarship. Clinical judgment traceability is therefore treated as one important subtype within the broader concept of domain judgment traceability. This formulation is consistent with recent peer-reviewed arguments that responsible

governance should preserve provenance, verification, and identifiable human responsibility rather than treating disclosure or prohibition as sufficient ends in themselves [4,7-9].

This viewpoint does not claim that AI-assisted writing is inherently defective or that disclosure policies are unimportant. Instead, it identifies a narrower governance gap. Disclosure documents whether and how a tool contributed to manuscript preparation; judgment traceability asks whether the resulting claims are sufficiently supported, limited, and accountable. This distinction allows the manuscript to engage AI-related risks while preserving a balanced position toward legitimate AI-assisted language, organization, and revision support [3-5,10]. Its primary contribution is to operationalize emerging claim accountability arguments for health sciences publishing through domain judgment traceability: a way of asking whether major scholarly claims remain connected to evidence, interpretation, uncertainty, context, and accountable authorship [7,8]. The framing deliberately avoids the stronger assertion that judgment traceability has already been empirically shown to reduce research waste. Instead, it treats judgment traceability as a provisional governance aid whose usefulness, burden, fairness, and effects on editorial decision-making require empirical evaluation [4,7,8].

Ethics Approval and Informed Consent

Ethics approval and informed consent were not applicable because this viewpoint did not involve human participants, personal data, or identifiable clinical information.

Why Disclosure Is Necessary but Insufficient

Disclosure is a necessary part of responsible AI-assisted publishing because it informs editors, reviewers, and readers that a nonhuman system contributed to manuscript development. Peer-reviewed guidance increasingly recommends disclosures that specify not only the tool but also the task, extent of assistance, verification process, and location of AI-supported work [3,10-14]. Frameworks such as GAIDeT (Generative Artificial Intelligence Delegation Taxonomy) further provide structured language for documenting which research or writing tasks were delegated to GAI while preserving human oversight [16]. These approaches strengthen transparency because “AI was used” can otherwise conceal materially different practices, ranging from grammar correction to literature summarization, conceptual brainstorming, data analysis, or recommendation drafting. However, disclosure is a description of process rather than a direct test of claim quality. A manuscript may transparently disclose AI use and still contain overextended interpretations, unsupported clinical implications, fabricated or erroneous citations, or recommendations whose strength exceeds the underlying evidence [17-19]. Conversely, AI may be used for language editing, organization, or perspective checking,

while authors preserve rigorous source interpretation, contextual limits, and responsibility for every substantive claim [3-5].

This focus on claim quality addresses a practical weakness of tool-centered governance. A disclosure statement can be accurate yet uninformative about whether a recommendation is justified by the cited evidence. For example, an author may disclose AI-assisted drafting or summarization but still fail to show why a limited observational finding warrants a broad clinical implication; conversely, an author may use AI to improve readability while maintaining rigorous source interpretation and cautious recommendations [4,7-9,15]. A citation should therefore not merely appear relevant: the cited source should directly support the claim attributed to it, and authors should be able to explain how the evidence justifies the level of inference or recommendation being made. This distinction is particularly important because disclosure practices remain heterogeneous across journals, disciplines, and publishers [3,13,14]. Some policies request only the name of a tool, whereas others ask authors to describe the task, location of use, or degree of human verification [3,10,13,14]. Disclosure can support transparency, but it remains procedural unless connected to the claims the manuscript asks readers to accept. A claim-centered standard therefore complements rather than competes with disclosure by asking whether consequential interpretations and recommendations remain attributable to responsible human judgment [7,8].

Expression-Rich but Judgment-Thin Scholarship

Expression-rich but judgment-thin scholarship is defined as scholarly work in which the language, structure, and apparent completeness of a manuscript exceed the traceability of the judgment supporting its major claims. The concept does not refer simply to poor writing, weak scholarship, or the use of GAI. Rather, it identifies a mismatch between scholarly expression and accountable reasoning: a manuscript may read as fluent, balanced, clinically relevant, and well organized while leaving unclear how its central claims were framed, interpreted, limited, and made defensible by responsible authors. This concern is compatible with broader warnings that GAI can create persuasive or seemingly authoritative outputs without reliable epistemic grounding and that apparent fluency should not be mistaken for understanding [4,6,20]. It also aligns with recent claim accountability and provenance arguments that the central governance question is whether a human can reconstruct and defend the reasoning, evidence, and verification behind what enters the scholarly record [7,8].

Judgment thinness is best understood as a property of the relationship among text, authorial process, and claims. A manuscript can be judgment thin without AI because conventional academic writing can also be formulaic, overgeneralized, densely cited but poorly interpreted, or disconnected from clinical and methodological realities. A manuscript can also be AI assisted and judgment rich if authors use AI to improve expression, identify missing perspectives, challenge assumptions, or organize reasoning while retaining responsibility for the final claims [3-5,9]. AI may amplify

judgment thinness, not because it uniquely causes weak reasoning, but because it can increase the speed and scale at which plausible scholarly text is produced and can make weakly examined reasoning appear polished enough to pass superficial review. The risk is therefore not the presence of machine-generated language itself; it is the possibility that fluency masks unresolved evidentiary, methodological, or contextual problems that responsible authors have not adequately examined [6-8,19].

This concept overlaps with, but is not reducible to, several existing concerns. Hallucination concerns fabricated or inaccurate content, paper mill production concerns fraudulent or industrialized paper generation, automation bias concerns overreliance on machine output [21,22], and epistemic opacity concerns difficulty understanding how algorithmic systems produce outputs [23]. Paper mills represent a documented threat to research integrity [24,25], and work on AI in science has warned that AI systems may produce illusions of understanding [6]. Concerns about fluent but ungrounded language model output, including the “stochastic parrots” critique, are relevant background but do not by themselves define the claim-level failure mode addressed here [20]. Judgment thinness specifically concerns whether the reasoning behind major claims is visible, proportionate, and defensible.

We propose several observable warning signs that may prompt closer review: broad clinical or policy implications that exceed the evidence cited; generic recommendations that do not specify population, setting, resource, or equity limits; literature summaries that list findings without interpreting their strength or relevance; conclusions that convert limited evidence into action-oriented recommendations without explaining inferential steps; citations that exist but do not support the sentence in which they are placed; and statements that appear authoritative but cannot be defended by a responsible author when challenged [7,8,17-19]. These features should prompt clarification of claim-to-evidence reasoning, not accusations of AI use. They are intentionally framed as review prompts rather than diagnostic criteria because AI detection tools have documented limitations and may produce inequitable misclassification, including for nonnative English writers [2,26]. Nor should the prompts penalize concise writing. Their purpose is to identify where a manuscript may require clearer explanation of evidence strength, inferential steps, contextual limits, or authorial responsibility. In this sense, judgment traceability functions as a quality and accountability standard rather than an AI detection strategy [7,8].

From Research Waste to AI-Induced Research Waste

Research waste is commonly discussed in relation to the production, conduct, reporting, dissemination, and usability of empirical research [27-29]. In health sciences, this framing has ethical as well as economic significance because low-value research can consume participant contributions, public funding, institutional resources, reviewer labor, editorial attention, and reader trust. Contemporary research integrity literature also documents how fraudulent publication systems and paper mills

can introduce large volumes of unreliable work into the scholarly record and create downstream burdens for journals and evidence synthesis [24,25,30]. This literature provides the conceptual foundation for asking whether AI-assisted scholarly production may create an additional pathway to low-value output while also requiring caution: AI-induced research waste is proposed here as a conceptual extension, not as a measured prevalence claim.

Empirical evidence supports the plausibility of this concern without proving its prevalence. Large-scale lexical analysis suggests detectable traces of LLM-assisted writing in biomedical abstracts [1]. Separate studies demonstrate risks, including fraudulent-looking AI-generated medical manuscripts, fabricated or erroneous AI-generated citations, and citation errors that can appear bibliographically plausible [17-19]. These findings do not show that AI-assisted manuscripts are inherently lower quality or judgment thin. They do show why authors, reviewers, and editors cannot rely on surface fluency, citation density, or procedural disclosure alone when evaluating whether a manuscript's claims add proportionate knowledge value. The problem is compounded when unreliable material enters later scholarship: empirical work on paper mill contamination has shown that retracted fraudulent papers can be incorporated into systematic reviews and continue to be cited even after retraction [30]. Although paper mills and AI-assisted writing are distinct phenomena, both illustrate why scholarly systems need safeguards that evaluate the integrity of the claim-source relationship rather than only the appearance of a completed manuscript [24,25,30].

On this basis, AI-induced research waste should be understood as a possible pathway by which AI-assisted scholarly outputs consume attention without adding proportionate epistemic, clinical, methodological, or social value. The concern includes obvious problems, such as fabricated references and unreliable summaries, but it also includes a subtler failure: manuscripts that are fluent and rhetorically complete yet leave reviewers and readers to reconstruct missing judgment behind their major claims [6-8]. This formulation is intended to prevent overclaiming. The available empirical evidence supports component risks (detectable AI-assisted writing, fabricated or erroneous citations, fraudulent-looking text, detector bias, automation-related overreliance, and broader threats to provenance), but it does not measure the prevalence of judgment-thin scholarship as an empirical category [1,2,6,8,17-19,21,22,26,31]. The proposed contribution is therefore conceptual and governance-oriented: it names a failure mode that may become more scalable under AI-assisted writing conditions and specifies what would need to be visible for major claims to remain accountable. This emphasis also avoids equating productivity with value; the ethical concern is not that AI makes writing faster but that increased output can become wasteful when the additional text does not carry commensurate evidentiary or interpretive value [9,27-29].

From Clinical Judgment to Domain Judgment Traceability

Because health sciences publishing includes diverse forms of inquiry, clinical judgment traceability should not be treated as the only relevant standard. Evidence-based practice has long emphasized that clinical decisions require integration of the best-available evidence, professional expertise, and patient values [32]. Yet, manuscripts in public health, epidemiology, implementation science, health informatics, qualitative inquiry, health services research, policy research, and biomedical science may not make direct clinical recommendations while still depending on domain-specific judgment. Recent discussions of claim accountability and provenance similarly emphasize that the defensibility of scholarly claims depends on identifiable human responsibility for evidence selection, interpretation, and verification regardless of discipline [7,8]. For this reason, this viewpoint uses domain judgment traceability as the broader governance concept.

Domain judgment traceability refers to the ability of readers, reviewers, and editors to identify how responsible authors interpreted evidence, handled uncertainty, considered context, and justified the level of claim being made. Clinical judgment traceability is one subtype that becomes especially important when a manuscript makes claims about patient care, professional practice, clinical workflow, health system operations, or policy decisions with foreseeable implications for care. Other forms include methodological, statistical, ethical, policy, and implementation judgment. Judgment types describe the substantive domains in which authors exercise judgment, whereas the traceability domains in Table 1 describe what must be made visible for a major claim to remain accountable. Where prior claim accountability work establishes that a named human must be able to reconstruct and defend each scholarly claim [7], domain judgment traceability specifies what should be visible for that defense in health sciences manuscripts: the evidence base, author interpretation, uncertainty, population and setting context, equity and harms, recommendation strength, and accountable authorship. Its additional contribution is to connect those elements to health sciences-specific concerns about research waste; transferability; patient or population consequences; and the proportionality of clinical, policy, and implementation recommendations. This approach extends rather than replaces existing publication ethics expectations and reporting standards [4,5,7-14]. The CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) extension and DECIDE-AI (Developmental and Exploratory Clinical Investigation of Decision Support Systems Driven by Artificial Intelligence), for example, emphasize intended use, clinical setting, human-AI interaction, input and output handling, implementation context, and safety evaluation [33,34]. Judgment traceability is not proposed as a replacement reporting guideline; rather, these reporting concerns illustrate why claims about AI-supported care, workflow, risk prediction, or implementation should make the evidence base, setting, uncertainty, harms, and level of action justified visible.

Table 1. Proposed claim-centered framework for domain judgment traceability in AI-assisted health sciences manuscripts.

Traceability domain	What authors should make visible	Potential failure mode	Example of adequate traceability
Evidence base	Studies, guidance, data, or conceptual sources supporting the claim	Sources are cited but not interpreted; weak evidence supports strong claims.	The recommendation is based on two observational studies and should be interpreted cautiously.
Author interpretation	How responsible authors moved from evidence to meaning and why the claim follows	Findings are listed or paraphrased without an identifiable inferential step.	We interpret these findings as relevant to workflow redesign because the observed effect concerns response timing while recognizing that alternative explanations remain possible.
Uncertainty	What remains uncertain, including limitations, competing explanations, and confidence in the inference	The manuscript presents a qualified finding as settled or omits material uncertainty.	The direction of effect is consistent, but uncertainty remains because evidence is heterogeneous and external validation is limited.
Population, setting, and context	Who the claim applies to, where it applies, and under what resource or workflow conditions	A narrow finding is generalized to all patients, settings, or systems.	These implications apply primarily to adult EDs ^a with integrated electronic health record alert systems and defined response protocols.
Equity and harms	Potential differential effects, exclusions, burdens, or unintended consequences	The manuscript assumes equal benefit across groups or ignores possible harms.	Performance should be assessed across relevant demographic, clinical, language, and access-related subgroups, when appropriate.
Recommendation strength	Whether the recommended action is proportionate to the strength, consistency, and applicability of the evidence	A tentative implication is written as a universal implementation mandate.	Current evidence supports further evaluation in appropriately resourced settings, not immediate universal adoption.
Accountable authorship and AI assistance	Which authors are responsible for interpretation and recommendation; whether AI materially supported major claims; and how relevant outputs were checked	Responsibility is diffused into passive wording or AI-generated generality.	AI assisted drafting of this section; the authors checked the cited sources, revised the inference, and accept responsibility for the final recommendation.

^aED: emergency department.

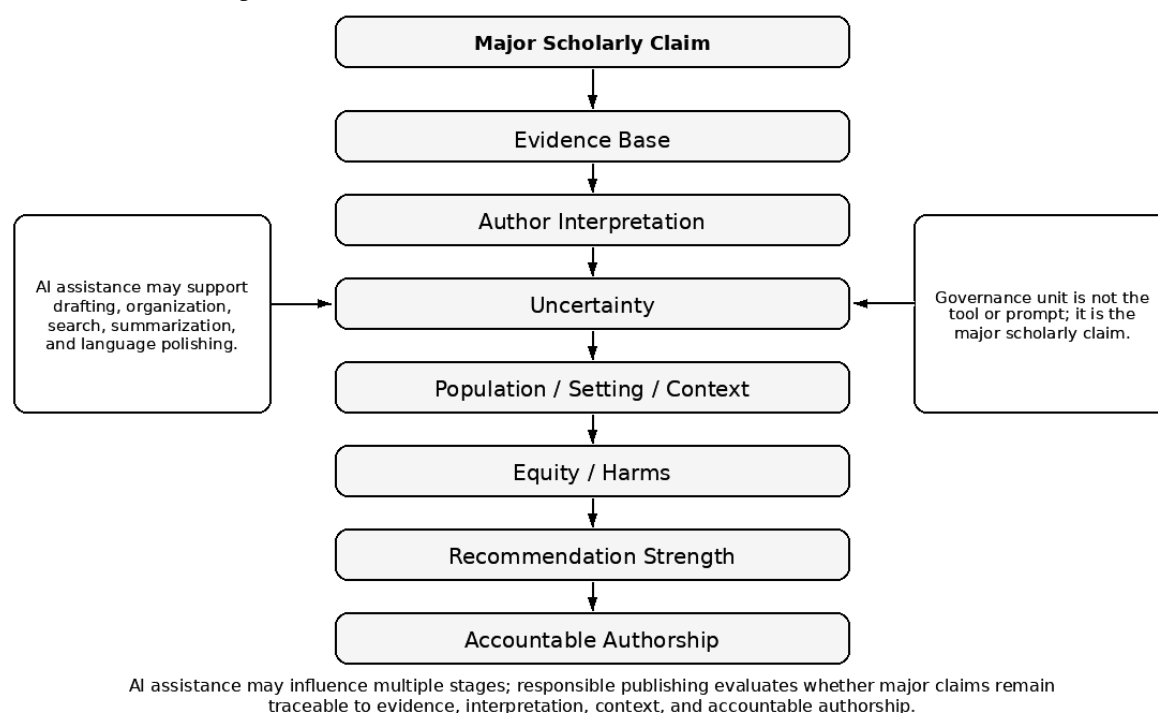
Different forms of domain judgment require different forms of traceability. Clinical judgment concerns whether claims about care, risk, diagnosis, treatment, triage, monitoring, or professional action are appropriate for the relevant population and setting. Methodological judgment concerns whether the study design, sampling, measurement, comparison, and interpretation are aligned with the question being asked. Statistical judgment concerns whether analytic choices, uncertainty, effect estimates, assumptions, and limitations are interpreted proportionately. Ethical judgment concerns whether benefits, burdens, risks, consent, privacy, equity, and potential harms are adequately considered. For the purposes of the framework proposed here, policy and implementation claims should additionally make feasibility, stakeholders, resources, workflows, workforce capacity, and unintended consequences visible rather than presenting these considerations as if they were automatically established by a generic AI reporting standard. The framework is therefore not restricted to bedside clinical decisions: A statistical claim may require traceability of analytic assumptions and uncertainty; an implementation claim may require traceability of staffing, infrastructure, and local feasibility; and a policy claim may require traceability of stakeholders, unintended consequences, and the level of action justified. These distinctions matter because health sciences

manuscripts often combine empirical evidence, interpretation, and recommendations within the same paper, and the strength of each layer should remain visible rather than being compressed into a single authoritative sentence [4,7,8].

What Should Be Traceable?

Judgment traceability should be claim centered. It should apply primarily to major interpretive, clinical, methodological, ethical, policy, and implementation claims rather than to every sentence. A major claim is one that asks readers to accept an interpretation, change a practice, adopt a policy, cite a conclusion, prioritize an intervention, or treat a proposed implication, as warranted. Stronger claims require stronger traceability. Figure 1 therefore places the major scholarly claim, rather than the AI tool or prompt, at the center of governance. AI assistance may influence multiple stages of scholarly work, but the unit of governance is the claim and whether it remains traceable to an evidence base, author interpretation, uncertainty, population and setting context, equity and harms, recommendation strength, and accountable authorship [7,8]. This claim-centered orientation is also consistent with peer-reviewed calls to prioritize provenance and verification over attempts to govern GAI through prohibition alone [8].

Figure 1. Claim-centered judgment traceability in AI-assisted health sciences publishing. The figure presents a conceptual framework, not an empirically validated model. AI: artificial intelligence.



The practical value of a claim-centered approach is proportionality. A minor wording edit does not require extensive documentation, whereas a claim that implies a change in clinical practice, health policy, professional responsibility, or implementation priority should make its evidentiary basis and inferential pathway visible [4,7,8]. Before applying the traceability domains, authors and reviewers should first classify the claim as an evidence summary, inference, implication, recommendation, or speculation because the required level of support depends on what the statement asks readers to accept. Table 1 then translates judgment traceability into seven domains that correspond to Figure 1: evidence base, author interpretation, uncertainty, population and setting context, equity and harms, recommendation strength, and accountable authorship. These domains are not intended to create a scoring instrument; they provide a vocabulary for asking what must be visible for a major claim to be defensible. Existing reporting guidelines, such as CONSORT-AI and DECIDE-AI, demonstrate the broader value of making the intended use, setting, human-AI interaction, safety, and implementation context explicit, even though judgment traceability is not proposed as a replacement reporting guideline [33,34].

The framework also clarifies the difference between empirical support and conceptual contribution. Empirical studies can support component concerns, such as fabricated citations, detector bias, language barriers, publication policy heterogeneity, paper mill contamination, and automation bias [1-3,13,14,17-19,21,22,24-26,30,31]. They cannot, by themselves, prove the full construct of expression-rich but judgment-thin scholarship, because that construct is proposed here as an analytic lens. Accordingly, this viewpoint uses empirical evidence to support the plausibility and relevance of the problem while using explicit proposal language when advancing judgment traceability as a governance response [7,8].

This separation is important for metascience: a useful conceptual framework should generate testable questions without presenting its proposed constructs as though their prevalence, reliability, or effectiveness had already been established.

Table 1 summarizes the seven traceability domains shown in Figure 1. It is not a validated assessment instrument. Rather, it is a conceptual governance aid intended to help authors, reviewers, and editors identify what should be made visible for consequential scholarly claims to remain defensible.

When AI Contributes an Idea: a Worked Example

A further question is whether a recommendation initially suggested by an AI tool can count as being supported by human judgment if the authors had not independently conceived it before the AI output was produced. The answer depends not on the origin of the idea but on the authorial work that follows. Accountability rests on whether responsible authors can reconstruct the reasoning, evaluate the evidence, consider alternative interpretations, identify limits, and defend the final claim as their own scholarly judgment [7-9]. This position is consistent with publication ethics arguments that AI systems cannot bear authorship responsibility and with peer-reviewed guidance that verification and human responsibility remain necessary even when GAI materially assists writing or idea development [4,5,9-12,14]. The practical distinction is between assistance that stimulates accountable scholarly work and delegation that leaves no responsible human able to defend the resulting claim.

This issue is also relevant to honest AI disclosure. If authors believe that acknowledging AI use will automatically trigger disproportionate procedural burdens, they may be less willing

to disclose assistance; conversely, generic disclosures may become performative if they are treated as sufficient evidence of responsible use [3,4,7,10]. A claim-centered approach reduces that risk by focusing scrutiny on the consequences of the claim rather than the mere presence of AI support. The same standard can therefore be applied to AI-assisted and non-AI-assisted manuscripts: are the major claims proportionate, source grounded, context sensitive, and attributable to accountable authors? This preserves the transparency value of disclosure while avoiding a governance model in which every prompt or stylistic edit receives the same scrutiny as a recommendation that could influence care, policy, or research priorities [7,8].

The worked example here illustrates this distinction in a clinical context, but the same logic can be adapted to other health sciences claims. We propose that a qualitative manuscript should make visible how interpretations were derived from data and context; an informatics manuscript should clarify intended users, workflow assumptions, model limitations, and potential harms; and a policy viewpoint should distinguish evidence supporting a proposed action from normative or speculative elements. These are applications of the proposed traceability framework rather than claims that one existing reporting guideline already covers all paper types [7,8]. Consider a hypothetical AI-assisted manuscript discussing an emergency department (ED) sepsis alert intervention. A judgment-thin implication might state, “AI-enabled sepsis alerts should be implemented across EDs to improve early recognition, accelerate treatment, and enhance patient safety.” This sentence is fluent and clinically plausible, but it does not show whether the supporting evidence came from adult or pediatric populations, academic or community hospitals, high-resource or underresourced settings, or prospective trials or early-stage evaluations or whether the alert system’s positive predictive value, false-positive burden, workflow effects, and performance across patient subgroups were adequately assessed. Empirical sepsis alert literature illustrates why these details matter: ED alert systems vary substantially across studies, and recent evidence shows both potential clinical benefits and persistent concerns about the alert burden, generalizability, and sociodemographic performance [35-37].

A judgment-traceable version would make the claim more proportionate: “Available evidence may support further evaluation of AI-enabled sepsis alerts in EDs with sufficient informatics infrastructure; clinician response protocols; and monitoring for false positives, alert burdens, workflow effects, and subgroup performance. The current evidence should not be interpreted as a universal implementation mandate. Before adoption, journals, institutions, and clinical leaders should examine whether the evaluated population, setting, workflow, AI system version, input data, output interpretation, and outcome measures correspond to the intended local use case.” The difference is not merely tone. The judgment-traceable version distinguishes evidence from inference and recommendation, identifies contextual limits, makes uncertainty visible, and preserves accountability for the level of action being proposed [7,8,33-37]. It also demonstrates why judgment traceability is not reducible to adding more citations: the same evidence base

can support very different claims, depending on how authors interpret scope, uncertainty, harms, and transferability.

Implications for Authors, Reviewers, Editors, and Journals

For authors, judgment traceability means identifying major claims and ensuring that each is supported by an appropriate evidence base, an explicit interpretive step, and a proportionate level of recommendation. Authors should distinguish evidence summaries from inferences, implications, recommendations, and speculation and should verify that cited sources actually support the statements attributed to them [7,8,18,19]. Transparent AI-assisted writing can support this process when authors document human intellectual contribution and verification without turning governance into indiscriminate prompt-level surveillance [3,4,9,10,14-16]. The goal is not to record every prompt but to ensure that claims likely to influence interpretation, practice, policy, or implementation can be reconstructed and defended by responsible authors. A useful author-facing question is therefore not “Did AI write this sentence?” but “If challenged, can we show why this claim follows from the cited evidence, why its limits are appropriate, and who accepts responsibility for it?” [7,8].

For reviewers and editors, judgment traceability should function as an interpretive aid rather than a punitive label. Reviewers should not be asked to determine whether text is AI generated, particularly because detection tools may be unreliable and may disadvantage nonnative English writers [2,26]. Instead, they can examine whether cited evidence directly supports major claims, whether population and setting limits are specified, whether competing interpretations are acknowledged, and whether recommendation strength is proportionate to the evidence [7,8]. Editors can use this approach selectively when claims appear clinically, methodologically, ethically, or policy relevant but insufficiently supported. Journals could pilot structured traceability prompts before treating them as settled standards and evaluate outcomes, such as author burden, reviewer agreement, editorial usefulness, citation accuracy, correction requests, and perceived fairness. The staged approach also clarifies what journals should not routinely require: complete prompt histories, private drafting records, or speculative reconstructions of every AI interaction may be burdensome, difficult to standardize, and poorly aligned with the integrity risk [3,4,8,10]. A more feasible approach is to request clarification when a consequential claim lacks adequate evidence, context, uncertainty, or accountability, preserving the integrity function of peer review while avoiding a governance model that rewards paperwork over reasoning.

For journals and publishers, a practical starting point is not to require a new universal form but to identify where current editorial workflows already ask for evidence, limitation, and accountability. Traceability prompts could be integrated selectively into reviewer forms, revision letters, or editorial checks for manuscripts containing high-consequence clinical, policy, or implementation claims. For example, reviewers could be asked whether a major recommendation is directly supported by the cited evidence, whether the target population and setting

are explicit, and whether uncertainty or foreseeable harms have been acknowledged. Such prompts would operationalize claim accountability without requiring editors to infer how much text was generated by AI or to police every stage of drafting [7,8]. Existing disclosure taxonomies and peer-reviewed guidance can continue to document the role of AI in manuscript production, while judgment traceability addresses the separate question of whether the resulting claims are defensible [3,4,10,14,16]. Importantly, implementation should be evaluated rather than assumed beneficial. Journals would need to monitor whether additional prompts improve citation accuracy, reviewer consistency, and recommendation proportionality and whether they create new burdens, delays, or inequities for authors and reviewers. A governance intervention that increases paperwork without improving claim quality would itself risk becoming a form of editorial waste. The proposed framework therefore favors targeted clarification of consequential claims over blanket expansion of submission requirements, and it treats feasibility, fairness, and workload as empirical outcomes for future testing.

Equity, Subjectivity, and Gatekeeping

Any framework that asks reviewers to evaluate judgment carries risks of subjectivity. Reviewers may mistake familiar disciplinary language, institutional prestige, seniority, native English fluency, or agreement with their own clinical preferences for stronger judgment. Evidence from peer-reviewed research suggests that author and affiliation information can influence reviewer behavior in some settings [31]. Research on AI text detection raises a separate equity concern: GPT detectors have misclassified nonnative English writing as AI generated in some evaluation settings [26], while nonnative English-speaking researchers face measurable burdens in reading, writing, publishing, and presenting science [2]. Domain judgment traceability should therefore be applied through observable claim features rather than impressions of author competence. A judgment-rich manuscript does not need ornate prose or extensive rhetorical display; it needs alignment among claim, evidence, inference, uncertainty, context, and accountability [7,8]. Authors should have opportunities to clarify allegedly insufficiently traceable claims, and reviewer concerns should be framed as requests for claim evidence clarification rather than accusations about AI use or author capability. These safeguards are essential if claim-centered governance is to improve accountability without reproducing linguistic or institutional gatekeeping [2,3,26,31].

Boundaries of Judgment Traceability and Future Research Priorities

Judgment traceability should be understood as a proposed conceptual governance aid rather than a validated assessment instrument. Its domains and examples require refinement through stakeholder consultation, piloting, and empirical evaluation. Future work should examine whether claim-centered traceability improves reviewer agreement, author experience, editorial usefulness, citation accuracy, correction rates, and proportionality between evidence and recommendations while also measuring burden and unintended consequences [7,8].

Validation should proceed in stages: Qualitative work can first examine how authors, reviewers, and editors understand traceability; consensus methods can then refine candidate domains; pilot studies can assess interviewer agreement and feasibility; and comparative evaluations can test whether traceability prompts improve claim-source alignment or recommendation proportionality without creating inequitable barriers. The emerging literature on claim accountability, provenance, and responsible AI-assisted writing is developing rapidly, so the contribution of judgment traceability should be interpreted within this evolving conversation rather than as a claim of conceptual exclusivity [4,7,8,15].

The concept should also be applied proportionately across paper types. A clinical guideline, empirical study, qualitative analysis, policy commentary, laboratory report, implementation report, and philosophical essay make different kinds of claims and require different levels of traceability. Judgment traceability should not become a universal checklist applied mechanically to all scholarly writing, nor should it be used as a proxy for seniority, institutional prestige, English fluency, or disciplinary conformity [2,26,31]. Its purpose is to make consequential scholarly claims more accountable while preserving equitable access to legitimate AI-assisted language and writing support [2-4]. A boundary-focused implementation should also distinguish private reasoning from accountable scholarly justification. Judgment traceability does not require authors to reveal private chain-of-thought, informal notes, or every intermediate draft; it requires that consequential scholarly judgments—such as why evidence was interpreted in a particular way, why a recommendation is limited to a certain population or setting, and who accepts responsibility for the final claim—be sufficiently visible for review, revision, and accountability [7-9]. This boundary is important because responsible publishing needs verifiable justification, not surveillance of cognition.

Conclusion

GAI does not make scholarship irresponsible by default; it may improve clarity, accessibility, organization, and language equity when used with transparent human verification [2-5]. The integrity problem begins when fluent scholarly expression becomes disconnected from accountable judgment. In health sciences publishing, where claims may influence care, policy, education, research priorities, and public trust, the crucial question is not only whether AI was used but also whether major claims remain traceable to evidence, interpretation, uncertainty, context, and responsible authorship [7,8]. This viewpoint proposes expression-rich but judgment-thin scholarship as a conceptual lens for recognizing that mismatch and domain judgment traceability as a claim-centered governance response. Detection asks whether text can be identified as machine generated; disclosure asks whether AI use has been reported; and judgment traceability asks whether the claims carried by the text can be reconstructed and defended by responsible authors. The framework does not replace disclosure, reporting guidelines, peer review, or research integrity standards. It adds a proportional question at the point where those mechanisms converge: does the strength of a consequential claim remain matched to its evidence, interpretation, limits, and accountable

human judgment? In an era when scholarly expression can be generated at scale, responsible publishing will increasingly depend not only on knowing how text was produced but also on whether the claims it carries are proportionate, defensible, and traceable to accountable human judgment [4,6-9,15].

Acknowledgments

The authors thank the editors and reviewers for their constructive comments, which strengthened the conceptual precision, evidentiary grounding, and practical orientation of this manuscript.

Using the GAIDeT (Generative Artificial Intelligence Delegation Taxonomy) declaration framework [16], the authors report that generative AI (GAI) tools, including ChatGPT by OpenAI, were used during manuscript development and revision in 2026 to support conceptual discussion, structural editing, manuscript revision, literature search support, summarization support, reference-checking support, language refinement, proofreading, and language polishing. These delegated tasks correspond to GAIDeT categories related to conceptualization, literature review support, writing and editing, and publication support. The authors did not use GAI to replace scholarly, methodological, clinical, or editorial judgment and did not treat GAI output as authoritative evidence.

Funding

No financial support was received for this work.

Data Availability

Data sharing is not applicable to this viewpoint because no empirical datasets were generated or analyzed.

Authors' Contributions

TH conceptualized the manuscript, developed the central argument and proposed framework, drafted the manuscript, revised the intellectual content, verified references and claim-to-source correspondence, and approved the final version for submission. YTC contributed to conceptual refinement, interpretation of the literature, critical revision of the manuscript for important intellectual content, reference verification, and approval of the final version for submission. All substantive claims, interpretations, recommendations, source selection, and final decisions were reviewed and revised by the authors. Final verification of bibliographic details, publication status, DOI information, author names, and claim-to-source correspondence was performed by the authors using publisher pages, indexing records, PubMed records, arXiv or SSRN records, or full-text sources, when available. No confidential patient information, unpublished data, or identifiable personal data were entered into GAI tools. Both authors reviewed the final manuscript, can take responsibility for the central claims and cited sources relevant to their contributions, and accept full responsibility for the final manuscript, including all claims, interpretations, references, and conclusions.

Conflicts of Interest

None declared.

References

1. Kobak D, González-Márquez R, Horvát EÁ, Lause J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Sci Adv* 2025 Jul 04;11(27):eadt3813 [FREE Full text] [doi: [10.1126/sciadv.adt3813](https://doi.org/10.1126/sciadv.adt3813)] [Medline: [40601754](https://pubmed.ncbi.nlm.nih.gov/40601754/)]
2. Amano T, Ramírez-Castañeda V, Berdejo-Espinola V, Borokini I, Chowdhury S, Golivets M, et al. The manifold costs of being a non-native English speaker in science. *PLoS Biol* 2023 Jul 18;21(7):e3002184 [FREE Full text] [doi: [10.1371/journal.pbio.3002184](https://doi.org/10.1371/journal.pbio.3002184)] [Medline: [37463136](https://pubmed.ncbi.nlm.nih.gov/37463136/)]
3. Tang A, Li K, Kwok KO, Cao L, Luong S, Tam W. The importance of transparency: declaring the use of generative artificial intelligence (AI) in academic writing. *J Nurs Scholarsh* 2024 Mar;56(2):314-318. [doi: [10.1111/jnu.12938](https://doi.org/10.1111/jnu.12938)] [Medline: [37904646](https://pubmed.ncbi.nlm.nih.gov/37904646/)]
4. Cohen JF, Moher D. Generative artificial intelligence and academic writing: friend or foe? *J Clin Epidemiol* 2025 Mar;179:111646 [FREE Full text] [doi: [10.1016/j.jclinepi.2024.111646](https://doi.org/10.1016/j.jclinepi.2024.111646)] [Medline: [39706536](https://pubmed.ncbi.nlm.nih.gov/39706536/)]
5. Bagenal J, Biamis C, Boillot M, Brierley R, Chew M, Dehnel T, et al. Generative AI: ensuring transparency and emphasising human intelligence and accountability. *Lancet* 2024 Nov 30;404(10468):2142-2143 [FREE Full text] [doi: [10.1016/S0140-6736\(24\)02615-1](https://doi.org/10.1016/S0140-6736(24)02615-1)] [Medline: [39615987](https://pubmed.ncbi.nlm.nih.gov/39615987/)]
6. Messeri L, Crockett MJ. Artificial intelligence and illusions of understanding in scientific research. *Nature* 2024 Mar 06;627(8002):49-58 [FREE Full text] [doi: [10.1038/s41586-024-07146-0](https://doi.org/10.1038/s41586-024-07146-0)] [Medline: [38448693](https://pubmed.ncbi.nlm.nih.gov/38448693/)]
7. van Zoonen W, Morgan-Thomas A, Tursunbayeva A. Beyond AI disclosure: claim accountability and responsible research in scholarly publishing. *Eur Manag J* 2026 Aug;44(4):550-556 [FREE Full text] [doi: [10.1016/j.emj.2026.06.001](https://doi.org/10.1016/j.emj.2026.06.001)]

8. Johnson RF. Provenance, not prohibition: a framework for AI in scholarly publishing. *Laryngoscope Investig Otolaryngol* 2026 Aug;11(4):e70487 [FREE Full text] [doi: [10.1002/lio2.70487](https://doi.org/10.1002/lio2.70487)] [Medline: [42382923](https://pubmed.ncbi.nlm.nih.gov/42382923/)]
9. Moffatt B, Hall A. Is AI my co-author? The ethics of using artificial intelligence in scientific publishing. *Account Res* 2025 Nov;32(8):1313-1329. [doi: [10.1080/08989621.2024.2386285](https://doi.org/10.1080/08989621.2024.2386285)] [Medline: [39109816](https://pubmed.ncbi.nlm.nih.gov/39109816/)]
10. Cleland J, Driessen E, Masters K, Lingard L, Maggio LA. When and how to disclose AI use in academic publishing: AMEE guide no.192. *Med Teach* 2025 Dec 30;48(4):542-553 [FREE Full text] [doi: [10.1080/0142159x.2025.2607513](https://doi.org/10.1080/0142159x.2025.2607513)]
11. Zielinski C, Winker MA, Aggarwal R, Ferris LE, Heinemann M, Lapeña JF, et al. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Colomb Med (Cali)* 2023 Dec 01;54(3):e1015868 [FREE Full text] [doi: [10.25100/cm.v54i3.5868](https://doi.org/10.25100/cm.v54i3.5868)] [Medline: [38089825](https://pubmed.ncbi.nlm.nih.gov/38089825/)]
12. Flanagan A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting use of AI in research and scholarly publication—JAMA Network guidance. *JAMA* 2024 Apr 02;331(13):1096-1098 [FREE Full text] [doi: [10.1001/jama.2024.3471](https://doi.org/10.1001/jama.2024.3471)] [Medline: [38451540](https://pubmed.ncbi.nlm.nih.gov/38451540/)]
13. Ganjavi C, Eppler MB, Pekcan A, Biedermann B, Abreu A, Collins GS, et al. Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *BMJ* 2024 Jan 31;384:e077192 [FREE Full text] [doi: [10.1136/bmj-2023-077192](https://doi.org/10.1136/bmj-2023-077192)] [Medline: [38296328](https://pubmed.ncbi.nlm.nih.gov/38296328/)]
14. Yoo JH. Defining the boundaries of AI use in scientific writing: a comparative review of editorial policies. *J Korean Med Sci* 2025 Jun 16;40(23):e187 [FREE Full text] [doi: [10.3346/jkms.2025.40.e187](https://doi.org/10.3346/jkms.2025.40.e187)] [Medline: [40524628](https://pubmed.ncbi.nlm.nih.gov/40524628/)]
15. Park J. Towards a transparent and reproducible AI-assisted research paper writing. *Genomics Inform* 2025 Dec 02;23(1):26 [FREE Full text] [doi: [10.1186/s44342-025-00057-0](https://doi.org/10.1186/s44342-025-00057-0)] [Medline: [41331701](https://pubmed.ncbi.nlm.nih.gov/41331701/)]
16. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. *Account Res* 2026 Apr;33(3):2544331 [FREE Full text] [doi: [10.1080/08989621.2025.2544331](https://doi.org/10.1080/08989621.2025.2544331)] [Medline: [40781729](https://pubmed.ncbi.nlm.nih.gov/40781729/)]
17. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res* 2023 May 31;25:e46924 [FREE Full text] [doi: [10.2196/46924](https://doi.org/10.2196/46924)] [Medline: [37256685](https://pubmed.ncbi.nlm.nih.gov/37256685/)]
18. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep* 2023 Sep 07;13(1):14045 [FREE Full text] [doi: [10.1038/s41598-023-41032-5](https://doi.org/10.1038/s41598-023-41032-5)] [Medline: [37679503](https://pubmed.ncbi.nlm.nih.gov/37679503/)]
19. Resnik DB, Hosseini M. Hallucinated citations produced by generative artificial intelligence may constitute research misconduct when citations function as data in scholarly papers. *Account Res* 2026 Oct 15;33(7):2645390 [FREE Full text] [doi: [10.1080/08989621.2026.2645390](https://doi.org/10.1080/08989621.2026.2645390)] [Medline: [41833014](https://pubmed.ncbi.nlm.nih.gov/41833014/)]
20. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? New York, NY: Association for Computing Machinery; 2021 Presented at: 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21); March 3-10, 2021; Canada p. 610-623. [doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)]
21. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012;19(1):121-127 [FREE Full text] [doi: [10.1136/amiainl-2011-000089](https://doi.org/10.1136/amiainl-2011-000089)] [Medline: [21685142](https://pubmed.ncbi.nlm.nih.gov/21685142/)]
22. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017 Mar 01;24(2):423-431 [FREE Full text] [doi: [10.1093/jamia/ocw105](https://doi.org/10.1093/jamia/ocw105)] [Medline: [27516495](https://pubmed.ncbi.nlm.nih.gov/27516495/)]
23. Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc* 2016 Jan 06;3(1):1-12 [FREE Full text] [doi: [10.1177/2053951715622512](https://doi.org/10.1177/2053951715622512)]
24. Parker L, Boughton S, Bero L, Byrne JA. Paper mill challenges: past, present, and future. *J Clin Epidemiol* 2024 Dec;176:111549 [FREE Full text] [doi: [10.1016/j.jclinepi.2024.111549](https://doi.org/10.1016/j.jclinepi.2024.111549)] [Medline: [39389277](https://pubmed.ncbi.nlm.nih.gov/39389277/)]
25. Candal-Pedreira C, Ross JS, Ruano-Ravina A, Egilman DS, Fernández E, Pérez-Ríos M. Retracted papers originating from paper mills: cross sectional study. *BMJ* 2022 Nov 28;379:e071517 [FREE Full text] [doi: [10.1136/bmj-2022-071517](https://doi.org/10.1136/bmj-2022-071517)] [Medline: [36442874](https://pubmed.ncbi.nlm.nih.gov/36442874/)]
26. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. *Patterns (N Y)* 2023 Jul 14;4(7):100779 [FREE Full text] [doi: [10.1016/j.patter.2023.100779](https://doi.org/10.1016/j.patter.2023.100779)] [Medline: [37521038](https://pubmed.ncbi.nlm.nih.gov/37521038/)]
27. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009 Jul 04;374(9683):86-89. [doi: [10.1016/S0140-6736\(09\)60329-9](https://doi.org/10.1016/S0140-6736(09)60329-9)] [Medline: [19525005](https://pubmed.ncbi.nlm.nih.gov/19525005/)]
28. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: increasing value, reducing waste. *Lancet* 2014 Jan 11;383(9912):101-104. [doi: [10.1016/S0140-6736\(13\)62329-6](https://doi.org/10.1016/S0140-6736(13)62329-6)] [Medline: [24411643](https://pubmed.ncbi.nlm.nih.gov/24411643/)]
29. Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PMM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? *Lancet* 2016 Apr 09;387(10027):1573-1586. [doi: [10.1016/S0140-6736\(15\)00307-4](https://doi.org/10.1016/S0140-6736(15)00307-4)] [Medline: [26423180](https://pubmed.ncbi.nlm.nih.gov/26423180/)]
30. Tang G, Cai H. Citation contamination by paper mill articles in systematic reviews of the life sciences. *JAMA Netw Open* 2025 Jun 02;8(6):e2515160 [FREE Full text] [doi: [10.1001/jamanetworkopen.2025.15160](https://doi.org/10.1001/jamanetworkopen.2025.15160)] [Medline: [40504528](https://pubmed.ncbi.nlm.nih.gov/40504528/)]
31. Tomkins A, Zhang M, Heavlin WD. Reviewer bias in single- versus double-blind peer review. *Proc Natl Acad Sci U S A* 2017 Nov 28;114(48):12708-12713 [FREE Full text] [doi: [10.1073/pnas.1707323114](https://doi.org/10.1073/pnas.1707323114)] [Medline: [29138317](https://pubmed.ncbi.nlm.nih.gov/29138317/)]
32. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996 Jan 13;312(7023):71-72 [FREE Full text] [doi: [10.1136/bmj.312.7023.71](https://doi.org/10.1136/bmj.312.7023.71)] [Medline: [8555924](https://pubmed.ncbi.nlm.nih.gov/8555924/)]

33. 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 Sep 09;26(9):1364-1374 [[FREE Full text](#)] [doi: [10.1038/s41591-020-1034-x](https://doi.org/10.1038/s41591-020-1034-x)] [Medline: [32908283](#)]
34. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 2022 May 18;28(5):924-933 [[FREE Full text](#)] [doi: [10.1038/s41591-022-01772-9](https://doi.org/10.1038/s41591-022-01772-9)] [Medline: [35585198](#)]
35. Kim HJ, Ko RE, Lim SY, Park S, Suh GY, Lee YJ. Sepsis alert systems, mortality, and adherence in emergency departments: a systematic review and meta-analysis. *JAMA Netw Open* 2024 Jul 01;7(7):e2422823 [[FREE Full text](#)] [doi: [10.1001/jamanetworkopen.2024.22823](https://doi.org/10.1001/jamanetworkopen.2024.22823)] [Medline: [39037814](#)]
36. Wang H, Li Y, Naidech A, Luo Y. Comparison between machine learning methods for mortality prediction for sepsis patients with different social determinants. *BMC Med Inform Decis Mak* 2022 Jun 16;22(Suppl 2):156 [[FREE Full text](#)] [doi: [10.1186/s12911-022-01871-0](https://doi.org/10.1186/s12911-022-01871-0)] [Medline: [35710407](#)]
37. Hwang MI, Bond WF, Powell ES. Sepsis alerts in emergency departments: a systematic review of accuracy and quality measure impact. *West J Emerg Med* 2020 Aug 24;21(5):1201-1210 [[FREE Full text](#)] [doi: [10.5811/westjem.2020.5.46010](https://doi.org/10.5811/westjem.2020.5.46010)] [Medline: [32970576](#)]

Abbreviations

AI: artificial intelligence

CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence

DECIDE-AI: Developmental and Exploratory Clinical Investigation of Decision Support Systems Driven by Artificial Intelligence

ED: emergency department

GAI: generative AI

GAIDeT: Generative Artificial Intelligence Delegation Taxonomy

LLM: large language model

Edited by S Brini; submitted 02.Jul.2026; peer-reviewed by D Barrett, U Omar; comments to author 22.Jul.2026; revised version received 26.Sep.2026; accepted 02.Oct.2026; published 19.Oct.2026.

Please cite as:

Hsieh T, Chen YT

From AI Disclosure to Judgment Traceability: Addressing Research Waste in AI-Assisted Health Sciences Publishing
J Meta and Res Int 2026;1:e106068

URL: <https://metascience.jmir.org/2026/1/e106068>

doi: [10.2196/106068](https://doi.org/10.2196/106068)

PMID:

©Tino Hsieh, Ya-Ting Chen. Originally published in the JMIR Metascience and Research Integrity (<https://metascience.jmir.org>), 19.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Metascience and Research Integrity, is properly cited. The complete bibliographic information, a link to the original publication on <https://metascience.jmir.org>, as well as this copyright and license information must be included.

Original Paper

Human-AI Agreement in Citation Validation of an AI-Generated Subspecialty Anesthesiology Knowledge Base: Cross-Sectional Paired Validation Study

Juan Li^{1,2}, MD; Shun Huang³, MD, PhD; Yue Qiu⁴, MD; Wei Chen⁵, MD; Yang Zhang⁶, MD; Xuechun Zhao⁷, MD; Ge Qu⁸, MD; Yifan Bu^{1,2}, MD; Xiaomei Feng⁹, MD; Jiawei Xu¹⁰, MD; Haobo Ma^{2,11}, MD, MPH; Jingping Wang^{2,12}, MD, PhD

¹Department of Anesthesia, Critical Care and Pain Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

²Harvard Medical School, Boston, MA, United States

³University of Louisville Hospital, Louisville, KY, United States

⁴University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

⁵Tufts Medical Center, Boston, MA, United States

⁶Stanford University School of Medicine, Stanford, CA, United States

⁷University of California San Francisco, San Francisco, CA, United States

⁸Emory University School of Medicine, Atlanta, GA, United States

⁹University of Utah School of Medicine, Salt Lake City, UT, United States

¹⁰UMass Chan Medical School – Baystate, Springfield, MA, United States

¹¹Department of Anesthesia and Pain Management, Brigham and Women's Hospital, Boston, MA, United States

¹²Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, United States

Corresponding Author:

Juan Li, MD

Department of Anesthesia, Critical Care and Pain Medicine

Beth Israel Deaconess Medical Center

330 Brookline Avenue

Boston, MA, 02215

United States

Phone: 1 6176673112

Email: likely08@gmail.com

Abstract

Background: Evaluating AI-generated citations requires checking article existence, clinical relevance, scientific accuracy, and translation fidelity. Agreement alone cannot explain judgments or their temporal stability. Retained records and repeat assessments can support transparent citation review in bilingual educational resources.

Objective: This study quantifies paired AI-human agreement at binary, 3-category, and criterion levels; evaluates AI alert rules against human FAIL ratings; compares original and repeat AI ratings; and examines evaluation records.

Methods: This cross-sectional study systematically sampled 570 OpenEvidence-generated citations across 19 subspecialties of a bilingual anesthesiology knowledge base. Each citation received 1 AI and 1 human overall verdict; 4 domains assessed existence, relevance, scientific accuracy, and translation fidelity. Ten validators represented 9 institutions; validator and subspecialty effects were inseparable. The primary post hoc binary comparison combined PASS and CONDITIONAL. Human verdicts served as a reference, not as truth. Records documented search and recovery. Applicable 95% CIs were calculated using tier-preserving subspecialty-cluster bootstrapping (B=10,000). AI reassessed 112 citations.

Results: Binary agreement was 489 out of 570 (85.8%; 95% CI 79.12%-91.75%); Cohen κ was 0.0645 (95% CI -0.0103 to 0.1639), Gwet first-order agreement coefficient was 0.8334 (95% CI 0.7388-0.9084), and Brennan-Prediger agreement was 0.7158 (95% CI 0.5825-0.8351). Three-category agreement was 271 out of 570 (47.5%; 95% CI 41.93%-53.16%). Agreement was lower for relevance (323/569, 56.8%; 95% CI 51.75%-61.62%) and scientific accuracy (440/568, 77.5%; 95% CI 72.63%-81.83%). Flagging AI CONDITIONAL/FAIL identified 274 out of 570 citations (48.1%; 95% CI 42.98%-52.98%) and captured 46 out of 74 human FAIL ratings (62%; 95% CI 48.27%-74.42%); the positive predictive value was 46 out of 274

(16.8%; 95% CI 7.89%-26.27%). Among unflagged citations, 28 out of 296 (9.5%) received human FAIL ratings (95% CI 4.84%-14.81%). In a later AI and retrieval run, unweighted agreement with original AI ratings was 91 out of 112 (81.3%) for 3 categories and 107 out of 112 (95.5%) for 2 categories. Original AI FAIL ratings persisted for 12 out of 17 citations overall and 2 out of 4 citations for article existence. According to human records, 389 citations were assigned no issue by rule without individual rereview; the principal investigator individually reviewed the remaining 181. Post hoc classification assigned 154 out of 570 (27%) citations to primary issue categories 1-5, including 33 out of 154 verification gaps; this classification is not an estimate of error prevalence.

Conclusions: This study combines paired AI-human judgments, citation checks, repeat AI assessment, and human review within a bilingual knowledge base. The framework highlights concerns requiring specialist review of source support, summaries, and translations. Retained records make citation evaluation inspectable, while public structured ratings enable statistical recalculation and reuse, supporting metascience and research integrity as AI-assisted evidence use expands.

(*J Meta and Res Int* 2026;1:e105947) doi:[10.2196/105947](https://doi.org/10.2196/105947)

KEYWORDS

AI; citation validation; interrater agreement; anesthesiology; large language models; retrieval; auditability; medical education

Introduction

Evaluating citations in a clinical knowledge resource requires checking both the cited article and the information attached to it. We compared AI and human judgments on 570 citation records across 19 subspecialties of a bilingual anesthesiology knowledge base. Earlier studies have assessed reference existence and bibliographic details [1-6], including mismatches in cited content [5,6], or compared citation credibility and accuracy [7]. Statement-support research has examined whether source text supports a claim and compared automated assessments with physicians' judgments [8]. Scripted synthetic examples have distinguished apparent evidence support from correctness [9]. A guideline study assessed citation relevance, metadata, existence, and publication in peer-reviewed journals [10]. In our resource, each record connects an article to a clinical concept and summaries in English and Chinese. A source may be real yet fail to answer the assigned clinical question, or a summary may alter an outcome, qualification, or translated term. We therefore distinguished article existence, relevance, scientific accuracy, and translation fidelity.

We used OpenEvidence to obtain candidate references and bilingual summaries for this resource. Retrieval-augmented generation supplies retrieved material to a language model, but generated answers can still contain errors [11]. A biomedical citation-recommendation system combines retrieval and reranking [12]. Previous research describes OpenEvidence as a biomedical literature retrieval-augmented generation system [13], although our records did not identify its backend implementation. Clinical evaluations have tested document or online support in neurology [14] and the relevance and quality of evidence in clinical answers [15]. Other tasks include guideline concordance [16], cardiology questions with and without retrieval support [17], treatment plans [18], repeated management recommendations [19], and anesthesia decisions [20]. Physician assessments have also included citation characteristics [21], while a pediatric question study compared input formats [22]. A spine-surgery citation audit tested OpenEvidence version 2.7.12 in October 2025 and scored article existence and bibliographic accuracy [23]. Our citations were generated in 2026, and the OpenEvidence version was not

recorded. Evidence-summary evaluations have separately assessed main-point coverage, errors, and discussion of quality and context [24]. These tasks inform our evaluation, but their percentages describe different outputs and rating rules rather than a common benchmark for judgments on this corpus.

We paired overall verdicts from a separate web-enabled AI checking workflow with human subspecialty reviewers' judgments. Human ratings provided a reference for comparison, not independently established correct answers. We first compared whether AI and human reviewers agreed on retaining a citation, treating PASS and CONDITIONAL as one group and FAIL as the other. We then examined whether both assessors gave exactly the same overall rating: PASS, CONDITIONAL, or FAIL. For example, an AI PASS paired with a human CONDITIONAL counted as agreement in the 2-group comparison but as disagreement in the 3-category comparison. We also compared AI and human ratings for article existence, relevance, scientific accuracy, and translation fidelity wherever both ratings were available. To inform specialist review, we compared 2 AI alert rules: flagging only AI FAIL ratings or flagging both AI CONDITIONAL and FAIL ratings. For each rule, we counted how many citations it flagged and how many citations rated FAIL by human reviewers it left unflagged. We also repeated the AI assessment on a sample of 112 citations and compared the repeat AI ratings with the original AI ratings for the same citations. To support metascience and research integrity, we documented evaluation choices, retained source locations, search events, and recovery records, and released structured ratings for statistical recomputation. Reporting guidance addresses AI tools, prompts, roles, and content checks [25], and raters, sampling, procedures, and statistics in agreement studies [26]. These records can inform specialist checks of source support and the reasons for judgments.

Methods

Study Design, Setting, Temporal Structure, and Units of Analysis

This cross-sectional paired study compared AI and human ratings of citations from the Chinese American Society of

Anesthesiology Digital Library, a member-authenticated English-Chinese anesthesiology platform. The knowledge base was organized into 28 subspecialties, each containing clinical topics termed concepts. For each concept, a standardized query to OpenEvidence requested 10 PubMed-indexed papers and a structured synthesis. Each citation record linked an article reference to its clinical concept and summaries in English and Chinese.

We selected 19 of the 28 subspecialties across 4 study-defined tiers and sampled 30 citation records from each, yielding 570 records. For each sampled record, we paired the overall rating from the AI checking workflow with the overall rating from the assigned human reviewer, producing 570 matched AI-human rating pairs. We also compared the paired ratings available for article existence, relevance, scientific accuracy, and translation fidelity. Stage 1 (AI) and stage 2 (human) denote the 2 assessment streams and do not imply that one preceded the other.

Ten stage 1 batches began on 5 local dates between April 28 and May 3, 2026. Eighteen stage 2 forms covering 19 subspecialties were dated between April 10 and June 7, 2026; form dates neither identify scoring starts nor establish a complete scoring window. The ranges overlapped; 5 subspecialties had form dates before stage 1, and scoring starts were unrecorded. Item 5 of the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist ([Multimedia Appendix 1](#)) maps these dates and limitations. The 30-per-subspecialty target reflected a feasible volunteer workload, without prospective end point-specific power or precision calculations; sample size was separate from random-start selection. The study was not prospectively registered; no dated design-phase plan documented prespecification of the binary end point or coefficient set before data collection or outcome inspection. Institutional record 2026D000620 documents the ethics determination, not registration or an analysis protocol.

Knowledge-Base Construction, Citation Generation, and Sampling

Concept Generation

From December 2025 to March 2026, the principal investigator (PI) developed subspecialty matrices with Google Gemini, OpenAI ChatGPT, and Anthropic Claude over 3-5 iterative rounds. The 28-subspecialty inventory comprised 1667 concepts. Concept-generation rules required topics to fall within a subspecialty's 6-module taxonomy and the scope of clinical anesthesiology and to be addressable in peer-reviewed medical literature; basic science without clinical application was excluded. Retrospective PI clarification defined structural completeness as 10 recognizable, nonempty primary citation-summary entries, regardless of numbering or file size. Historical size/marker screens were not equivalent, so no full-library completion rate is reported. This definition concerns structure, not validity. PI review addressed 6-module adherence, broad subspecialty scope, and within-module duplicates, but not concept-by-concept clinical review; verbatim transcripts were not retained.

Citation Generation

A script sequentially submitted the standardized bilingual queries through the investigator's authenticated OpenEvidence user-facing session and saved outputs locally. Retained records identify neither the automation framework, application programming interface access, nor the provider backend. Sampling-frame files were generated March 20-April 3, 2026; library-wide querying continued through April 18, 2026. Later regeneration in 4 sampled subspecialties left the frozen frame unchanged. With no retained provider version, user-selected mode, or underlying model identifier, we identify the corpus by interface and query period. [Multimedia Appendix 2](#) reports the retained query and deployment metadata.

Citation-Level Sampling

Nineteen of 28 subspecialties were selected across 4 study-defined tiers: all tier 1 high-acuity (4/4) and tier 2 large surgical (5/5), 6 out of 13 tier 3 medium surgical, and 4 out of 6 tier 4 small or cross-disciplinary subspecialties. The historical sampler used 19 retained operational N values that summed to 1166, giving 11,660 nominal requested slots (10 per concept count). These sampling counts neither estimate revision-stage completeness nor identify the highest original file number. Within each subspecialty, $k = \text{round}(10N/30)$; a retained random start r from 1 through k generated positions $r+jk$ for $j=0$ through 29. Five sequences exceeded nominal $10N$ but identified existing files because original numbering differed from operational N counts. These overruns were distinct from substitutions. The tier 3/4 subspecialty draw cannot be independently reproduced because no seed, script, or time stamp was retained. [Multimedia Appendix 2](#) lists N, k, r, and overruns for all 19 subspecialties. If a planned position failed the operational usability check, selection advanced to the next usable concept, retaining the citation slot. Retained records and revision-stage source-file readback identified 57 forward substitutions: 1 missing file, 53 outputs with fewer than 10 primary citation-summary entries, and 3 with 10 entries whose numbering of the historical check was deemed ineligible. Frozen replacements were not reversed; no backward or wraparound fallback occurred. Rounded intervals, uncompressed original file numbers, and forward-only substitutions produced unequal selection probabilities.

AI Citation-Validation Workflow and Tool-Use Audit

Model and Access

Each of the 570 citations received one final AI verdict from a web-enabled command-line workflow. All frozen production records from April 28 to May 3, 2026, recorded the requested/selected identifier `claude-opus-4-7`, corresponding to Anthropic Claude Opus 4.7 [27]. Without a retained interface version or provider-reported runtime identity, this documents the configuration, not an independently verified backend identity.

Production Schedule and Calibration

Production comprised nine 60-citation batches and one 30-citation batch; 4 batches totaling 210 citations began May 3, 2026. A separate 10-citation AI-only calibration run on April 27, 2026, lacked human ratings, was excluded from the 570

pairs, and yielded neither an agreement statistic nor a calibration threshold.

Rubric

The workflow assessed existence, relevance, scientific accuracy, and translation fidelity. AI criterion 1 was binary (PASS or

FAIL); a FAIL fixed the AI overall rating at FAIL. Criteria 2-4 were still recorded when criterion 1 failed but could neither reverse the overall FAIL rating nor establish existence. Written key-issue fields held comments, not additional criteria. Multimedia Appendix 3 shows the sequence. Table 1 compares shared domains and stream-specific options.

Table 1. Shared 4-domain framework for paired human and AI validation of 570 citations across 19 anesthesiology subspecialties. Instruments differed in wording and interface: AI criterion 1 was binary; the human grid displayed MINOR, and 1 such rating was retained without recoding.

Criterion	Domain	Permitted ratings under the rubric	Operational focus
Criterion 1	Citation existence	PASS or FAIL; MINOR not permitted	Article existence and consistency of the DOI, title, and first author
Criterion 2	Relevance	PASS, MINOR, or FAIL	Applicability of the article to the assigned concept
Criterion 3	Scientific accuracy	PASS, MINOR, or FAIL	Agreement of key data and claims with the cited source
Criterion 4	Translation fidelity	PASS, MINOR, or FAIL	Equivalence of clinical meaning in the Chinese and English summaries

Prompt Configuration

All 10 production batches referenced 1 system-prompt file with the same recorded Secure Hash Algorithm 256-bit hash, establishing cross-batch file stability only. Multimedia Appendix 4 reproduces it. Per-citation records document tool use and recovery but contain no prompt or independent proof of prompt delivery.

Tool-Use Audit

Each evaluation wrote a JSON record with invocation/result counts, WebSearch (ie, the web-search tool) queries, first-attempt compliance, and enforcement-retry status. The workflow specified 1 enforcement retry after a parseable response without WebSearch; if neither attempt invoked WebSearch, it assigned ERROR. Separately, parse recovery allowed up to 4 total attempts for unreadable structured output. Enforcement outcomes, recovered parse retries, and duration outliers were recorded separately. These records identify workflow steps, not retrieval completeness or judgment correctness.

Human Validation, Assignment, and Blinding

Panel and Assignment

Ten US-based validators (9-institution final-affiliation count) completed 18 forms covering 19 subspecialties. JL and YB shared a Beth Israel Deaconess Medical Center affiliation; the other 8 validators represented 8 additional institutions. HM and JW were senior authors and did not serve as stage 2 validators. Thus, the 12 authors included 10 validators representing 9 institutions. Convenience recruitment opened April 7, 2026; the first external validator agreed on April 10, 2026. Written instructions were provided without shared training or calibration. Each validator received 1-4 subspecialties within their stated expertise (30-120 citations). Each citation belonged to 1 subspecialty and had 1 assigned validator, making their effects inseparable and validator summaries descriptive.

Ratings and Reconciliation

Each citation had one human overall verdict; none had a second human rating, so human-human reliability could not be

estimated. Multimedia Appendix 5 reproduces the blank Expert Citation Validation Form, version 2.1 (April 2026). Revision-stage reconciliation addressed variation in instrument use through a documented hierarchy that prioritized explicit citation-level verdicts. One validator returned workbook comments without verdict symbols for 60 citations in 2 subspecialties. The PI derived analytic verdicts under this hierarchy; each derived non-PASS verdict had a supporting comment. These were reconciled comments, not independently recorded form ratings; no retrospective form was created. Later PI-framed color clarification only reconciled existing annotations.

Blinding

Non-PI validators covered 17 subspecialties (510 citations). Five (150 citations) had form dates preceding stage 1, indicating that citation-level AI outputs were unavailable on those dates. For the other 12 (360 citations), nondisclosure of AI criterion/overall ratings rests on retrospective PI confirmation; scoring-start dates and assignment-package transmission, receipt, and acknowledgment dates were not retained. The PI rated “pediatric anesthesia and pharmacology” (60 citations) after reviewing stage 1 outputs and was unblinded. Neither the AI-FAIL audit, whose sampling frame used AI verdicts, nor revision-stage issue adjudication was independent or blinded.

Administrative Sign-Off

On August 13, 2026, 1 validator signed 2 previously dated forms; the dates remained unchanged, and comparison with retained pre-sign-off copies found no scoring changes.

Outcomes and Timing of Analysis Decisions

Rating Framework

The 4 domains were shared, but the instructions and recording procedures differed. The human form included definitions, random-start/substitution instructions, 30 scoring blocks, notes, qualitative assessment, and sign-off; the AI prompt required per-citation WebSearch, evidence addresses, structured output, and tool-compliance handling. Criterion 4 asked both streams to compare the clinical meaning of the Chinese and English summaries for each sampled output. Human ratings reflected

subspecialty judgment, without a prospective terminology-control protocol or separate independent adjudication to establish preservation of all nuance in the original concepts. Human criterion 1 specified PASS/FAIL, although the generic grid displayed MINOR for every criterion. One submitted criterion 1 MINOR was retained without recoding or exclusion in 570 records. The form defined overall PASS using criteria 1-3 and called translation optional in its citation log, although the scoring instructions included criterion 4. The unchanged blank form preserves these inconsistencies; recorded ratings were unchanged.

Citation-Use Decisions and Timing

Operational curation treated PASS/CONDITIONAL as USABLE and FAIL as requiring correction or replacement, without a written standard operating procedure. Binary USABLE versus NOT-USABLE became the primary post hoc end point during the initial presubmission analysis, after collection and inspection. Three-category, criterion-level, and alert-rule analyses had no prospectively specified priority. The April 2026 Expert Citation Validation Form documented 4 domains. A protocol added to Google Drive on April 21, 2026, with an April 9, 2026, source-file modification time stamp, named Cohen κ as secondary before AI evaluation and paired-analysis completion, but not before the earliest human rating or any data inspection. This superseded document was neither a registration nor a statistical analysis plan; neither artifact fixed the binary end point or later coefficient set. An April 28, 2026, manuscript retained κ while paired results were pending. Gwet first-order agreement coefficient (AC1) was added after recognizing unequal category use. In the retained manuscript sequence ordered by source-file modification time stamps, the first reviewed file containing AC1 and Brennan-Prediger carries a June 7, 2026, time stamp. Revision-stage plan version 4.1 records an August 13, 2026, freeze but contains August 14-15, 2026, amendments; it is not an untouched August 13, 2026, byte snapshot and governs revision reanalysis only. [Multimedia Appendix 6](#) gives this bounded chronology.

Statistical Analysis

Statistics Reported

Observed agreement is the proportion of citations receiving identical ratings from both assessment streams. Binary analyses report observed agreement, unweighted Cohen κ , Brennan-Prediger agreement, and AC1, all calculated from the same ratings. Three-category analyses report exact agreement, unweighted and linearly weighted Cohen κ , Brennan-Prediger agreement, and AC1. As the coefficients account for chance differently, we avoid strong/weak agreement labels. Unweighted Cohen κ follows the original 2-rater formulation [28]. We also calculated criterion-specific unweighted Cohen κ ; each available pair contained 1 AI and 1 human rating, and no Fleiss κ was computed. CIs used the tier-preserving subspecialty-cluster percentile bootstrap described below, rather than a large-sample analytic standard-error formula.

Testing and Intervals

We estimated agreement and uncertainty without null-hypothesis tests, a prespecified α level, or P values. For agreement,

coefficient, and alert-rule estimates requiring inferential intervals, we used 2-sided 95% CIs from the specified percentile-bootstrap procedures. Limits were the 2.5th and 97.5th percentiles of the resampled estimates. Complete-corpus counts, leave-one-subspecialty/validator-out analyses, and temporal reproducibility summaries have no inferential intervals. No causal exposures, predictors, or confounding-adjustment models were specified.

Available Pairs

All 570 AI records had 4 criterion ratings. Pairs were missing only for blank or AMBIGUOUS human entries; other available ratings remained analyzed. Criteria 1-4 denominators were 570, 569, 568, and 565, respectively. The single human criterion 1 MINOR rating, disallowed by the rubric, was retained as submitted.

Alert Rules

Threshold A flagged AI CONDITIONAL/FAIL and cleared PASS; threshold B flagged only AI FAIL. Sensitivity was the proportion of 74 human-FAIL citations that were flagged. Specificity was the proportion of 496 human not-FAIL citations that were unflagged (421 PASS and 75 CONDITIONAL). Neither threshold was evaluated as a standalone decision rule.

Grouped Uncertainty

Citations from the same concept stayed together within their subspecialty; concept effects were not separately estimated. Primary and most sensitivity intervals used a tier-preserving subspecialty-cluster percentile bootstrap ($B=10,000$; seed 20260814). Each replicate sampled subspecialties with replacement within tiers, retaining counts of 4/5/6/4, carried all rows from each selected subspecialty, and recalculated the statistics. With only 19 subspecialty groups, these approximate uncertainty intervals account for grouping but do not guarantee nominal coverage or support population-wide inference. A separate validator-level sensitivity percentile bootstrap used the same B and seed, resampling 10 observed validators with replacement and all their assigned subspecialties/rows. It cannot separate validator from subspecialty effects; convenience recruitment precludes population-wide rater inference.

Sensitivity Analyses

Post hoc analyses compared human category use with AI as a common comparator, resampled validators, omitted 1 subspecialty/validator at a time, excluded the PI, compared substituted and nonsubstituted citations, and counted each validator's human-overall CONDITIONAL ratings. Revision analyses repeated binary, 3-category, criterion-level, and threshold A/B outcomes in 2 subsets: excluding all 120 WC/YB assignments as a proxy for exception-style recording, and restricting to validators with at least one final human-overall rating in each category: PASS, CONDITIONAL, and FAIL. Two-sided 95% CIs used the same tier-preserving subspecialty-cluster percentile bootstrap ($B=10,000$; seed 20260814), preserving each subset's observed subspecialties and tier counts. Undefined and degenerate replicates were audited. Validator/specialty effects remain inseparable; AI was a comparator, not truth, and population-rater inference was not supported.

Citation-Issue Classification

Revision assigned each citation either no issue or a primary category from 8 post hoc levels: (1) nonexistent or fabricated citation; (2) identifier mismatch or metadata cross-assignment; (3) wrong-article or concept mismatch; (4) verification gap; (5) content misrepresentation; (6) translation error; (7) outdated source or temporal-currency concern; and (8) upstream concept-generation concern. When multiple categories applied, all individual labels were retained, and the lowest numbered category became primary. Labels could overlap and were not summed to estimate prevalence.

The post hoc classification began with the ratings and comments from the original human reviewers. Of the 570 citations, 389 had a human overall PASS rating, human PASS ratings on all 4 criteria, and no retained human comments identifying an issue. These 389 citations were assigned “no issue” by rule; the PI did not individually recheck their sources during this classification. The PI individually reviewed the remaining 181 records because at least one of these conditions was not met. Of these, 154 were assigned to primary categories 1-5 and included in the combined count, 16 were assigned only to primary categories 6-8, and 11 had no assigned issue. The complete-frozen-corpus count has no inferential CI. It includes verification gaps, and adjudication was neither independent nor blinded, so it does not estimate objectively established error or hallucination prevalence. Revision review of retained annotations selected examples without changing classifications; it was not a prespecified qualitative analysis with independent coding and cannot establish frequencies, prevalence, consensus, or causes.

Temporal Reproducibility and Process Audit

During revision, we used the recovered AI citation-checking workflow to reassess 112 citations. We compared these repeat AI ratings with the original AI ratings for the same citations. The sample was selected by original AI overall rating: 50 PASS, 45 CONDITIONAL, and all 17 FAIL citations. Including all 17 FAIL citations allowed us to examine whether every original AI FAIL rating was repeated. The ordered sampling lists and selection record were retained. We compared whether the original and repeat AI ratings matched exactly across PASS, CONDITIONAL, and FAIL, and whether they matched after combining PASS and CONDITIONAL into 1 group. We reported unweighted results for the 112 citations. As the sample included all original FAIL citations but only some PASS and CONDITIONAL citations, we also reported secondary estimates weighted to the proportions of these rating groups in the full 570-citation corpus. These summaries were descriptive, without inferential CIs.

We requested the same claude-opus-4-7 model identifier for the repeat assessment. However, missing interface-version and provider runtime records prevented confirmation of the model version actually used. Web search results and rankings were neither frozen nor controlled. The comparison therefore assessed the repeatability of the AI and retrieval workflow under later conditions, rather than the model alone or an exact reproduction of the original run.

Ethical Considerations

On June 30, 2026, Beth Israel Deaconess Medical Center’s Committee on Clinical Investigations determined under record 2026D000620 that this activity was not human research, requiring no further committee review or approval. This was not an exemption; participant-informed consent and a waiver were inapplicable. Citation-level records and professional assessments contain no patient data or clinical records. Validators are named in the manuscript, reader-facing supplements, and the released row-level dataset. These intentionally identifiable professional-assessment data were released with documented authorization and exact-file approval from all 10 validators [29]. Completed forms, signatures, contact information, institution fields, form dates, raw free-text comments, and provider/tool records were excluded; [Multimedia Appendix 5](#) contains only the blank instrument. Validators received no study-specific compensation. No patient or platform user is identifiable in the analytic figure or reader-facing supplements.

Results

Cohort and Analytic Completeness

The cohort comprised 570 citations with paired AI-human evaluations, 30 per sampled subspecialty (19 total). Ten US-based validators from 9 institutions recorded assessments on 18 forms. One validator covered 4 subspecialties, 6 covered 2 each, and 3 covered 1 each. The PI rated 2 subspecialties, and the other validators rated the remaining 17. Each subspecialty had 1 validator, preventing separation of validator and subspecialty effects.

Human overall verdicts were 421 PASS, 75 CONDITIONAL, and 74 FAIL (total 570); AI verdicts were 296 PASS, 257 CONDITIONAL, and 17 FAIL (total 570). Six citations lacked complete 4-criterion records—4 had a blank human entry and 2 had an AMBIGUOUS entry—but retained their overall verdicts. Three of these 6 had determinate overall outcomes because another criterion was FAIL. Paired-rating denominators were 570, 569, 568, and 565 for criteria 1 through 4, respectively.

Post hoc revision-stage classification assigned 154 out of 570 (27.0%) citations to primary levels 1-5: 1 nonexistent or fabricated citation, 1 identifier mismatch or metadata cross-assignment, 106 wrong-article or concept mismatches, 33 verification gaps, and 13 content misrepresentations. Of 181 individually adjudicated records, 16 received only levels 6-8, and 11 received no assigned issue. Of the 400 citations with no assigned issue, 389 were classified from existing records without individual source rereview because they had no retained human-validator issue text, a human overall PASS, and PASS on all 4 criteria; 11 arose from adjudication. The classification includes verification gaps, was neither independent nor blinded, and is not an error-prevalence estimate; no inferential CI is reported. [Tables 2](#) and [3](#) summarize cohort composition and classifications; overall verdict counts are reported above. [Figure 1](#) shows overall verdicts by subspecialty.

Table 2. Cohort and analytic completeness for 570 citations across 19 anesthesiology subspecialties^a.

Item	Value, n	Details ^b
Full-library designed concepts	Not expressed as a percentage	1667
Citations evaluated	570	30 per subspecialty × 19 subspecialties
Subspecialty clusters	19	Tier 1: 4; tier 2: 5; tier 3: 6; and tier 4: 4
Human validation forms	18	Recorded form dates: April 10-June 7, 2026
Human validators	10	9 distinct institutions
Criterion 1 pairable	570	570/570
Criterion 2 pairable	569	569/570
Criterion 3 pairable	568	568/570
Criterion 4 pairable	565	565/570
Records not fully resolved on all 4 criteria	6	6/570
Of these, overall outcome not determined by another recorded FAIL	3	3/570
Strict complete-record sensitivity cohort	564	564/570
Determinate-outcome sensitivity cohort	567	567/570; post hoc definition

^aThe 1667 figure is the 28-subspecialty design-inventory sum. No full-library completion rate is reported because historical operational screens were not equivalent to the retrospectively clarified definition. Unresolved human entries were excluded only from the affected criterion.

^bData are presented as numerator/denominator (n/N) or design detail.

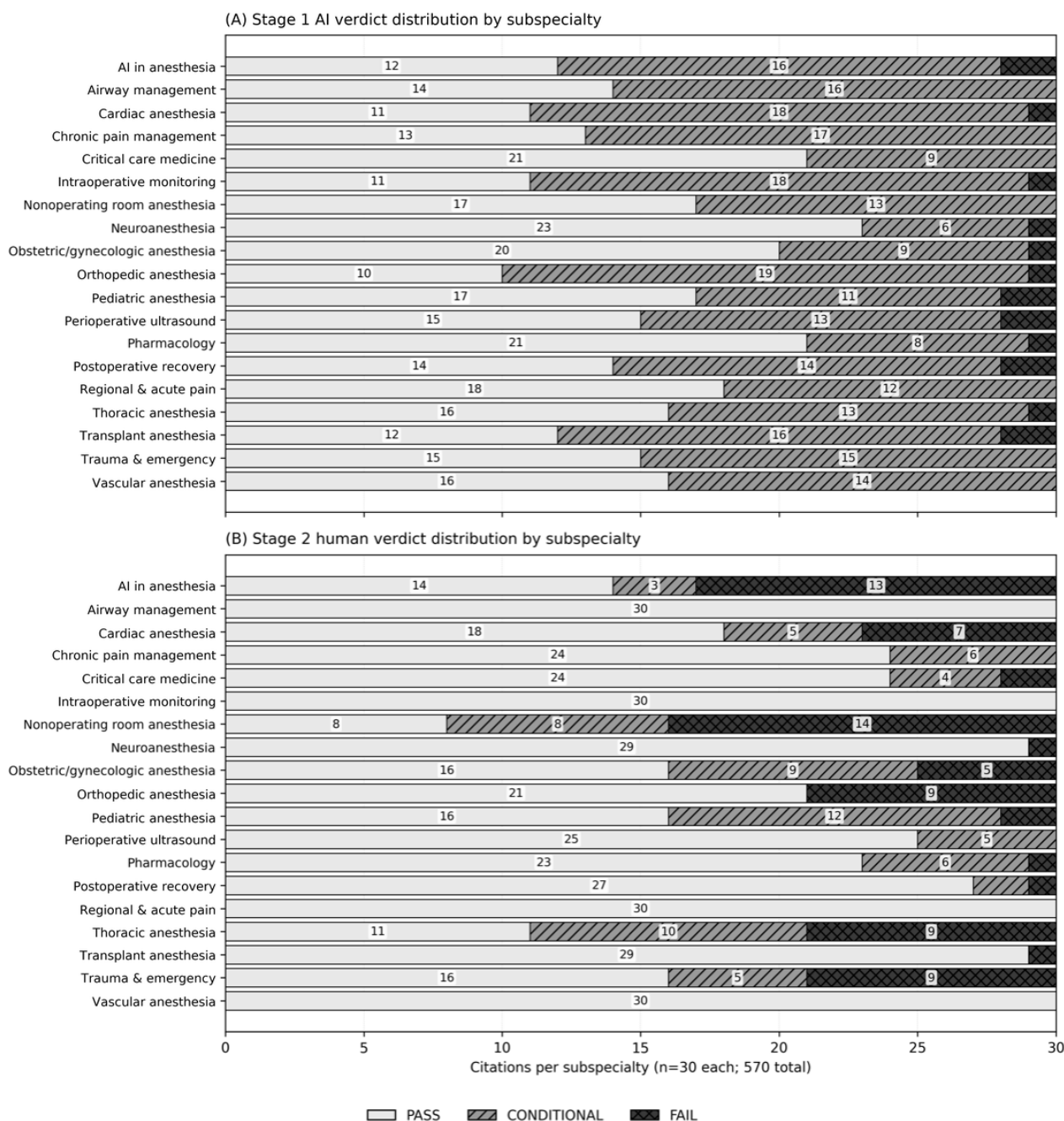
Table 3. Post hoc revision-stage citation-issue classification for 570 citations across 19 anesthesiology subspecialties^a.

Classification or review item	Count, n	Numerator/denominator, n/N (%)	Individual issue labels (can overlap), n
Classification outcomes (570 citations)			
Levels 1-5 composite	154	154/570 (27.0)	N/A ^b
Level 1: nonexistent or fabricated citation	1	1/570 (0.2)	1
Level 2: identifier mismatch or metadata cross-assignment	1	1/570 (0.2)	1
Level 3: wrong-article or concept mismatch	106	106/570 (18.6)	106
Level 4: verification gap	33	33/570 (5.8)	46
Level 5: content misrepresentation	13	13/570 (2.3)	46
Level 6: translation error	14	14/570 (2.5)	19
Level 7: outdated source or temporal-currency concern	1	1/570 (0.2)	1
Level 8: upstream concept-generation concern	1	1/570 (0.2)	20
No assigned issue (389 rule-assigned+11 individually reviewed)	400	400/570 (70.2)	N/A
Review pathway: 570=389+181; 181 reviewed=170 with assigned issues+11 without			
Rule-assigned no issue: no human-validator issue text; overall and all 4 criteria PASS; and no rereview	389	389/570 (68.2)	N/A
Individually reviewed by the principal investigator	181	181/570 (31.8)	N/A
Of the 181 individually reviewed records: no assigned issue	11	11/181 (6.1)	N/A

^aEach citation had at most 1 primary issue category but could have several individual issue labels: 170 citations had at least one label—109 had 1, 53 had 2, 7 had 3, and 1 had 4—yielding 240 label assignments; 400 had none. The 240 assignments cannot be used as a prevalence numerator because 1 citation may contribute more than 1 label. This complete-corpus, human evidence-based post hoc classification includes verification gaps and is not an estimate of objective error or hallucination prevalence.

^bN/A: not applicable.

Figure 1. AI and human overall verdict distributions for 570 citations across 19 anesthesiology subspecialties (30 citations each). (A) AI (296 PASS, 257 CONDITIONAL, and 17 FAIL); (B) human (421 PASS, 75 CONDITIONAL, and 74 FAIL). These marginal distributions do not explain individual paired disagreements or distinguish validator effects from subspecialty effects.



Overall Human-AI Agreement

For the binary end point, PASS and CONDITIONAL were classified as USABLE and FAIL as NOT-USABLE. Cross-classifying AI rows against human columns gave 484 citations rated USABLE by both streams, 69 rated USABLE by AI and NOT-USABLE by the human stream, 12 rated NOT-USABLE by AI and USABLE by the human stream, and 5 rated NOT-USABLE by both. The proportion rated USABLE by AI was 553 out of 570 (97.0%), and the proportion rated USABLE by human validators was 496 out of 570 (87.0%). Binary exact agreement was 489 out of 570 (85.8%; 95% CI 79.12%-91.75%). Cohen κ was 0.0645 (95% CI -0.0103 to 0.1639), AC1 was 0.8334 (95% CI 0.7388-0.9084), and

Brennan-Prediger agreement was 0.7158 (95% CI 0.5825-0.8351). The coefficients differed because they account for chance in different ways when the assessment streams use rating categories in unequal proportions. They are reported together, without privileging one coefficient or attaching verbal strength labels.

Three-category exact agreement was 271 out of 570 (47.5%; 95% CI 41.93%-53.16%). Combining PASS and CONDITIONAL counted 218 mismatched pairs among the 489 binary agreements. Three-category coefficients were unweighted Cohen κ 0.0519 (95% CI -0.0112 to 0.1121), linearly weighted Cohen κ 0.0828 (95% CI 0.0087-0.1523), Brennan-Prediger 0.2132 (95% CI 0.1289-0.2974), and AC1 0.2944 (95% CI

0.2043-0.3841). Tables 4-6 present the binary matrix and both coefficient sets. Binary disagreements were unequal in direction (69 vs 12 citations), but their causes were not determined; this

does not establish that AI or human validators were systematically more cautious or lenient.

Table 4. Binary contingency table for 570 paired citation evaluations across 19 anesthesiology subspecialties; rows are AI verdicts, and columns are human verdicts.

AI verdict	USABLE, n	NOT-USABLE, n	Row total, N
USABLE ^a	484	69	553
NOT-USABLE	12	5	17
Column total	496	74	570

^aUSABLE combines PASS and CONDITIONAL; NOT-USABLE is FAIL.

Table 5. Binary end point agreement for 570 paired citation evaluations across 19 anesthesiology subspecialties.

Statistic	Estimate	Numerator/denominator, n/N	95% CI ^a
Exact agreement	85.79%	489/570	79.12% to 91.75%
Unweighted Cohen κ	0.0645	N/A ^b	-0.0103 to 0.1639
Gwet first-order agreement coefficient	0.8334	N/A	0.7388 to 0.9084
Brennan-Prediger agreement	0.7158	N/A	0.5825 to 0.8351

^aAll intervals are 2.5th-97.5th percentile limits from the tier-preserving subspecialty-cluster bootstrap (B=10,000; seed 20260814). Coefficients are agreement measures, not accuracy measures. The human stream is an operational reference, not ground truth.

^bN/A: not applicable.

Table 6. Three-category end point agreement for 570 paired citation evaluations across 19 anesthesiology subspecialties^a.

Statistic	Estimate	Numerator/denominator, n/N	95% CI
Exact agreement	47.54%	271/570	41.93% to 53.16%
Unweighted Cohen κ	0.0519	N/A ^b	-0.0112 to 0.1121
Linearly weighted Cohen κ	0.0828	N/A	0.0087 to 0.1523
Brennan-Prediger agreement	0.2132	N/A	0.1289 to 0.2974
Gwet first-order agreement coefficient	0.2944	N/A	0.2043 to 0.3841

^aAll intervals are 2.5th-97.5th percentile limits from the tier-preserving subspecialty-cluster bootstrap (B=10,000; seed 20260814). Coefficients are agreement measures, not accuracy measures. The human stream is an operational reference, not ground truth.

^bN/A: not applicable.

Criterion-Level Agreement and Disagreement Pattern

Observed agreement was numerically highest for criteria 1 and 4 and lower for criteria 2 and 3 (Table 7). Nearly all criterion 1 ratings were PASS, and the FAIL sets did not overlap; its agreement (ie, 563/570, 98.8%) therefore does not establish concordance on rare FAIL ratings. Exact agreement was 563 out of 570 (98.8%) for criterion 1 (95% CI 97.72%-99.65%);

323 out of 569 (56.8%) for relevance (95% CI 51.75%-61.62%); 440 out of 568 (77.5%) for scientific accuracy (95% CI 72.63%-81.83%); and 529 out of 565 (93.6%) for translation fidelity (95% CI 91.23%-95.95%). Corresponding unweighted Cohen κ estimates were -0.0050 (95% CI -0.0097 to 0.0000), 0.1329 (95% CI 0.0737 to 0.1910), 0.0508 (95% CI -0.0165 to 0.1149), and 0.0678 (95% CI -0.0310 to 0.1773), respectively.

Table 7. Criterion-level exact agreement and unweighted Cohen κ for paired evaluations of 570 citations across 19 anesthesiology subspecialties.

Criterion	Exact agreement, n/N (%)	95% CI	Cohen κ	95% CI
1. Citation existence ^a	563/570 (98.77)	97.72%-99.65%	-0.0050	-0.0097 to 0.0000
2. Relevance ^b	323/569 (56.77)	51.75%-61.62%	0.1329	0.0737 to 0.1910
3. Scientific accuracy ^b	440/568 (77.46)	72.63%-81.83%	0.0508	-0.0165 to 0.1149
4. Translation fidelity ^b	529/565 (93.63)	91.23%-95.95%	0.0678	-0.0310 to 0.1773

^aCriterion 1 was binary, but the generic human grid displayed MINOR; 1 such rating was retained without recoding.

^bAll AI records contained 4 ratings; human blank or AMBIGUOUS entries reduced the denominators for criteria 2-4. The κ values are unweighted Cohen estimates; criterion 1 κ is descriptive only because ratings were nearly constant. Intervals use the tier-preserving subspecialty-cluster percentile bootstrap (B=10,000; seed 20260814).

Criterion 1 was binary, although the generic human grid displayed MINOR for every criterion. AI recorded 566 PASS, 0 MINOR, and 4 FAIL (total 570); humans recorded 567 PASS, 1 out-of-rubric MINOR, and 2 FAIL (total 570). The FAIL sets did not overlap. The 7 disagreements comprised 4 AI FAIL/human PASS, 2 AI PASS/human FAIL, and 1 AI PASS/human MINOR. The out-of-rubric MINOR was retained unchanged and was not treated as a hallucination flag. Criterion 1 κ is descriptive only: of 10,000 bootstrap replicates, 29 were undefined, and 623 were defined but degenerated because ratings were nearly constant. It is not interpreted as meaningful below-chance agreement. Mapping ratings to the revision-stage classification cannot establish which stream was correct because that classification used human-validation evidence as its operational reference.

Vascular anesthesia, airway management, regional and acute pain, and intraoperative monitoring each had 30 human-PASS citations (120 total). Their zero κ values and zero-to-zero intervals reflect constant human ratings, not an absence of agreement. Using AI as a common comparator, PASS, CONDITIONAL, and FAIL correlations across 19 subspecialties were $r=-0.004$, -0.297 , and -0.003 , respectively; the CONDITIONAL correlation was weakly negative. These comparisons are descriptive. [Multimedia Appendix 7](#) reports signed gaps and validator-pooled summaries. Human-overall CONDITIONAL counts and proportions, in neutral source order, were WC 8 out of 60 (13%), SH 0 out of 120 (0%), YQ 14 out of 60 (23%), GQ 11 out of 60 (18%), XZ 4 out of 30 (13%), YZ 18 out of 60 (30%), XF 0 out of 30 (0%), JX 0 out of 30 (0%), JL (PI) 18 out of 60 (30%), and YB 2 out of 60 (3%),

totaling 75 out of 570. These are not rankings; validator and subspecialty effects remain inseparable. Retained annotations illustrated plausible locations of citation issues but were not a prespecified, independently coded qualitative dataset; they neither changed classifications nor supported frequency, consensus, or causal inference.

Citations Flagged or Missed by the Alert Rules

Using human FAIL as the operational reference, threshold A flagged AI CONDITIONAL or FAIL: 274 out of 570 citations were flagged, and 296 were cleared. Sensitivity was 46 out of 74 (62%; 95% CI 48.27%-74.42%); 28 out of 74 (a false-negative rate of 38%) were missed (95% CI 25.58%-51.73%). Specificity was 268 out of 496 (54.0%; 95% CI 49.35%-58.90%), with all 421 human PASS and 75 CONDITIONAL citations in the denominator. Positive predictive value was 46 out of 274 (16.8%; 95% CI 7.89%-26.27%); negative predictive value was 268 out of 296 (90.5%; 95% CI 85.19%-95.16%). Of the 296 cleared citations, 28 (9.5%) were human FAIL (95% CI 4.84%-14.81%). The flag rate was 274 out of 570 (48.1%; 95% CI 42.98%-52.98%).

Threshold B flagged only AI FAIL: 17 out of 570 (3%) citations (95% CI 1.93%-4.04%). It captured 5 out of 74 human-FAIL records and missed 69 out of 74 (93%; 95% CI 86.36%-97.70%). Sensitivity was 6.76% (95% CI 2.30%-13.64%) and specificity was 484 out of 496 (97.6%; 95% CI 96.30%-98.80%). In each bootstrap replicate, the missed proportion equaled one minus sensitivity; its interval therefore complements the sensitivity end points in reverse order. [Table 8](#) compares citations flagged and human-FAIL citations missed. Neither threshold was a standalone decision rule or citation-accuracy measure.

Table 8. Citations flagged or missed by 2 AI alert rules using human FAIL as the operational reference, across 570 citations from 19 anesthesiology subspecialties^a.

Metric	Threshold A: flag AI CONDITIONAL or FAIL, n/N (%); 95% CI	Threshold B: flag AI FAIL only, n/N (%); 95% CI
Flagged citations	274/570 (48.1); 42.98%-52.98%	17/570 (2.98); 1.93%-4.04%
Sensitivity	46/74 (62.2); 48.27%-74.42%	5/74 (6.76); 2.30%-13.64%
Missed human-FAIL citations	28/74 (37.8); 25.58%-51.73%	69/74 (93.24); 86.36%-97.70%
Specificity	268/496 (54.0); 49.35%-58.90%	484/496 (97.58); 96.30%-98.80%
Positive predictive value	46/274 (16.8); 7.89%-26.27%	Not reported
Negative predictive value	268/296 (90.5); 85.19%-95.16%	Not reported
Human FAIL among citations cleared by threshold A	28/296 (9.5); 4.84%-14.81%	N/A ^b

^aSensitivity uses 74 human-FAIL citations; specificity uses all 496 human not-FAIL citations (421 PASS plus 75 CONDITIONAL), not 421 PASS alone. Threshold A flags AI CONDITIONAL or FAIL; threshold B flags only AI FAIL without treating CONDITIONAL as cleared. The missed proportion is the replicate-wise complement of sensitivity; interval end points follow that identity. Human FAIL is an operational reference, not ground truth.

^bN/A: not applicable.

Sensitivity Analyses, Temporal Reproducibility, and Process Audit

Excluding all 120 WC/YB assignments left 450 citations in 15 subspecialty clusters. Validator assignment was a proxy for the exception-style recording workflow, not a citation-level identification of its use. Binary agreement was 396 out of 450 (88%; 95% CI 81.33%-93.56%), with Cohen $\kappa=0.1155$ (95% CI 0.0269-0.2253); 3-category exact agreement was 225 out of 450 (50%; 95% CI 44.22%-55.56%), with $\kappa=0.0723$ (95% CI -0.0031 to 0.1421). Restricting to validators with at least one final human-overall PASS, CONDITIONAL, and FAIL retained WC, YQ, XZ, YZ, JL (PI), and YB (330 citations; 11 clusters). Binary agreement was 263 out of 330 (79.7%; 95% CI 72.12%-86.36%), with $\kappa=0.0315$ (95% CI -0.0356 to 0.1012); 3-category exact agreement was 159 out of 330 (48.2%; 95% CI 41.21%-54.85%), with $\kappa=0.1010$ (95% CI 0.0254-0.1783). In both revision-stage subsets, 3-category agreement remained near the full-corpus estimate, while coefficients diverged; binary

agreement varied with subgroup composition. Agreement depended on whether 2 or 3 categories were compared, and no single binary percentage described every subgroup. [Multimedia Appendix 7](#) provides complete coefficient families, criterion and threshold A/B results, intervals, and bootstrap accounting. [Multimedia Appendix 8](#) maps reporting items from the 3 cited reporting frameworks [25,26,30] to manuscript locations without evaluating adherence or assigning a grade.

[Tables 9](#) and [10](#) summarize the remaining sensitivity and influence analyses. Resampling validators produced wider uncertainty intervals without changing the point estimates. Excluding the PI and omitting 1 subspecialty or validator at a time did not identify a single assignment whose removal eliminated the observed binary pattern. These analyses still cannot separate validator effects from the mix of specialties or eliminate bias from overlapping investigator roles. Future studies could examine whether substitutions affect validation results. The comparison of 57 substituted with 513 nonsubstituted citations could not isolate such an effect.

Table 9. Post hoc sensitivity analyses for 570 paired citation evaluations across 19 anesthesiology subspecialties^a.

Analysis and statistic	Estimate	Numerator/denominator, n/N	95% CI
Validator-resampling percentile bootstrap			
Binary agreement	85.8%	489/570	75.49% to 93.83%
Cohen κ	0.0645	N/A ^b	-0.0216 to 0.1872
Gwet first-order agreement coefficient	0.8334	N/A	0.6825 to 0.9338
Brennan-Prediger agreement	0.7158	N/A	0.5098 to 0.8767
Exclusion of 2 principal investigator-rated subspecialties			
Binary agreement	85.3%	435/510	77.84% to 91.96%
Cohen κ	0.0752	N/A	0.0016 to 0.1818
Gwet first-order agreement coefficient	0.8264	N/A	0.7177 to 0.9099
Substituted citations			
Binary agreement	78.9%	45/57	65.79% to 90.91%
Cohen κ	-0.0857	N/A	-0.1343 to -0.0279
Gwet first-order agreement coefficient	0.7406	N/A	0.5225 to 0.9005
Nonsubstituted citations			
Binary agreement	86.5%	444/513	80.00% to 92.43%
Cohen κ	0.0855	N/A	0.0006 to 0.1919
Gwet first-order agreement coefficient	0.8432	N/A	0.7509 to 0.9170

^aThis table uses percentile intervals (B=10,000; seed 20260814); validator-resampling rows resample the 10 validators with all assigned rows; principal investigator-exclusion rows use tier-preserving subspecialty-cluster resampling across the 17 retained clusters (3/5/5/4); and substituted and nonsubstituted rows use tier-preserving subspecialty-cluster resampling across all 19 clusters (4/5/6/4).

^bN/A: not applicable.

Table 10. Leave-one-out influence diagnostics for 570 paired citation evaluations across 19 anesthesiology subspecialties^a.

Diagnostic	Range of binary agreement	Minimum, n/N	Maximum, n/N
Leave one subspecialty out	0.8500-0.8778	459/540; airway management omitted	474/540; AI in anesthesia omitted
Leave one validator out	0.8244-0.8882	371/450; Shun Huang omitted	453/510; Wei Chen omitted

^aValues are diagnostics, not CIs.

In the stratified 112-citation rerun, sample-unweighted exact 3-category and binary reproducibility were 91 out of 112 (81.3%) and 107 out of 112 (95.5%), respectively; secondary estimates weighted to the original verdict distribution were 82.84% and 99.12% (analyzed n=112), respectively (Table 11). Original verdicts persisted for 43 out of 50 PASS, 36 out of 45 CONDITIONAL, and 12 out of 17 FAIL citations; criterion 1

FAIL persisted in 2 out of 4. These descriptive summaries have no inferential CIs. The audit (Table 12) recorded first-attempt WebSearch for 570 out of 570 citations, 0 enforcement retries, 0 ERROR verdicts, 20 recovered parse retries, 3 duration outliers, and 1108 invocations (mean 1.9439). These records demonstrate tool use and processing recovery, not retrieval correctness, completeness, or correct model interpretation.

Table 11. Temporal reproducibility in a 112-citation AI rerun on August 14-15, 2026 (analyzed n=112)^a.

Summary	Value, %	Numerator/denominator, n/N
Sample-unweighted exact 3-category/binary reproducibility	81.3/95.5	91/112 (3-category); 107/112 (binary)
Design-weighted exact 3-category/binary reproducibility	82.84/99.12	Weighted estimates; analyzed n=112
Original FAIL ratings still FAIL at rerun	70.6	12/17
Criterion 1 FAIL ratings still FAIL at rerun	Not expressed as a percentage	2/4

^aThis table reports how often the combined workflow reproduced its ratings under later retrieval conditions, not model-only behavior or accuracy. FAIL citations were sampled at a higher rate, and the weighted estimates do not have observed sample numerators.

Table 12. Process audit for 570 paired citation evaluations across 19 anesthesiology subspecialties^a.

Metric	Value ^b	Numerator/denominator, n/N
First-attempt WebSearch compliance	100	570/570
WebSearch enforcement retries	0	0/570
ERROR verdicts	0	0/570
Recovered parse-layer retries	20	20/570
Wall-clock duration outliers	3	3/570
Total WebSearch invocations	1108	Mean 1.9439 per citation

^aThis table separates enforcement, parse recovery, and duration outliers; recording a tool invocation does not establish that the retrieved evidence was correct.

^bThe first value is a percentage; the remaining ones in the “Value” column are counts. Mean invocations per citation are reported separately.

Discussion

Principal Findings

AI and human reviewers could agree to retain a citation while disagreeing about whether it still needed checking. Our binary comparison grouped PASS and CONDITIONAL together. For example, an AI PASS paired with a human CONDITIONAL counted as agreement in the binary analysis but as disagreement when PASS, CONDITIONAL, and FAIL were compared separately. Higher binary agreement therefore does not mean that both assessors considered a citation acceptable without further checking. This distinction remained in subset analyses, although binary agreement varied with the citations included. Individual checks helped identify what needed attention: relevance and scientific accuracy showed more disagreement than article existence and translation fidelity. Relevance asks whether the article addresses the clinical question; scientific accuracy asks whether the summary preserves its findings and qualifications. An existing article may provide unsuitable support for a statement, or its findings may be summarized misleadingly. Disagreements could reflect different interpretations of evidence or applications of rating rules; our comparisons cannot identify the explanation for an individual citation. Specialists can examine the source passage, the linked clinical statement, and the clinical context. Reporting these checks alongside overall ratings helps direct that examination. The retained examples below illustrate what such examination can reveal.

The choice of agreement measure affects how readers interpret the same AI-human ratings. Most binary matches were citations that both streams classified as USABLE. Percent agreement counts how often ratings matched, without adjusting for category use [31]. Cohen κ compares observed agreement with the agreement expected under its chance model, based on the marginal use of each rating category by the 2 streams. This reference level can be high when both streams often assign the same category, so many matching ratings can coexist with a low κ . AC1 uses a different chance-agreement calculation and responds differently when 1 rating is much more common [31,32]. A higher AC1 should not be interpreted as a more accurate measurement of the same quantity. We also reported Brennan-Prediger agreement, without choosing a preferred coefficient. Several κ CIs included 0, leaving uncertainty about

agreement beyond the coefficient’s reference level; this does not mean that the assessors never agreed. None of these measures establishes citation accuracy because human judgments were not independently established as correct answers [33,34]. Reporting rating definitions, category frequencies, measures, and CIs together lets readers examine how conclusions depend on analysis choices. Such reporting also helps future studies explain why agreement estimates differ.

The alert-rule comparison asks which citations AI would bring to a specialist’s attention and which would remain outside the review list. Flagging only AI FAIL produced a shorter list but left most human FAIL ratings unflagged. Adding AI CONDITIONAL captured more human FAIL ratings, although not all, and selected more citations for checking. The number selected must therefore be considered alongside what remains unflagged. An AI flag identifies a citation for examination, not a confirmed error; an unflagged citation is not confirmed correct. AI flags may help organize specialist review if unflagged citations also receive appropriate scrutiny. We did not measure the time needed to review either list or determine which rule would be more efficient. Future evaluations should assess the effort required and the problems found and corrected, including among unflagged citations.

Comparison With Previous Studies

Comparisons with earlier citation studies must begin with what each study checked. An overall score cannot distinguish confirmation of article identity from judgment of whether content supports a clinical statement. Earlier audits included article existence and citation details [1-7], while statement-source research examined whether the source text supported an attached claim [8]. Scripted conceptual examples separated apparent evidence support from correctness [9]. Other studies assessed relevance, metadata, existence, and peer-reviewed publication status [10], or main points, errors, quality assessment, and context in evidence summaries [24]. Building on these distinctions, our paired evaluations applied different checks to the same bilingual citation records. The paired ratings show which aspects of a citation AI and human reviewers judged differently.

Retained comments and citation materials preserve details that a rating label alone cannot convey. One human comment noted that a paper addressed transversus abdominis plane block safety

and median dosage rather than the assigned question about maximum dose. In a stored bilingual summary, “stroke volume” appeared as “卒中容积” rather than the cardiac term “每搏量.” The summary for US-013-03 described tricuspid annular plane systolic excursion ≥ 1.7 cm as suggesting right-ventricular dysfunction, whereas the cited guideline defines normal M-mode tricuspid annular plane systolic excursion as >1.70 cm [35]. These materials allow reviewers to examine the specific question, terminology, or threshold, including a meaningful translation concern despite higher overall translation agreement. They identify concrete checks, not issue frequencies, the cause of the overall disagreement pattern, or which assessment stream was superior.

The examples above illustrate why the type of concern matters for further citation review. Harada’s [9] conceptual audit used scripted examples to distinguish clinical correctness, evidence support, and numeric consistency. Colder Carras et al [24] documented inaccurate terminology and a summary of the wrong publication in AI-generated annotations. Our post hoc classification addressed related concerns in retained records from a working bilingual knowledge base. The PI distinguished article or concept mismatch, insufficient evidence for verification, and misrepresentation of source content, alongside concerns about translation, outdated sources, and the original clinical concept. These distinctions can help specialists decide whether to recheck the chosen article, seek supporting passages, or examine a summary or translation for correction. Retaining overlapping labels preserves details that an overall verdict alone cannot convey. The classification was unblinded and not independently repeated, so it describes identified review concerns rather than confirmed error prevalence. Its contribution is to make the reasons for further review explicit and provide candidate criteria for future evaluation by independent reviewers.

The generation setting and evaluation unit provide another basis for comparison. Previous studies used restricted website searches [13], supplied documents or online material [14,17], and citation-recommendation systems [12]. Others assessed clinical answers alongside source usefulness or citation characteristics [15,21]; pediatric perioperative work separately assessed answer quality, citation authenticity, and clinical safety [36]. Here, OpenEvidence generated candidate citations and bilingual summaries, while a separate AI checking workflow and human reviewers evaluated the same sampled citation records. Each record linked an article to a clinical concept and English-Chinese summaries, with article existence explicitly included among the checks. This adds evidence about judgments made during citation curation in a working resource; it does not make different retrieval and clinical-answer tasks directly comparable as performance tests.

Interpretation within this study also depends on how the assessments were arranged. The 2 streams assigned criterion 1 FAIL to different citations, providing candidates for further checking rather than evidence of complementary detection of true errors. Each subspecialty had 1 assigned human reviewer, so reviewer and subspecialty differences could not be separated, consistent with methodological work on nested rater designs [37]. Comparisons of category use against the AI ratings may

help identify assignments for rubric review but cannot rank reviewer strictness or determine which stream was correct.

Methodological Innovation and Practical Implications

The methodological contribution combines evaluations of the same 570 citation records from a working bilingual anesthesiology knowledge base covering 19 subspecialties. Each citation received 1 AI and 1 assigned human overall judgment. We compared whether both assessors would retain a citation and whether they assigned the same PASS, CONDITIONAL, or FAIL rating. We also compared their available ratings for article existence, relevance, scientific accuracy, and translation fidelity. The separate post hoc classification described recorded review concerns, while alert rules identified citations selected for further checking and human FAIL ratings left outside that selection. Together, these comparisons describe how often judgments agree and which aspects of a citation may warrant closer examination.

Repeat AI assessment and retained process records add information about whether judgments remain stable and how the workflow was executed. In the later run, decisions to retain or remove citations repeated more often than exact PASS, CONDITIONAL, or FAIL ratings, and some original FAIL ratings changed. Following individual ratings therefore helps identify judgments that may warrant another look when an AI workflow is used again. The requested model identifier was unchanged, but the model actually used could not be independently confirmed, and search results were not held fixed. The comparison therefore describes the combined AI and retrieval workflow at a later time. Related research found that more repeatable antibiotic recommendations were not necessarily more clinically correct [38]. Repeated requests for laboratory reference intervals produced responses whose consistency depended on the model version and laboratory test [39]. We also retained checks confirming that the saved prompt file was unchanged across batches, source locations, search records, and records of how output problems were handled. These records give later reviewers a starting point for examining the documented steps associated with a changed or disputed judgment.

The framework can guide a review process in which AI identifies citations for attention and specialists judge their source support. Brini and Leung’s [40] tutorial emphasizes human oversight suited to the task, checking whether tools are valid for their intended use, and transparent documentation. Screening studies show that selection can change with prompt wording [41,42] and the way eligibility criteria are expressed [43]. Reported efficiency findings included measured reductions in screening time [41], estimated human time savings [42], and calculated reductions in a second reviewer’s workload [43]. Disagreements with reference screening judgments remained [41-43]. Clinical studies also show why assistance needs to be evaluated in its intended setting: a physician trial found no significant improvement in diagnostic reasoning with large language model access [44], radiology research found different effects across readers and tasks [45], and a systematic review found variable outcomes from human-large language model collaboration [46]. Applied to citation curation, our framework

would retain AI ratings and recorded concerns alongside each citation. A specialist could use these records to locate the relevant source passage, check the clinical claim and bilingual summary, and decide whether to retain, correct, or replace the citation.

Implications for Metascience and Research Integrity

This study has a metascientific contribution by illustrating how rating definitions and agreement measures influence the interpretation of the same AI-human citation evaluations. Applying established methods in a working bilingual knowledge base makes these evaluation choices and their consequences visible. For research integrity, explicit criteria and retained reasons and process records provide a basis for examining documented judgments. Public structured ratings allow others to recalculate the reported statistics. These resources support scrutiny of the evaluation but do not independently establish citation correctness. The public release excludes raw comments and provider/tool records, and the available records cannot reproduce the original AI and retrieval conditions. Within these limits, the study offers a practical example of documenting citation evaluation and sharing its structured results, which may help other researchers plan and interpret similar assessments.

Limitations

First, each citation had only 1 human overall rating, with no independently double-scored subset, consensus panel, or human-human agreement estimate. Human ratings were a comparison reference, not independent confirmation of citation correctness [33]. One convenience-recruited validator per subspecialty prevents separation of reviewer and subspecialty effects. Comparisons of human and AI category use remain descriptive because AI performance may differ by subspecialty. Records do not confirm that criteria left unmarked on forms recording only exceptions were checked as explicitly as criteria with recorded PASS ratings. Excluding 120 assignments therefore uses validator assignment as a proxy for workflow differences, without confirming how each citation was assessed. Overlapping citation sets assessed by multiple validators across subspecialties would help separate sources of disagreement [37]. Consensus would still require source-passage checks and rules for resolving disputes, without guaranteeing correctness.

Second, for 12 subspecialties assessed by other validators (360 citations), evidence that AI ratings were withheld rests on the PI's later confirmation. Records of scoring-start dates and when review materials were sent were not retained. The PI saw stage 1 outputs before rating 2 subspecialties (60 citations). These ratings, the AI-FAIL audit, and revision-stage classification were unblinded; the latter 2 were not independent. Excluding PI ratings cannot remove potential bias from the PI's design, corpus construction, workflow operation, validation, audit, and classification roles. Original file numbers did not directly match nominal sampling counts; the tier 3/4 selection seed, script, and time stamp were missing. No registration or preregistration record was found. The binary end point and full coefficient set were not fixed before data inspection. Volunteer workload determined sample size, without agreement-specific power or precision calculations. With only 19 subspecialty groups and convenience recruitment, bootstrap intervals may not achieve

their intended coverage and cannot establish generalizability to all subspecialties or validators.

Third, 1 unblinded investigator classified the expanded retained records during revision; no independent investigator repeated the classification. Most no-issue assignments used existing ratings without individually rechecking sources, so concerns may remain unrecorded. The combined count includes verification gaps, and 1 primary category may summarize several overlapping concerns. The 4 domains did not systematically assess outdated sources or poorly specified clinical concepts; passing ratings therefore leave these questions unresolved. Selected examples were not part of a prespecified qualitative study or independent coding process and cannot establish frequencies or causes.

Fourth, retained records allow statistical recalculation but cannot recreate the original run. Missing information includes provider confirmation of the actual model, original provider responses, complete command-line and search-result records, the software environment, and a frozen search index. The later assessment tested the AI and retrieval workflow together under uncontrolled search conditions. A newer model was released after the primary AI evaluation but within the dates on some human review forms [47]; these dates do not establish the full scoring period. Frozen records identify the requested/selected claude-opus-4-7 identifier, but provide no later-model identifier or verified runtime identity. Findings may not apply to newer systems. Repeat evaluations should document the requested generation system and retrieval environment.

Fifth, the bilingual prompt and criterion 4 specified Chinese-English summary comparison without prespecified terminology rules or separate independent bilingual adjudication. The human form described translation's role in the overall rating inconsistently and called it optional in the citation log. Ratings were preserved, but consistent interpretation across validators cannot be confirmed. The rubric incompletely covered subspecialty nuance, overlooked better sources, and failed to account for the effects of omitted literature on summaries. We did not measure how much relevant literature the searches retrieved; concept-generation transcripts and complete transmission timing were unavailable. Upstream review checked structure, not every clinical concept. We assessed citation curation, not correction time, prospective review burden, learner knowledge, clinical decisions, or patient outcomes. One account, interface, and anesthesiology corpus limit generalizability. Future work needs multilingual expert adjudication, independently assembled relevant sources, system versions documented over time, and prospective workflow outcomes across institutions, specialties, and interfaces.

Conclusions

AI-generated citations need to be evaluated both as bibliographic references and as evidence supporting clinical statements. This study brings together overall AI-human judgments and specific checks of article existence, clinical relevance, scientific accuracy, and translation fidelity in a working bilingual anesthesiology knowledge base. These complementary comparisons help distinguish agreement on a broad decision from specific concerns that require further review. Repeat AI

assessment adds information about rating stability, while subsequent human review and a separate post hoc 8-category classification describe concerns that the 3 overall verdicts alone cannot explain. Together, these components can help specialists determine what needs checking, whether source passages support the attached claims, and whether summaries or translations require correction. Retained assessment and process

records support later examination of documented judgments, and public structured ratings enable statistical recalculation and further analysis. The contribution to metascience and research integrity lies in making evaluation choices and their supporting records inspectable and potentially useful for future work as AI-assisted citation use expands.

Acknowledgments

The authors declare the use of generative AI (GAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy) guidelines (2025), the following tasks were delegated to GAI tools under full human supervision: idea generation, defining the research objective, literature search and systematization, writing the literature review, research design, development of experimental or research protocols, selection of research methods, code generation, creation of algorithms for data analysis, data analysis, text generation, proofreading and editing, formulation of conclusions, and preparation of press releases and outreach materials. The GAI tools used were Google Gemini 3.0 Pro; OpenAI ChatGPT 5.5, 5.6 Sol, and 6 Astra; and Anthropic Claude Opus 4.7, Opus 4.8, and Opus 5. Some historical model details rely on JL's recollection. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

The frozen 570-row, 15-column named professional-assessment dataset is publicly available in the Zenodo repository under CC BY 4.0 [29]. It contains stable citation identifiers, subspecialty, validator, 4 criterion ratings, and overall verdicts from each assessment stream, and the post hoc issue-label fields `primary_issue_category` and `atomic_issue_labels`. The accompanying data dictionary, README, license notice, and integrity manifest define the released files. The package supports recomputation of the reported agreement statistics but not regeneration of the original model-plus-retrieval executions. Completed forms, signatures, contact information, institution fields, form dates, raw free-text comments, and provider/tool records are excluded.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

Conceptualization: JL, HM, JW

Data curation: JL

Formal analysis: JL

Investigation: SH, YQ, WC, YZ, XZ, GQ, YB, XF, JX

Methodology: JL

Project administration: JL

Software: JL

Supervision: HM, JW

Validation: JL, SH, YQ, WC, YZ, XZ, GQ, YB, XF, JX

Visualization: JL

Writing – original draft: JL

Writing – review & editing: JL, SH, YQ, WC, YZ, XZ, GQ, YB, XF, JX, HM, JW

Conflicts of Interest

None declared.

Multimedia Appendix 1

STROBE checklist for the cross-sectional study, indicating the corresponding locations in the manuscript.

[[DOCX File, 28 KB](#) - [metascience_v1i1e105947_app1.docx](#)]

Multimedia Appendix 2

OpenEvidence citation-generation prompt, provenance, and sampling parameters.

[[DOCX File, 28 KB](#) - [metascience_v1i1e105947_app2.docx](#)]

Multimedia Appendix 3

Selected AI interaction examples, including OpenEvidence citation generation and a worked example of stage 1 citation validation.
[DOCX File, 30 KB - [metascience_v1i1e105947_app3.docx](#)]

Multimedia Appendix 4

Stage 1 validation system prompt and workflow.

[DOCX File, 29 KB - [metascience_v1i1e105947_app4.docx](#)]

Multimedia Appendix 5

Expert citation validation form: 4-criterion information retrieval evaluation framework, version 2.1, April 2026.

[PDF File (Adobe PDF File), 698 KB - [metascience_v1i1e105947_app5.pdf](#)]

Multimedia Appendix 6

Revision-stage analysis chronology.

[DOCX File, 27 KB - [metascience_v1i1e105947_app6.docx](#)]

Multimedia Appendix 7

Human citation assessment results by validator and subspecialty, with corresponding AI ratings and post hoc sensitivity analyses.

[DOCX File, 37 KB - [metascience_v1i1e105947_app7.docx](#)]

Multimedia Appendix 8

Manuscript locations corresponding to the GAMER, GRRAS, and METRICS reporting items.

[DOCX File, 29 KB - [metascience_v1i1e105947_app8.docx](#)]

References

1. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *J Med Internet Res* 2024 Apr 05;26:e52935 [FREE Full text] [doi: [10.2196/52935](#)] [Medline: [38578685](#)]
2. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. *J Med Internet Res* 2024 May 22;26:e53164 [FREE Full text] [doi: [10.2196/53164](#)] [Medline: [38776130](#)]
3. Linardon J, Jarman HK, McClure Z, Anderson C, Liu C, Messer M. Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: experimental study. *JMIR Ment Health* 2025 Nov 12;12:e80371 [FREE Full text] [doi: [10.2196/80371](#)] [Medline: [41223407](#)]
4. Pastucha M, Skarżyński H, Kochanek K, Jędrzejczak WW. Reference accuracy in large language model chatbots: a metric for inherent misinformation? *Med Sci Monit* 2026 Jan 17;32:e950916. [doi: [10.12659/msm.950916](#)]
5. Sawamura S, Bito T, Ando T, Masuda K, Kameyama S, Ishida H. Evaluation of the accuracy of ChatGPT's responses to and references for clinical questions in physical therapy. *J Phys Ther Sci* 2024 May;36(5):234-239 [FREE Full text] [doi: [10.1589/jpts.36.234](#)] [Medline: [38694019](#)]
6. Civelekler M, Çitirik M. The role of artificial intelligence in academic citation: a study on lens, cataract, and anterior segment research. *Indian J Ophthalmol* 2026 Jul 01;74(7):1073-1076 [FREE Full text] [doi: [10.4103/IJO.IJO_3155_25](#)] [Medline: [42378572](#)]
7. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of OpenAI's GPT-4 and Google's Gemini. *Comput Biol Med* 2025 Feb;185:109545 [FREE Full text] [doi: [10.1016/j.combiomed.2024.109545](#)] [Medline: [39667055](#)]
8. Wu K, Wu E, Wei K, Zhang A, Casasola A, Nguyen T, et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nat Commun* 2025 Apr 16;16(1):3615 [FREE Full text] [doi: [10.1038/s41467-025-58551-6](#)] [Medline: [40240349](#)]
9. Harada Y. Detecting citation veneer in guideline-grounded clinical and public health large language model (LLM) outputs: a clinical evidence audit grid using CENHOV tags. *Cureus* 2026 Apr;18(4):e107729. [doi: [10.7759/cureus.107729](#)] [Medline: [42205634](#)]
10. Şengül HB, Akın B, Kayaalp ME, Sezgin EA. Deep research capabilities in GPT-5 thinking and Gemini 2.5 Pro improve citation integrity and concordance with American Academy of Orthopaedic Surgeons anterior cruciate ligament and rotator cuff guidelines. *Knee Surg Sports Traumatol Arthrosc* 2026 Apr 06;34(4):1504-1512. [doi: [10.1002/ksa.70315](#)] [Medline: [41649196](#)]
11. Gilbert S, Kather JN, Hogan A. Augmented non-hallucinating large language models as medical information curators. *NPJ Digit Med* 2024 Apr 23;7(1):100 [FREE Full text] [doi: [10.1038/s41746-024-01081-0](#)] [Medline: [38654142](#)]

12. Xie Q, Zhang J, Wang Y, Huang J, Lin F, Weng R, et al. CiteSure: retrieval-augmented large language models for faithful biomedical citation recommendation. *J Am Med Inform Assoc* 2026 Oct 01;33(10):1934-1942 [[FREE Full text](#)] [doi: [10.1093/jamia/ocag122](https://doi.org/10.1093/jamia/ocag122)] [Medline: [42551844](#)]
13. Masanneck L, Epping PZ, Meuth SG, Pawlitzki M. Evaluating web retrieval-assisted large language models with and without whitelisting for evidence-based neurology: comparative study. *J Med Internet Res* 2025 Oct 29;27:e79379 [[FREE Full text](#)] [doi: [10.2196/79379](https://doi.org/10.2196/79379)] [Medline: [41159599](#)]
14. Masanneck L, Meuth SG, Pawlitzki M. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. *NPJ Digit Med* 2025 Mar 04;8(1):137 [[FREE Full text](#)] [doi: [10.1038/s41746-025-01536-y](https://doi.org/10.1038/s41746-025-01536-y)] [Medline: [40038423](#)]
15. Low YS, Jackson ML, Hyde RJ, Brown RE, Sanghavi NM, Baldwin JD, et al. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digit Health* 2025 Jun 09;11:20552076251348850 [[FREE Full text](#)] [doi: [10.1177/20552076251348850](https://doi.org/10.1177/20552076251348850)] [Medline: [40510193](#)]
16. Hsu W, Chuang H, Wang Y, Hsu K, Kuan F, Su W, et al. Concordance of ChatGPT, Gemini, Claude, and OpenEvidence with the 2024 AAOS guidelines on acute isolated meniscal pathology. *Knee* 2026 Aug;61:104427. [doi: [10.1016/j.knee.2026.104427](https://doi.org/10.1016/j.knee.2026.104427)] [Medline: [41865615](#)]
17. Tarabanis C, Khurshid S, Karamanou A, Piperaki R, Mavromatis LA, Hatzimemos A, et al. Cardiology knowledge assessment of retrieval-augmented open versus proprietary large language models. *PLOS Digit Health* 2026 Mar 12;5(3):e0001029 [[FREE Full text](#)] [doi: [10.1371/journal.pdig.0001029](https://doi.org/10.1371/journal.pdig.0001029)] [Medline: [41818295](#)]
18. Freudenberg J, Knitza J, Gremke N, Amann N, Deutsch TM, Tauber N, et al. AI-enabled clinical decision support in breast cancer care: a blinded multicenter benchmarking study comparing medically specialized with a general-purpose system. *J Med Syst* 2026 Jul 04;50(1):107. [doi: [10.1007/s10916-026-02434-w](https://doi.org/10.1007/s10916-026-02434-w)] [Medline: [42400697](#)]
19. Landon S, Savage T, Greysen SR, Bressman E. Variation in large language model recommendations in challenging inpatient management scenarios. *J Gen Intern Med* 2026 Apr;41(5):1286-1291. [doi: [10.1007/s11606-025-09888-7](https://doi.org/10.1007/s11606-025-09888-7)] [Medline: [41055682](#)]
20. Chen R, Warburton A, Do R, Chen DD, Katz D, Burnett GW. Large-scale evaluation of five large language models in anesthesia decision-making for hip fracture surgery. *Anesth Analg* 2026 Jun 10 Online ahead of print. [doi: [10.1213/ANE.0000000000008148](https://doi.org/10.1213/ANE.0000000000008148)] [Medline: [42268736](#)]
21. Ben Shitrit I, Idan D, Volevich M, Goldenberg HS, Vaknin D, Degany O, et al. Real world human-LLM interactions - prospective blinded versus unblinded expert physician assessments of LLM responses to complex medical dilemmas. *PLOS Digit Health* 2026 Mar 12;5(3):e0001278 [[FREE Full text](#)] [doi: [10.1371/journal.pdig.0001278](https://doi.org/10.1371/journal.pdig.0001278)] [Medline: [41818246](#)]
22. James V, Caronia C, Savargaonkar R. Input structure-driven instability and convergence in large language model clinical reasoning: a formative study using pediatric residency-level MCQs. *Sci Rep* 2026 Apr 17;16(1):17829 [[FREE Full text](#)] [doi: [10.1038/s41598-026-48326-4](https://doi.org/10.1038/s41598-026-48326-4)] [Medline: [41998244](#)]
23. McLaughlin ND, Srinivas AN, Lowe ZF, Botterbush KS, Patel MS, Avila MJ. Large language model hallucinations in spine surgery: a comparative analysis of clinician vs patient-level prompts. *Neurosurg Pract* 2026 Jun;7(3):e000244. [doi: [10.1227/neuprac.0000000000000244](https://doi.org/10.1227/neuprac.0000000000000244)] [Medline: [42232526](#)]
24. Colder Carras M, Qureshi R, Naaman K, Aldayel F, Date M, AlJuboori D, et al. Using large language models to summarize evidence in biomedical articles: exploratory comparison between AI- and human-annotated bibliographies. *JMIR Form Res* 2026 Feb 12;10:e69707 [[FREE Full text](#)] [doi: [10.2196/69707](https://doi.org/10.2196/69707)] [Medline: [41678657](#)]
25. Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, GAMER Working Group. Reporting guideline for the use of generative artificial intelligence tools in MEDical Research: the GAMER Statement. *BMJ Evid Based Med* 2025 Dec 01;30(6):390-400 [[FREE Full text](#)] [doi: [10.1136/bmjebm-2025-113825](https://doi.org/10.1136/bmjebm-2025-113825)] [Medline: [40360239](#)]
26. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011 Jan;64(1):96-106. [doi: [10.1016/j.jclinepi.2010.03.002](https://doi.org/10.1016/j.jclinepi.2010.03.002)] [Medline: [21130355](#)]
27. Claude Opus 4.7 System Card. Anthropic. 2026 Apr 16. URL: <https://www.anthropic.com/claude-opus-4-7-system-card> [accessed 2026-10-05]
28. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960 Apr 01;20(1):37-46. [doi: [10.1177/001316446002000104](https://doi.org/10.1177/001316446002000104)]
29. Li J, Wang J, Ma H, Huang S, Qiu Y, Chen W, et al. JMIR 105947 Named Professional Assessments Dataset. Zenodo. 2026. URL: <https://zenodo.org/records/22177917> [accessed 2026-10-05]
30. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: development study involving a literature review. *Interact J Med Res* 2024 Feb 15;13:e54704 [[FREE Full text](#)] [doi: [10.2196/54704](https://doi.org/10.2196/54704)] [Medline: [38276872](#)]
31. Geijer M, Båth M, Wessman C. Some common statistical methods for assessing rater agreement in radiological studies. *Acta Radiol* 2025 Jun 23;66(6):675-683 [[FREE Full text](#)] [doi: [10.1177/02841851251319666](https://doi.org/10.1177/02841851251319666)] [Medline: [39988909](#)]
32. Vach W, Gerke O. Gwet's AC1 is not a substitute for Cohen's kappa - a comparison of basic properties. *MethodsX* 2023;10:102212 [[FREE Full text](#)] [doi: [10.1016/j.mex.2023.102212](https://doi.org/10.1016/j.mex.2023.102212)] [Medline: [37234937](#)]

33. Trikalinos TA, Balion CM. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". J Gen Intern Med 2012 Jun;27(Suppl 1):S67-S75 [FREE Full text] [doi: [10.1007/s11606-012-2031-7](https://doi.org/10.1007/s11606-012-2031-7)] [Medline: [22648677](https://pubmed.ncbi.nlm.nih.gov/22648677/)]
34. Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007 Dec;11(50):iii, ix-iii, 51 [FREE Full text] [doi: [10.3310/hta11500](https://doi.org/10.3310/hta11500)] [Medline: [18021577](https://pubmed.ncbi.nlm.nih.gov/18021577/)]
35. Mukherjee M, Rudski LG, Addetia K, Afilalo J, D'Alto M, Freed BH, et al. Guidelines for the echocardiographic assessment of the right heart in adults and special considerations in pulmonary hypertension: recommendations from the American Society of Echocardiography. J Am Soc Echocardiogr 2025 Mar;38(3):141-186. [doi: [10.1016/j.echo.2025.01.006](https://doi.org/10.1016/j.echo.2025.01.006)] [Medline: [40044341](https://pubmed.ncbi.nlm.nih.gov/40044341/)]
36. Kang T, Yuan C, Hu X, Huang W. Performance of 5 large language models in perioperative consultation for pediatric hypospadias: cross-sectional comparative study. J Med Internet Res 2026 Jul 29;28:e93393 [FREE Full text] [doi: [10.2196/93393](https://doi.org/10.2196/93393)] [Medline: [42525870](https://pubmed.ncbi.nlm.nih.gov/42525870/)]
37. Stora B, Hagtvet KA, Heyerdahl S. Reliability of observers' subjective impressions of families: a generalizability theory approach. Psychother Res 2013 Jul;23(4):448-463 [FREE Full text] [doi: [10.1080/10503307.2012.733830](https://doi.org/10.1080/10503307.2012.733830)] [Medline: [23066691](https://pubmed.ncbi.nlm.nih.gov/23066691/)]
38. De Vito A, Geremia N, Russo A, Colpani A, Vaira LA, Madeddu G. Reproducibility of AI-generated antibiotic recommendations in standardised clinical scenarios: a proof-of-concept experimental study. Int J Antimicrob Agents 2026 Oct;67(10):107881 [FREE Full text] [doi: [10.1016/j.ijantimicag.2026.107881](https://doi.org/10.1016/j.ijantimicag.2026.107881)] [Medline: [42285312](https://pubmed.ncbi.nlm.nih.gov/42285312/)]
39. Meyer A, Schömig E, Streichert T. ChatGPT and reference intervals: a comparative analysis of repeatability in GPT-3.5 Turbo, GPT-4, and GPT-4o. Front Artif Intell 2025;8:1681979. [doi: [10.3389/frai.2025.1681979](https://doi.org/10.3389/frai.2025.1681979)] [Medline: [41459577](https://pubmed.ncbi.nlm.nih.gov/41459577/)]
40. Brini S, Leung TI. Value and credibility of meta-analysis: tutorial on enhancing methodological rigor and AI-powered efficiency. J Med Internet Res 2026 Jul 02;28:e92132 [FREE Full text] [doi: [10.2196/92132](https://doi.org/10.2196/92132)] [Medline: [42390911](https://pubmed.ncbi.nlm.nih.gov/42390911/)]
41. Oami T, Okada Y, Nakada T. Performance of a large language model in screening citations. JAMA Netw Open 2024 Jul 01;7(7):e2420496 [FREE Full text] [doi: [10.1001/jamanetworkopen.2024.20496](https://doi.org/10.1001/jamanetworkopen.2024.20496)] [Medline: [38976267](https://pubmed.ncbi.nlm.nih.gov/38976267/)]
42. Cao C, Sang J, Arora R, Chen D, Kloosterman R, Cecere M, et al. Development of prompt templates for large language model-driven screening in systematic reviews. Ann Intern Med 2025 Mar;178(3):389-401. [doi: [10.7326/annals-24-02189](https://doi.org/10.7326/annals-24-02189)]
43. Delgado-Chaves FM, Jennings MJ, Atalaia A, Wolff J, Horvath R, Mamdouh ZM, et al. Transforming literature screening: the emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A 2025 Jan 14;122(2):e2411962122 [FREE Full text] [doi: [10.1073/pnas.2411962122](https://doi.org/10.1073/pnas.2411962122)] [Medline: [39761403](https://pubmed.ncbi.nlm.nih.gov/39761403/)]
44. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024 Oct 01;7(10):e2440969 [FREE Full text] [doi: [10.1001/jamanetworkopen.2024.40969](https://doi.org/10.1001/jamanetworkopen.2024.40969)] [Medline: [39466245](https://pubmed.ncbi.nlm.nih.gov/39466245/)]
45. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med 2024 Mar 19;30(3):837-849 [FREE Full text] [doi: [10.1038/s41591-024-02850-w](https://doi.org/10.1038/s41591-024-02850-w)] [Medline: [38504016](https://pubmed.ncbi.nlm.nih.gov/38504016/)]
46. Wang G, Zhang K, Jiang J, Wang C, Bi H, Liang H, et al. Human-large language model collaboration in clinical medicine: a systematic review and meta-analysis. NPJ Digit Med 2026 Jan 28;9(1):195 [FREE Full text] [doi: [10.1038/s41746-026-02382-2](https://doi.org/10.1038/s41746-026-02382-2)] [Medline: [41606089](https://pubmed.ncbi.nlm.nih.gov/41606089/)]
47. Introducing Claude Opus 4.8. Anthropic. URL: <https://www.anthropic.com/news/claude-opus-4-8> [accessed 2026-10-05]

Abbreviations

AC1: Gwet first-order agreement coefficient

PI: principal investigator

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 30.Jun.2026; peer-reviewed by S Bhetwal, K Jaiswal, H Maheshwari; comments to author 10.Aug.2026; revised version received 29.Sep.2026; accepted 02.Oct.2026; published 19.Oct.2026.

Please cite as:

Li J, Huang S, Qiu Y, Chen W, Zhang Y, Zhao X, Qu G, Bu Y, Feng X, Xu J, Ma H, Wang J

Human-AI Agreement in Citation Validation of an AI-Generated Subspecialty Anesthesiology Knowledge Base: Cross-Sectional Paired Validation Study

J Meta and Res Int 2026;1:e105947

URL: <https://metascience.jmir.org/2026/1/e105947>

doi:[10.2196/105947](https://doi.org/10.2196/105947)

PMID:

©Juan Li, Shun Huang, Yue Qiu, Wei Chen, Yang Zhang, Xuechun Zhao, Ge Qu, Yifan Bu, Xiaomei Feng, Jiawei Xu, Haobo Ma, Jingping Wang. Originally published in the JMIR Metascience and Research Integrity (<https://metascience.jmir.org>), 19.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the JMIR Metascience and Research Integrity, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

<https://www.jmirpublications.com/>