

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 Ngwu, 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.

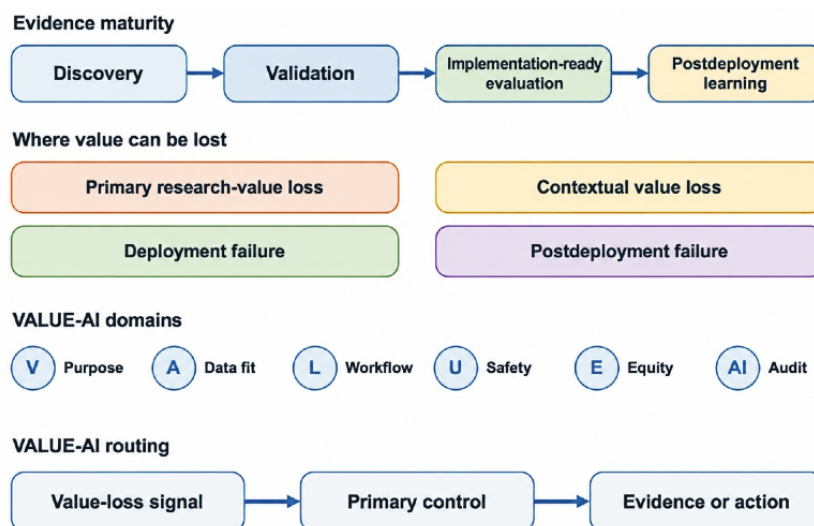
At validation, the evidentiary burden shifts toward reproducibility and performance under meaningful variation [15,18]. Independent temporal, geographic, or institutional evaluation should be selected based on the proposed use rather than treated as a ceremonial exercise. Calibration (agreement between 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 learning 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*. Apr 2016;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*. Jan 2014;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*. Jan 11, 2014;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*. Jun 26, 2024;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*. Jul 04, 2025;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*. Jan 13, 2022;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*. Oct 20, 2021;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*. Jan 06, 2025;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*. Mar 25, 2020;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*. Jul 01, 2025;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*. Aug 18, 2016;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*. Jun 07, 2025;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*. Mar 2022;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*. Oct 29, 2019;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*. Sep 2020;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*. May 2023;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*. Sep 2016;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*. Mar 24, 2025;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, et al. DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. May 18, 2022;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*. Feb 5, 2025;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; Jan 26, 2023. 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*. Apr 08, 2022;22(1):101. [FREE Full text] [doi: [10.1186/s12874-022-01577-x](https://doi.org/10.1186/s12874-022-01577-x)] [Medline: [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*. Nov 01, 2017;19(11):e367. [FREE Full text] [doi: [10.2196/jmir.8775](https://doi.org/10.2196/jmir.8775)] [Medline: [29092808](#)]
24. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. *Implement Sci*. Oct 29, 2022;17(1):75. [FREE Full text] [doi: [10.1186/s13012-022-01245-0](https://doi.org/10.1186/s13012-022-01245-0)] [Medline: [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*. Dec 01, 2024;22(12):2423-2446. [doi: [10.11124/JBIES-24-00042](https://doi.org/10.11124/JBIES-24-00042)] [Medline: [39658865](#)]
26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. Oct 25, 2019;366(6464):447-453. [FREE Full text] [doi: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)] [Medline: [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*. Apr 18, 2024;385:q902. [FREE Full text] [doi: [10.1136/bmj.q902](https://doi.org/10.1136/bmj.q902)] [Medline: [38636956](#)]
28. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. Geneva, Switzerland. World Health Organization; Jun 28, 2021. URL: <https://www.who.int/publications/i/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*. May 09, 2024;7(1):120. [FREE Full text] [doi: [10.1038/s41746-024-01085-w](https://doi.org/10.1038/s41746-024-01085-w)] [Medline: [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*. Oct 2025;170:104902. [doi: [10.1016/j.jbi.2025.104902](https://doi.org/10.1016/j.jbi.2025.104902)] [Medline: [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*. May 2024;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*. Apr 07, 2025;8(1):190. [FREE Full text] [doi: [10.1038/s41746-025-01587-1](https://doi.org/10.1038/s41746-025-01587-1)] [Medline: [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*. Jul 2020;18(7):e3000737. [FREE Full text] [doi: [10.1371/journal.pbio.3000737](https://doi.org/10.1371/journal.pbio.3000737)] [Medline: [32673304](#)]
34. Chambers CD, Tzavella L. The past, present and future of registered reports. *Nat Hum Behav*. Jan 2022;6(1):29-42. [doi: [10.1038/s41562-021-01193-7](https://doi.org/10.1038/s41562-021-01193-7)] [Medline: [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*. Sep 09, 2020;370:m3210. [FREE Full text] [doi: [10.1136/bmj.m3210](https://doi.org/10.1136/bmj.m3210)] [Medline: [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*. Jan 01, 2019;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*. Sep 09, 2020;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](#)]
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](#)]
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*. Mar 19, 2011;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](#)]
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; Aug 18, 2025. 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*. Feb 22, 2024;15(1):1619. [FREE Full text] [doi: [10.1038/s41467-024-45355-3](https://doi.org/10.1038/s41467-024-45355-3)] [Medline: [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.* Aug 01, 2021;181(8):1065-1070. [FREE Full text] [doi: [10.1001/jamainternmed.2021.2626](https://doi.org/10.1001/jamainternmed.2021.2626)] [Medline: [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.* Apr 2026;64:100746. [FREE Full text] [doi: [10.1016/j.ahjo.2026.100746](https://doi.org/10.1016/j.ahjo.2026.100746)] [Medline: [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.* Apr 2022;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.* Jan 2026;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.* Jan 23, 2024;7(1):14. [FREE Full text] [doi: [10.1038/s41746-023-00986-6](https://doi.org/10.1038/s41746-023-00986-6)] [Medline: [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.* May 01, 2026;7(5):e261204. [FREE Full text] [doi: [10.1001/jamahealthforum.2026.1204](https://doi.org/10.1001/jamahealthforum.2026.1204)] [Medline: [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.* Oct 03, 2024;7(1):273. [FREE Full text] [doi: [10.1038/s41746-024-01270-x](https://doi.org/10.1038/s41746-024-01270-x)] [Medline: [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.* Aug 19, 2026;5(8):e0001597. [FREE Full text] [doi: [10.1371/journal.pdig.0001597](https://doi.org/10.1371/journal.pdig.0001597)] [Medline: [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.* Aug 02, 2017;358:j3453. [FREE Full text] [doi: [10.1136/bmj.j3453](https://doi.org/10.1136/bmj.j3453)] [Medline: [28768629](#)]
51. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med.* Feb 16, 2010;7(2):e1000217. [FREE Full text] [doi: [10.1371/journal.pmed.1000217](https://doi.org/10.1371/journal.pmed.1000217)] [Medline: [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.