Original Article

Measuring Pharmacy Artificial Intelligence by Avoided Harm, Recovered Time, and Better Decisions

Ana Rodrigues[1]*, Tiago Martins[1], Bruno Lopes[2]

[1]Department of Pharmacy AI Outcome Measurement, Faculty of Pharmacy, University of Lisbon, Lisbon, Portugal. [2]Department of Value-Based AI Assessment, Faculty of Pharmacy, University of Porto, Porto, Portugal.

Abstract

Artificial intelligence in pharmacy is often evaluated through discrimination, accuracy, alert yield, or task completion, yet these measures do not establish whether medication-related harm was avoided, professional time was genuinely recovered, or decisions became better. This article proposes a non-empirical value-evaluation framework for connecting technical performance to pharmacy-practice consequences without treating any single metric as sufficient. Value is defined as attributable, net, distributed, and sustained benefit relative to a specified medication-use problem, comparator, perspective, and time horizon. The framework separates seven evaluative layers: baseline need, technical fitness, pharmacy-task performance, human–AI and workflow interaction, consequence measurement, attribution, and net value. Three consequence domains are distinguished. Avoided harm requires evidence linking an AI signal to professional review, action, changed medication use, and a credible counterfactual clinical consequence. Recovered time requires deduction of verification, correction, training, maintenance, and coordination work, followed by assessment of how released capacity is used. Better decisions require evaluation of interpretation, uncertainty handling, appropriateness, timeliness, and final action rather than agreement with AI alone. Validation should combine task-specific performance studies, workflow observation, human-factors assessment, comparative outcome designs, economic evaluation, subgroup analysis, and postimplementation monitoring. The proposed framework does not supply universal thresholds, combine heterogeneous outcomes into a single score, or establish safety, cost-effectiveness, regulatory acceptability, or deployment readiness. Its original contribution is an evidence-bounded architecture for designing, interpreting, and governing pharmacy-AI value claims.

Keywords: Digital pharmacy, Artificial intelligence, Pharmacy practice, Medication safety, Human–AI collaboration, Clinical decision support

Introduction

Artificial intelligence may classify prescriptions, prioritize medication alerts, predict clinical deterioration, identify anomalous orders, or generate medication-related recommendations. These functions are commonly described through model-level measures such as discrimination, sensitivity, specificity, calibration, or predictive accuracy. Such measures establish aspects of technical performance, but clinical impact also depends on the intended task, data provenance, user behavior, workflow integration, implementation conditions, and relationship between predictions and patient-relevant action [1].

Comparisons between algorithms and health professionals can further encourage the interpretation that superior benchmark performance establishes clinical value. However, comparative studies have frequently been limited by retrospective designs, weak external validation, incomplete reporting, and comparisons that do not reproduce real clinical work [2]. A model can therefore outperform a professional on an isolated classification task without improving a medication decision, reducing preventable harm, or creating usable professional capacity.

Safety also cannot be inferred from average performance. Responsible clinical machine learning requires attention to problem formulation, data quality, subgroup performance, human interaction, deployment context, monitoring, and consequences arising after implementation [3]. In pharmacy, these concerns are especially important because model outputs often influence high-frequency decisions involving medicine selection, dose, route, monitoring, interaction management, and escalation. False reassurance, excessive

Address for correspondence: Ana Rodrigues, Department of Pharmacy AI Outcome Measurement, Faculty of Pharmacy, University of Lisbon, Lisbon, Portugal.
ana.rodrigues@ff.ulisboa.pt
Received: 29 July 2025; **Accepted:** 25 November 2025

How to cite this article: Rodrigues A, Martins T, Lopes B. Measuring Pharmacy Artificial Intelligence by Avoided Harm, Recovered Time, and Better Decisions. Arch Pharm Pract. 2025;16(4):64-72. https://doi.org/10.51847/zwjUAgneRK

interruption, inappropriate reliance, or displacement of professional attention may create risks that are not visible in conventional validation.

Patient-benefit claims consequently require more than technical reproducibility. They require transparent intended use, an appropriate comparator, clinically meaningful outcomes, evidence that the output changed care, and evaluation of effectiveness under realistic conditions [4]. Even areas reporting strong diagnostic accuracy have often lacked prospective testing, representative populations, workflow evaluation, and direct measurement of patient consequences [5].

Comparative performance reports, methodological appraisal, and diagnostic-accuracy synthesis collectively show why model-centred evidence should not be treated as sufficient evidence of patient or practice value [2, 4, 5]. This article therefore proposes that pharmacy AI be evaluated through three connected but non-equivalent consequence domains: avoided harm, recovered time, and better decisions. These domains are placed within a wider architecture incorporating baseline need, technical fitness, pharmacy-task performance, human–AI interaction, implementation burden, attribution, equity, and sustained value. The contribution is conceptual rather than empirical: it organizes what would need to be demonstrated before an AI-related change could credibly be described as valuable.

Defining Value in Pharmacy AI

A pharmacy-specific definition of value must begin with the consequences that matter to medication use rather than with the availability of an algorithm. Current pharmacy applications frequently target workflow, productivity, prescription review, dispensing, and information-processing tasks, whereas direct patient-health outcomes are less commonly positioned as the principal purpose of evaluation [6]. This imbalance does not make workflow outcomes unimportant; it means that workflow improvement should not be represented as patient benefit without evidence of the intervening pathway.

Value also depends on the evaluative perspective. Patients may prioritize reduced preventable harm, access, understanding, continuity, and treatment burden. Pharmacists may value safer prioritization, reduced repetitive work, improved information synthesis, and preserved professional control. Organizations may consider capacity, cost, workforce sustainability, implementation burden, and service reliability. Health-economic accounts likewise recognize that value can include several forms of clinical, experiential, organizational, and uncertainty-related benefit [7]. The relevant elements must nevertheless be stated explicitly rather than aggregated without justification.

This article defines pharmacy-AI value as the attributable, net, equitably distributed, and sustained benefit produced by an AI-enabled change relative to a defined medication-use problem, comparator, perspective, and time horizon. “Attributable” requires a defensible connection between the intervention and the observed consequence. “Net” requires subtraction of implementation, verification, correction, training, maintenance, coordination, and opportunity costs. “Distributed” requires examination of who benefits and who carries risk or burden. “Sustained” requires evidence that value persists after initial adoption.

A staged validation approach is necessary because technical, clinical, usability, and economic validation answer different questions [8]. Adoption and early use likewise do not establish scale-up, sustainability, or protection against abandonment when technologies become difficult to integrate into complex services [9]. Economic claims additionally require a stated perspective, comparator, resource consequences, outcomes, time horizon, and treatment of uncertainty [10].

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses for the article Measuring Pharmacy Artificial Intelligence by Avoided Harm, Recovered Time, and Better Decisions.

Table 1. Definitions, components, relationships, and intended uses for the article Measuring Pharmacy Artificial Intelligence by Avoided Harm, Recovered Time, and Better Decisions.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Baseline medication-use need	AI may be introduced without a sufficiently important or measurable problem	Harm burden, task volume, delay, variation, unmet need	**Proposed:** value evaluation begins from documented baseline burden	Establishes the denominator and comparator	Local baseline data and problem definition [1]	Solution-first procurement; unstable baseline	Need does not establish that AI is the appropriate response
Technical fitness	Model metrics may be mistaken for clinical value	Discrimination, calibration, threshold performance, missingness	Technical performance supports but does not complete task evaluation	Determines whether further evaluation is justified	Internal, external, temporal, and subgroup validation [8]	Dataset shift; miscalibration; unsuitable thresholds	Technical fitness is necessary but insufficient

Pharmacy-task performance	Prediction quality may not translate into better prescription or medication review	Task definition, user, setting, output, decision threshold	**Proposed:** model output must improve or preserve a defined pharmacy task	Connects model behavior to professional work	Realistic task-performance evidence [6]	Artificial test conditions; unsafe false negatives	Task improvement does not establish patient benefit
Human–AI interaction	Outputs may be misunderstood, ignored, or over-relied upon	Presentation, uncertainty, expertise, workload, authority	**Proposed:** professional interpretation mediates consequences	Identifies reliance, correction, escalation, and override pathways	Human-factors and workflow evaluation [3]	Automation bias; alert fatigue; ambiguity	AI output is not equivalent to professional judgment
Avoided harm	Detected errors or accepted alerts may be counted as prevented harm	Signal, review, action, changed medication use, counterfactual risk	**Proposed:** every causal link must be evidenced	Supports credible medication-safety claims	Process, outcome, severity, and attribution evidence	Counting alerts as harms prevented	A process event is not an avoided clinical outcome
Recovered time	Gross task-time reduction may conceal new work	Time saved, verification, correction, training, maintenance	**Proposed:** new AI-related work is deducted from gross savings	Estimates usable professional capacity	Time-and-motion, log, and activity-destination evidence [9]	Burden transfer; hidden work; unused capacity	Faster completion is not necessarily recovered time
Better decisions	Agreement with AI may be mistaken for decision improvement	Interpretation, uncertainty, action, appropriateness, timeliness	**Proposed:** decision process and final decision are evaluated separately	Tests whether care choices improve	Independent decision assessment and outcome follow-up [4]	Harmful reliance; confirmation bias	Concordance with AI is not a quality endpoint
Net and sustained value	Initial benefit may be offset or decline after implementation	All benefits, costs, equity effects, drift, workarounds	**Proposed:** value is reassessed across the lifecycle	Supports continuation, modification, restriction, or withdrawal decisions	Economic and longitudinal implementation evidence [10]	Selective accounting; benefit decay	No single domain establishes overall value

The proposed architecture treats model performance as an upstream condition, pharmacy-task performance and human–AI interaction as mediating levels, and avoided harm, recovered time, and better decisions as consequence domains. Attribution then determines whether observed changes can reasonably be assigned to the AI-enabled intervention, while net-value assessment considers implementation burden, opportunity cost, uncertainty, equity, and sustainability.

Figure 1 presents the original conceptual synthesis for multidimensional value-evaluation framework for pharmacy ai, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

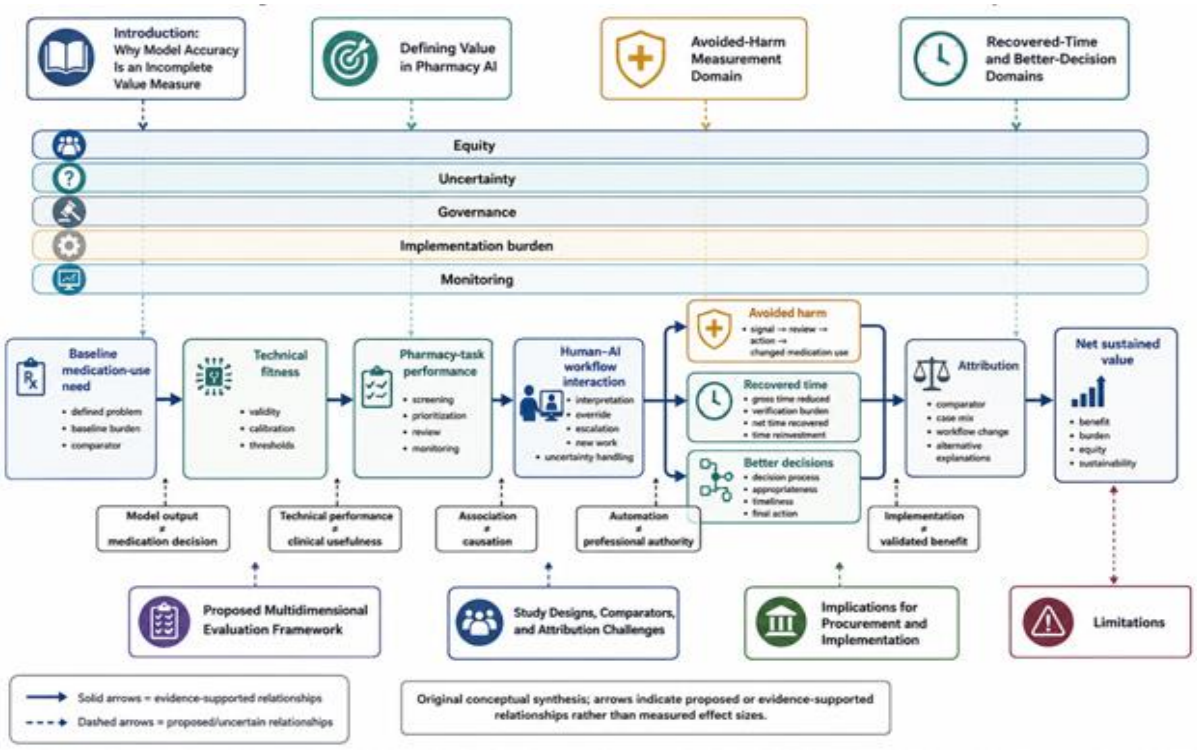

Figure 1. Multidimensional Value-Evaluation Framework for Pharmacy AI

The framework is intended to structure evaluation, not to collapse heterogeneous consequences into a composite score. Critical failure in safety, equity, professional control, or attribution cannot be cancelled by favorable performance elsewhere. Evidence must instead be traced across layers, with uncertainty retained where the transition from model output to consequence has not been demonstrated.

Avoided-Harm Measurement Domain

Avoided harm is the most clinically compelling but methodologically difficult value claim. AI-based medication-alert systems may prioritize alerts, identify potentially inappropriate prescriptions, or reduce low-value interruption, but available evaluations frequently emphasize alert acceptance, alert burden, or classification performance rather than confirmed patient outcomes [11]. These measures can support an avoided-harm hypothesis, yet they do not establish the complete causal pathway.

The proposed pathway begins with an AI-generated signal that identifies a clinically relevant medication risk. A pharmacist or other authorized professional must then review the output, interpret it in context, and decide whether to accept, modify, defer, or reject the implied action. Prescribing-error systems have identified clinically valid alerts and estimated possible cost consequences, but an identified or corrected error remains different from an observed adverse event that was prevented [12]. The term "avoided harm" should therefore be reserved for claims supported by evidence extending beyond detection.

Prescription-triage systems illustrate the distinction. A system may classify prescriptions as unlikely to require detailed review, thereby changing how pharmacists allocate attention. Such an intervention may release capacity, but false-negative decisions can remove professional review from prescriptions that contain clinically meaningful problems [13]. Evaluation must consequently examine both the work avoided and the risks created by omission. A high proportion of correctly deprioritized prescriptions cannot compensate automatically for a small number of severe missed hazards.

The next transition is from professional action to changed medication use. A recommendation that is viewed but not acted upon cannot plausibly prevent harm. Even an accepted intervention may change dose, medicine, timing, monitoring, or counselling without producing a measurable clinical consequence. Conversely, the absence of an adverse event does not prove prevention because the event may not have occurred without intervention. The counterfactual question is therefore central: what would probably have happened to a comparable patient, prescription, or medication-use episode without the AI-influenced action?

Severity must also be represented. Medication errors have heterogeneous clinical and economic consequences, ranging from negligible effects to prolonged morbidity, hospitalization, or death [14]. Counting all corrected errors equally can exaggerate value by treating minor documentation changes and prevention of serious injury as equivalent. The framework does not prescribe numerical

weights; it requires transparent classification of the potential harm, evidentiary basis, and uncertainty attached to the counterfactual estimate.

Four levels of avoided-harm evidence are proposed. First, signal evidence establishes that the system detected a relevant risk. Second, action evidence establishes that the signal changed professional or patient behavior. Third, medication-use evidence establishes that exposure, monitoring, access, or administration changed. Fourth, outcome and attribution evidence evaluates whether clinically meaningful harm was reduced relative to an appropriate comparator. Claims should identify the highest level actually observed rather than presenting downstream benefit as established.

This approach also preserves professional responsibility. AI may direct attention or reveal patterns, but it does not independently determine the appropriateness of a medication decision. Contextual information, competing risks, treatment goals, uncertainty, and patient preferences may justify rejecting an apparently accurate recommendation. Avoided-harm evaluation should therefore examine appropriate acceptance, appropriate rejection, harmful acceptance, and harmful rejection. Without these distinctions, high alert acceptance may reward compliance rather than safe judgment.

Recovered-Time and Better-Decision Domains

Recovered time should be distinguished from theoretical automation, faster model execution, reduced clicks, shorter task duration, workforce reduction, and financial saving. Gross task-time reduction is only the first component. Verification of outputs, correction of errors, escalation, documentation, staff training, maintenance, monitoring, and coordination may create new work that is distributed across different people or periods. Net recovered time is the remaining usable capacity after these burdens have been deducted.

Human-assistance studies demonstrate why task consequences cannot be assumed from model performance. AI support can alter professional classification performance, but the magnitude and direction of change may differ by user, case difficulty, and baseline expertise [15]. Human–computer collaboration may outperform either component in some conditions, although performance depends on interaction design and the relative strengths of the user and system [16]. These relationships support evaluation of the combined work system rather than independent comparison of pharmacist and algorithm.

Better decisions require an additional distinction between decision process and final decision. A process may improve when relevant information is detected earlier, uncertainty is communicated more clearly, priorities become more consistent, or escalation becomes more appropriate. The final decision may nevertheless remain unchanged. Conversely, a decision may change without becoming better. Clinicians can be influenced by erroneous advice, including recommendations that contradict available evidence [17]. Agreement with AI, override rate, or recommendation acceptance cannot therefore function alone as a decision-quality measure.

Randomized evidence involving large language model assistance similarly indicates that access to an AI tool and improvement in diagnostic reasoning or final performance are separable outcomes [18]. For pharmacy evaluation, better decisions should therefore be assessed through appropriateness, safety, timeliness, consistency with treatment goals, uncertainty handling, and patient-relevant consequences. The assessor should be sufficiently independent of the intervention, and the evaluation should include cases in which the model is wrong, uncertain, unavailable, or discordant with professional judgment.

Observed human–AI performance depends on the task, user, interaction design, and correctness of the advice rather than arising automatically from access to an AI assistant [15–17]. The proposed framework accordingly distinguishes four time states: gross time reduced, new work introduced, net time recovered, and time destination. Time may be absorbed by rising workload, removed from staffing, reserved as resilience, or reinvested in medication review, counselling, monitoring, coordination, or access. These destinations have different consequences and should not be represented by one generic efficiency claim.

Recovered time and better decisions may interact but should remain analytically separate. Faster prioritization can give pharmacists more time for complex decisions; it can also compress review, increase throughput pressure, or shift attention away from cases misclassified as low risk. Better information can reduce search time while increasing deliberation because uncertainty becomes more visible. The framework therefore proposes parallel measurement of time, workload, decision quality, safety, and destination of released capacity.

Neither domain alone establishes net value. Recovered time is valuable only when its use is known and when new burdens and safety effects are acceptable. A better decision process is valuable only when it improves or preserves appropriate action without generating disproportionate delay, cognitive burden, inequity, or harmful reliance. Both domains require empirical validation in the intended pharmacy task and setting.

Proposed Multidimensional Evaluation Framework

The proposed framework treats pharmacy AI as a sociotechnical intervention rather than an isolated model. Human and machine capabilities may be complementary, but value depends on how responsibilities, information, uncertainty, and authority are distributed across the combined system [19]. The framework therefore evaluates a sequence

of connected layers rather than asking whether the algorithm performs well in isolation.

The first layer is baseline need: a defined medication-use problem, affected population, existing workflow, burden, and comparator. The second is technical fitness, including validity for the intended data, setting, and decision threshold. The third is pharmacy-task performance, where evaluation asks whether the output improves or preserves a specified activity such as screening, prioritization, verification, review, monitoring, or counselling. The fourth layer concerns human–AI and workflow interaction, including output presentation, uncertainty communication, professional interpretation, override, escalation, and newly generated work. Early clinical evaluation should document these users, interactions, errors, and contextual conditions rather than report technical performance alone [20].

The fifth layer comprises the three consequence domains: avoided harm, recovered and reinvested time, and better decisions. Evidence should identify the furthest demonstrated point in each pathway. The sixth layer is attribution, which examines whether observed consequences could instead reflect staffing changes, education, workflow redesign, case mix, secular trends, or concurrent interventions. The seventh is net, distributed, and sustained value, incorporating implementation burden, opportunity cost, subgroup effects, unintended consequences, and benefit persistence.

For confirmatory evaluation, intervention characteristics, human involvement, errors, outcomes, and analytic methods require transparent reporting [21]. Prospective protocols should also prespecify intended use, output handling, user interaction, failure management, and outcome assessment [22]. These requirements support three proposed rules. First, favorable evidence at an upstream layer cannot substitute for missing downstream evidence. Second, severe failure in safety, equity, or professional control is not compensable by efficiency elsewhere. Third, the strength of a value claim should not exceed the weakest unverified transition on which it depends.

Seven proposed transition gates follow: problem legitimacy, technical fitness, task safety, workflow fitness, consequence evidence, attribution credibility, and net sustained value. Failure at a gate signifies an unresolved evidence requirement rather than automatic rejection. The framework is not a certification system, and its layers, gates, and relationships require prospective testing in pharmacy practice.

Study Designs, Comparators, and Attribution Challenges

Study design should be selected according to the consequence claimed. Randomized evaluations offer strong protection against measured and unmeasured confounding, but randomized evidence for clinical machine-learning interventions remains limited and heterogeneous [23]. Cluster randomization or stepped-wedge designs may be appropriate when an intervention changes team-level workflow, although learning, contamination, and temporal change must still be addressed.

Pharmacy AI also behaves as a complex intervention: effects emerge through interaction among the system, users, organizational routines, implementation strategy, and local context [24]. Silent-mode studies can establish technical and workflow feasibility before outputs affect decisions. Prospective task studies can evaluate pharmacist performance and reliance. Controlled before–after designs may compare matched services, while interrupted time-series analysis may estimate changes in outcome level or trend when randomization is impractical [25].

The comparator must represent the decision actually faced. Relevant alternatives may include usual practice, an existing rule-based alert, an additional pharmacist, redesigned workflow, or no intervention. Historical controls are vulnerable when medication mix, staffing, documentation, or safety initiatives change. Generalisability also cannot be inferred from one dataset, institution, professional group, or implementation period [26].

Credible attribution requires attention to intervention–context interaction, temporal alternative explanations, and transportability across populations and settings [24–26]. Evaluation should consequently document co-interventions, case mix, staffing, implementation intensity, learning effects, and workflow changes. No design removes every source of uncertainty; the permissible value claim must remain proportional to the comparator, measurement quality, and causal assumptions.

Implications for Procurement and Implementation

Procurement should begin with the medication-use problem and required consequence rather than a vendor's performance metric. End-to-end implementation frameworks emphasize that problem selection, data, model behavior, workflow, governance, and monitoring are interdependent [27]. A procurement dossier should therefore specify intended users, task boundaries, baseline burden, comparator, evidence maturity, uncertainty, human authority, total resource requirements, monitoring arrangements, and exit conditions.

Implementation may redesign the service in which the technology operates. Benefits attributed to AI may partly arise from new staffing, standardized processes, additional training, or altered communication pathways [28]. These changes should be measured rather than treated as background conditions. Health technology assessment of adaptive AI also requires attention to organizational, ethical, lifecycle, and evidentiary characteristics that may not be captured by conventional one-time appraisal [29]. Predictive performance should therefore be linked to clinical benefit and governance rather than accepted as a stand-alone purchasing justification [30].

Figure 2 presents the original conceptual synthesis for linking avoided harm, recovered time, better decisions, and attribution, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

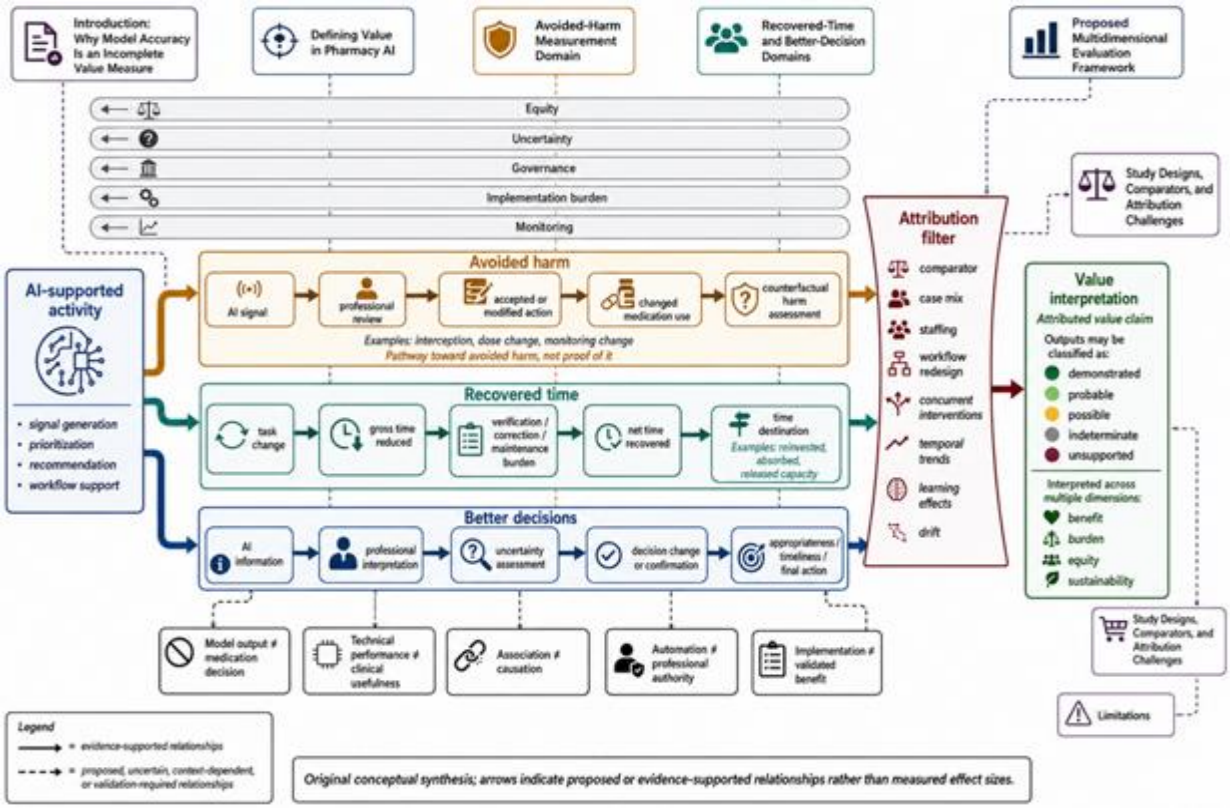

Figure 2. Linking Avoided Harm, Recovered Time, Better Decisions, and Attribution

Postimplementation governance should compare observed performance and consequences with the original intended use, baseline, and procurement claims. Reassessment is warranted when task boundaries change, outputs drift, subgroup performance deteriorates, workarounds emerge, professional reliance changes, or verification burden offsets expected benefit. Continuation, modification, restriction, or discontinuation should remain available responses rather than assuming that initial implementation creates permanent value.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for governance responsibilities, implementation conditions, and monitoring requirements for the article Measuring Pharmacy Artificial Intelligence by Avoided Harm, Recovered Time, and Better Decisions.

Table 2. Governance responsibilities, implementation conditions, and monitoring requirements for the article Measuring Pharmacy Artificial Intelligence by Avoided Harm, Recovered Time, and Better Decisions.

Governance domain	Responsible actor	Required authority	Documentation requirement	Monitoring indicator	Escalation trigger	Corrective action	Accountability boundary
Problem and intended-use definition	Pharmacy, clinical, patient, and operational representatives	Approve problem, population, task, and exclusions	Baseline burden, comparator, intended consequence	Continued relevance of need and use case	Use expands beyond documented scope	Restrict scope or repeat evaluation	AI availability does not establish legitimate need
Technical oversight	Developer and deploying organization	Approve version, data inputs, and thresholds	Validation, calibration, subgroup results, version history	Drift, missingness, calibration, output failure	Material deterioration or undocumented change	Recalibrate, suspend, or revalidate	Technical approval does not establish clinical value

Professional-use governance	Pharmacy leadership and authorized users	Define review, override, escalation, and final authority	Workflow map, training, uncertainty and override rules	Reliance, overrides, harmful reversals, alert burden	Unsafe reliance or loss of effective professional control	Redesign interface, retrain, or restrict use	Professional accountability cannot be transferred implicitly to the model
Safety and avoided-harm review	Medication-safety and clinical-governance bodies	Investigate incidents and approve safety responses	Signal-to-action and outcome evidence	Missed risks, harmful recommendations, intervention consequences	Serious event or repeated unsafe pattern	Incident review, restriction, or suspension	Detected errors are not automatically harms avoided
Time and workload review	Pharmacy operations and workforce leads	Approve workload and capacity claims	Gross time, new work, net time, time destination	Verification burden, delays, workload transfer	New work offsets or exceeds released time	Redesign task allocation or revise value claim	Task acceleration does not establish workforce saving
Equity review	Clinical, patient, data, and equality governance	Require subgroup evaluation and mitigation	Population coverage, subgroup errors, access effects	Unequal errors, exclusions, or benefit distribution	Material subgroup disadvantage	Investigate data, thresholds, workflow, or access	Positive average results do not establish equitable value
Procurement and economic review	Procurement, pharmacy, finance, and governance	Approve purchase, renewal, and contractual evidence terms	Total cost, comparator, uncertainty, maintenance and exit terms	Realized consequences and lifecycle costs	Vendor claim is unsupported or value declines	Renegotiate, restrict, or discontinue	Procurement is not clinical validation
Lifecycle monitoring and change control	Multidisciplinary AI-governance body	Require reassessment after material change	Monitoring plan, incidents, changes, decisions	Drift, new failure modes, sustainability	Model, data, task, population, or workflow changes	Revalidate, modify, suspend, or withdraw	Initial acceptance does not authorize indefinite unchanged use

The responsibilities in **Table 2** are proposed governance functions rather than fixed legal assignments. Local structures may distribute them differently, but responsibility for evidence review, monitoring, escalation, and corrective action should remain explicit.

Limitations

The framework is conceptual and has not been prospectively validated. Proxy targets may generate apparent benefit while distributing resources inequitably, as demonstrated when an algorithm's cost-based target failed to represent clinical need equally across racial groups [31]. Explainability tools may also create unwarranted confidence because post hoc explanations do not necessarily reveal faithful model reasoning or establish safety [32]. Aggregate performance may conceal systematic underdiagnosis in underserved populations [33].

The framework may be difficult to operationalize where baseline data, reliable comparators, patient outcomes, or workflow measurements are unavailable. Some consequences cannot be reduced to a common unit, and monetization may be ethically or methodologically contested. AI can also introduce complacency, deskilling, distribution shift, and unforeseen dependencies [34]. The proposed domains should therefore remain separate where aggregation would conceal critical trade-offs.

Conclusion

Pharmacy AI should not be valued by model accuracy, alert volume, adoption, or task speed alone. Its value depends on whether a defined medication-use problem is addressed through technically fit and contextually appropriate support that improves pharmacy work without creating disproportionate safety, cognitive, organizational, or equity burdens.

This article proposes an integrated framework linking baseline need, technical fitness, pharmacy-task performance, human–AI interaction, avoided harm, recovered and reinvested time, better decisions, attribution, and net sustained value. It distinguishes signals from actions, corrected errors from avoided harm, gross time reduction from usable capacity, decision-process improvement from better final decisions, and observed change from attributable benefit.

The framework also places professional authority, uncertainty, subgroup effects, implementation burden, and lifecycle monitoring inside the value assessment rather than treating them as peripheral considerations. Procurement and continuation decisions should therefore be based on evidence extending as far along the relevant consequence pathway as the claim being made.

The proposed framework does not determine universal thresholds, combine all benefits into a single score, certify technologies, or establish deployment readiness. Its contribution is a structured and testable basis for asking what pharmacy AI changes, for whom, through which mechanism,

at what cost and risk, and with what degree of causal confidence.

ACKNOWLEDGMENTS: None
CONFLICT OF INTEREST: None
FINANCIAL SUPPORT: None
ETHICS STATEMENT: None

References

1. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
2. 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;368. doi:10.1136/bmj.m689
3. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
4. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368. doi:10.1136/bmj.l6927
5. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digit Health. 2019;1(6). doi:10.1016/S2589-7500(19)30123-2
6. Hatzimanolis J, Riley B, El-Den S, Aslani P, Zhou J, Chaar BB. Applications of artificial intelligence in current pharmacy practice: A scoping review. Res Social Adm Pharm. 2025;21(3):134-41. doi:10.1016/j.sapharm.2024.12.007
7. Lakdawalla DN, Doshi JA, Garrison LP Jr, Phelps CE, Basu A, Danzon PM. Defining elements of value in health care—a health economics approach: An ISPOR Special Task Force report [3]. Value Health. 2018;21(2):131-9. doi:10.1016/j.jval.2017.12.007
8. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: A path to validation. NPJ Digit Med. 2019;2(1):38. doi:10.1038/s41746-019-0111-3
9. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res. 2017;19(11). doi:10.2196/jmir.8775
10. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: Updated reporting guidance for health economic evaluations. Value Health. 2022;25(1):3-9. doi:10.1016/j.jval.2021.11.1351
11. Graafsma J, Murphy RM, van de Garde EMW, Karapinar-Çarkit F, Derijks HJ, Hoge RHL, et al. The use of artificial intelligence to optimize medication alerts generated by clinical decision support systems: A scoping review. J Am Med Inform Assoc. 2024;31(6):1411-22. doi:10.1093/jamia/ocae076
12. Rozenblum R, Rodriguez-Monguio R, Volk LA, Forsythe KJ, Myers S, McGurrin M, et al. Using a machine learning system to identify and prevent medication prescribing errors: A clinical and cost analysis evaluation. Jt Comm J Qual Patient Saf. 2020;46(1):3-10. doi:10.1016/j.jcjq.2019.09.008
13. Levivien C, Cavagna P, Grah A, Buronfosse A, Courseau R, Bézie Y, et al. Assessment of a hybrid decision support system using machine learning with artificial intelligence to safely rule out prescriptions from medication review in daily practice. Int J Clin Pharm. 2022;44(2):459-65. doi:10.1007/s11096-021-01366-4
14. Elliott RA, Camacho E, Jankovic D, Sculpher MJ, Faria R. Economic analysis of the prevalence and clinical and economic burden of medication error in England. BMJ Qual Saf. 2021;30(2):96-105. doi:10.1136/bmjqs-2019-010206
15. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med. 2020;3(1):23. doi:10.1038/s41746-020-0232-8
16. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229-34. doi:10.1038/s41591-020-0942-0
17. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: Susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 2021;4(1):31. doi:10.1038/s41746-021-00385-9
18. 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;7(10). doi:10.1001/jamanetworkopen.2024.40969
19. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
20. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9
21. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan AW, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med. 2020;26(9):1364-74. doi:10.1038/s41591-020-1034-x
22. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafian H, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nat Med. 2020;26(9):1351-63. doi:10.1038/s41591-020-1037-7
23. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: A systematic review. JAMA Netw Open. 2022;5(9). doi:10.1001/jamanetworkopen.2022.33946
24. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. BMJ. 2021;374. doi:10.1136/bmj.n2061
25. Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: A tutorial. Int J Epidemiol. 2017;46(1):348-55. doi:10.1093/ije/dyw098
26. 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;2(9). doi:10.1016/S2589-7500(20)30186-2
27. van der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Implementation frameworks for end-to-end clinical AI: Derivation of the SALIENT framework. J Am Med Inform Assoc. 2023;30(9):1503-15. doi:10.1093/jamia/ocad088
28. Shaw J, Agarwal P, Desveaux L, Palma DC, Stamenova V, Jamieson T, et al. Beyond "implementation": Digital health innovation and service design. NPJ Digit Med. 2018;1(1):48. doi:10.1038/s41746-018-0059-8
29. Farah L, Borget I, Martelli N, Vallee A. Suitability of the current health technology assessment of innovative artificial intelligence-based medical devices: Scoping literature review. J Med Internet Res. 2024;26. doi:10.2196/51514
30. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science. 2019;363(6429):810-2. doi:10.1126/science.aaw0029
31. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. doi:10.1126/science.aax2342
32. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9
33. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. Nat Med. 2021;27(12):2176-82. doi:10.1038/s41591-021-01595-0
34. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8. doi:10.1001/jama.2017.7797