Original Article

Why Better Predictions Can Still Produce Worse Pharmacy Decisions

Emma Wilson[1]*, Frederik De Jong[2], Lara Peters[2]

[1]Department of Pharmacy Decision Science and AI, Faculty of Pharmacy, University of Edinburgh, Edinburgh, United Kingdom. [2]Department of Predictive Analytics and Decision Quality, Faculty of Pharmaceutical Sciences, Utrecht University, Utrecht, Netherlands.

Abstract

Artificial intelligence in pharmacy is frequently evaluated through prediction-centred measures, including discrimination, calibration, sensitivity, specificity, and alert reduction. These measures are necessary for characterizing model behaviour but do not establish whether the model improves medication-related decisions. A prediction becomes consequential only when it is interpreted by a professional, connected to an available action, introduced at an appropriate time, and applied within a workflow in which errors have asymmetric consequences. This article critically examines the prediction-to-decision translation gap: the possibility that technically improved predictions may produce unchanged or worse pharmacy decisions. Six mechanisms are distinguished: target mismatch, automation bias, treatment leakage, poor timing, actionability failure, and unequal error consequences. Decision usefulness is proposed as a conditional property arising from alignment among the prediction target, patient and medication context, uncertainty communication, professional interpretation, action feasibility, and the consequences of acting or not acting. The article develops a proposed Prediction–Interpretation–Action Framework that separates model output from professional interpretation and medication action through three translation gates: target validity, interpretation integrity, and action suitability. Evaluation should therefore extend beyond model performance to include human–AI interaction, decision consequences, workflow burden, medication safety, subgroup effects, governance, and post-implementation change. The framework is an original conceptual synthesis rather than a validated clinical model, procurement standard, or regulatory instrument. Its propositions require prospective testing across pharmacy settings, medication tasks, professional roles, patient populations, and technology configurations before decision readiness can be inferred.

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

Introduction

Separating Predictive Performance from Decision Quality

Artificial intelligence has expanded the range of medication-related events that can be predicted, classified, prioritized, or converted into clinical decision-support alerts. Pharmacy applications may identify prescribing anomalies, estimate adverse-event risk, rank patients for review, detect potential interactions, or reduce apparently low-value alerts. Yet predictive success is not equivalent to decision success. Clinical impact depends on whether a model is incorporated into an appropriate care pathway, reaches the intended user, changes a relevant action, and produces consequences that are preferable to those arising without the model [1]. A technically better prediction can therefore remain clinically neutral or become harmful when the surrounding decision process is poorly specified.

This distinction requires the analytical unit to expand beyond the algorithm. Responsible health-care machine learning depends on choices made across problem formulation, data generation, model development, validation, implementation, monitoring, and stakeholder governance [2]. In pharmacy practice, these choices interact with professional judgment, medication histories, treatment intent, patient preferences, workload, communication pathways, and the authority to intervene. The resulting decision cannot be attributed solely to either the model or pharmacist. It is produced by a sociotechnical arrangement in which model behaviour, human interpretation, workflow design, and organizational controls may amplify or counteract one another.

Address for correspondence: Emma Wilson, Department of Pharmacy Decision Science and AI, Faculty of Pharmacy, University of Edinburgh, Edinburgh, United Kingdom.
emma.wilson@ed.ac.uk
Received: 01 February 2025; **Accepted:** 16 May 2025

How to cite this article: Wilson E, De Jong F, Peters L. Why Better Predictions Can Still Produce Worse Pharmacy Decisions. Arch Pharm Pract. 2025;16(2):44-52. https://doi.org/10.51847/699h8jFRoX

Even within technical evaluation, apparently strong performance may conceal limitations relevant to action. Discrimination concerns the ranking or separation of patients with different outcomes, whereas calibration concerns agreement between predicted and observed probabilities [3]. A model may discriminate well while systematically overstating or understating risk. Conversely, an adequately calibrated population estimate may remain unhelpful when the available action is ineffective, burdensome, delayed, or associated with substantial false-positive consequences. Predictive performance should therefore be treated as one input into decision evaluation rather than a substitute for it.

The distinction is especially important in medication-alert systems. Artificial intelligence has been used to prioritize or suppress clinical decision-support alerts, but the available literature contains heterogeneous objectives, outcome definitions, validation practices, and implementation evidence [4]. Reducing alert volume or improving classification does not establish that important medication problems are detected, that inappropriate interventions are avoided, or that pharmacists can respond effectively. The central problem addressed in this article is consequently not whether prediction can be improved, but under which conditions an improved prediction can be translated into a better pharmacy decision.

This article proposes the concept of a prediction-to-decision translation gap and develops an original Prediction–Interpretation–Action Framework for examining it. The framework does not assume that artificial intelligence is intrinsically beneficial or harmful. Instead, it identifies the conditions, failure mechanisms, evaluation domains, and governance boundaries that determine whether predictive information can support medication-related judgment without displacing context, accountability, or professional authority.

The Prediction-to-Decision Translation Gap

The prediction-to-decision translation gap is proposed as the discordance between measured predictive performance and the quality of the decision produced when a prediction is introduced into pharmacy practice. The gap can emerge even when the reported performance estimate is technically correct within its development or evaluation setting. Independent assessment of a widely implemented clinical prediction model has shown that deployment and broad adoption do not guarantee that reported performance will be reproduced in another population or institution [5]. The relevant pharmacy question is therefore not only whether a model predicts accurately somewhere, but whether it supplies dependable information for a defined medication decision in the setting where it will be used.

Generalisability should not be treated as an unrestricted property that a model either possesses or lacks. Clinical usefulness is conditional on the relationship among the development data, local patient population, documentation practices, available treatments, professional users, and care processes [6]. A prediction may be statistically transportable but operationally irrelevant because the receiving pharmacy uses different thresholds, lacks access to the required contextual information, or cannot perform the action assumed by the model developers. Conversely, a model with limited geographic transportability may still be useful within a narrowly defined and continuously monitored local workflow.

Variation across settings can also arise because algorithms learn signals associated with institutional practice rather than stable clinical relationships. A model evaluated across hospitals may perform differently when local acquisition processes, coding practices, patient characteristics, or treatment patterns change [7].

The translation gap is further widened when the model is treated as if it directly produces the decision. Clinical decision-support systems operate through multiple connected elements, including a knowledge or model base, an inference process, an interface, a user, and a workflow in which information is presented and acted upon [8]. A prediction can be lost, misunderstood, delayed, duplicated, overridden, or accepted without sufficient verification at any point in this chain.

Decision quality is therefore proposed as the expected appropriateness of an action after considering medication benefit, preventable harm, delay, workload, patient burden, opportunity cost, and the consequences of error. It is distinct from clinical outcome because a reasonable decision can still produce an adverse outcome under uncertainty, and an inappropriate decision can occasionally be followed by a favourable outcome. It is also distinct from adherence to an alert or recommendation. Acceptance may reflect appropriate agreement, automation bias, time pressure, or absence of a feasible alternative, whereas override may reflect either expert correction or unsafe dismissal.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses for the article Why Better Predictions Can Still Produce Worse Pharmacy Decisions.

Table 1. Definitions, components, relationships, and intended uses for the article Why Better Predictions Can Still Produce Worse Pharmacy Decisions.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Predictive performance	A single model metric may be mistaken for decision quality.	Target definition, discrimination, calibration, threshold, evaluation population	Evidence-supported distinction: discrimination and calibration characterize different aspects of prediction [3]. Proposed relationship: both remain upstream of decision usefulness.	Clarifies what technical evaluation can and cannot establish.	Internal, temporal, and external performance assessment using decision-relevant data	Strong ranking may coexist with unreliable risk estimates or unsuitable thresholds.	Predictive performance does not establish medication benefit, safety, or workflow value.
Prediction-to-decision translation gap	Reported performance may not be reproduced or converted into useful action locally.	Local population, data systems, workflow, users, interventions, implementation conditions	Evidence indicates that implemented models may perform differently under independent evaluation [5]. The formal translation-gap construct is proposed in this article.	Directs evaluation toward the pathway between output and action.	Local validation followed by prospective evaluation of decisions and consequences	The gap may be attributed incorrectly to the model when workflow or implementation is responsible.	The construct does not imply that every performance difference causes harm.
Contextual usefulness	General performance claims may obscure setting-specific requirements.	Patient mix, clinical practices, resources, professional roles, intended use	Evidence-supported relationship: usefulness is conditional on local context rather than universal generalisability [6].	Makes setting, user, and decision explicit.	Assessment across relevant populations, institutions, roles, and workflow conditions	A locally useful system may fail elsewhere; a broadly validated system may still be locally unsuitable.	Local usefulness does not establish universal applicability.
Data and institutional shift	Models may learn setting-associated signals that change across sites.	Documentation, formulary, coding, acquisition processes, treatment patterns	Cross-site performance variation can arise when data-generating processes differ [7].	Supports transportability and drift analysis.	Multisite, temporal, and subgroup evaluation with examination of changing data sources	Apparent clinical prediction may partly reflect institutional practice.	Evidence from one task or institution cannot determine pharmacy performance elsewhere.
Decision-support pathway	Treating the algorithm as the whole intervention obscures human and workflow effects.	Model, inference engine, interface, user, timing, communication, workflow	Evidence-supported relationship: decision support is a multicomponent system [8]. Proposed relationship: failures in any component may interrupt or distort translation.	Establishes the combined human–technology workflow as the unit of analysis.	Usability, human-factors, workflow, safety, and implementation evaluation	Interface design, workload, or role ambiguity may dominate model performance.	A validated component does not validate the complete system.
Actionability	A correct prediction may identify a problem for which no beneficial response is available.	Available interventions, authority, timing, reversibility, resources	Proposed relationship: decision usefulness is conditional on the existence of a feasible and beneficial action.	Prevents risk detection from being equated with improved care.	Prospective evidence connecting outputs to appropriate actions and consequences	Additional alerts may increase workload, delay, or unnecessary intervention.	Actionability must be established for the specific user and setting.
Decision quality	Alert acceptance or outcome occurrence may be misclassified as evidence of a good decision.	Patient context, evidence, uncertainty, alternatives, error consequences	Proposed definition: expected appropriateness of the selected action considering benefits, harms, burden, delay, and opportunity cost.	Supplies the outcome construct for the proposed framework.	Transparent decision criteria, expert review, patient-relevant outcomes, and consequence-sensitive evaluation	Reasonable decisions may have adverse outcomes; favourable outcomes may follow poor decisions.	Decision quality is context-dependent and is not reducible to adherence to AI advice.
Governance and feedback	Initial validation may be treated as permanent authorization for use.	Ownership, monitoring, incidents, updates, workflow change, model change	Evidence supports lifecycle attention beyond initial development [2]. Proposed relationship: observed consequences should trigger reassessment of the prediction, interpretation,	Connects implementation to accountability and change control.	Ongoing performance, safety, equity, workflow, and reliance monitoring	Unmonitored changes can invalidate prior evidence without obvious technical failure.	Governance does not itself prove clinical effectiveness or regulatory acceptability.

action pathway, or
continued use.

Mechanisms through Which Accurate Predictions Cause Harm

Target Mismatch

A model can predict its specified target accurately while supporting the wrong decision because the modeled quantity is not equivalent to the underlying patient need. This failure is especially important when a convenient proxy replaces a clinically meaningful construct. A population-management algorithm, for example, produced unequal allocation because expenditure was used as a proxy for health need, even though access and spending patterns differed across groups [9]. In pharmacy, analogous mismatch could arise when the model predicts prior intervention, medication cost, alert acceptance, documentation activity, or health-care utilization rather than preventable medication harm or expected benefit from pharmacist action. Improving prediction of the proxy may then strengthen the wrong allocation rule.

Target mismatch differs from ordinary model error. The output may be statistically accurate for the selected label, yet the label may be inappropriate for the decision. The relevant validation question is therefore not only whether the outcome is predicted correctly, but whether predicting that outcome identifies patients for whom the contemplated medication action is beneficial. This relationship is task-specific and must not be inferred merely because the target is clinically associated with risk.

Automation Bias

Automation bias occurs when users give disproportionate weight to computerized advice, resulting in omission errors when the system fails to identify a problem or commission errors when incorrect advice is followed. Verification becomes more difficult when the underlying reasoning is complex, the user is under time pressure, or the system appears consistently reliable [10]. Better average performance can paradoxically increase this risk if it encourages users to reduce independent scrutiny while rare but consequential errors remain.

Treatment Leakage

Treatment leakage arises when model inputs contain information generated by treatment, clinical response, or decisions occurring after the point at which the prediction is intended to support action. Leakage cannot be identified solely by checking whether a variable has a timestamp before the recorded outcome. Its meaning depends on prediction cadence, observer perspective, causal position, and intended applicability [11]. A variable may be legitimate for monitoring an evolving patient but inappropriate for a model intended to guide the earlier treatment decision that produced it.

In pharmacy AI, leakage may involve laboratory tests ordered after clinical deterioration, medication changes initiated in response to suspected risk, pharmacist documentation created after intervention, or utilization patterns that encode previous professional concern. Such variables can produce impressive retrospective performance while being unavailable, altered, or causally downstream at the intended decision point. Deployment under those conditions may generate unstable predictions, false reassurance, or delayed recognition that the model has lost the information on which its apparent performance depended.

Poor Timing

A clinically meaningful target can still be predicted at the wrong time. Outputs delivered after dispensing, after irreversible administration, or after the relevant monitoring window may be accurate but non-actionable. Predictions generated too early may also be weakly connected to the later decision because treatment, physiology, or patient circumstances can change. Timing should therefore be evaluated as alignment among the prediction horizon, information-availability horizon, professional review point, and action horizon. A model cannot be considered decision-useful merely because the predicted event eventually occurs.

Actionability Failure

Some systems identify elevated risk without establishing what the pharmacist can do differently. Alert optimization research illustrates that improved prioritization or reduced alert burden does not by itself demonstrate that retained alerts lead to appropriate interventions or better medication outcomes [4]. Actionability may fail because no effective response exists, the user lacks authority, the patient cannot be contacted, the prescriber is unavailable, required monitoring is inaccessible, or the recommended response imposes greater burden than the predicted risk justifies.

Unequal Error Consequences and Reliance-Mediated Harm

Equivalent error rates do not imply equivalent harm. False-positive alerts may create disproportionate treatment interruption, surveillance, cost, or burden for some patients, whereas false negatives may deny review to those already facing limited access. Target mismatch can magnify these differences by allocating pharmacy attention according to variables shaped by existing inequities [9]. Evaluation should consequently examine who receives the benefit of correct predictions and who bears the consequences of error.

Human interpretation can further amplify unequal or biased outputs. Experimental evidence indicates that clinicians exposed to systematically biased artificial-intelligence advice can make less accurate decisions, and the presence of an explanation does not necessarily neutralize that effect [12].

This finding does not establish the magnitude of harm in pharmacy practice, but it shows why model evaluation cannot assume that a professional user will consistently detect or correct erroneous assistance. Accurate average predictions may therefore coexist with worse individual decisions when rare errors are persuasive, context is incomplete, or error consequences are asymmetric.

Conditions Required for Decision Usefulness

Decision usefulness is proposed as the conditional capacity of a prediction to improve a defined medication-related decision for a particular user, patient, setting, time point, and feasible action set. It is not an intrinsic attribute of an algorithm. The same prediction may be useful during inpatient medication review, irrelevant after discharge, and harmful when delivered to a user who lacks the information or authority required to interpret it. Decision usefulness must therefore be assessed at the level of the combined prediction–user–workflow–action system.

Meaningful Uncertainty

The first condition is that uncertainty must be represented in a form that supports action. A point estimate or categorical label can create an appearance of certainty that the available evidence does not warrant. Communicating uncertainty can help users identify cases requiring additional information, specialist review, deferral, or non-automated assessment [13]. However, merely displaying a probability interval is not necessarily useful. The representation must correspond to the source of uncertainty, be understandable at the decision point, and connect to a permissible response.

Uncertainty also includes limits that are not captured by a numerical confidence measure, such as unfamiliar patient characteristics, missing contextual data, suspected distribution shift, or disagreement between the model and medication history. A useful system should allow the pharmacist to recognize these conditions and should not convert every uncertain estimate into a mandatory binary action.

Interpretable Meaning without Explanation Overreach

The second condition is that users must understand what the output represents, which decision it is intended to inform, and which conclusions cannot be drawn from it. Post-hoc explanations may highlight variables associated with a prediction, but they do not automatically establish that the model is reliable, that the highlighted features are causal, or that the proposed action is safe [14]. Explanation should therefore support interrogation rather than serve as a visual assurance that the model is trustworthy.

For pharmacy decisions, interpretability includes clarity about the prediction target, reference population, time horizon, required inputs, treatment context, threshold, and foreseeable failure modes. It also requires access to patient-specific evidence that may contradict the output. An explanation that is technically plausible but disconnected from the medication question can strengthen inappropriate reliance rather than improve interpretation.

Calibrated Trust and Contestability

The third condition is calibrated trust. Trust should correspond to evidence of capability in the relevant task and setting rather than to novelty, visual polish, vendor reputation, or generalized confidence in artificial intelligence. Both excessive and insufficient trust may undermine performance: overtrust can produce passive acceptance, whereas blanket distrust can cause useful information to be ignored [15]. The appropriate objective is therefore not maximum adoption but context-sensitive reliance.

Contestability is proposed as a related requirement. Pharmacists should be able to challenge the output, obtain the underlying patient information, document disagreement, select an alternative action, and escalate uncertainty without being treated automatically as non-compliant. Override rates alone cannot distinguish expert correction from alert fatigue, just as acceptance rates cannot distinguish valid agreement from automation bias.

Target Congruence, Timing, and Action Suitability

A decision-useful prediction must address the quantity that matters for the contemplated action. Reliable risk estimation is insufficient when risk does not identify who can benefit from intervention. The target should therefore correspond to a medication-relevant decision, and the prediction horizon should match the period during which the pharmacist can act. Input variables must also be available without relying on information generated after the intended decision.

Action suitability requires at least one feasible response whose expected benefits justify its burdens and error consequences. The response must be available within the workflow, compatible with professional authority, and sufficiently reversible or monitorable when uncertainty is substantial. Deliberate non-action should remain available when intervention would produce greater harm than observation or reassessment.

Context, Accountability, and Evidential Sufficiency

Patient and medication context must remain visible throughout interpretation. Indication, treatment goals, duration, previous response, organ function, interacting therapies, adherence, patient preferences, and access constraints can change the meaning of the same prediction. Digital pharmacy systems that remove these factors from view may increase standardization while decreasing clinical relevance.

Accountability must also be allocated across development, procurement, implementation, professional use, monitoring, and system change. Human oversight should not be invoked

as a general solution when the pharmacist lacks time, information, training, or authority to correct the model. Conversely, the presence of automation does not remove professional responsibility for contextual judgment. The proposed requirement is a documented distribution of responsibility that specifies who validates the target, monitors system behaviour, reviews incidents, manages updates, and decides whether use should be restricted or suspended.

Finally, evidence must match the claim being made. Technical validation can support statements about prediction under specified conditions. Human-factors studies can support statements about comprehension or reliance. Prospective workflow evaluation can examine whether the system changes actions, workload, delays, or error recovery. Medication-safety and patient-outcome claims require evidence at those corresponding levels. No single metric, interface assessment, expert review, or retrospective analysis can establish the complete chain from better prediction to better pharmacy decision.

The conditions developed here are proposed as necessary evaluative questions rather than empirically proven sufficient criteria. Their relative importance will differ across medication tasks. A low-risk prioritization aid may tolerate different uncertainty and oversight arrangements from a system that influences dispensing, treatment interruption, dose adjustment, or escalation of urgent care. Validation must consequently be tied to the intended decision, affected population, pharmacy setting, professional role, and consequences of being wrong.

Proposed Prediction–Interpretation–Action Framework

Clinical decision support should position artificial intelligence as one component of a broader professional decision process rather than as an independent decision-maker [16]. The proposed Prediction–Interpretation–Action Framework accordingly separates three functions that are often compressed into a single claim of "AI performance." Prediction concerns what the model estimates. Interpretation concerns how that estimate is related to medication and patient context. Action concerns what a pharmacist or another authorized professional can appropriately do in response.

The relevant unit of evaluation is the combined human–AI system. Evidence from human–computer collaboration shows that performance depends not only on algorithmic capability but also on how assistance is represented and incorporated into human judgment [17]. The framework therefore rejects the assumption that adding a pharmacist to an automated pathway automatically creates meaningful oversight. Oversight requires access to relevant information, sufficient time, authority to disagree, and a workable escalation route.

Professional authority and accountability must remain explicit because algorithmic assistance can redistribute influence without clearly redistributing responsibility [18]. The framework proposes three translation gates:

1. Target-validity gate: Does the model predict the quantity required for the intended medication decision, using information legitimately available at that time?
2. Interpretation-integrity gate: Can the pharmacist understand, verify, challenge, defer, or contextualize the output without being induced into inappropriate reliance?
3. Action-suitability gate**:** Is a timely, feasible, authorized, and proportionate response available, and are the consequences of action and non-action acceptable?

Failure at any gate may interrupt translation even when predictive performance is strong. The framework also includes a feedback pathway through which overrides, medication outcomes, workload, incidents, subgroup effects, and workflow changes can trigger reassessment of the model, interface, action pathway, or continued use.

Figure 1 presents the original conceptual synthesis for prediction–interpretation–action framework for pharmacy decisions, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

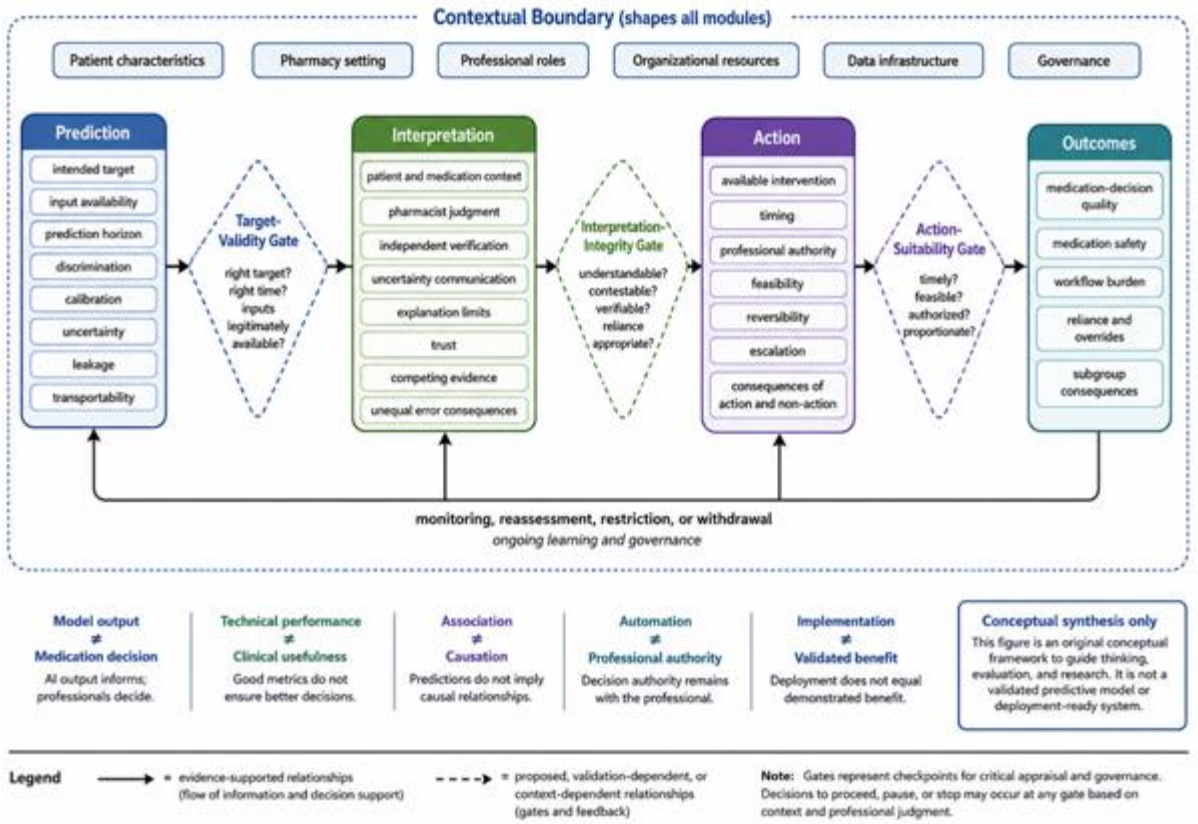

Figure 1. Prediction–Interpretation–Action Framework for Pharmacy Decisions.

The figure is an original conceptual synthesis developed for "Why Better Predictions Can Still Produce Worse Pharmacy Decisions." A prediction layer containing the target, inputs, horizon, performance, calibration, uncertainty, leakage, and transportability is separated from professional interpretation by a proposed target-validity gate. The interpretation layer incorporates patient and medication context, verification, uncertainty, explanations, trust, competing evidence, and error consequences. A proposed interpretation-integrity gate separates interpretation from an action layer containing available interventions, timing, authority, feasibility, reversibility, escalation, and consequences of action or non-action. A proposed action-suitability gate leads to medication, workflow, safety, and equity outcomes. Feedback from these outcomes supports monitoring, reassessment, restriction, or withdrawal. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, clinical recommendation, regulatory determination, or deployment-ready system.

The proposed framework does not establish that passage through the three gates guarantees a better decision. Instead, it supplies a falsifiable structure for determining where prediction-to-decision translation may fail and what evidence would be required to investigate that failure.

Evaluation beyond Discrimination and Calibration

Evaluation should assess whether model outputs lead to decisions whose expected benefits outweigh the consequences of false-positive and false-negative actions. Decision-curve analysis provides one method for relating prediction thresholds to clinical net benefit rather than relying only on discrimination or calibration [19]. It does not, however, capture every medication-safety, workload, access, or equity consequence.

Prospective evaluation should examine the system with its intended users and workflow. Early-stage clinical evaluation guidance emphasizes human factors, workflow integration, safety, and actual interaction with artificial-intelligence decision support [20]. For pharmacy applications, relevant outcomes may include verification behaviour, intervention appropriateness, delay, alert burden, escalation, missed medication problems, reliance, recovery from erroneous advice, and patient-relevant consequences.

The underlying prediction evidence should also be appraised for risk of bias and applicability across participants, predictors, outcomes, and analysis [21]. The proposed evaluation structure therefore includes seven linked domains: technical performance, interpretation quality, action appropriateness, workflow effects, medication safety, equity, and governance. No domain can substitute for the others.

The framework generates testable propositions. Decision quality should improve only when the target corresponds to

the intended action, information is available at the relevant time, uncertainty can be interpreted, users can contest the output, and a beneficial response is feasible. It predicts worse performance when biased or leaked outputs are persuasive, when action is unavailable, or when aggregate gains conceal severe subgroup consequences. These propositions require prospective and task-specific validation.

Implications for Pharmacy AI Procurement and Use

Procurement should evaluate the complete prediction-to-decision pathway rather than comparing products primarily through headline accuracy. Governance should specify ownership, accountability, performance monitoring, incident review, and reassessment across the technology lifecycle [22]. Evidence should identify the intended users, medication task, population, setting, prediction horizon, action pathway, contraindicated uses, update process, and circumstances requiring restriction or withdrawal.

Responsible clinical decision-support implementation also requires documentation, validation, user preparation, safety oversight, and post-deployment monitoring [23]. Vendor claims should therefore be separated into technical, workflow, medication-safety, patient-outcome, and equity claims, with evidence assessed at the corresponding level.

Local clinical-pharmaceutical review remains necessary because alert content can differ in appropriateness and patient relevance when applied in practice [24]. Procurement should consequently ask whether pharmacists can access the evidence underlying an output, determine why it applies to the patient, override it without unreasonable burden, and initiate a suitable response.

The article proposes a procurement evidence contract organized around six questions: What is predicted? Who interprets it? What action becomes possible? When can that action occur? Who bears the consequences of error? Who monitors and governs change? Failure to answer any question should narrow the permitted use rather than be compensated for by higher reported model performance.

Limitations

The evidence informing this analysis spans heterogeneous clinical tasks, professional groups, and settings. Comparative artificial-intelligence studies may contain design, reporting, and claim-inflation problems that limit conclusions about clinical superiority [25]. Transparent reporting can improve appraisal of intended use, data, modelling, and evaluation, but reporting compliance does not establish decision benefit [26].

Several human–AI and equity mechanisms were derived from non-pharmacy settings. Aggregate performance can conceal unequal underdiagnosis or error burdens across underserved populations [27], but the form and magnitude of these effects require direct pharmacy investigation. The proposed framework has not been validated as a measurement instrument, causal model, procurement standard, or clinical decision rule. Its applicability may differ across community, hospital, ambulatory, specialty, regulatory, and patient-facing pharmacy contexts.

Conclusion

Better prediction and better pharmacy decision-making are related but non-equivalent objectives. Prediction becomes useful only through interpretation and action within a particular patient, medication, professional, workflow, and governance context. Target mismatch, automation bias, treatment leakage, poor timing, actionability failure, and unequal error consequences explain how technically improved outputs may produce neutral or harmful decisions.

The proposed Prediction–Interpretation–Action Framework makes this translation pathway explicit through target-validity, interpretation-integrity, and action-suitability gates. It shifts evaluation from asking whether a model predicts accurately to asking whether the combined human–AI system supports an appropriate, timely, contestable, equitable, and accountable medication action.

The framework is intended to organize research, procurement scrutiny, implementation evaluation, and governance discussion. It does not establish clinical effectiveness, prescribe pharmacy practice, allocate legal responsibility, or certify a system for deployment. Prospective validation must test its constructs and propositions across medication tasks, patient groups, pharmacy settings, professional roles, and changing technological environments.

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. 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
3. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group "Evaluating Diagnostic Tests and Prediction Models" of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
4. 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
5. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70. doi:10.1001/jamainternmed.2021.2626

6. 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
7. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15(11). doi:10.1371/journal.pmed.1002683
8. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17. doi:10.1038/s41746-020-0221-y
9. 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
10. Lyell D, Coiera E. Automation bias and verification complexity: A systematic review. J Am Med Inform Assoc. 2017;24(2):423-31. doi:10.1093/jamia/ocw105
11. Davis SE, Matheny ME, Balu S, Sendak MP. A framework for understanding label leakage in machine learning for health care. J Am Med Inform Assoc. 2024;31(1):274-80. doi:10.1093/jamia/ocad178
12. Jabbour S, Fouhey D, Shepard S, Valley TS, Kazerooni EA, Banovic N, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: A randomized clinical vignette survey study. JAMA. 2023;330(23):2275-84. doi:10.1001/jama.2023.22295
13. Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4. doi:10.1038/s41746-020-00367-3
14. 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
15. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: Focus on clinicians. J Med Internet Res. 2020;22(6). doi:10.2196/15154
16. Shortliffe EH, Sepulveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199-200. doi:10.1001/jama.2018.17163
17. 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
18. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-11. doi:10.1136/medethics-2019-105586
19. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and interpreting decision curve analysis: A guide for investigators. Eur Urol. 2018;74(6):796-804. doi:10.1016/j.eururo.2018.08.038
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. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8. doi:10.7326/M18-1376
22. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc. 2020;27(3):491-7. doi:10.1093/jamia/ocz192
23. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: Recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc. 2024;31(11):2730-9. doi:10.1093/jamia/ocae209
24. Bauer J, Busse M, Koch S, Schmid M, Sommer J, Fromm MF, et al. Clinical-pharmaceutical assessment of medication CDSS alerts: Content appropriateness and patient relevance in clinical practice. Front Pharmacol. 2025;16:1510425. doi:10.3389/fphar.2025.1510425
25. 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
26. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385. doi:10.1136/bmj-2023-078378
27. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-82. doi:10.1038/s41591-021-01595-0