Original Article

Why Pharmacy AI Evaluation Must Begin with Plausible Harm Scenarios

Andrei Popescu[1]*, Mihai Ionescu[1], Elena Stan[2], Cristina Radu[3]

[1]Department of AI Harm Scenario Evaluation, Faculty of Pharmacy, University of Bucharest, Bucharest, Romania. [2]Department of Pharmacy AI Safety Assessment, Faculty of Pharmacy, University of Agricultural Sciences Cluj-Napoca, Cluj-Napoca, Romania. [3]Department of Plausible Harm Analysis in Pharmacy, Faculty of Pharmacy, Polytechnic University of Bucharest, Bucharest, Romania.

Abstract

Artificial intelligence evaluation in pharmacy commonly begins with model-centred measures such as discrimination, sensitivity, specificity, calibration, alert acceptance, or processing efficiency. These measures are necessary, but aggregate performance can obscure how errors affect particular patients, professionals, medication-use tasks, and organizational conditions. This article develops an original, non-empirical Plausible-Harm Scenario Framework for Pharmacy AI Evaluation. The proposed approach begins by specifying who may be affected, which medication-use task is involved, how an AI-related failure could propagate through data, model output, interface presentation, professional interpretation, and action or inaction, and what medication-related consequence could follow. Each scenario is then examined through five distinct dimensions: severity, exposure, detectability, reversibility, and recovery. These dimensions are not combined into a numerical score. Instead, they organize the selection of technical, clinical-task, human–AI, workflow, medication-safety, equity, implementation, and lifecycle evidence. The framework further proposes governance gates for scenario completeness, evidence adequacy, residual-risk deliberation, bounded permission, monitoring, rollback, and requalification. Its principal contribution is to reposition performance metrics as scenario-dependent evidence rather than sufficient indicators of safety or readiness. Validation would require prospective and post-deployment testing across relevant users, settings, populations, workflows, model versions, and failure conditions. The framework cannot guarantee complete hazard discovery, establish universal thresholds, replace professional judgment, or demonstrate clinical benefit. It is intended as a conceptual structure for making the consequences and evidentiary assumptions of pharmacy AI evaluation more explicit.

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

Introduction

Artificial intelligence is increasingly positioned to support medication selection, prescribing review, alert prioritization, dispensing, monitoring, adherence assessment, pharmacovigilance, and other pharmacy-related decisions. Evaluation often begins with a model-level question: how accurately does the system classify, predict, rank, or retrieve? Yet clinical impact depends on more than technical performance. It also depends on whether the output corresponds to a meaningful medication-use task, reaches an appropriate professional at a useful time, is interpreted correctly, fits the workflow, and can be acted upon safely [1].

Population-average performance can conceal clinically important differences in consequence. Two systems with similar discrimination may expose different patients to error, operate at different points in the medication-use process, or create different opportunities for detection and correction. A false-negative alert during routine review may permit later interception, whereas the same class of error in an autonomous triage pathway could remove a prescription from professional scrutiny. Likewise, a false-positive recommendation may cause only a brief interruption in one setting but contribute to treatment delay, alert fatigue, or inappropriate medication change in another. Strong aggregate performance can therefore coexist with restricted clinical evaluation, lifecycle hazards, and clinically important unintended consequences [2-4].

The central problem is not that conventional performance measures are irrelevant. Rather, they are frequently interpreted without an explicit account of the pathway

Address for correspondence: Andrei Popescu, Department of AI Harm Scenario Evaluation, Faculty of Pharmacy, University of Bucharest, Bucharest, Romania.
andrei.popescu@usamv.ro

Received: 28 March 2026; **Accepted:** 02 August 2026

How to cite this article: Popescu A, Ionescu M, Stan E, Radu C. Why Pharmacy AI Evaluation Must Begin with Plausible Harm Scenarios. Arch Pharm Pract. 2026;17(3):28-35. https://doi.org/10.51847/7eukyIeyrZ

connecting model behaviour to medication-related outcomes. Sensitivity, calibration, processing time, and alert acceptance answer different questions, and their importance depends on the users, task, setting, failure mode, and potential consequence. Evaluation becomes incomplete when a metric is treated as a property of the system that carries the same meaning across contexts.

This article proposes that pharmacy AI evaluation should begin with plausible harm scenarios. A plausible-harm scenario is a structured account of how an AI system, its inputs, its interface, or its use could contribute to an adverse medication-related consequence under credible conditions. It identifies the affected patient or population, intended and actual users, task, setting, decision authority, failure chain, possible consequence, opportunities for detection, and capacity for correction and recovery.

The contribution is conceptual rather than empirically validated. It integrates established concerns from clinical AI evaluation, human–AI interaction, medication safety, uncertainty, workflow, equity, and governance into one proposed pharmacy-specific architecture. It does not assume that all hazards can be anticipated, that every plausible scenario will occur, or that describing a scenario proves the associated risk is controlled.

Hazard-Based versus Metric-First Evaluation

Metric-first evaluation begins by selecting measurements—such as accuracy, sensitivity, specificity, discrimination, calibration, alert volume, acceptance, or time saved—and then infers usefulness or safety from the resulting values. This approach is attractive because metrics are reproducible, comparable, and often available before clinical integration. However, evaluations comparing AI with clinicians have frequently been limited by study design, reporting quality, external validation, and the distance between experimental performance and actual clinical use [5].

A hazard-based evaluation begins with a different question: how could this system contribute to a medication-related adverse consequence in its intended or reasonably foreseeable context? Metrics are selected only after the user, task, setting, failure pathway, and consequence have been specified. This reverses the conventional evaluative sequence. Instead of asking whether the system performs well and subsequently considering possible uses, the evaluator first defines the medication-use scenario and then determines what evidence would be required to assess it.

The distinction does not oppose qualitative safety analysis to quantitative model testing. Early-stage clinical evaluation already requires attention to users, workflow integration, human factors, errors, system modifications, and safety events in addition to technical performance [6]. Hazard-based evaluation extends this logic by making the potential consequence the organizing unit. A sensitivity estimate, for example, becomes meaningful only after identifying what a false negative means, who is exposed, whether the error is detectable, and whether recovery remains possible.

Scenario construction also depends on transparent specification of intended use, target population, input data, model development, validation, bias assessment, and implementation context [7]. Without these details, the evaluator cannot determine whether the test population represents the intended users or patients, whether the outcome reflects the medication decision, or whether a favourable metric transfers to the proposed workflow.

Calibration illustrates why no single metric is sufficient. Discrimination describes relative separation, whereas calibration concerns agreement between predicted values and observed frequencies. A model may discriminate adequately while producing risk estimates that are unsuitable for threshold-based decisions [8]. Conversely, a calibrated estimate does not establish that the predicted outcome is clinically appropriate, that the user will interpret it correctly, or that intervention improves medication safety. Hazard-based evaluation therefore treats calibration, discrimination, uncertainty, usability, and workflow performance as complementary evidence selected for a defined scenario.

Alternative explanations must also be preserved. Poor clinical performance may arise from a technically weak model, but it may also result from incompatible workflows, misunderstood outputs, incomplete data, inappropriate thresholds, automation bias, or insufficient organizational response. Similarly, a favourable outcome may reflect expert compensation for system errors rather than intrinsic model reliability. The unit of evaluation must consequently extend from the isolated model to the model–user–workflow configuration.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, components, relationships, and intended uses for the article Why Pharmacy AI Evaluation Must Begin with Plausible Harm Scenarios.

Table 1. Definitions, components, relationships, and intended uses for the article Why Pharmacy AI Evaluation Must Begin with Plausible Harm Scenarios.

Component or scientific dimension	Problem addressed	Inputs or determinants	Proposed mechanism or relationship	Expected contribution	Evidence required	Failure or uncertainty risk	Boundary statement
Metric-first evaluation	Aggregate measures may be mistaken for	Selected outcomes, thresholds,	Metrics are interpreted before the medication-use	Efficient technical comparison and	Design quality, external validation, calibration, error	Good average performance may conceal	Metrics remain necessary but do not independently

	clinical usefulness.	comparator, dataset, averaging method.	consequence is specified.	model characterization.	distributions, and reporting completeness [5, 8].	consequential failure conditions.	establish safety or readiness.
Hazard-based evaluation	Evaluation may omit how failure reaches a patient or workflow.	User, task, setting, authority, failure pathway, affected population.	**Proposed:** begin with a plausible consequence and select evidence for each pathway link.	Align evaluation with medication-related consequences.	Early clinical, workflow, human-factor, and safety evidence [6].	Scenario omission or speculative pathways may distort priorities.	Hazard identification does not prove probability, causality, or control.
Plausible-harm scenario	Generic risk lists lack task and context specificity.	Patient, professional, medication, setting, timing, model role, fallback.	**Proposed:** combine a scenario anchor with a traceable failure chain.	Make evaluative assumptions and affected parties explicit.	Intended-use, population, data, validation, and bias documentation [7].	Important actors, dependencies, or indirect consequences may be excluded.	Plausibility is not evidence that the scenario has occurred.
Model and clinical-task performance	Prediction quality may not correspond to decision quality.	Target definition, reference standard, decision threshold, workflow timing.	Model output contributes to, but does not determine, professional or organizational action.	Separate technical validity from task usefulness.	Task-relevant discrimination, calibration, error analysis, and clinical evaluation [5, 6, 8].	Proxy outcomes may be mistaken for meaningful medication decisions.	Technical success cannot be equated with patient benefit.
Human–AI configuration	Identical outputs may produce different actions across users.	Expertise, workload, interface, explanation, confidence, authority.	**Proposed:** evaluate the combined model–user–workflow unit.	Reveal reliance, override, misunderstanding, and coordination failures.	User-centred observation, error analysis, and workflow testing [6].	Experimental users may not represent actual pharmacy teams.	Human oversight is not automatically effective merely because it exists.
Medication-use consequence	Model errors may have unequal clinical implications.	Medication, dose, indication, patient vulnerability, timing, interception opportunities.	**Proposed:** connect action or inaction to a defined consequence.	Prioritize tests by consequence rather than metric convenience.	Medication-safety and task-specific clinical evidence.	Consequences may be delayed, indirect, or difficult to attribute.	The framework does not assign universal severity rankings.
Evidence bundle	A single evidence source cannot characterize a sociotechnical system.	Technical, clinical-task, human, workflow, safety, equity, implementation, lifecycle data.	**Proposed:** evidence is assembled around each scenario and failure-chain link.	Expose missing or weak evidence before a permission decision.	Transparent reporting and early-stage clinical evaluation [6, 7].	Evidence may remain fragmented or non-transferable.	Completeness of documentation is not equivalent to validation.
Governance boundary	Performance reports may not specify who may use the system or under what conditions.	Approved users, population, setting, autonomy, fallback, monitoring, escalation.	**Proposed:** permission is bounded to the tested configuration.	Prevent unrestricted inference from narrow evaluation.	Accountable governance, monitoring, and change control.	Boundaries may become obsolete after workflow or model change.	A local permission decision is not regulatory approval or universal endorsement.

Proposed Plausible-Harm Scenario Framework

The proposed framework treats pharmacy AI evaluation as a sequence connecting a scenario anchor, failure chain, consequence characterization, evidence bundle, and governance decision. Its purpose is not to calculate a universal risk score but to make explicit what could go wrong, who could be affected, what evidence is needed, and where organizational permission should stop.

The first layer is the scenario anchor. It records the intended and actual users, affected patients or populations, medication-use task, setting, timing, workload, decision authority, degree of automation, fallback process, and critical data dependencies. These fields prevent the AI system from being evaluated as a context-free product. A model may support a pharmacist's review, prioritize prescriptions for later assessment, generate medication advice, or suppress alerts; each role creates a different relationship among output, authority, and potential consequence.

The second layer is the failure chain. Clinical decision-support systems operate through interactions among knowledge or model logic, data, interfaces, users, workflows, and maintenance processes [9]. The proposed chain therefore follows an input or data condition through model behaviour, expressed uncertainty or abstention, interface presentation, professional interpretation, action or inaction, and a medication-use consequence. This structure distinguishes a model error from the downstream decision. It also permits scenarios in which technically correct output contributes to harm because it is mistimed, misunderstood, overtrusted, or inserted into an incompatible workflow.

The third layer characterizes the consequence through severity, exposure, detectability, reversibility, and recovery. These dimensions remain separate. A highly severe outcome may be rare and readily intercepted, whereas a less severe outcome may affect many patients and accumulate through repeated workflow exposure. Detectability concerns whether

the problem can be recognized before consequence propagation; reversibility concerns whether an erroneous recommendation or medication action can be corrected; recovery concerns whether the patient and organization can return to an acceptably controlled state.

The fourth layer assembles a scenario-linked evidence bundle. Governance scholarship indicates that responsible AI use requires multidisciplinary accountability across technical development, clinical use, implementation, and monitoring [10]. Patient-safety literature likewise indicates that AI can affect multiple safety functions while generating new risks and evidence gaps [11]. The proposed bundle therefore includes technical performance, clinical-task validity, human–AI team performance, workflow and workload effects, medication-safety consequences, subgroup equity, implementation capability, and lifecycle stability. Evidence is judged by its relevance to the specified scenario, not solely by methodological prestige.

The final layer applies proposed governance gates. An incomplete scenario returns for clarification. An inadequately tested pathway requires additional evidence or redesign. Residual uncertainty is then considered against the intended users, autonomy, fallback, and recovery capability. Any permission is expressed as a bounded envelope specifying who may use the system, for which task, in which population and setting, with what monitoring and escalation. Drift, incidents, workflow change, or model modification can trigger restriction, suspension, rollback, or requalification. These gates organize deliberation but do not establish validated thresholds.

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

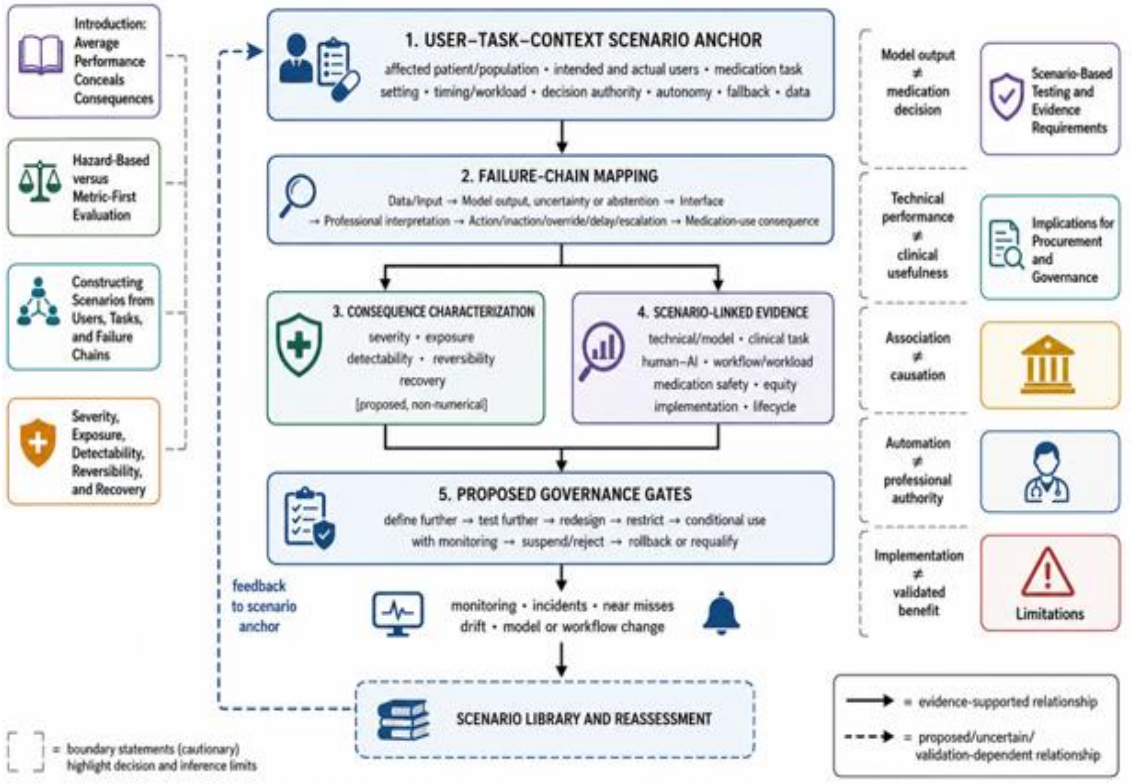

Figure 1. Plausible-Harm Scenario Framework for Pharmacy AI Evaluation

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for evaluation criteria, evidence requirements, and failure modes for the article Why Pharmacy AI Evaluation Must Begin with Plausible Harm Scenarios.

Table 2. Evaluation criteria, evidence requirements, and failure modes for the article Why Pharmacy AI Evaluation Must Begin with Plausible Harm Scenarios.

Architecture layer or process stage	Core function	Information flow	Interaction with other components	Evaluation criterion	Potential failure mode	Human or organizational responsibility	Readiness boundary

Scenario specification	Define the intended use and foreseeable conditions.	Patient, user, task, setting, timing, authority, autonomy, fallback, and data enter the framework.	Determines which failure chains and tests are relevant.	Intended use, population, inputs, outputs, and workflow are explicitly documented [7].	Evaluation uses a generic or idealized scenario that omits actual users or dependencies.	Pharmacy, clinical, informatics, safety, and patient representatives define scope.	No progression when the scenario is materially incomplete.
Data and input layer	Establish whether inputs represent the scenario.	Clinical, medication, laboratory, administrative, or behavioural data enter the model.	Constrains model validity and subgroup exposure.	Provenance, completeness, measurement, and scenario relevance are assessable.	Missing, delayed, shifted, or selectively documented data produce misleading output.	Data stewards and local clinical owners verify suitability and change control.	Dataset availability alone does not establish fitness for the task.
Model-behaviour layer	Characterize predictions, recommendations, uncertainty, and abstention.	Inputs are transformed into scores, classifications, rankings, or advice.	Supplies the interface but does not determine the medication decision.	Discrimination, calibration, error distribution, robustness, and uncertainty match the scenario [5, 8].	Acceptable average performance conceals high-consequence errors.	Developers document limitations; evaluators test clinically relevant conditions.	Technical performance does not establish clinical usefulness.
Interface and communication layer	Present output in an interpretable, timely form.	Model output is displayed, explained, prioritized, or suppressed.	Shapes professional attention, confidence, and response.	Users can identify the recommendation, evidence limits, urgency, and required action [6].	Misleading confidence, poor timing, or alert design induces misuse or neglect.	Human-factors specialists and pharmacy users test the interface in workflow.	Usability does not prove appropriate reliance or safety.
Human–AI decision layer	Convert information into professional judgment and action.	Output is accepted, rejected, modified, escalated, or ignored.	Mediates the transition from model behaviour to medication consequence.	Team performance and error patterns are evaluated, including incorrect advice.	Automation bias, under-reliance, responsibility diffusion, or expertise mismatch.	Professionals retain defined authority; organizations provide training and escalation.	Nominal human oversight is insufficient without evidence that oversight functions.
Medication-use consequence layer	Identify what follows from action or inaction.	Medication decision affects prescribing, verification, dispensing, administration, monitoring, or counselling.	Supplies the harm dimensions and recovery requirements.	Consequences are linked to the exact task, patient, timing, and interception opportunities [11].	A proxy outcome is mistaken for medication-safety benefit.	Medication-safety and clinical teams assess downstream pathways.	No readiness claim may rest solely on model or process outcomes.
Scenario-linked evidence bundle	Assemble evidence for every material pathway link.	Technical, clinical, human, workflow, safety, equity, implementation, and lifecycle evidence converge.	Informs governance gates and exposes evidence gaps.	Evidence is relevant to the intended users, population, workflow, and model version [6, 7].	Strong evidence in one domain compensates rhetorically for missing evidence elsewhere.	Multidisciplinary reviewers document unresolved uncertainty.	Evidence domains are complementary and not interchangeable.
Governance and lifecycle layer	Bound permission and respond to change.	Evaluation findings become restrictions, monitoring duties, escalation triggers, or rejection.	Feeds incidents and changes back into the scenario library.	Accountability, monitoring, fallback, suspension, and requalification are specified [10].	Permission expands beyond tested conditions or persists after material change.	Named institutional owners maintain the permission envelope and reassessment process.	A governance decision is local and conditional, not universal approval.

Constructing Scenarios from Users, Tasks, and Failure Chains

A plausible-harm scenario should begin with the medication-use activity rather than a generic catalogue of algorithmic errors. The same technical model may be used to prioritize prescriptions, advise a pharmacist, support patient counselling, suppress low-priority alerts, or trigger autonomous workflow actions. Each use redistributes attention, authority, and opportunities for interception. Scenario construction must therefore identify not only what the model predicts, but what work the prediction is expected to change.

The first element is the affected party. This includes the patient receiving the medication-related decision, but may also include caregivers, pharmacists, prescribers, technicians, nurses, and organizational units whose work is redirected. Effects can be direct, such as an incorrect dose recommendation, or indirect, such as deprioritizing a prescription that would otherwise have received timely

review. Equity assessment also begins here because population averages can obscure concentrated exposure among patients with incomplete records, uncommon conditions, complex regimens, or limited access to follow-up.

The second element is the user configuration. Intended users may differ from actual users in expertise, workload, training, or authority. Human–computer collaboration studies demonstrate that joint performance depends on the user, task, and form of assistance rather than on model quality alone [12]. A system assessed with specialists under controlled conditions may operate differently when used by trainees, technicians, generalist pharmacists, or time-pressured clinicians. The scenario should state who sees the output, who may act, who can override it, and who remains accountable for escalation.

The third element is the task definition. "Medication review" is too broad to specify an evaluation. A scenario should identify whether the system detects contraindications, prioritizes prescriptions, proposes dose changes, identifies monitoring gaps, reconciles records, predicts non-adherence, or drafts patient-facing advice. It should also identify the consequence of delayed, absent, incorrect, or overconfident output. The intended action must be distinguished from a model's predictive target because a statistically valid prediction may still be insufficient for the medication decision.

The fourth element is the failure chain. A scenario may begin with missing renal-function data, an outdated medication list, a shifted patient population, an ambiguous instruction, or a model operating outside its development conditions. The output may then be incorrect, poorly calibrated, overconfident, or appropriately uncertain but presented without a usable warning. The professional may accept, reject, overlook, misunderstand, or delay action. Experimental evidence indicates that AI assistance can alter expert error patterns and that incorrect recommendations can induce inappropriate reliance under some conditions [12-14]. These findings establish a mechanism requiring evaluation, not a numerical estimate of pharmacist susceptibility.

The fifth element is the medication-use consequence. Scenario construction should continue beyond the immediate click, override, or recommendation. It should ask whether a prescription is dispensed, withheld, delayed, modified, inadequately monitored, or transferred without escalation. It should also identify later interception points, including pharmacist review, patient counselling, laboratory monitoring, administration checks, or follow-up. A failure chain can terminate without patient harm when another safeguard functions, but that does not make the upstream failure irrelevant; repeated near misses may indicate dependence on compensating work.

Scenarios should be sufficiently concrete to direct testing but not so narrow that they exclude foreseeable variations. Multiple scenarios may be needed for the same system because users, settings, medications, autonomy, and fallback processes differ. The proposed scenario library should therefore include anticipated failures, observed near misses, workflow changes, model revisions, and newly recognized patient groups. It should be periodically reviewed rather than treated as a one-time predeployment document.

Finally, scenario construction must preserve uncertainty. A plausible pathway is not necessarily probable, and a failure observed in one setting may not transfer unchanged to another. Conversely, absence of observed harm may reflect limited exposure, incomplete detection, or effective professional compensation. The framework therefore treats each scenario as a testable proposition about a sociotechnical medication-use pathway, not as proof of causation or a predetermined argument against AI use.

Severity, Exposure, Detectability, Reversibility, and Recovery

Medication-alert research illustrates why harm cannot be characterized by severity alone. Evaluation must also consider alert burden, inappropriate-prescription detection, external validation, workflow effects, and whether clinically important problems remain identifiable after prioritization [15]. Systems that reduce review workload or suppress low-priority alerts may create value, but their safety depends on which prescriptions are excluded, which problems are missed, and whether another safeguard can intercept the error. Medication-review triage and wrong-drug alerting therefore show why alert burden, triage exclusions, missed drug-related problems, and clinically meaningful interception must be assessed together [15-17].

The proposed framework uses five distinct dimensions. Severity concerns the potential clinical or organizational consequence. Exposure concerns how often or how broadly the triggering conditions arise, including concentration within a subgroup or workflow. Detectability concerns whether the failure can be recognized before the consequence propagates. Reversibility concerns whether an incorrect recommendation, decision, or medication action can be corrected after it begins. Recovery concerns whether the patient, professional, and organization can return to an acceptably controlled state and what delay, expertise, or resources this requires.

These dimensions should not be multiplied into an unvalidated score. A detectable error is not necessarily acceptable, and reversibility does not erase the burden or inequity of repeated failure. Their purpose is to prevent a single model metric or severity label from substituting for structured examination of the complete medication-use consequence.

Scenario-Based Testing and Evidence Requirements

Each material link in a plausible-harm scenario should generate a corresponding test. Technical testing may examine discrimination, calibration, robustness, uncertainty, and subgroup errors. Human–AI testing should examine interpretation, reliance, override, workload, and response to incorrect advice. Workflow testing should assess timing, interruption, escalation, fallback, and compensating work. Medication-safety testing should follow errors beyond the output to prescribing, verification, dispensing, administration, monitoring, or counselling consequences.

Testing must also extend across time. Calibration may deteriorate after deployment even when development performance was acceptable [18]. Changes in patient populations, documentation, practice patterns, or causal relationships can produce dataset shift and invalidate previous assumptions [19]. Outputs used in consequential medication decisions should therefore communicate uncertainty and support abstention or escalation when evidence is inadequate [20].

Passing a scenario test supports only the evaluated population, users, workflow, model version, autonomy, and fallback conditions. It does not establish universal safety. Monitoring should be linked to explicit reassessment triggers, including drift, incidents, near misses, workflow redesign, changed data sources, or model modification.

Implications for Procurement and Governance

Procurement should begin with the intended medication-use scenario rather than a feature list or headline performance value. Ethical implementation requires attention to transparency, responsibility, bias, and the distribution of authority between the system, professionals, and organization [21]. Purchasers should therefore require an evidence dossier that identifies intended users, affected populations, task boundaries, model limitations, failure scenarios, monitoring duties, escalation routes, and accountable owners.

Equity must be examined within each scenario. Electronic-health-record data can encode unequal selection, missingness, measurement, and documentation practices, producing concentrated errors that remain hidden in aggregate performance [22]. Procurement review should consequently ask which populations are absent, underrepresented, differently measured, or more dependent on unavailable follow-up and recovery resources.

Regulatory authorization or market availability does not by itself establish local workflow compatibility, pharmacy-task validity, or control of the institution's plausible harm scenarios [23]. The proposed governance output is therefore a conditional permission envelope defining approved users, tasks, populations, settings, autonomy, fallback, monitoring, review dates, and suspension criteria. This envelope supplements rather than replaces legal, regulatory, cybersecurity, clinical-engineering, and professional obligations.

Limitations

The framework cannot ensure complete hazard discovery. Generalizability remains conditional on the population, setting, measurement process, and use context [24]. Ethical and equity failures may arise at multiple stages that a scenario library does not fully capture [25]. Transparent documentation improves appraisal but does not establish clinical effectiveness, safety, or deployment readiness [26].

Scenario construction also involves judgment. Stakeholders may disagree about plausibility, severity, acceptable uncertainty, or required evidence. Rare events may be difficult to test, and detailed evaluation may demand substantial organizational resources. The proposed dimensions and governance gates therefore require empirical validation and should not be treated as universal thresholds.

Conclusion

Pharmacy AI evaluation should not begin by asking whether a model performs well in the abstract. It should begin by asking who could be affected, which medication-use task is being changed, how failure could propagate, what consequence could follow, and whether the system and organization can detect, reverse, and recover from that consequence.

The proposed Plausible-Harm Scenario Framework connects a user–task–context anchor with failure-chain mapping, five non-numerical consequence dimensions, scenario-linked evidence, and conditional governance gates. Its central scientific contribution is to reposition metrics as evidence selected for a defined harm scenario rather than as independent proof of clinical usefulness or safety.

The framework preserves professional authority by separating model output from medication decisions and organizational permission. It also makes equity, workflow, uncertainty, monitoring, and recovery part of evaluation rather than secondary implementation concerns. However, it remains a conceptual synthesis. Its categories, decision gates, and proposed permission envelope require prospective and post-deployment evaluation across different pharmacy tasks, populations, institutions, and human–AI configurations before any claims of effectiveness, safety, or readiness can be justified.

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. 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):e271-e97. doi:10.1016/S2589-7500(19)30123-2
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. 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
5. 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:m689. doi:10.1136/bmj.m689
6. 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
7. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27(12):2011-5. doi:10.1093/jamia/ocaa088
8. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
9. 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
10. 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
11. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: A scoping review. NPJ Digit Med. 2021;4(1):54. doi:10.1038/s41746-021-00423-6
12. 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
13. 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;3:23. doi:10.1038/s41746-020-0232-8
14. 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
15. 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
16. 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
17. Chen CY, Chen YL, Scholl J, Yang HC, Li YCJ. Ability of machine-learning based clinical decision support system to reduce alert fatigue, wrong-drug errors, and alert users about look alike, sound alike medication. Comput Methods Programs Biomed. 2024;243:107869. doi:10.1016/j.cmpb.2023.107869
18. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052-61. doi:10.1093/jamia/ocx030
19. Subbaswamy A, Saria S. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345-52. doi:10.1093/biostatistics/kxz041
20. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell. 2019;1(1):20-3. doi:10.1038/s42256-018-0004-1
21. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981-3. doi:10.1056/NEJMp1714229
22. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544-7. doi:10.1001/jamainternmed.2018.3763
23. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): A comparative analysis. Lancet Digit Health. 2021;3(3):e195-e203. doi:10.1016/S2589-7500(20)30292-2
24. 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):e489-e92. doi:10.1016/S2589-7500(20)30186-2
25. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123-44. doi:10.1146/annurev-biodatasci-092820-114757
26. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y