



Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

AI-Assisted Pharmacovigilance Should Prioritize Calibration Before Automation

Chen Hao^{1*}, Liu Fang¹, Zhao Lin², Wei Zhang³

¹Department of AI-Assisted Pharmacovigilance and Calibration, Faculty of Pharmacy, Zhejiang University, Hangzhou, China.

²Department of Automated Drug Safety Systems, Faculty of Pharmaceutical Sciences, Nanjing University, Nanjing, China.

³Department of Pharmacovigilance Informatics, Faculty of Pharmacy, University of Melbourne, Melbourne, Australia.

*Email: chen.hao@zju.edu.cn

ABSTRACT

Artificial intelligence is increasingly positioned to accelerate pharmacovigilance by extracting adverse-event information, classifying cases, ranking candidate signals, and directing human attention. Yet these functions are often evaluated primarily through discrimination, retrieval, or benchmark accuracy, even though operational pharmacovigilance decisions depend on a different question: whether an output can be interpreted reliably in the population, data source, time period, and decision context in which it is used. This Current Opinion argues that calibration should therefore precede automation as the governing design principle for consequential AI-assisted pharmacovigilance. Calibration is interpreted broadly but precisely: not merely as post-hoc probability adjustment, but as alignment among model output, event prevalence, data-generating conditions, evidential maturity, clinical seriousness, and the human action triggered by that output. Rare-event imbalance, dataset shift, heterogeneous information sources, changing reporting practices, and reviewer interaction can all disrupt this alignment. We propose that AI should progress from classification or ranking to increasingly autonomous workflow roles only when its outputs have been calibrated for their intended interpretation and when escalation pathways preserve human review where uncertainty remains consequential. This approach reframes automation as an outcome of demonstrated calibration rather than as the default objective of pharmacovigilance AI development.

Keywords: Pharmacovigilance, Artificial intelligence, Probability calibration, Adverse drug reactions, Signal detection, human-AI interaction

Received: 20 July 2026; **Revised:** 28 September 2026; **Accepted:** 02 October 2026

INTRODUCTION

Pharmacovigilance is an attractive setting for artificial intelligence because the work combines high information volume with persistent scarcity of expert attention. Individual case safety reports, clinical narratives, spontaneous-report databases, biomedical literature, and other post-marketing sources can create more candidate information than can be examined manually with equal intensity. The resulting temptation is to treat automation as the principal endpoint of AI development. That framing is incomplete. For rare events, apparent classifier performance can change markedly when the evaluation prevalence differs from the prevalence encountered in operation, making apparently strong results difficult to translate directly into useful case-level or signal-level decisions [1]. The relevant question is therefore not simply whether an algorithm can automate a

pharmacovigilance task, but whether its output has an interpretation that remains reliable when embedded in the intended safety workflow.

Contemporary reviews describe AI moving from experimental pharmacovigilance applications toward potential routine use while emphasizing persistent challenges involving consistency, transparency, bias, interpretability, data quality, and regulatory integration [2]. These limitations are not merely obstacles surrounding an otherwise complete prediction problem. They help determine what a model output means. Earlier mapping of machine-learning research in pharmacovigilance similarly showed that methodological maturity was uneven: among 393 included publications, only a minority exhibited the characteristics the review associated with contemporary good machine-learning practice [3]. That historical finding should not be projected numerically onto current systems, but it cautions against equating the volume of AI research with evidentiary readiness for autonomous deployment. Regulatory experience reinforces the distinction. AI may assist selected components of individual case processing and evaluation, but human quality assurance and task-specific assessment remain important when outputs contribute to consequential safety judgments [4]. The position advanced here is therefore narrower than either enthusiasm for automation or resistance to it. We propose calibration before automation: the degree of automation should be earned by demonstrating that model outputs, their uncertainty, and the actions attached to them are appropriately aligned with the pharmacovigilance context in which they operate.

Automation without calibration

“Pharmacovigilance automation” often combines tasks that have different inferential meanings. Signal detection, case intake, adverse-event extraction, case classification, prioritization, signal validation, and signal evaluation are not interchangeable. A recent review of AI in signal management found that published applications were concentrated disproportionately in detection and that transparency and use of suitable control data remained inconsistent [5]. A model that identifies records for inspection may therefore be valuable even when it would be inappropriate to treat its score as the probability that a causal adverse drug reaction exists.

This distinction creates a calibration hierarchy. At the lowest level, an algorithm may be calibrated only to an information-processing task: for example, identifying text likely to contain an adverse event. At a higher level, the output may be intended to rank cases for review. More demanding still is use of a score to support escalation, resource allocation, or a substantive safety interpretation. Each step requires additional evidence because the target being calibrated changes. Ethical considerations reinforce this requirement. Fairness, non-discrimination, accountability, privacy, and responsibility are not external decorations on an AI system; they determine whether similar scores can legitimately prompt similar actions across groups and settings [6].

Deployment also exposes an epistemic problem. FDA experience with AI-assisted case-based reasoning has highlighted uncertainty, reproducibility, trust, and the difficulty of formalizing expert causal reasoning as important barriers between model development and dependable safety use [7]. Explainable output may assist inspection, but explanation does not itself establish calibrated probability, causal attribution, or actionability.

The distinction is especially visible in natural-language processing. Reviews of clinical-text adverse-event extraction show that deep-learning and transformer-based approaches can perform strongly on named-entity recognition, relation classification, and end-to-end benchmark tasks [8]. Yet extraction confidence answers whether a text span or relation has been detected; it does not establish whether a reported event is caused by a medicine, whether it represents a new signal, or whether regulatory escalation is warranted. A systematic review of adverse-event extraction from clinical notes also demonstrates substantial heterogeneity in datasets, extraction targets, and algorithmic architectures [9]. Scores generated in such settings cannot be assumed to share a common evidentiary scale.

Calibration before automation therefore requires an explicit declaration of the inference target. Automation is least hazardous where model output triggers reversible information-processing steps and most demanding where it changes substantive safety interpretation or suppresses human review. The objective is not to prohibit automation, but to prevent confidence at one analytical level from being silently transferred to another.

Classification and ranking tasks

Classification and ranking are natural early uses of AI because they can organize large safety-information streams without requiring every score to be interpreted as an absolute probability. Even here, however, performance must be judged against the intended workflow. In a pilot predictive-surveillance study involving two products, machine-learning models identified some known signals and generated candidate signals for expert assessment, including an example of earlier detection. Sensitivity and positive predictive value were nevertheless imperfect,

leaving both missed signals and candidates that did not become confirmed signals [10]. The appropriate interpretation is not that machine learning succeeded or failed categorically, but that ranking may alter the distribution of expert attention while preserving substantial uncertainty.

Feature enrichment can change this ranking function. Models combining demographic information from FAERS with chemical, molecular, and biological drug characteristics have been investigated for prediction of selected adverse reactions [11]. Such work supports the possibility that safety classification can exploit information beyond report frequency alone. It does not convert spontaneous-report predictions into causal estimates, because reporting processes, case composition, indication, exposure, and unmeasured confounding remain relevant.

Clinical-document automation presents the same distinction at a different level. Deep-learning natural-language processing has achieved strong discrimination for identifying adverse drug reactions in discharge summaries, with institution-specific pretraining improving performance [12]. Yet a highly discriminative document classifier can still produce errors when deployed at scale. If every positive classification automatically becomes a report or escalation, the operational consequence depends on positive predictive value, prevalence, duplicate handling, competing drug-related events, and the cost of unnecessary review—not merely on area under the receiver-operating-characteristic curve.

Multimodal models provide another illustration. Combining text representations with molecular drug embeddings improved performance in selected multilingual social-media adverse-event classification tasks [13]. This is valuable evidence that representation choice changes ranking ability, but it also emphasizes source dependence. A score trained on patient-generated social-media text and a score derived from structured or clinically adjudicated information are not necessarily commensurate. A calibration-first workflow should therefore preserve the distinction between ordering candidates and interpreting the absolute meaning of the score used to order them.

Probability calibration

Probability calibration becomes critical when an AI output is treated as more than a relative ranking. Discrimination describes whether higher-risk observations tend to receive higher scores than lower-risk observations. Calibration instead concerns agreement between predicted probabilities and observed event frequencies. A model can discriminate well while systematically overestimating or underestimating probability [14]. In pharmacovigilance, this distinction matters whenever a model score is linked to a review threshold, escalation rule, or assertion about the likelihood of a safety-relevant state.

Calibration is also not permanently acquired at model development. A deployed clinical prediction model has demonstrated that calibration can differ across subgroups and that recalibration may materially improve agreement between predicted and observed risks [15]. The clinical endpoint in that study was not pharmacovigilance, so its quantitative results should not be imported into drug-safety practice. Its methodological lesson is directly relevant: a model may retain technical functionality while its probability scale becomes inappropriate for the population in which it is being used.

Class-balancing procedures add a particularly important hazard. Random undersampling and oversampling alter the event proportion seen during training. Mathematical analysis, simulation, and a clinical-data example have shown that naïvely using probabilities from models trained on artificially balanced data can produce poor calibration even when discrimination does not improve; probability correction can restore interpretability [16]. For rare pharmacovigilance outcomes, therefore, rebalancing may facilitate learning but should never be assumed to preserve the meaning of the resulting probability.

Calibration assessment should consequently be multidimensional. Calibration-in-the-large can detect systematic overprediction or underprediction, while calibration slope and flexible calibration curves interrogate other forms of mismatch [17]. No single statistic should serve as a universal release criterion. For AI-assisted pharmacovigilance, the appropriate calibration target depends on what the output purports to mean and what decision follows it.

The practical distinction is best represented as decision states rather than as a single “good versus poor calibration” label. The table below therefore separates what is known about an AI output from what action its evidentiary state can reasonably support.

Table 1. Decision states for interpreting calibrated and uncalibrated AI outputs in pharmacovigilance

Decision context	Evidence and limits	Permitted action
------------------	---------------------	------------------

Clinical-seriousness context	Information available / Unresolved uncertainty	Decision state / Permissible AI role / Calibration test and reassessment trigger
Seriousness may guide review priority but cannot repair an uncalibrated score	Information available: Ordering performance is established, but the score lacks a validated probability interpretation Unresolved uncertainty: Absolute event likelihood and threshold meaning remain uncertain	Decision state: Relative ranking only Permissible AI role: Prioritize candidates for human assessment Calibration test and reassessment trigger: Reassess before attaching probability language or autonomous action
More consequential outcomes justify stronger evidence before reducing oversight	Information available: Predicted probabilities agree acceptably with observed outcomes in the evaluated setting Unresolved uncertainty: Transportability to new populations, sources, or periods is unresolved	Decision state: Calibrated within the development setting Permissible AI role: Support threshold testing under controlled use Calibration test and reassessment trigger: Reassess after population, prevalence, source, or workflow change
High seriousness increases the cost of interpreting distorted probabilities as reassurance	Information available: Ranking remains useful after class rebalancing, but raw probabilities do not reflect target prevalence Unresolved uncertainty: Magnitude and direction of probability distortion	Decision state: Discriminative but prevalence-distorted Permissible AI role: Candidate ranking only until probability correction and validation Calibration test and reassessment trigger: Reassess after prevalence restoration or independent recalibration
Consequences are greatest where underestimation affects vulnerable or poorly represented groups	Information available: Overall calibration appears acceptable but differs across relevant strata Unresolved uncertainty: Whether local errors reflect data scarcity, bias, or genuine population heterogeneity	Decision state: Subgroup-dependent calibration Permissible AI role: Decision support with subgroup-specific safeguards Calibration test and reassessment trigger: Reassess when subgroup calibration moves outside the locally justified operating range
Serious signals should not be suppressed on the basis of stale calibration	Information available: Earlier validation was adequate but the current data-generating process may have changed Unresolved uncertainty: Whether deterioration reflects drift, reporting change, case mix, or coding change	Decision state: Temporally unstable calibration Permissible AI role: Maintain human review and restrict autonomous filtering Calibration test and reassessment trigger: Reassess when drift indicators or outcome monitoring show material change
Review intensity can reflect seriousness without pretending that seriousness changes probability	Information available: Probability interpretation, intended population, review action, and evidentiary limits are jointly specified Unresolved uncertainty: Residual uncertainty remains explicit rather than hidden in the score	Decision state: Decision-calibrated use Permissible AI role: Use AI at the validated level of autonomy only Calibration test and reassessment trigger: Reassess whenever the inference target, action, source, prevalence, or workflow changes

Note: The decision states are a proposed synthesis. They do not define universal probability cutoffs or regulatory thresholds.

Class imbalance

Rare adverse events make class imbalance a structural property of many pharmacovigilance tasks rather than a nuisance that can simply be engineered away. Importantly, event rarity and population heterogeneity should not be collapsed into the same problem. Work on imbalanced clinical prediction has shown that low outcome incidence can interact with demographic heterogeneity and worsen subgroup performance if both are not considered explicitly [18]. For pharmacovigilance, an overall event prevalence can therefore conceal clinically important differences in data availability or prediction quality across age groups, indications, comorbid states, or other relevant populations.

The difficulty is amplified where evidence is intrinsically sparse. Pregnancy drug-safety prediction provides an example in which restricted evidence and rare adverse outcomes have motivated synthetic-data and generative approaches to class imbalance [19]. Such methods can increase the information available to an algorithm, but

synthetic observations do not create empirical clinical safety evidence. Any probability intended for decision support still requires validation against an appropriate target population and outcome definition.

The broader adverse-event prediction literature shows why this distinction matters. A 2026 systematic review reported frequent reliance on internally evaluated discrimination, whereas only a minority of included studies explicitly addressed class imbalance and an even smaller proportion conducted external validation [20]. Thus, a seemingly impressive classifier may remain uncertain precisely where pharmacovigilance is most demanding: rare outcomes, changing populations, and transport to new data.

Calibration before automation treats imbalance as an interpretation constraint. Resampling, weighting, synthetic augmentation, or alternative loss functions may be legitimate learning strategies, but they do not by themselves justify a changed level of autonomy. Before a rare-event score can suppress review, trigger escalation, or be interpreted probabilistically, its relationship to the operating prevalence and the relevant subgroup must be demonstrated rather than assumed.

Dataset and concept drift

Even a well-calibrated model can become poorly calibrated when the system generating its data changes. A systematic review of dataset-shift management in health prediction found substantial methodological diversity, with temporal shift and concept drift among the most commonly addressed forms, statistical tests and model-based monitoring among frequently used detection strategies, and retraining or feature engineering among common responses. No single strategy emerged as broadly generalizable, and prospective implementation remained limited [21].

This problem is especially important in pharmacovigilance because the data-generating environment is not stationary. Reporting intensity can change after communications or media attention; prescribing patterns can shift; indications can expand; new formulations or combinations can enter practice; coding systems and case-processing conventions can change; and the mix of spontaneous reports, clinical records, literature-derived cases, and other sources can evolve. These examples define plausible drift channels, but the exact effect of each on a given model must be measured rather than presumed.

Historical analysis of clinical algorithms demonstrates that dataset shift predates modern machine learning: algorithms can lose validity when populations, practices, measurement processes, or relationships between predictors and outcomes change [22]. The implication for pharmacovigilance is that calibration should be regarded as a time-indexed property. A validation result establishes evidence for a specified setting and period; it is not a permanent certificate.

The proposed calibration-first approach therefore requires a feedback channel. Monitoring should look for changes in inputs, prevalence, missingness, case composition, label definitions, prediction distributions, and—where outcome ascertainment permits—calibration itself. Detection of change should not automatically mandate retraining, because some shifts may be benign and retraining can introduce new instability. Instead, material drift should trigger investigation of whether the model's probability interpretation or threshold remains valid. Automation is maintained only while that interpretation remains defensible.

Human review thresholds

The phrase “human in the loop” can create false reassurance if the human component is treated as a fail-safe rather than as part of the system being evaluated. In a controlled study of physicians interacting with safe and unsafe AI recommendations, unsafe recommendations attracted greater visual attention, yet the accompanying explanations did not reliably rescue decision makers from the unsafe advice [23]. This was not a pharmacovigilance experiment and should not be transferred as a quantitative estimate of reviewer error. It nevertheless demonstrates a general high-stakes principle: human oversight can itself be altered by AI output.

For pharmacovigilance, the review threshold should therefore answer two separate questions. First, when does uncertainty require expert review? Second, what action can a reviewer reasonably take with the available evidence? User-centred work on intelligent systems for drug safety identified actionability as an important determinant of adoption [24]. A threshold that sends large numbers of cases to reviewers without changing what they can decide merely transfers computational uncertainty into workload. Conversely, an aggressive filtering threshold may create efficiency while concealing cases for which human interpretation would have been consequential.

This article therefore proposes that human review thresholds should be calibrated to the *decision state*, not optimized solely against a classifier metric. Lower-confidence outputs may still be useful for broad candidate

generation. Higher-confidence outputs may support prioritization. Reduction of human oversight requires stronger evidence because the cost of a miscalibrated probability changes when the model can prevent a case or signal from being examined. Clinical seriousness should modify the tolerance for uncertainty, but seriousness does not itself make the underlying probability more accurate.

Guidance for early clinical evaluation of AI decision-support systems similarly emphasizes that evaluation must extend beyond standalone algorithm metrics to workflow, human factors, safety, and real interaction with users [25]. In pharmacovigilance, this means evaluating the human–AI pair: what reviewers see, how uncertainty is displayed, whether model output anchors interpretation, whether discordant evidence can override a score, and whether workload alters the effective threshold used in practice.

The relationships developed through the preceding sections are compressed in **Figure 1**. The figure makes the review threshold a boundary rather than an endpoint: ranking, probability calibration, prevalence distortion, drift, and reviewer action remain distinguishable, while feedback channels return discordant or changing evidence to the analytical level that requires reassessment.

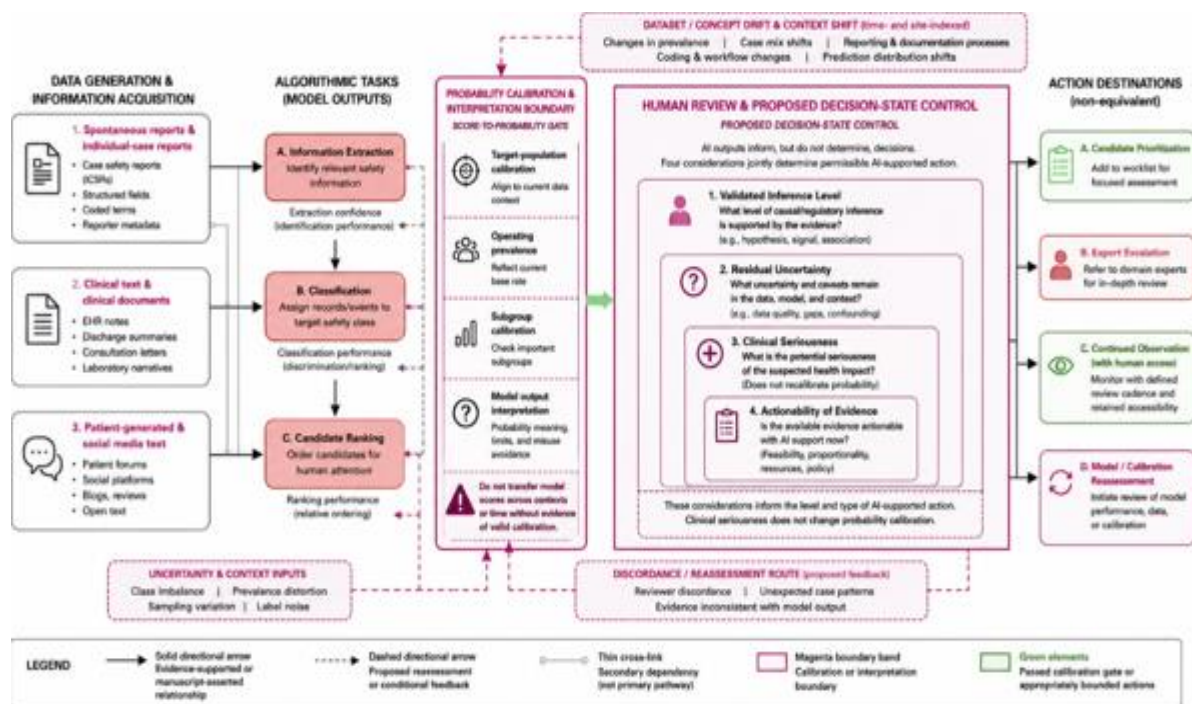


Figure 1. Where a Pharmacovigilance Score Becomes a Review Decision: Calibration, Prevalence, Drift, and Human Escalation Boundaries

False-positive and false-negative costs

A review threshold converts statistical error into an operational consequence. False positives can consume specialist attention, duplicate work, dilute genuinely informative signals, and encourage escalation unsupported by mature evidence. False negatives create a different hazard: potentially informative cases or emerging patterns may be delayed, deprioritized, or never presented for expert interpretation. Prediction-model methodology therefore emphasizes that performance metrics must be interpreted against the intended decision and its consequences rather than as isolated numerical achievements [26].

This asymmetry is particularly important in pharmacovigilance because the cost of error is conditional. A false-positive classification attached to a reversible information-retrieval step may be inexpensive. The same error attached to an automated signal-escalation process may consume substantial analytical resources or create unwarranted concern. Conversely, a false negative involving a nonspecific, readily reversible event does not have the same consequence as failure to surface a potentially serious, difficult-to-reverse event. Clinical seriousness should consequently influence the tolerance for classification uncertainty, while remaining separate from the estimated probability that the underlying relationship is real.

Error costs also cannot be interpreted independently of study validity. PROAST+AI emphasizes that risk of bias and applicability depend on the data, predictors, outcome definition, and analytical design used to develop and evaluate a prediction model [27]. A favorable false-positive rate generated from a narrow or poorly representative

dataset may therefore be less reassuring than a numerically less impressive result obtained under conditions close to intended deployment. The same principle applies to false negatives: absence of detected errors in an inadequately challenging validation set cannot establish reliable omission of human review.

For calibration-first pharmacovigilance, threshold selection should therefore be treated as a boundary-condition problem rather than an optimization problem with one universal objective. The proposed decision logic asks what type of error is plausible, which data-generating mechanism could produce it, how serious the missed or unnecessarily escalated event would be, and whether the downstream action is reversible. These distinctions are difficult to recover from a single sensitivity, specificity, or accuracy value, which is why **Table 2** organizes them as cross-source boundary states.

Table 2. Boundary conditions that change the meaning of false-positive and false-negative AI outputs in pharmacovigilance

Source state and scale	Transfer conditions	Qualification endpoint
Clinical-seriousness condition	Dominant bias or error mechanism / Evidentiary posture at the boundary	Cross-source decision state / Preventability or reversibility consequence / Calibration-first framework endpoint
Greater seriousness justifies review but does not convert duplicated evidence into stronger probability	Dominant bias or error mechanism: Shared reporting stimulus, duplicated content, coding conventions, or correlated extraction errors Evidentiary posture at the boundary: Apparent concordance may overstate independent evidence	Cross-source decision state: Repeated positives from low-specificity sources Preventability or reversibility consequence: Premature escalation is usually more reversible than suppression of subsequent review Calibration-first framework endpoint: Preserve expert assessment while testing source independence and recalibration need
Serious or poorly reversible events warrant low tolerance for dismissing the discordant source	Dominant bias or error mechanism: Unequal ascertainment, missing detail, or source-specific sensitivity Evidentiary posture at the boundary: Numerical majority may be less informative than evidence quality	Cross-source decision state: One clinically rich source conflicts with multiple weak negative sources Preventability or reversibility consequence: A missed credible case may be difficult to recover if automated filtering removes visibility Calibration-first framework endpoint: Escalate discordance for human assessment rather than average scores mechanically
Seriousness can raise review priority while causal interpretation remains unresolved	Dominant bias or error mechanism: Low individual specificity with potentially informative cross-source consistency Evidentiary posture at the boundary: Evidence strengthens only if source-generation mechanisms are sufficiently independent	Cross-source decision state: Weak positive patterns recur across genuinely different sources Preventability or reversibility consequence: Early review can be reversible; causal or regulatory action requires additional evidence Calibration-first framework endpoint: Continue calibrated accumulation and targeted expert review
High-consequence disagreement increases the value of explicit adjudication	Dominant bias or error mechanism: Model misspecification, reviewer anchoring, incomplete phenotype representation, or unmodeled context Evidentiary posture at the boundary: Neither model nor reviewer should automatically dominate	Cross-source decision state: Model score and expert interpretation diverge Preventability or reversibility consequence: Suppression of disagreement can hide both model failure and reviewer error Calibration-first framework endpoint: Trigger structured reassessment of model, evidence, and reviewer reasoning
Serious events require conservative handling while current validity is uncertain	Dominant bias or error mechanism: Dataset shift, concept drift, changing prevalence, coding, case mix, or reporting behavior Evidentiary posture at the boundary: Historical calibration may	Cross-source decision state: Previously stable output becomes discordant after deployment change Preventability or reversibility consequence: Continuing stale automation can propagate repeated errors

no longer support present
interpretation

Calibration-first framework endpoint: Restrict
autonomy and reassess calibration before restoring
prior workflow

Note: These boundary states are a proposed synthesis. They do not assign numerical error costs, universal action thresholds, or regulatory categories.

Calibration across data sources

Calibration becomes more difficult when pharmacovigilance integrates sources that were created for different purposes. Spontaneous reports reflect voluntary or stimulated reporting processes; clinical records reflect care and documentation; biomedical literature contains selected and interpreted cases; and patient-generated data encode symptoms and experiences using different vocabularies and levels of clinical specificity. A probability derived from one source therefore does not acquire the same meaning merely because another model outputs a numerically similar value.

The influence of representation can be substantial. In patient-generated text, adverse-event-related information extraction has shown sensitivity to ontology granularity, with measured performance changing when highly detailed labels were collapsed into broader categories [28]. The implication is broader than social-media analysis itself: the target being predicted is partly defined by how a phenotype or information state is represented. Calibration against a broad symptom category cannot be assumed to establish calibration for a clinically adjudicated adverse-event phenotype.

Explainability does not solve this problem. Explainable machine-learning approaches can identify features that contributed strongly to an adverse-outcome prediction [29]. Such attribution may assist review, debugging, or hypothesis generation, but it remains distinct from calibration and from causal evidence. A model may be interpretable yet systematically overpredict, and a strongly weighted feature may be predictive because of clinical context, documentation practice, confounding, or selection rather than because it causes the adverse event.

We therefore propose source-local calibration before cross-source synthesis. Each source should first be evaluated against the outcome definition, prevalence, ascertainment process, and population appropriate to that source. Only after the meaning of its output is established should concordance across sources be interpreted. Cross-source agreement can strengthen prioritization when evidence streams are genuinely informative and sufficiently independent; it should not be created artificially by averaging scores whose targets or calibration scales differ. Discordance is likewise informative. Instead of being treated automatically as noise, disagreement between sources can reveal phenotype mismatch, reporting bias, temporal change, or a failure of transportability that requires reassessment.

A calibrated human–AI pharmacovigilance architecture

The preceding sections imply that calibration is not one final statistical correction but a control architecture spanning the AI lifecycle. International trustworthy-AI guidance similarly treats reliable deployment as multidimensional, emphasizing fairness, universality, traceability, usability, robustness, and explainability rather than a single performance criterion [30]. Pharmacovigilance requires an additional emphasis on evidentiary maturity because AI outputs often sit upstream of expert causal reasoning and regulatory judgment.

We therefore propose a calibrated human–AI pharmacovigilance architecture with five linked controls. The first is source control, which preserves the data-generating context of each input rather than converting heterogeneous sources immediately into one undifferentiated score. The second is task control, which identifies whether the algorithm is extracting, classifying, ranking, estimating probability, or supporting a decision. The third is calibration control, which tests whether the output has the interpretation required at its intended population, prevalence, subgroup, and time point. The fourth is review control, which connects residual uncertainty and clinical seriousness to an actionable human workflow. The fifth is lifecycle control, which monitors drift, discordance, subgroup performance, and evidence that the calibration state has changed.

These controls are deliberately non-substitutable. Explainability cannot compensate for unmeasured calibration. High discrimination cannot compensate for a poorly defined outcome. Human review cannot compensate reliably for overwhelming false-positive burden. Recalibration cannot repair a fundamentally inappropriate inference target. Similarly, cross-source concordance cannot compensate for correlated reporting biases if the apparently independent sources share the same underlying information.

Traceability is therefore essential. TRIPOD+AI emphasizes transparent reporting of model development, evaluation, intended use, and analytical decisions [31]. In the proposed architecture, every change that may alter

score interpretation—including retraining, rebalancing, outcome redefinition, source addition, threshold modification, and substantial workflow change—should be treated as potentially calibration relevant. The architecture is not presented as an empirically validated regulatory standard. It is a proposed organizing principle derived from the convergent limitations identified across pharmacovigilance and prediction-model evidence.

Prospective evaluation

A calibration-first architecture ultimately requires prospective testing in the workflow for which it is intended. External-validation methodology stresses that evaluation in new data should examine calibration, discrimination, and decision-relevant performance in populations representative of actual use [32]. For pharmacovigilance AI, this means moving beyond retrospective benchmark comparison to evaluation of what happens when model output changes the allocation of expert attention.

Prospective assessment should begin with a clearly specified inference target. A system designed to retrieve candidate cases should not be evaluated as though it predicts causality; a ranking algorithm should not be judged against requirements appropriate to absolute probability estimation. When probability is intended to support a threshold, the evaluation should establish calibration in the target setting and examine whether clinically relevant subgroups or source strata behave differently.

Human–AI performance should be assessed as a system property. Reviewers may respond differently when AI confidence, explanations, or rankings are visible, and workflow load can change how nominal thresholds operate in practice. Evaluation should therefore record whether AI changes case-review time, prioritization patterns, disagreement, escalation behavior, and recovery of discordant cases without assuming in advance that any particular change is beneficial. Where serious events are involved, safety evaluation should preserve routes by which low-confidence or model-discordant evidence can still reach expert attention.

Temporal evaluation is equally necessary. Calibration observed during development or initial deployment may change as prevalence, reporting behavior, coding, prescribing, or source composition evolves. Monitoring should consequently be linked to prespecified evidence of drift or calibration deterioration. The appropriate response may be investigation, recalibration, retraining, restriction of automation, or no change when observed distributional differences are shown not to affect the intended inference. Fixed universal retraining intervals are not supported by the evidence synthesized here.

The decisive prospective comparison is therefore not “AI versus no AI” in the abstract. It is whether a specific calibrated human–AI workflow, under specified data and decision conditions, performs its intended pharmacovigilance function with acceptable error trade-offs, transparent uncertainty, and recoverable failure modes. Autonomous operation should be regarded as one possible result of that evidence—not as the assumption with which evaluation begins.

CONCLUSION

AI can make pharmacovigilance faster without necessarily making its inferences more reliable. Classification, extraction, and ranking are useful capabilities, but they should not be allowed to acquire stronger meanings merely because they are automated or technically sophisticated. The central requirement is alignment: between the score and the event it represents, between training prevalence and operating prevalence, between source and phenotype, between probability and action, and between model output and the human review process surrounding it.

Calibration before automation provides a practical principle for maintaining those distinctions. It treats automation as conditional, preserves human review where uncertainty remains consequential, and requires renewed assessment when populations, data sources, workflows, or reporting processes change. The calibrated human–AI architecture proposed here is therefore not a claim that one universal pharmacovigilance system has been validated. It is a framework for deciding what evidence would be necessary before progressively stronger forms of automation could be justified.

The most useful pharmacovigilance AI may ultimately be highly automated. But autonomy should be the consequence of demonstrated, context-specific reliability rather than the starting assumption of system design.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Norén GN, Meldau EL, Ellenius J. Critical appraisal of artificial intelligence for rare-event recognition: Principles and pharmacovigilance case studies. *Drug Saf.* 2026;49(7):747-60. doi:10.1007/s40264-026-01649-7
2. Nagar A, Gobburu J, Chakravarty A. Artificial intelligence in pharmacovigilance: Advancing drug safety monitoring and regulatory integration. *Ther Adv Drug Saf.* 2025;16:20420986251361435. doi:10.1177/20420986251361435
3. Kompa B, Hakim JB, Palepu A, Kompa KG, Smith M, Bain PA, et al. Artificial intelligence based on machine learning in pharmacovigilance: A scoping review. *Drug Saf.* 2022;45(5):477-91. doi:10.1007/s40264-022-01176-1
4. Ball R, Dal Pan G. "Artificial intelligence" for pharmacovigilance: Ready for prime time? *Drug Saf.* 2022;45(5):429-38. doi:10.1007/s40264-022-01157-4
5. Warner J, Prada Jardim A, Albera C. Artificial intelligence: Applications in pharmacovigilance signal management. *Pharmaceut Med.* 2025;39(3):183-98. doi:10.1007/s40290-025-00561-2
6. Jain A, Salas M, Aimer O, Adenwala Z. Safeguarding patients in the AI era: Ethics at the forefront of pharmacovigilance. *Drug Saf.* 2025;48(2):119-27. doi:10.1007/s40264-024-01483-9
7. Ball R, Talal AH, Dang O, Muñoz M, Markatou M. Trust but verify: Lessons learned for the application of AI to case-based clinical decision-making from postmarketing drug safety assessment at the US Food and Drug Administration. *J Med Internet Res.* 2024;26:e50274. doi:10.2196/50274
8. Li Y, Tao W, Li Z, Sun Z, Li F, Fenton S, et al. Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. *J Biomed Inform.* 2024;152:104621. doi:10.1016/j.jbi.2024.104621
9. Modi S, Kasmiran KA, Sharef NM, Sharum MY. Extracting adverse drug events from clinical notes: A systematic review of approaches used. *J Biomed Inform.* 2024;151:104603. doi:10.1016/j.jbi.2024.104603
10. De Abreu Ferreira R, Zhong S, Moureaud C, Le MT, Rothstein A, Li X, et al. A pilot, predictive surveillance model in pharmacovigilance using machine learning approaches. *Adv Ther.* 2024;41(6):2435-45. doi:10.1007/s12325-024-02870-5
11. Farnoush A, Sedighi-Maman Z, Rasoolian B, Heath JJ, Fallah B. Prediction of adverse drug reactions using demographic and non-clinical drug characteristics in FAERS data. *Sci Rep.* 2024;14(1):23636. doi:10.1038/s41598-024-74505-2
12. McMaster C, Chan J, Liew DFL, Su E, Frauman AG, Chapman WW, et al. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. *J Biomed Inform.* 2023;137:104265. doi:10.1016/j.jbi.2022.104265
13. Sakhovskiy A, Tutubalina E. Multimodal model with text and drug embeddings for adverse drug reaction classification. *J Biomed Inform.* 2022;135:104182. doi:10.1016/j.jbi.2022.104182
14. 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
15. Liou L, Scott E, Parchure P, Ouyang Y, Egorova N, Freeman R, et al. Assessing calibration and bias of a deployed machine learning malnutrition prediction model within a large healthcare system. *NPJ Digit Med.* 2024;7(1):149. doi:10.1038/s41746-024-01141-5
16. Piccininni M, Wechsung M, Van Calster B, Rohmann JL, Konigorski S, van Smeden M. Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models. *J Biomed Inform.* 2024;155:104666. doi:10.1016/j.jbi.2024.104666
17. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc.* 2020;27(4):621-33. doi:10.1093/jamia/ocz228
18. Hong C, Liu M, Wojdyla DM, Hickey J, Pencina M, Henao R. Trans-Balance: Reducing demographic disparity for prediction models in the presence of class imbalance. *J Biomed Inform.* 2024;149:104532. doi:10.1016/j.jbi.2023.104532

19. Chaurasia A, Kumar D, Yogita. PregAN-NET: Addressing class imbalance with GANs in interpretable computational framework for predicting safety profile of drugs considering adverse reactions during pregnancy. *J Biomed Inform.* 2025;166:104832. doi:10.1016/j.jbi.2025.104832
20. Chalabianloo N, Ahmadi F, Omrani MA, Abdullah SS, Rostamzadeh N, Jafari A, et al. Machine learning methods for predicting adverse drug events: A systematic review. *Br J Clin Pharmacol.* 2026;92(2):422-44. doi:10.1002/bcp.70377
21. Silva GFDSS, Barcellos Filho FN, Wichmann RM, Costa da Silva Junior F, Chiavegatto Filho ADP. Strategies for detecting and mitigating dataset shift in machine learning for health predictions: A systematic review. *J Biomed Inform.* 2025;170:104902. doi:10.1016/j.jbi.2025.104902
22. Lea AS, Jones DS. Mind the gap—machine learning, dataset shift, and history in the age of clinical algorithms. *N Engl J Med.* 2024;390(4):293-5. doi:10.1056/NEJMp2311015
23. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. *NPJ Digit Med.* 2024;7(1):202. doi:10.1038/s41746-024-01200-x
24. Gavrilidis GI, Dimitriadis VK, Jaulent MC, Natsiavas P. Identifying actionability as a key factor for the adoption of ‘intelligent’ systems for drug safety: Lessons learned from a user-centred design approach. *Drug Saf.* 2021;44(11):1165-78. doi:10.1007/s40264-021-01103-w
25. 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
26. Collins GS, Dhiman P, Ma J, Schluskel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ.* 2024;384:e074819. doi:10.1136/bmj-2023-074819
27. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ.* 2025;388:e082505. doi:10.1136/bmj-2024-082505
28. Dirkson A, Verberne S, van Oortmerssen G, Gelderblom H, Kraaij W. How do others cope? Extracting coping strategies for adverse drug events from social media. *J Biomed Inform.* 2023;139:104228. doi:10.1016/j.jbi.2022.104228
29. Ward IR, Wang L, Lu J, Bennamoun M, Dwivedi G, Sanfilippo FM. Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes? *Comput Methods Programs Biomed.* 2021;212:106415. doi:10.1016/j.cmpb.2021.106415
30. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554. doi:10.1136/bmj-2024-081554
31. 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:e078378. doi:10.1136/bmj-2023-078378
32. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ.* 2024;384:e074820. doi:10.1136/bmj-2023-074820