Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Explainable Machine Learning for Classifying Documented Death Outcomes in FDA Adverse-Event Reports: A Multi-Quarter Exploratory Study and Pharmacovigilance Framework

**Nitin Triphati**[1]

Nitin.tripathipwn@gmail.com

Dr. Ram Manohar Lohia Avadh University

**Prashant Kumar Pandey**[2]

College of Technical University, Uttar Pradesh, Lucknow, India

**Research Article**

**Keywords:** pharmacovigilance, adverse event reporting, FAERS, AEMS, death outcome, explainable artificial intelligence, machine learning

**Posted Date:** November 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11369456/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Explainable Machine Learning for Classifying Documented Death Outcomes in FDA Adverse-Event Reports: A Multi-Quarter Exploratory Study and Pharmacovigilance Framework

Nitin Triphati[1]; Prashant Kumar Pandey[2]

[1] Institute of Engineering and Technology, Dr. Ram Manohar Lohia Avadh University, Ayodhya, Uttar Pradesh, India

[2] College of Technical University, Uttar Pradesh, Lucknow, India

Corresponding author: Nitin Triphati | Email: Nitin.tripathipwn@gmail.com | ORCID: https://orcid.org/0009-0006-4102-8407

Author email: Prashant Kumar Pandey | pandey.prash@gmail.com | ORCID: https://orcid.org/0009-0009-2713-7473

Author metadata: Nitin Triphati — Master in Computer Applications and Mathematics; Technical Assistant. Prashant Kumar Pandey — Master of Computer Applications (MCA); Tech Lead, Sopra Steria India.

**Highlights**

- Uses 21 FDA AEMS/FAERS quarterly extracts spanning 2020 Q1–2025 Q1 to study documented death outcomes in spontaneous adverse-event reports.
- Compares calibrated logistic regression and histogram-based gradient boosting with a chronological holdout and a prespecified calibration-based threshold-selection procedure.
- Adds subgroup evaluation and SHAP-based global interpretation while explicitly documenting limitations of same-report predictors.
- Separates retrospective report classification from prospective patient-level ADR prediction and drug-causality assessment.
- Provides an explainable-AI pharmacovigilance framework and a reproducible exploratory analysis, not a clinically deployable prediction tool.

Keywords: pharmacovigilance; adverse event reporting; FAERS; AEMS; death outcome; explainable artificial intelligence; machine learning

**Abstract**

Objective: To describe an explainable artificial intelligence pharmacovigilance framework and evaluate whether machine-learning models can classify FDA spontaneous adverse-event reports that document a death outcome, while clearly distinguishing retrospective report classification from prospective patient-level prediction.

Methods: We analyzed official FDA AEMS/FAERS ASCII quarterly extracts from 2020 Q1 through 2025 Q1 (21 quarters). Reports without a recorded outcome-table entry were excluded. A report was labelled positive when at least one OUTC row recorded the DE (death) code; other reports with recorded outcomes formed the comparison class. We excluded case identifiers listed in the supplied quarterly deletion files and retained one report version per CASEID using the highest CASEVERSION, then FDA receipt date and PRIMARYID as tie-breakers. A disk-

backed SQLite workflow resolved versions across quarters. A bounded sample was drawn within chronological fit (2020–June 2023), calibration (July–December 2023), and test (2024 Q1–2025 Q1) windows. Logistic regression and histogram-based gradient boosting were fitted, sigmoid-calibrated on the separate calibration sample, and assessed on the held-out test sample. The operating threshold for each model was selected by maximizing F1 on the calibration sample. Subgroup metrics and TreeSHAP explanations were evaluated on sampled test records.

Results: After outcome filtering, deletion-list exclusion, case-version resolution and removal of records without usable FDA receipt dates, the analysis cohort contained 4,534,679 report versions/cases, including 667,062 reports with a recorded DE outcome and 3,867,617 reports with other recorded outcomes. The model sample included 250,000 fit reports, 75,000 calibration reports and 150,000 chronological test reports. In the test sample, gradient boosting achieved ROC-AUC 0.722 and average precision 0.343, compared with 0.657 and 0.247 for logistic regression. At thresholds selected on the calibration sample to maximize F1, gradient boosting achieved death-class precision 0.304, recall 0.469, F1 0.369 and balanced accuracy 0.645; its Brier score was 0.110. SHAP summaries ranked reaction-row count, age, sex, reporter type, occupation code and country-related fields among the leading model features. These are report-level associations and may reflect reporting and documentation processes.

Conclusion: The multi-quarter analysis demonstrates that a carefully defined retrospective report-classification task can be evaluated reproducibly across FDA quarterly extracts. It does not establish prospective prediction of ADRs or patient deaths, drug causality, or clinical utility. Same-report reaction and drug fields may be recorded after or alongside the outcome, and the study used a sampled test set rather than a full-census holdout. Independent validation and a prospectively timed feature design would be required before any early-warning claim or clinical deployment.

## 1. Introduction

Pharmacovigilance is the science and set of activities concerned with detecting, assessing, understanding and preventing adverse effects and other medicine-related problems. The contemporary pharmacovigilance environment is no longer limited to manually reviewed spontaneous reports. Regulators and health systems increasingly work with large, heterogeneous streams of individual case safety reports (ICSRs), EHRs, claims, medication administration records, laboratory results, clinical narratives, registries, literature and patient generated information. The World Health Organization (WHO) emphasizes that pharmacovigilance exists to improve patient care and safety and to support assessment of the benefit-risk profile of medicines. This mandate becomes more difficult as the volume and complexity of safety information increase [1].

The scale of contemporary reporting illustrates the computational challenge (Figure 1). The European Medicines Agency (EMA) reported that more than 1.7 million suspected ADR reports were submitted to EudraVigilance in 2025 and that 1,201 potential signals were reviewed during the year. Approximately 68% of those potential signals originated from monitoring EudraVigilance. In the United States, a 2026 analysis of more than 32 million pharmacovigilance reports found annual reporting volumes above two million in several recent years, while cautioning that recent years are incomplete because reports continue to be processed and added.[2],[3]

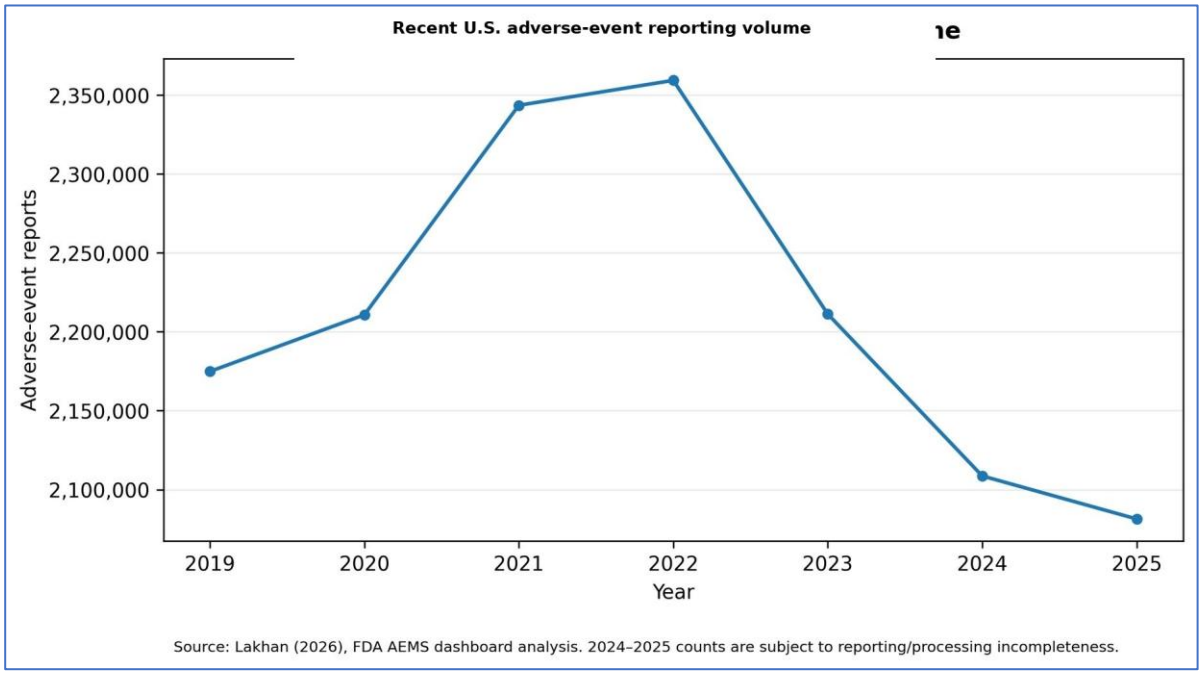



*Figure 1. Recent U.S. adverse-event reporting volume. Annual counts for 2019–2025 are reproduced from Lakhan (2026); recent-year counts are subject to reporting and processing incompleteness.*

AI is attractive in this environment because it can learn nonlinear relationships, prioritize high-risk cases, extract safety information from free text and integrate high-dimensional patient characteristics. Recent systematic reviews show that machine learning has been applied to both specific adverse drug-event (ADE) prediction tasks and multiple-ADE prediction using EHR data. A 2024 systematic review of specific ADEs identified 59 studies covering 15 drugs and 15 ADE categories and reported 38 machine-learning algorithms, with random forest the most frequently used approach. A separate 2024 systematic review and meta-analysis of multiple-ADE prediction included 10 studies and reported a mean summary area under the receiver-operating-characteristic curve of 0.76, with substantial heterogeneity between studies.[4–5]

The technical promise does not automatically translate into safe clinical use. An ADR prediction model can be statistically accurate while being clinically misleading if it learns proxies for healthcare utilization, documentation patterns or treatment intensity rather than patient vulnerability. A model may also produce an alert without providing enough information for a pharmacist or clinician to decide whether the medicine should be continued, stopped, monitored or investigated. Explainability is therefore not a cosmetic property. It can support error analysis, clinical review, model debugging, identification of data leakage and communication of uncertainty. The FDA and its international regulatory partners now explicitly connect transparency and explainability with safe and effective use of machine-learning-enabled medical devices, emphasizing intended use, workflow integration, performance, risks, limitations, bias and the logic underlying outputs [6–8].

The research gap is consequently not simply a shortage of predictive algorithms. It is the absence of a sufficiently integrated framework that connects prospective patient-level ADR prediction to trustworthy explanation, temporal validation, clinical action, human oversight and post-deployment surveillance. Existing evidence reviews have identified promising AI applications but also substantial weaknesses in external validation, reproducibility, clinical implementation and explainability. Dimitsaki et al. [9] found that only 6 of 36 studies in their structured-RWD pharmacovigilance review evaluated models in clinical environments and only 4 published code in public registries. Their review also highlighted the relative underuse of explainable and causal AI.

Dsouza et al. [10] similarly concluded that AI-based ADR prediction in hospitalized patients remains comparatively limited and fragmented.

This paper addresses that gap by proposing an Explainable AI Pharmacovigilance (EAI-PV) framework for future prospective patient-level ADR prediction. The objective is not to claim that a new algorithm has already been prospectively validated. Instead, the contribution is a methodologically explicit framework grounded in recent evidence and designed to guide future model development, validation and implementation. The paper asks four questions: (1) what data and AI approaches are most appropriate for prospective patient-level ADR prediction; (2) what forms of explanation are clinically meaningful; (3) what validation and governance controls are necessary before deployment; and (4) how should predictions and explanations be incorporated into the pharmacovigilance workflow?[11]

## 2. Methods: Evidence Synthesis and Framework Development

This manuscript combines (1) a structured narrative synthesis and framework-development exercise and (2) an exploratory retrospective classification analysis of 21 official FDA AEMS/FAERS quarterly ASCII extracts from 2020 Q1 through 2025 Q1. The study was designed to evaluate report-level classification, not patient-level clinical prediction. The data are spontaneous individual case safety reports and do not supply a population exposure denominator or independently adjudicated causal outcomes.

The evidence synthesis organized sources into five domains: pharmacovigilance data and signal detection; machine-learning methods for adverse drug event detection; clinical-text processing; explainable AI; and reporting, regulatory and lifecycle-governance standards. The empirical pipeline used the official quarterly tables DEMO, DRUG, REAC and OUTC, together with the quarterly deleted-case lists. Quarterly row counts were audited before the report-level cohort was assembled.

*Table 1. Evidence domains informing the proposed Explainable AI Pharmacovigilance framework.*

| Evidence domain | Recent evidence | Implication for EAI-PV |
|---|---|---|
| Structured RWD | 36 studies; 28/36 used EHRs; RF 17/36; only 6/36 tested clinically | Prioritize interoperable data, temporal validation and implementation evaluation. |
| EHR-based ADE prediction | 59 studies across 15 drugs and 15 ADE categories; 38 ML algorithms; RF most frequent | Use strong tabular baselines; evaluate imbalance, calibration and transportability. |
| Multiple-ADE meta-analysis | 10 studies; mean summary AUROC 0.76; substantial heterogeneity | Avoid reporting a single headline metric without context. |
| Clinical-text extraction | BERT and deep-learning approaches perform strongly for end-to-end ADE extraction | Use NLP as an upstream event-extraction layer with uncertainty checks. |
| Hospital ADR prediction | 2025 review found limited and fragmented evidence | Prospective and external validation remain priorities. |
| Trustworthy AI | Bias, reproducibility, external validation and explainability remain inconsistent | Governance must be embedded throughout the lifecycle. |

The positive target was a report with at least one OUTC row coded DE (death). The comparison class consisted of reports with at least one recorded OUTC outcome but no DE code. Reports with no OUTC record were excluded, not labelled non-serious. A DE code indicates a death outcome recorded in the spontaneous report; it does not prove that a suspected medicine caused the death. This outcome definition was chosen to avoid treating missing outcome information as a negative label.

To resolve repeat versions, we combined quarterly DEMO records, excluded CASEIDs appearing in the available quarterly deletion lists, and retained one version per CASEID by descending CASEVERSION, then FDA receipt

date (FDA_DT), then PRIMARYID as a deterministic tie-breaker. The final cohort contained 4,534,679 records after outcome filtering, deletion-list exclusion, case-version resolution and exclusion of records without a usable FDA receipt date. The cohort included 667,062 reports with a recorded DE outcome and 3,867,617 reports with other recorded outcomes. Because the latest available version is not necessarily an independently adjudicated final clinical record, this is an operational deduplication rule and a study limitation.

For computational feasibility, we drew a bounded random sample from chronological windows: 250,000 reports dated from 2020 through 30 June 2023 for model fitting, 75,000 reports dated July–December 2023 for probability calibration and threshold selection, and 150,000 reports dated 2024 Q1–2025 Q1 for testing. The exact selected analytic sample is supplied as a supplementary CSV so that the reported model-stage results can be reproduced without repeating the database sampling query. Test sampling was not a census of all eligible reports. Predictors included age converted to years where a usable age unit was supplied, sex, age group, report type, reporter occupation, reporting and occurrence country fields, electronic submission/manufacturer fields, counts of drug rows, non-empty route and dose fields, drug-role counts, and reaction rows. These variables are recorded within the same report as the target; their availability before the outcome was not established. Accordingly, the experiment is retrospective classification only.

The evidence synthesis also recognizes a fundamental difference between spontaneous-report signal detection and patient-level clinical prediction. Spontaneous reports are primarily useful for detecting disproportional patterns after exposure and reporting, whereas EHR-based models can estimate an individual's future risk before an event is fully documented. These functions are complementary. The EAI-PV framework therefore treats regulatory signal detection and clinical risk prediction as connected but distinct layers. A signal emerging in EudraVigilance or FAERS may inform candidate adverse outcomes and drug-event pairs, while EHR models can estimate which currently treated patients are most vulnerable. Conversely, clinically confirmed events can improve signal interpretation and model retraining.

A second methodological implication is that the framework must distinguish detection from prediction. Detection asks whether an ADR has already occurred or is described in a record. Prediction asks whether a patient is likely to experience an ADR in a defined future window. These endpoints require different labels, different leakage controls and different clinical workflows. A model trained to detect an event in notes cannot automatically be represented as a prospective risk model. The manuscript therefore recommends maintaining separate evaluation tracks for event extraction, risk prediction and signal detection.

The framework also treats severity as a structured outcome dimension. Serious ADRs may require a lower false-negative tolerance than mild reactions. A useful implementation may therefore use a hierarchical design in which a broad model identifies possible safety events and a second model prioritizes serious outcomes. Such a design can reduce the temptation to use one threshold for all events. Severity classification should be aligned with recognized regulatory definitions and the intended clinical decision. The major pharmacovigilance data sources, their strengths and limitations, and their recommended roles are summarized in Table 2 in Section 3.

## 3. Current Pharmacovigilance Data Landscape

Modern pharmacovigilance depends on multiple data streams because no single source captures the complete clinical safety picture. Spontaneous-report systems provide broad post-marketing surveillance but are affected by under-reporting, stimulated reporting, duplicate cases, missing denominators and reporting changes over time. EHRs provide longitudinal patient context, including diagnoses, medication exposure, laboratory values, procedures and temporal clinical trajectories, but they may contain missingness, coding inconsistencies, documentation bias and institutional differences. Claims datasets provide longitudinal utilization and reimbursement information but may have limited clinical detail. Clinical notes can contain early safety clues that are absent from structured fields,

although extracting them requires natural-language processing. Registries and linked datasets can provide deeper phenotyping but may be difficult to access and harmonize.

*Table 2. Recommended roles of major data sources in an integrated ADR prediction architecture.*

| Source | Strengths | Principal limitations | Recommended role |
|---|---|---|---|
| EHR / clinical data | Longitudinal patient context; labs; medication exposure | Missingness; site-specific coding; documentation bias | Primary clinical prediction layer |
| Clinical notes | Rich temporal and contextual safety information | NLP errors; negation; ambiguity; privacy | Event extraction and explanation context |
| Claims | Large longitudinal populations; utilization patterns | Limited clinical granularity; coding delay | Population validation and utilization outcomes |
| FAERS/AEMS | Large post-marketing safety surveillance; broad product coverage | No reliable exposure denominator; reporting bias; duplicates | External signal context and hypothesis generation |
| EudraVigilance | Large regulatory ICSR system; signal-management integration | Spontaneous-report limitations; heterogeneous reporting | Regulatory signal detection and external validation |
| Registries / linked RWD | Deep phenotyping; outcome confirmation | Access, linkage and governance burden | Confirmatory validation and causal studies |

EudraVigilance illustrates the volume and regulatory relevance of spontaneous safety reports. EMA's 2025 annual reporting indicates that more than 1.7 million suspected ADR reports were submitted during 2025, with 64% originating outside the European Economic Area. EMA reviewed 1,201 potential safety signals and the Pharmacovigilance Risk Assessment Committee assessed 60 signals. These figures show both the opportunity and the limitation of AI: automated methods can prioritize and structure enormous volumes of evidence, but signal confirmation remains a scientific and regulatory process rather than a purely predictive task [2].

The U.S. system is similarly large. The FDA's Adverse Event Monitoring System (AEMS) public dashboard was introduced on March 11, 2026, providing near-real-time access to data from the FDA Adverse Event Reporting System (FAERS). A 2026 analysis reported approximately 2.08 million adverse-event reports in 2025, following 2.21 million in 2023 and 2.11 million in 2024, while noting that recent-year counts may be incomplete because of reporting delays and ongoing case processing. The practical implication is that models using the newest surveillance data must distinguish between genuine changes in reporting and administrative maturation of the database.

For early prediction, the most valuable data are those that precede the ADR. Medication start date, dose, route, treatment duration, concomitant medications, renal and hepatic laboratory values, age, sex, comorbidities, prior ADR history and recent changes in physiology are often more informative than a static diagnosis list. The framework also recommends representing time explicitly. For example, creatinine trajectory may matter more than a single creatinine value, and the addition of a nephrotoxic medication may be more informative when combined with a recent decline in renal function.

The data layer should therefore create an exposure-centered patient timeline. Every feature should be tagged as available before, at or after the prediction time. Features occurring after the clinical decision point should be excluded. For clinical notes, the system should preserve the date and provenance of the statement. An extracted phrase such as "rash after amoxicillin" is materially different from a retrospective note written months later. The

same principle applies to laboratory data: values obtained after an ADR may describe the consequence rather than predict the event.

Data interoperability is another major challenge. Dimitsaki et al. found that most structured real-world-data studies used proprietary or non-standardized datasets and that only a small proportion made code publicly available. The EAI-PV framework recommends mapping medications, conditions and laboratory observations to recognized standards where feasible, including RxNorm or ATC for medicines, SNOMED CT or ICD for clinical conditions, and LOINC for laboratory measurements. Standardization does not remove bias, but it improves portability and facilitates external validation.

## 4. AI Approaches for Early ADR Detection and Prediction

The AI layer of EAI-PV should be heterogeneous rather than algorithmically monolithic. ADR prediction involves tabular clinical variables, longitudinal sequences and unstructured language, and different model families offer different trade-offs between performance, transparency and operational complexity. A robust development strategy begins with interpretable statistical baselines and then adds more complex models only when they provide clinically meaningful improvement.

Logistic regression remains an important baseline because its coefficients can be communicated directly and because it supports probability estimation. Regularized logistic regression is particularly useful when the number of predictors is large relative to the number of events. Its limitations are the need for appropriate feature specification and difficulty representing complex nonlinear interactions without explicit engineering. Nevertheless, it provides a reference point against which more complex methods should be evaluated.

Tree-based ensemble models are especially attractive for structured EHR data. Random forests can capture nonlinear interactions and are relatively robust to variable scaling. Gradient boosting methods, including XGBoost and LightGBM, can provide strong performance on heterogeneous tabular data. Hu et al. [5] found random forest to be the most frequently reported method in the EHR-ADE literature, while Dimitsaki et al. [9] likewise found random forest to be the most common approach among structured-RWD pharmacovigilance studies. These findings support using tree ensembles as practical benchmark models rather than assuming that deep learning is automatically superior.

Deep learning becomes more attractive when the problem is strongly temporal or multimodal. Recurrent networks, temporal convolutional models and transformers can represent sequences of medications, laboratory values and clinical events. Their value is greatest when the temporal structure itself contains signal that cannot be captured adequately by summary statistics. However, more complex architectures introduce greater requirements for training data, calibration, interpretability and external validation. In safety-critical settings, complexity should be justified by incremental clinical value.

NLP provides a complementary pathway. Clinical narratives may contain early descriptions of symptoms, suspected reactions, medication changes and clinician assessments that do not yet appear as coded adverse event fields. Li et al. [12] reviewed AI-based clinical-text ADE extraction and found that deep-learning models performed strongly for named-entity recognition, while BERT-based approaches performed particularly well in end-to-end tasks. Recent work also indicates that large language models can extract ADR information from social media and clinical text, although noisy language and domain-specific terminology remain important limitations.[12–13]

A practical architecture should separate extraction from risk prediction. An NLP model can convert free text into structured candidate events, confidence scores and temporal relations. These extracted features can then be passed to a calibrated risk model. This separation makes validation easier because the performance of the extraction layer and prediction layer can be evaluated independently. It also makes explanation more meaningful: the system can show that a prediction was influenced by a documented prior reaction, a new medication exposure and a recent laboratory trajectory, rather than presenting an opaque text embedding.

Ensemble or stacked architectures may improve robustness but can complicate explanations. The recommended approach is to maintain a stable reference model and compare every additional component against it. A model should be retained only if it improves clinically relevant outcomes after calibration and subgroup analysis. The goal is not to maximize AUROC in isolation; it is to reduce missed serious ADRs without overwhelming clinicians with low-value alerts.

## 5. Explainability as a Clinical Safety Requirement

Explainability should answer a practical question: why did the system classify this patient as being at increased risk, and what evidence should the clinician verify? This is different from simply displaying a feature-importance chart. Global explanations describe how a model behaves across a population, whereas local explanations describe the factors that influenced an individual prediction. Both are needed. Global explanations support model validation and scientific understanding; local explanations support patient-level review.

SHAP is a particularly useful method for tree-based ADR models because it provides additive feature contributions that can be aggregated globally or inspected for individual cases. Lundberg and Lee [14] introduced SHAP as a unified approach to assigning feature importance to individual predictions, and Lundberg et al. [15] demonstrated how tree-specific SHAP methods can connect local explanations to global model understanding. In an ADR setting, SHAP can express that a patient's recent renal deterioration, polypharmacy and exposure to a particular medicine increased predicted risk, while a normal laboratory trajectory reduced it.

LIME provides a different local approximation strategy. Ribeiro et al. [16] proposed learning an interpretable model around an individual prediction. LIME can be useful when the underlying model is not naturally compatible with SHAP, including some text and black-box settings. However, local surrogate explanations can be sensitive to the perturbation scheme and feature representation. For clinical deployment, the explanation method should therefore be evaluated rather than assumed to be faithful.

Counterfactual explanations can be even more actionable. Instead of only saying that a patient's risk is high, a counterfactual can ask which plausible changes would lower risk. In an ADR system, the output might identify that risk would be materially lower under a different medication exposure, a lower dose or improved renal function. Such statements must be carefully constrained because a predictive counterfactual is not automatically a causal recommendation. The framework therefore prohibits language implying that an intervention will prevent an ADR unless causal evidence supports that conclusion.

Explanation should also include uncertainty. A risk score of 0.72 produced from a familiar patient population with well-measured features is not equivalent to a score of 0.72 produced for a patient whose medication, demographic characteristics or laboratory pattern is poorly represented in the training data. Prediction intervals, ensemble disagreement, conformal methods or out-of-distribution indicators can communicate when the system is less certain. The clinical interface should show uncertainty in a way that supports rather than paralyzes decision making.

A further requirement is explanation stability. If a tiny change in an irrelevant input causes the top explanation to change dramatically, the explanation may be unsuitable for clinical use. Stability can be assessed across bootstrap samples, repeated model runs and clinically plausible perturbations. The framework recommends that explanation quality be evaluated using at least four dimensions: fidelity to the model, stability, clinical plausibility and actionability. A model should not be considered explainable merely because it generates attractive visualizations.

The regulatory perspective supports this broader interpretation. FDA transparency guidance emphasizes communicating intended use, users, workflow integration, performance, benefits, risks, limitations, bias and the basis of outputs. This aligns with the proposed EAI-PV principle that explainability is a lifecycle property. It begins with documentation of data provenance and model purpose, continues through patient-level explanations and human-AI interaction, and ends with post-market monitoring for performance degradation and unintended consequences [6–8].

## 6. Proposed EAI-PV Framework

The proposed EAI-PV framework contains six connected layers: data and signal sources; data quality and temporalization; prediction engine; explainability; clinical safety gate; and action/learning. The architecture is intentionally modular. A health system can implement the framework using different algorithms while retaining the same safety controls. Figure 3 summarizes the architecture.

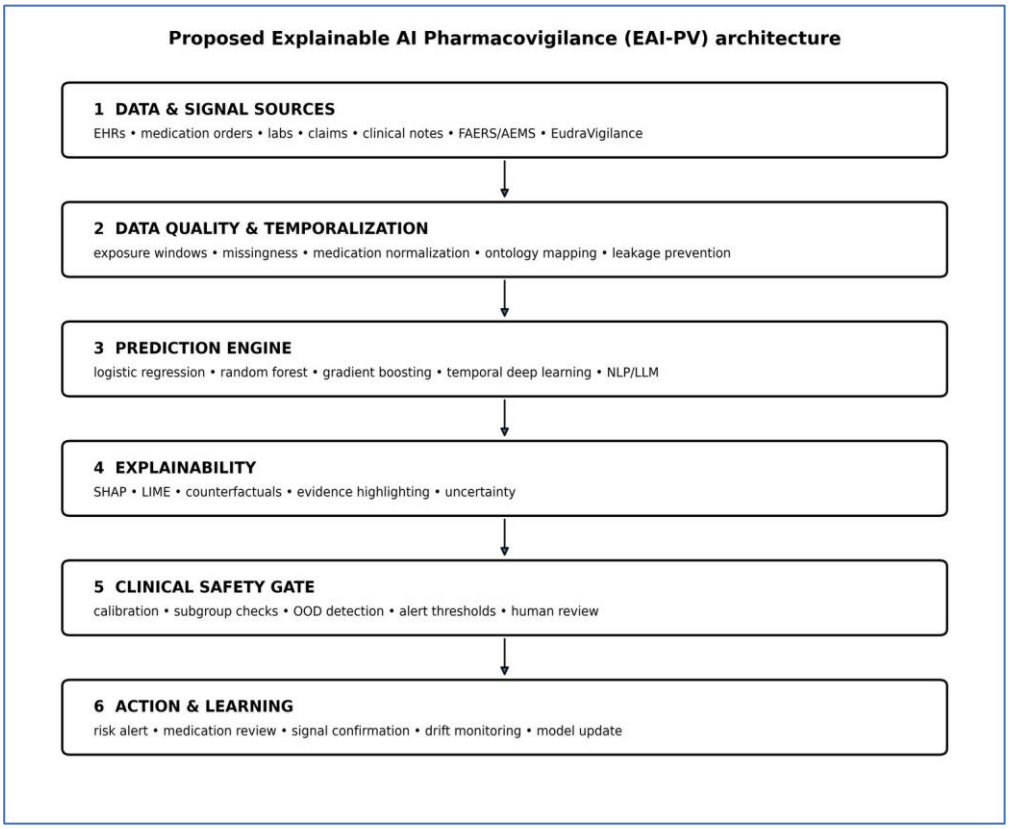



*Figure 3. Proposed Explainable AI Pharmacovigilance (EAI-PV) framework for early ADR detection and prediction.*

*Table 5. Operational components of the EAI-PV framework.*

| Layer | Core controls | Primary output | Responsible stakeholder |
|---|---|---|---|
| 1. Data sources | Provenance, consent/access, interoperability | Patient timeline + external signals | Data/clinical informatics |
| 2. Data quality | Temporal ordering, leakage checks, missingness, normalization | Validated feature set | Data science |
| 3. Prediction | Baseline + nonlinear model, calibration, thresholding | ADR risk probability | ML/clinical analytics |
| 4. Explanation | SHAP/LIME/counterfactuals, evidence highlighting, uncertainty | Patient-level explanation | ML + clinicians |
| 5. Safety gate | Fairness, OOD checks, alert governance, human review | Safe-to-display alert | Clinical governance |
| 6. Action/learning | Adjudication, feedback, drift monitoring, audit | Clinical action + validated labels | Pharmacovigilance team |

Layer 1—data and signal sources—collects structured and unstructured information. The framework recommends separating patient-level clinical data from population-level pharmacovigilance signals. Patient level data support individualized prediction, while FAERS/AEMS and EudraVigilance support external signal context and hypothesis generation. This separation prevents spontaneous-reporting artifacts from being treated as direct measures of individual risk.

Layer 2—data quality and temporalization—creates an auditable feature pipeline. It includes duplicate handling, missing-data characterization, medication normalization, unit harmonization, outlier checks, temporal ordering and leakage screening. Every feature should carry metadata indicating its source, timestamp, transformation and availability status. The model-development dataset should be frozen before the outcome window is observed. Data quality metrics should be monitored after deployment because changes in EHR templates, laboratory systems or medication coding can alter model inputs.

Layer 3—the prediction engine—uses a champion-challenger strategy. A transparent baseline is compared with one or more nonlinear models. The preferred model is not simply the one with the highest AUROC. It should demonstrate improvement in discrimination, calibration and decision utility while maintaining acceptable explanation quality. The framework recommends evaluating precision-recall curves because serious ADRs are often rare, and a high AUROC can coexist with poor positive predictive value in low prevalence settings.

*Table 3. Model families and recommended explainability approaches for EAI-PV.*

| Model family | Primary advantage | Key risk | Preferred XAI strategy |
|---|---|---|---|
| Regularized logistic regression | Transparent coefficients; stable baseline | Limited nonlinear representation | Coefficient plots; odds ratios; partial dependence |
| Random forest | Nonlinear interactions; robust tabular baseline | Complex local reasoning | TreeSHAP; permutation importance |
| Gradient boosting | High tabular predictive performance | Overfitting and calibration drift | TreeSHAP; calibration plots; counterfactuals |
| Temporal deep learning | Learns longitudinal patterns | Opacity; data hunger; drift | Integrated gradients; temporal attribution; attention inspection with caution |
| Transformer/NLP | Extracts ADR mentions and relations from text | Semantic errors; domain shift | Token/span attribution; evidence highlighting; confidence scoring |
| Hybrid/ensemble | Combines complementary signals | Complexity and explanation inconsistency | Layer-specific explanations + unified patient-level summary |

Layer 4—the explainability layer—generates both population and patient-level evidence. For a tree-based model, global SHAP summaries can identify dominant predictors while local SHAP values can explain an individual alert. For NLP, the system should highlight the clinical phrases or entities that triggered an extracted safety concept, with confidence and temporal context. The interface should distinguish evidence from inference: a statement documented in the record is evidence; a model's attribution is an explanation of its prediction; neither should be presented as proof of causality.

Layer 5—the clinical safety gate—is the central differentiator of the framework. Before an alert reaches a clinician, the system checks calibration, uncertainty, missingness, subgroup performance and explanation completeness. The safety gate can suppress alerts when the model is outside its validated population or when critical features are

missing. It can also route uncertain cases to pharmacist or pharmacovigilance review rather than generating a binary recommendation.

Layer 6—action and learning—connects the prediction to clinical workflow and future model improvement. Actions may include medication reconciliation, laboratory monitoring, review of alternative causes, adverse event documentation or escalation to a pharmacovigilance team. Human overrides should be recorded, but they should not automatically be interpreted as ground truth because overrides can reflect workflow constraints rather than model error. Confirmed outcomes should be curated through a clinically governed adjudication process before being used for retraining.

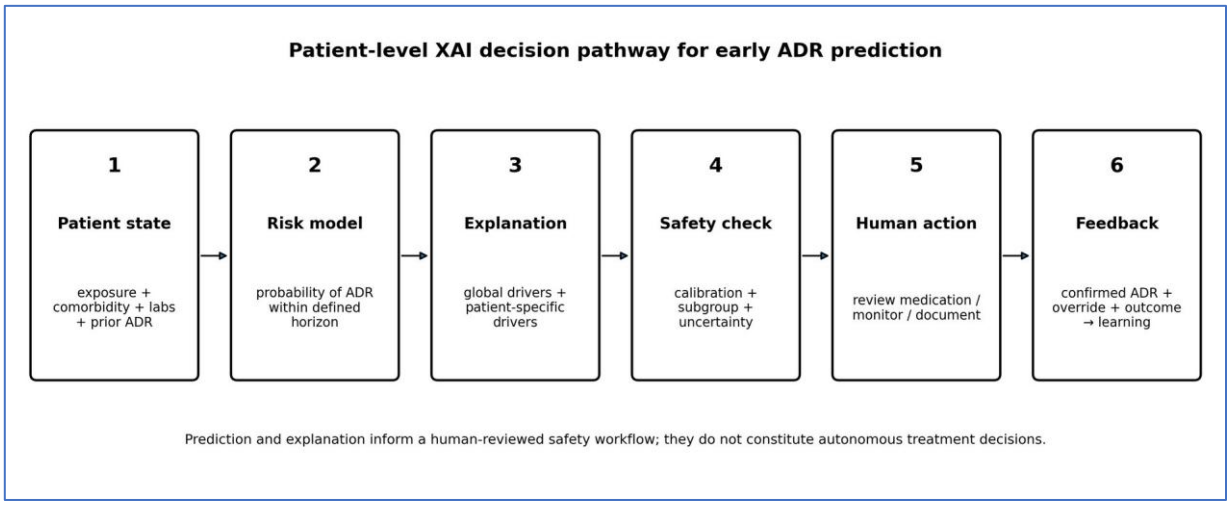



*Figure 2. Patient-level explainable-AI decision pathway for prospective patient-level ADR prediction. Prediction and explanation inform a human-reviewed safety workflow rather than an autonomous treatment decision.*

## 7. Validation Strategy and Evidence Standards

A clinically credible ADR prediction system requires validation beyond a random train-test split. Random splitting can place observations from the same patient or from the same clinical period in both training and test sets, producing optimistic estimates. The preferred design is temporal validation, in which earlier data are used for development and later data are held out. External validation should then evaluate the model in a different hospital, health system, population or time period.

Discrimination should be reported using AUROC and area under the precision-recall curve, but these metrics should not stand alone. A model can achieve a strong AUROC while being poorly calibrated. Calibration is particularly important because the output of an ADR model may be used to prioritize scarce clinical attention. If a model predicts 20% risk, clinicians need evidence that patients with similar predicted risks experience the outcome at approximately the expected rate, within the limitations of the setting.

Decision-curve analysis can estimate net benefit across clinically relevant thresholds. This is important because an alerting system has asymmetric consequences: missing a serious ADR can cause substantial harm, while excessive false positives can create alert fatigue and reduce trust. Thresholds should therefore be selected with clinical stakeholders and should be stratified by severity when appropriate. A model designed to identify life-threatening reactions should not use the same operating threshold as one designed to identify mild, self-limited reactions.

Subgroup evaluation is also essential. Performance should be examined across age groups, sex, race or ethnicity where legally and ethically appropriate, renal-function strata, care settings, medication classes and hospitals. The objective is not to guarantee identical performance in every group, which may be statistically unrealistic, but to identify clinically consequential disparities and determine whether recalibration or additional data are required. FDA transparency guidance specifically recognizes that transparency can help identify bias and support health equity.

Explanation validation should follow the same rigor as predictive validation. First, the system should test whether explanations are faithful to the model. Second, clinical experts should judge whether the explanations are plausible and useful. Third, stability should be measured across repeated samples and nearby cases. Fourth, the effect of explanations on clinician decisions should be evaluated prospectively. An explanation that clinicians find persuasive may still be misleading if it highlights correlated proxies rather than causal factors. Human-factors testing is therefore a necessary component of implementation.

TRIPOD+AI provides a reporting structure for clinical prediction model development and evaluation. CONSORT-AI and SPIRIT-AI provide complementary guidance for prospective evaluation of AI interventions, emphasizing clear descriptions of the AI system, human-AI interaction, input/output handling and error analysis. For EAI-PV, these reporting principles should be extended with explicit documentation of explanation generation, explanation limitations, data drift and alert management [11,17].

*Table 4. Minimum validation and governance standards for an EAI-PV implementation.*

| Evaluation dimension | Minimum recommendation | Why it matters for ADR prediction |
|---|---|---|
| Discrimination | AUROC + AUPRC with confidence intervals | Rare outcomes make precision-recall behavior important. |
| Calibration | Calibration plot, intercept/slope, Brier score | Risk probabilities must be clinically interpretable. |
| Temporal validity | Chronological holdout | Prevents future information leakage. |
| External validity | At least one independent site/time period | Tests portability and dataset shift. |
| Subgroup safety | Performance and calibration by clinically relevant strata | Detects unequal error and calibration. |
| Clinical utility | Decision-curve analysis + alert burden | Links statistics to workflow consequences. |
| Explainability | Fidelity, stability, plausibility, actionability | Prevents decorative explanations. |
| Implementation | Override rate, time-to-review, alert acceptance, outcomes | Tests real-world usefulness. |
| Lifecycle | Drift, recalibration and version control | Maintains safety after deployment. |

## 8. Results of the Evidence Synthesis and Exploratory FAERS Pilot

The evidence synthesis supports five broad conclusions. First, AI-based ADR/ADE detection and prediction are technically feasible, but study designs, outcomes, populations, and validation practices vary. Second, tree-based models are practical candidates for structured clinical data, although predictive performance must be interpreted alongside calibration and intended use. Third, clinical text can provide safety information but extraction is not equivalent to causality assessment. Fourth, external validation and reproducibility remain important gaps. Fifth, automation does not remove the need for pharmacovigilance expertise. These literature-derived conclusions are separate from the original pilot results reported below.

Multi-quarter cohort construction (Figure 4). The 21 quarterly extracts contained 9,138,968 DEMO rows in total before report-level filtering. Across the combined quarterly DEMO files, 3,700,445 rows had no matching recorded OUTC outcome and were excluded from the outcome-defined analysis. After excluding CASEIDs listed in the supplied deletion files, selecting the latest available CASEVERSION per CASEID and excluding reports without a parseable FDA_DT, 4,534,679 report-level records remained. Among these, 667,062 had a DE outcome code and

3,867,617 had recorded outcomes but no DE code. These counts refer to reports/case versions in the assembled files, not unique patients or population-level event rates.

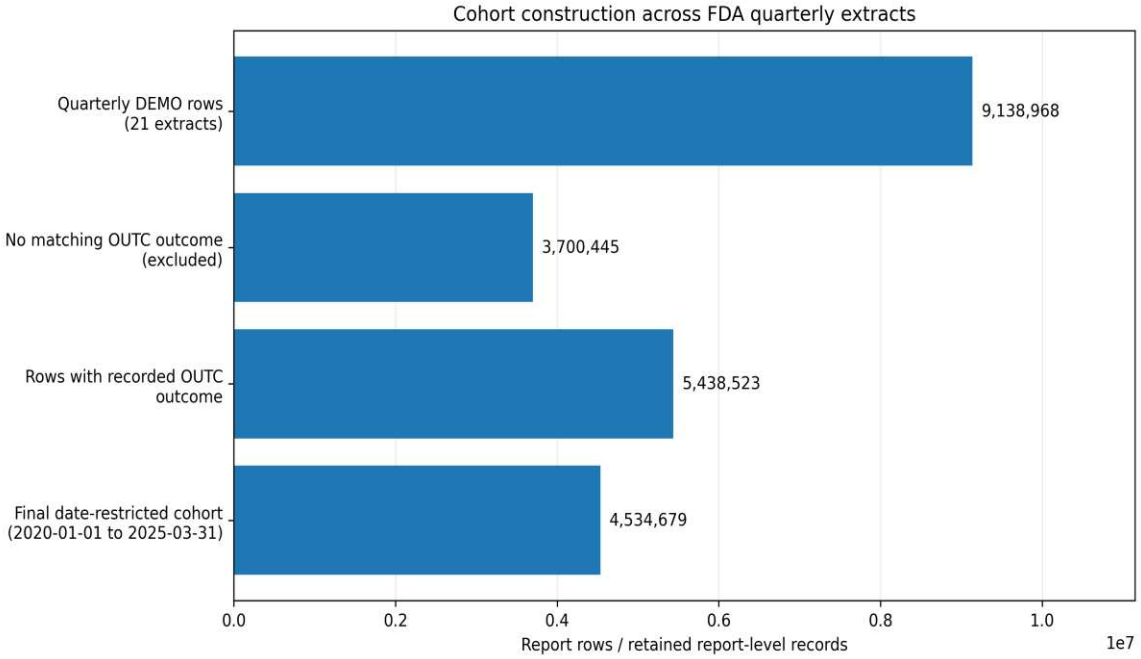



*Figure 4. Cohort construction across 21 quarterly extracts. Intermediate counts are not unique patients; the final cohort reflects additional deletion-list, case-version and date filters.*

Chronological model evaluation (Table 6 and Figure 5). The modelling sample comprised 250,000 fit reports, 75,000 calibration reports and 150,000 held-out test reports. On the test sample, histogram-based gradient boosting achieved ROC-AUC 0.722 and average precision 0.343; logistic regression achieved ROC-AUC 0.657 and average precision 0.247. For each model, the classification threshold was selected by maximizing F1 on the calibration sample, then applied without change to the test sample. Gradient boosting achieved precision 0.304, recall 0.469, F1 0.369 and balanced accuracy 0.645; logistic regression achieved precision 0.229, recall 0.528, F1 0.320 and balanced accuracy 0.617. The gradient-boosting Brier score was 0.110 versus 0.118 for logistic regression. These results indicate moderate discrimination in this sampled retrospective task, not clinical predictive performance.

*Table 6. Chronological test-set performance for classifying reports with a recorded death outcome (positive class).*

Table 6 reports discrimination, calibration and threshold-dependent classification metrics for the sampled chronological test set. The operating thresholds were selected on the calibration sample to maximize F1 and were then frozen for the test evaluation. This threshold-selection rule is a statistical operating choice, not a clinical utility threshold; no decision-curve analysis or prospective alert-burden assessment was performed. The reported threshold is the value selected on the calibration sample, not a clinically established decision threshold.

| Model | Test n | ROC-AUC | Average precision | Death precision | Death recall | Death F1 | Balanced accuracy | Threshold | Brier score |
|---|---|---|---|---|---|---|---|---|---|
| Logistic regression | 150,000 | 0.657 | 0.247 | 0.229 | 0.528 | 0.320 | 0.617 | 0.152 | 0.118 |
| HistGradientBoosting | 150,000 | 0.722 | 0.343 | 0.304 | 0.469 | 0.369 | 0.645 | 0.180 | 0.110 |

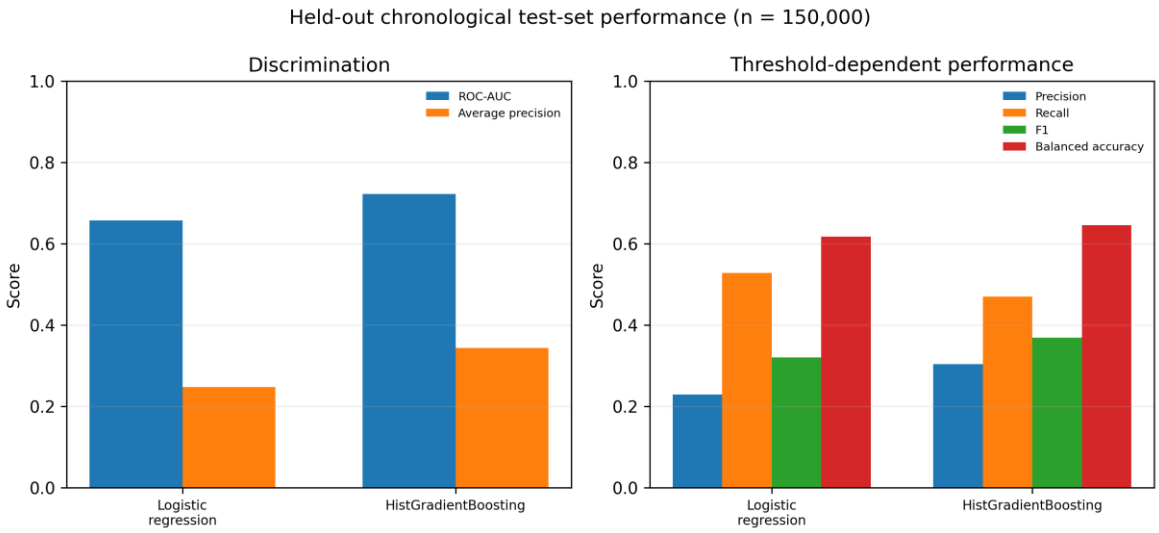



*Figure 5. Discrimination and threshold-dependent performance on the 150,000-report chronological test sample. Thresholds were selected on the separate calibration sample by maximizing F1.*

Subgroup evaluation. For the gradient-boosting model, discrimination, average precision, precision, recall, F1 and Brier score were calculated for groups with at least 100 sampled test records across sex, FDA age-group code, report type and reporter country. Performance varied across these groups, and group-specific sample sizes and outcome prevalence differed. The subgroup analysis is exploratory: it does not establish fairness, does not cover all clinically relevant strata, and can be affected by missing demographic information and reporting practices. Full values are supplied in the supplementary subgroup-metrics CSV.

SHAP interpretation (Figure 6). TreeSHAP was computed for a 1,000-report test subsample. The largest mean absolute SHAP values were observed for reaction-row count, age in years, sex, report type, occupation code, reporter country, an encoded drug-role count and age-group code. The gradient-boosting model used ordinal encoding for categorical inputs; therefore, SHAP values should be interpreted as explanations of this fitted model and encoding, not as causal effects or clinically validated risk factors. The results also highlight a central limitation: report completeness and documentation volume may be associated with the recorded outcome.

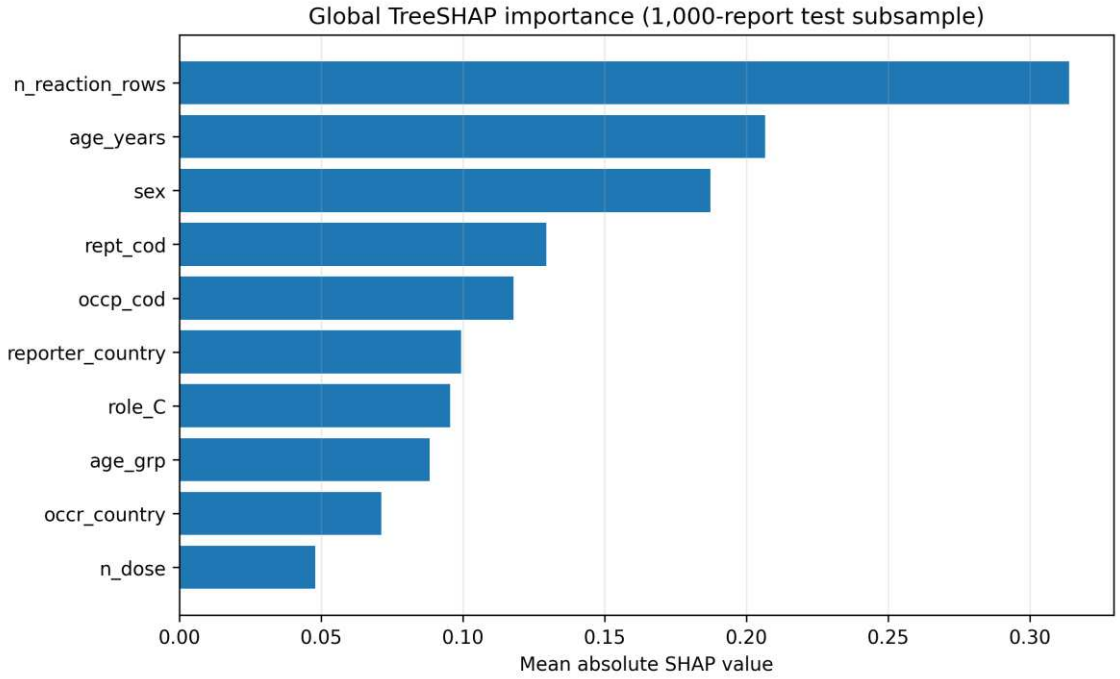



*Figure 6. Mean absolute TreeSHAP values for the ten highest-ranked features in a 1,000-report held-out test subsample. These are predictive attributions, not causal effects.*

Global SHAP rankings are provided in the supplementary materials. Reaction-row count was the leading feature by mean absolute SHAP value, followed by age, sex, report type, occupation code and country-related fields. These associations may arise from reporting intensity, case documentation, clinical severity, or the way spontaneous reports are completed. They must not be interpreted as causal determinants of death or as evidence that the model can identify risk before the outcome occurs.

## 9. Discussion

The framework and pilot address complementary but distinct problems. The EAI-PV framework describes how future patient-level ADR prediction systems should connect temporal data, explanation, calibration, human review, and lifecycle governance. The FAERS pilot instead classifies whether a spontaneous report records a death outcome. The latter cannot validate prospective clinical prediction, because the reports are submitted after an event and the available predictors were not proven to be available before the outcome.

A major implication is that explainability should influence model selection from the beginning. If a marginal performance improvement from a highly opaque model cannot be translated into a stable, clinically useful explanation, the model may be less suitable for deployment than a slightly less accurate but better calibrated and more interpretable model. This is particularly important in ADR prediction because clinical action is often conditional on context. A pharmacist may need to know not simply that risk is elevated, but whether the elevation is associated with renal impairment, polypharmacy, a prior reaction, dose escalation or a recent physiological change.

The gradient-boosting model outperformed the logistic-regression baseline in ROC-AUC, average precision, F1 and balanced accuracy on the sampled chronological test set, although logistic regression had higher death-class recall at the calibration-selected F1 threshold. The results show that nonlinear models can improve ranking in this report-level task, but do not establish clinical utility. The death outcome prevalence in the test sample was approximately 14.2%; precision and recall depend on this reporting-system-specific case mix and cannot be transported directly to clinical populations.

Label quality remains a central concern. Here, DE was defined from the FDA OUTC outcome code and missing OUTC entries were excluded rather than assumed negative. Even with this conservative rule, the label reflects what

was recorded in the spontaneous report, not independent adjudication. Spontaneous reports can be incomplete, duplicated across reporting periods, stimulated by publicity, and affected by differences in reporting practices. The pilot therefore estimates classification of recorded report content, not incidence, patient-level risk, or causal drug effects.

The use of LLMs creates additional opportunities and risks. LLMs can summarize clinical narratives, identify candidate ADR mentions and normalize free-text descriptions. However, the fluent nature of generated language can obscure uncertainty. The system should never permit an LLM to invent a drug-event relationship that is not grounded in source evidence. Evidence spans, timestamps and confidence scores should accompany every extracted safety claim. The LLM should function as a controlled information-extraction component, not an autonomous causal reasoner.[18]

Regulatory expectations reinforce the need for lifecycle governance. The FDA's current transparency principles emphasize intended users, workflow, performance, risks, limitations, bias and logic. The 2025 IMDRF Good Machine Learning Practice guidance frames safe AI development as a lifecycle activity. The EU AI Act establishes requirements for certain high-risk AI systems, including AI systems that function as safety components of regulated products, reinforcing the importance of risk management, testing, documentation and lifecycle oversight.[6–8,19]

The framework also has implications for health equity. ADR risk is not uniformly distributed, and differences in medication exposure, comorbidity, healthcare access, laboratory monitoring and documentation can create systematic disparities. If a model is trained predominantly in one population, an explanation that appears clinically reasonable may still encode historical patterns of unequal care. Subgroup calibration and error analysis should therefore be considered part of pharmacovigilance safety assessment rather than an optional fairness exercise.

Finally, the framework supports a feedback loop between clinical and regulatory pharmacovigilance. Clinical systems can identify patient-level patterns earlier, while regulatory databases can reveal population-level signals that are difficult to detect in a single health system. A mature EAI-PV ecosystem could exchange validated signals, ontology mappings and model monitoring information across institutions while maintaining privacy. Such an ecosystem would move pharmacovigilance from largely retrospective signal identification toward continuous, evidence-linked risk management.[20–21]

## 10. Practical Implementation Roadmap

Implementation should occur in stages. Stage 1 establishes the data foundation: medication exposure, laboratory measurements, diagnoses, demographics, prior reactions and relevant clinical narratives are harmonized into a longitudinal patient timeline. Stage 2 develops transparent baselines and defines a small number of clinically meaningful ADR endpoints. Stage 3 adds nonlinear models and explanation methods. Stage 4 performs temporal and external validation. Stage 5 conducts silent-mode prospective evaluation in which predictions are generated but do not affect care. Stage 6 begins controlled clinical deployment with monitoring. Stage 7 expands to multi-site learning and regulatory signal integration.

A silent-mode period is especially valuable because it reveals alert volume, subgroup performance, dataquality failures and explanation behavior without exposing clinicians or patients to unvalidated recommendations. During this period, the system can compare predictions against adjudicated outcomes and assess whether the chosen threshold produces a manageable alert burden. Human factors interviews should examine whether clinicians understand the risk score, whether they trust the explanation appropriately and whether the alert arrives at a useful point in the workflow.

Governance should assign explicit responsibilities. Data stewards should monitor data quality and provenance. Data scientists should monitor model performance, drift and explanation stability. Pharmacovigilance experts should adjudicate safety outcomes and interpret signals. Clinicians and pharmacists should assess workflow fit and

actionability. A governance committee should approve model releases, threshold changes and retraining. Version control should preserve the exact model, feature pipeline and explanation method used for every clinical alert.

The system should also include a rollback mechanism. If calibration deteriorates, a data source changes unexpectedly or an unexplained subgroup disparity appears, the model should be capable of being disabled or reverted to a validated prior version. This is consistent with the lifecycle orientation of good machine-learning practice and with the FDA's emphasis on maintaining safety and effectiveness over time.

Subgroup metrics are reported in Supplementary Table S2. In the sampled test set, model performance varied by sex, age-group code, report type and reporter country. These differences are descriptive and should not be interpreted as evidence of fairness or unfairness without accounting for missingness, outcome prevalence, sample size and reporting mechanisms.

## 11. Limitations

The empirical component is exploratory and has several limitations. First, the data are spontaneous reports and are subject to under-reporting, stimulated reporting, missingness, duplicates, follow-up updates and reporting-system changes. The target is the presence of a DE code in OUTC, not independently adjudicated cause-specific mortality; absence of DE among reports with other recorded outcomes is a comparison label for this task, not a guarantee that no death occurred. Second, reports without OUTC entries were excluded, limiting generalizability. Third, case-version resolution used the latest available CASEVERSION followed by FDA_DT and PRIMARYID, while all CASEIDs listed in supplied quarterly deletion files were excluded; this operational rule may not perfectly reproduce every FDA case-history convention.

Fourth, computational constraints required a sampled cohort for model fitting, calibration and testing. The chronological test was a reproducible 150,000-report sample from 2024 Q1–2025 Q1 rather than a full-census evaluation. Fifth, predictors were derived from the same report as the target, and their timestamps were not verified to precede the death outcome. The experiment therefore cannot support an early-warning or patient-level mortality-prediction claim. Sixth, gradient boosting used ordinal encoding for categorical variables, which can impose artificial ordering. Calibration was assessed with sigmoid calibration, Brier score and log loss, but calibration slope/intercept and decision-curve analysis were not estimated. Subgroup analyses were exploratory and no independent external dataset validation or clinician evaluation of explanations was performed. SHAP explains the fitted model, not causal mechanisms.

## 12. Conclusion

This study presents the EAI-PV framework and an exploratory multi-quarter analysis of documented death outcomes in FDA spontaneous adverse-event reports. Across 21 quarterly extracts from 2020 Q1 through 2025 Q1, the operationally deduplicated, outcome-defined cohort contained 4,534,679 report-level records, including 667,062 with a recorded DE code. In a sampled chronological test set of 150,000 reports, histogram-based gradient boosting achieved ROC-AUC 0.722 and average precision 0.343, outperforming logistic regression on these discrimination measures. At a threshold selected on the calibration sample to maximize F1, gradient boosting achieved precision 0.304, recall 0.469, F1 0.369, balanced accuracy 0.645 and Brier score 0.110.

The results should be interpreted as retrospective report classification, not early ADR prediction, prospective patient-level mortality prediction, or drug-causality estimation. Same-report reaction and drug fields may encode documentation and outcome-related information, and no external validation was performed. Before the framework can support early detection, future work should use patient-level longitudinal data with features explicitly timestamped before the prediction horizon, adjudicated outcomes, independent temporal/site validation, calibration and decision-curve assessment, clinically prespecified thresholds, subgroup safety review and prospective silent-mode evaluation.

## Author Contributions

Nitin Triphati: conceptualization, evidence synthesis, framework development, data curation, formal analysis, interpretation, and drafting. Prashant Kumar Pandey: methodological review, interpretation, manuscript review and editing. The authors should confirm that these contribution statements accurately reflect their work and approve the final submitted version.

Funding

No external funding was reported for this work.

## Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this manuscript. Both authors should confirm this declaration before submission.

## Data Availability

The analysis used official FDA AEMS/FAERS ASCII quarterly extracts from 2020 Q1 through 2025 Q1, obtained from https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html. The submitted supplementary package includes the processing and analysis scripts, quarter audit, sampled analysis cohort, model metrics, subgroup metrics, SHAP summary and reproducibility notes. Because the FDA source files are public and are not redistributed in full in the package, readers should download the same quarterly extracts from the FDA source. The model sample is a bounded reproducible sample rather than a full-census analytic dataset.

## Ethics Statement

This study analyzed publicly downloadable FDA spontaneous adverse-event reports and did not involve direct participant contact or intervention. The analytic fields used were report-level and did not include direct patient identifiers. The authors should confirm any institutional determination or exemption required by their institutions before submission; the manuscript does not claim a formal IRB determination.

Clinical trial number: not applicable.

Consent to Participate declaration: not applicable.

Consent to Publish declaration: not applicable.

## Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During manuscript preparation, the authors used OpenAI ChatGPT for language refinement, organization and formatting assistance, and used computational tools to support data processing and analysis. The authors are responsible for verifying all claims, citations, code, analyses and the final manuscript. The authors should ensure this statement conforms to the journal's current policy and disclose any additional AI-assisted work as required.

## References

Algarvio RC, Conceição J, Rodrigues PP, Ribeiro I, Ferreira-da-Silva R. Artificial intelligence in pharmacovigilance: a narrative review and practical experience with an expert-defined Bayesian network tool. Int J Clin Pharm. 2025;47(4):932-944. doi:10.1007/s11096-025-01975-3.

Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B. 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.

Dimitsaki S, Natsiavas P, Jaulent MC. Applying AI to structured real-world data for pharmacovigilance purposes: scoping review. J Med Internet Res. 2024;26:e57824. doi:10.2196/57824.

Dsouza VS, Leyens L, Kurian JR, Brand A, Brand H. Artificial intelligence (AI) in pharmacovigilance: a systematic review on predicting adverse drug reactions (ADR) in hospitalized patients. Res Soc Adm Pharm. 2025;21(6):453-462. doi:10.1016/j.sapharm.2025.02.008.

Elbiach O, Grissette H, Nfaoui EH. Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts. Results Eng. 2025;28:107362. doi:10.1016/j.rineng.2025.107362.

European Medicines Agency. Human medicines: EMA annual report 2025. Amsterdam: European Medicines Agency; 2026.

European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. 2024.

Hu Q, Chen Y, Zou D, He Z, Xu T. Predicting adverse drug event using machine learning based on electronic health records: a systematic review and meta-analysis. Front Pharmacol. 2024;15:1497397. doi:10.3389/fphar.2024.1497397.

Hu Q, Li J, Li X, Zou D, Xu T, He Z. Machine learning to predict adverse drug events based on electronic health records: a systematic review and meta-analysis. J Int Med Res. 2024;52(12):03000605241302304. doi:10.1177/03000605241302304.

International Medical Device Regulators Forum. Good machine learning practice for medical device development: guiding principles. IMDRF/AIML WG/N88 FINAL:2025. 2025.

Lakhan SE. Early evaluation of the Food and Drug Administration (FDA) Adverse Event Monitoring System (AEMS): an analysis of over 32 million pharmacovigilance reports. Cureus. 2026;18(4):e106259. doi:10.7759/cureus.106259.

Li Y, Tao W, Li Z, Sun Z, Li F, Fenton S, Xu H, Tao C. 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.

Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364-1374. doi:10.1038/s41591-020-1034-x.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56-67. doi:10.1038/s42256-019-0138-9.

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. 2017. p.4765-4774.

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.

Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p.1135-1144. doi:10.1145/2939672.2939778.

U.S. Food and Drug Administration. FDA Adverse Event Reporting System (FAERS) / Adverse Event Monitoring System (AEMS): latest quarterly data files. Available at: https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html. Accessed 12 October 2026.

U.S. Food and Drug Administration. FDA Adverse Event Reporting System (FAERS): questions and answers. Available at: https://www.fda.gov/drugs/surveillance/fda-adverse-event-reporting-system-faers-questions-and-answers. Accessed 12 October 2026.

U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles. 2021.

U.S. Food and Drug Administration. Transparency for machine learning-enabled medical devices: guiding principles. Silver Spring (MD): FDA; 2024.

World Health Organization. The WHO Programme for International Drug Monitoring. Geneva: World Health Organization; updated 2026.

Zitu MM, Owen D, Manne A, Wei P, Li L. Large language models for adverse drug events: a clinical perspective. J Clin Med. 2025;14(15):5490. doi:10.3390/jcm14155490.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DiscoverAIAuditedSupplementaryMaterials.zip