International Journal of Basic & Clinical Pharmacology
Dubey AD et al. Int J Basic Clin Pharmacol. 2026 Jul;15(4):775-780
https://www.ijbcp.com
pISSN 2319-2003 | eISSN 2279-0780

DOI: https://dx.doi.org/10.18203/2319-2003.ijbcp20261968

Short Communication

ADR•X: an interpretable, leakage-aware machine learning framework for sertraline adverse drug reaction signal detection using FAERS pharmacovigilance data

Adarsh Dheeraj Dubey*, Ranjana Mangesh Parab, Sermarani Nadar, Gursimran Kaur Uppal

Department of Bioinformatics, Guru Nanak Khalsa College of Arts, Science and Commerce (Autonomous), Matunga, Mumbai, Maharashtra, India

Received: 03 April 2026
Revised: 09 May 2026
Accepted: 19 May 2026

***Correspondence:**
Adarsh Dheeraj Dubey,
Email: g24.adarshdheeraj.dheerajdubey@gnkhalsa.edu.in

ABSTRACT

Adverse drug reactions (ADRs) are among the leading causes of preventable patient harm globally, and while the FDA Adverse Event Reporting System (FAERS) offers the most comprehensive post-marketing safety repository available, most published machine learning (ML) studies that work with this database introduce information leakage by incorporating outcome-derived disproportionality metrics—proportional reporting ratios (PRR) and reporting odds ratios (ROR)—directly as model features, thereby inflating performance estimates and undermining real-world generalisability. This study presents ADR•X, a LightGBM-based, leakage-aware framework designed to detect sertraline ADR signals from FAERS data using an approximately 208-variable feature space spanning patient demographics, physicochemical molecular descriptors, pharmacogenomic indicators, biology-guided multi-omics proxy variables, and mechanistic interaction terms, with all PRR-, ROR-, and frequency-derived variables explicitly excluded. Two model configurations were evaluated: an unweighted baseline and an inverse class-frequency-weighted variant. The baseline achieved an AUC-ROC of 0.53–0.54 and the imbalance-adjusted model reached 0.55–0.56. Global SHAP analysis identified dose mg, metabolic overload score, and polypharmacy flag as the three most influential predictors, while all remaining features clustered near zero, confirming the absence of leakage-driven dominance. The framework was deployed as a reproducible Streamlit research portal and is intended exclusively for population-level hypothesis generation, not individual clinical risk prediction. Modest AUC values reflect the bounded information content of voluntary reporting systems and represent honest signal estimation rather than model inadequacy. ADR•X demonstrates that biologically plausible and interpretable ADR signal detection is achievable from FAERS data without sacrificing methodological integrity.

Keywords: Adverse drug reactions, Pharmacovigilance, Sertraline, Machine learning, FAERS, Leakage-aware modelling, SHAP, LightGBM

INTRODUCTION

Adverse drug reactions (ADRs) are defined as unintended, harmful responses to medications occurring within the normal therapeutic dose range.[1] They account for a substantial proportion of preventable hospital admissions and medication-related deaths worldwide and a large fraction go undetected during pre-approval trials owing to restricted sample sizes, homogeneous study populations, and short follow-up windows.[2] post-marketing pharmacovigilance is therefore indispensable. The FDA Adverse Event Reporting System (FAERS), the largest

spontaneous reporting database in the world, contains millions of anonymised adverse event records submitted by healthcare professionals, manufacturers and patients, enabling population-level safety signal detection at a scale that no clinical trial could achieve.[3] Machine learning has been increasingly applied to FAERS data to surface signals that conventional disproportionality methods—proportional reporting ratios (PRR) and reporting odds ratios (ROR)—might otherwise miss.[4] A persistent methodological problem, however, is that many published studies include these very PRR and ROR values, or raw report counts, directly as model input features. Because such metrics are derived from adverse event frequencies, their inclusion transfers outcome information into the predictor space, constituting a classic case of target leakage.[5] The consequence is artificially elevated AUC scores that do not generalise to new data. Responsible application of ML in clinical and regulatory settings demands rigorous separation of covariate-derived information from outcome-derived statistics.[6] Sertraline, a selective serotonin reuptake inhibitor (SSRI) indicated for major depressive disorder, generalised anxiety disorder, and several related conditions, was chosen as the model compound because of its broad post-market use, well-documented pharmacology, and a wide adverse effect spectrum affecting the gastrointestinal, central nervous system (CNS), and cardiovascular systems.[7] Its primary hepatic metabolism via CYP2C19 makes it particularly well-suited for pharmacogenomic investigation.[8] This paper presents ADR•X, a fully leakage-aware LightGBM framework for sertraline ADR signal detection from FAERS data, deployed as an interactive Streamlit research portal with integrated SHAP-based interpretability.

METHODS

Data source and preprocessing

FAERS quarterly records linked to sertraline exposure were downloaded, deduplicated by primary case identifier, and processed through a schema-frozen deterministic pipeline. Binary ADR signal labels were constructed before any feature engineering step to prevent circular inference. All PRR-, ROR-, and report-count-derived variables were explicitly excluded from the feature matrix. Only adult case records were retained. The dataset was treated throughout as observational pharmacovigilance evidence rather than clinically confirmed diagnoses.

Feature engineering

A total of approximately 208 features were generated to comprehensively capture patient-, drug-, and biology-related factors potentially influencing outcomes. These features were grouped into six major categories. The first category included patient demographic characteristics, such as age and sex. The second comprised clinical exposure indicators, including drug dose category, the presence of polypharmacy and a proxy measure of hepatic impairment. The third category consisted of physicochemical properties of drugs derived using RDKit, including molecular weight, lipophilicity (logP), topological polar surface area (TPSA), the number of hydrogen bond donors and acceptors, and the number of rotatable bonds. The fourth category incorporated biology-guided multi-omics proxy variables designed to reflect key pathophysiological processes such as neuroinflammation, oxidative stress, blood–brain barrier integrity, cytokine activation and overall metabolic burden. The fifth category included pharmacogenomic indicators, representing CYP2C19 and CYP2D6 metaboliser phenotypes as well as serotonin transporter (SLC6A4) expression status, derived from curated genotype–phenotype knowledge bases. Finally, the sixth category comprised mechanistic interaction terms that captured biologically plausible interactions between variables, including age with dose, polypharmacy with liver disease, and oxidative stress with neuroinflammation. Together, these features provided a multidimensional and interpretable representation of factors relevant to the predictive modelling framework. All features were initialised at zero before selective population and the full schema was frozen prior to model training to ensure reproducible inference.

Model development and evaluation

Two LightGBM classifiers were developed: an unweighted baseline optimised for conservative probability estimation and an imbalance-adjusted variant employing inverse class-frequency weighting to improve sensitivity toward ADR-positive records.[10] Both configurations used 300 boosting rounds, a learning rate of 0.05, maximum tree depth of 6, and L1/L2 regularisation. An 80:20 stratified split with random seed 42 separated training from validation data.

Primary evaluation relied on AUC-ROC, supplemented by probability calibration assessment and sensitivity at a 0.5 decision threshold. Classification accuracy was deliberately deprioritised given the severe class imbalance characteristic of spontaneous reporting databases.

Explainability

SHAP Tree Explainer was applied to compute exact Shapley values for each prediction.[11,12] Global attribution was summarised as mean SHAP value, bar charts across the full validation set. Local attribution was examined through per-case waterfall plots. All SHAP outputs were cross-referenced against established sertraline pharmacology to assess biological plausibility.

Ethical considerations

All data originated from the publicly available and anonymised FAERS database maintained by the U.S. FDA. No personally identifiable patient information was accessed or stored. Institutional ethics approval was not required.

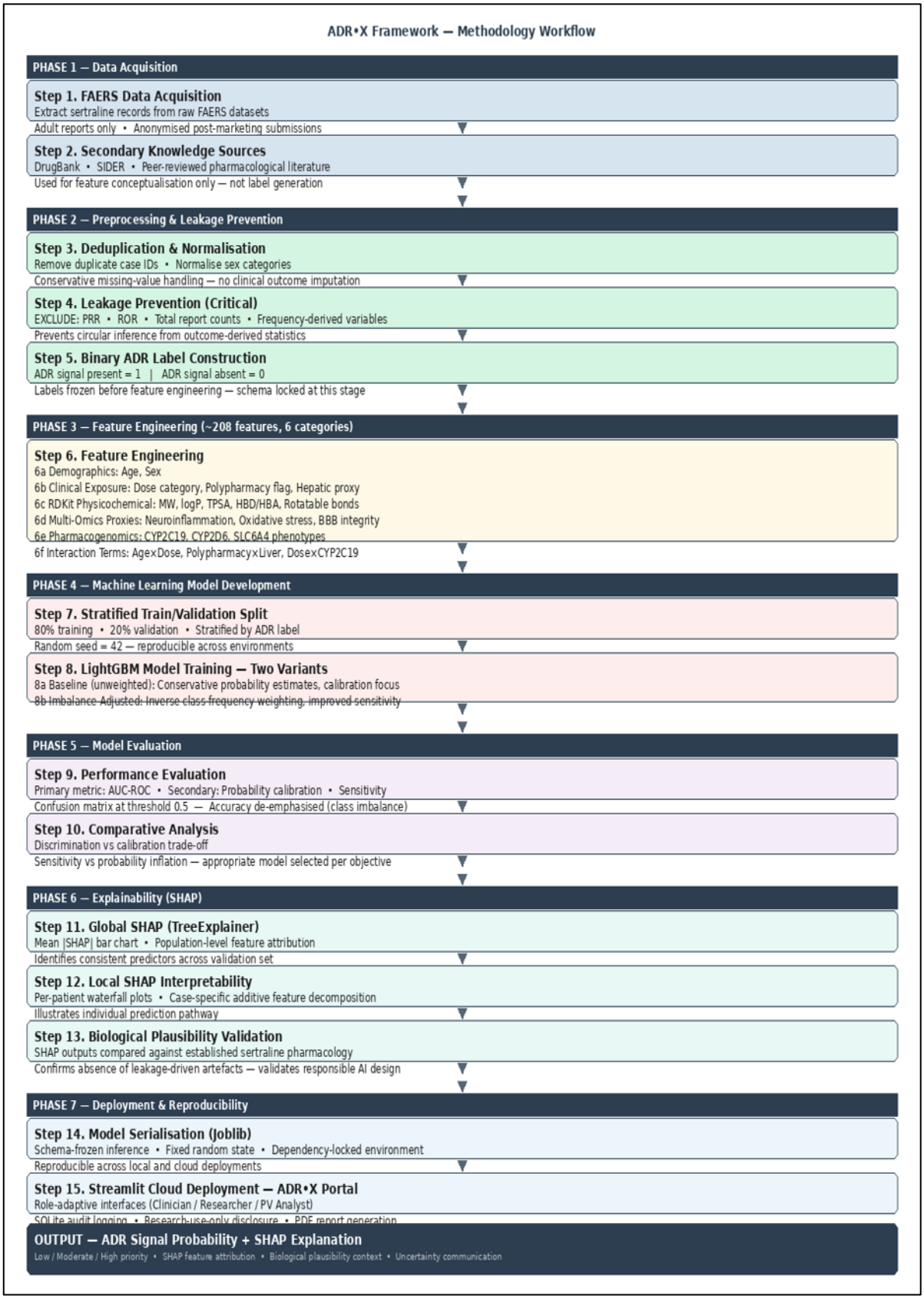

Figure 1: End-to-end methodology workflow of ADR•X. The pipeline spans FAERS data acquisition, deterministic preprocessing, feature engineering (~208 features across six categories), LightGBM model training (baseline and imbalance-adjusted variants), SHAP-based global and local explainability analysis and cloud deployment via streamlit.

RESULTS

Dataset characteristics

The FAERS-derived sertraline dataset exhibited pronounced class imbalance, with ADR-positive records comprising roughly 20–25% of the total (Table 1). Schema verification confirmed that no disproportionality-derived or frequency-count variables were present in the feature matrix.

Table 1: FAERS-derived sertraline dataset characteristics.

Characteristics	Value
ADR-positive reports	~20-25% (minority class)
ADR-negative reports	~75-80% (majority class)
Train/validation split	80:20, stratified
Total feature dimensions	~208 (six categories)
Leakage-prone variables included	None-PRR and ROR excluded

Model performance

The unweighted baseline model achieved an AUC-ROC of 0.53-0.54, with predicted probabilities clustering conservatively around the empirical class prior. The ROC curve (Figure 2A) closely tracked the diagonal reference, which is the expected signature of leakage-free modelling under the label noise inherent to spontaneous reporting systems. The imbalance-adjusted model yielded an AUC-ROC of 0.55-0.56 with higher sensitivity, though at the cost of reduced calibration stability (Table 2).

Table 2: Performance comparison of model variants.

Metric	Baseline model	Imbalance-adjusted
AUC-ROC	0.53-0.54	0.55-0.56
Mean ADR probability	0.20-0.25	0.30-0.40
Calibration	Stable	Reduced
Sensitivity at 0.5 threshold	Moderate	Higher
Overestimation risk	Low	Moderate

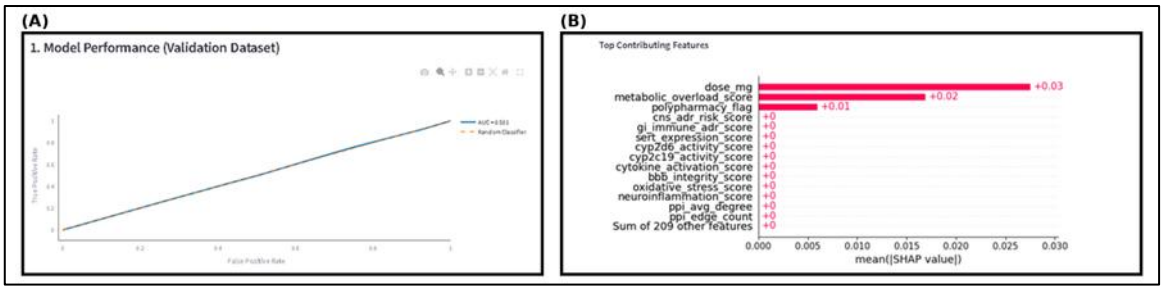

Figure 2: Leakage-free model performance and global feature importance: (A) receiver operating characteristic (ROC) curve of the unweighted baseline LightGBM model (AUC~0.53-0.54). The near-diagonal shape reflects methodologically conservative performance under leakage-free conditions and FAERS labelling uncertainty and (B) global SHAP feature attribution (mean absolute SHAP value). The three most important features are dose mg, metabolic overload score and polypharmacy flag; all remaining 209 features cluster near zero, confirming the absence of leakage-driven dominance.

SHAP feature attribution

Global SHAP analysis (Figure 2B, Table 3) revealed a clear three-tier attribution hierarchy. Sertraline dose mg was the primary predictor (mean Φ=0.030), followed by metabolic overload score (0.020) and polypharmacy flag (0.010). All remaining 209 features showed attribution values indistinguishable from zero, confirming that predictions arise from distributed feature contributions rather than any single leakage-prone variable. Local SHAP waterfall analysis (Figure 3) corroborated these findings at the individual case level. In a representative low-risk validation record (f(x)=−1.378), dose mg (−0.03) and metabolic overload score (−0.02) served as the principal contributors, with no feature displaying the disproportionate influence that would be expected if target leakage were present.

Table 3: Top global SHAP contributors with biological rationale.

Feature	Mean (φ)	Biological rationale
Dose (mg)	+0.030	Dose-dependent serotonergic exposure directly drives ADR risk
Metabolic overload score	+0.020	CYP2C19 saturation elevates plasma sertraline concentration
Poly pharmacy flag	+0.010	CYP-mediated drug–drug interaction amplifies ADR probability
All remaining 209 features	~0	Distributed contributions; no single leakage-driver identified

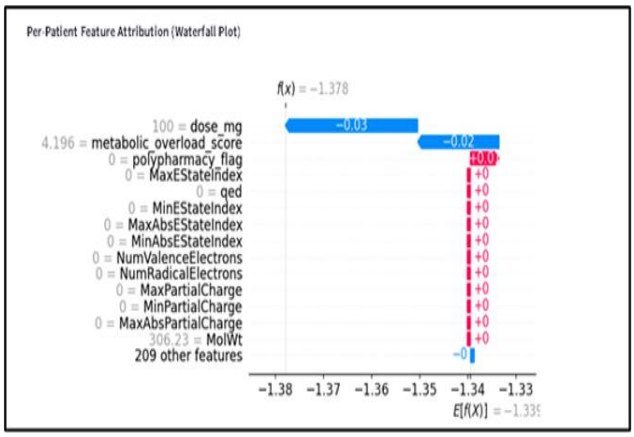

Figure 3: Local SHAP waterfall plot for a representative low-risk validation case (f(x)= −1.378, e(f(x))= −1.335).

Individual feature contributions are distributed across multiple predictors, with dose mg and metabolic overload score as principal contributors. The absence of any disproportionately dominant feature confirms that the leakage prevention strategy was effective.

DISCUSSION

The AUC-ROC range of 0.53-0.56 observed across both model variants, coupled with a near-diagonal ROC curve, should be understood as a direct reflection of FAERS structural properties rather than any limitation of the model itself. Binary ADR labels in spontaneous reporting systems are grounded in reported associations, not adjudicated causality and are compounded by systematic under-reporting, notoriety bias, and the complete absence of population denominator data.[13-15] Under such conditions, the theoretical discrimination ceiling for any leakage-free model is inherently modest. This aligns with classical pharmacovigilance signal detection literature: Bate and Evans demonstrated that even well-validated disproportionality methods applied to FAERS achieve only limited positive predictive value when benchmarked against confirmed safety signals.[13] Van Puijenbroek and colleagues similarly showed that PRR- and ROR-based analyses generate substantially different signal lists from the same dataset, underscoring the intrinsic noise embedded in voluntary reporting.[14]

Published ML studies that report AUC-ROC values above 0.80 on FAERS data consistently feature PRR, ROR, or raw report-count variables among their model inputs.[4,5] Because these statistics are directly derived from outcome frequencies, their inclusion constitutes target leakage that artificially inflates discrimination and invalidates generalisability.[5] Harpaz and colleagues explicitly cautioned that novel mining approaches must rigorously separate covariate information from signal-derived statistics to avoid circular inference.[4] Ryan and colleagues further documented that apparent performance advantages of pharmacovigilance models regularly disappear under external validation once outcome-correlated features are removed.[16] The conservative AUC values produced by ADR•X in the complete absence of such variables represent a methodologically credible and reproducible performance estimate not a deficiency.

The SHAP results carry clinically meaningful pharmacological content. Dose emerging as the strongest single predictor is consistent with the well-recognised dose-dependent pattern of serotonergic adverse effects, including nausea, insomnia, and serotonin syndrome.[7] Metabolic overload ranking second is pharmacokinetically coherent: sertraline is primarily metabolised via CYP2C19 and impaired enzymatic clearance in poor metaboliser phenotypes or in the presence of hepatic compromise leads to elevated plasma concentrations and heightened ADR probability.[8] Polypharmacy ranking third is equally expected given that co-administered CYP2C19 or CYP2D6 substrates substantially alter sertraline exposure and response.[9] The near-zero SHAP values for all remaining 209 features confirm that model predictions are not driven by artefactual signals, supporting the framework's biological plausibility and the success of the leakage prevention strategy.

The trade-off between the two model variants warrants careful consideration before any downstream application. The baseline model's tighter probability calibration makes it more appropriate for exploratory signal prioritisation, where overconfident estimates risk misleading case review. The imbalance-adjusted variant's higher sensitivity may better serve hypothesis-generation workflows where missing a candidate signal carries a greater cost than occasional false alarms, but its tendency to inflate predicted probabilities must be communicated transparently.[6] This mirrors the broader challenge in clinical AI of aligning model behaviour with specific downstream use cases rather than optimising a single aggregate metric.[17]

From a technical standpoint, LightGBM's histogram-based leaf-wise tree growth is well-matched to the high-dimensional, heterogeneous tabular structure of FAERS data and its built-in regularisation constrains overfitting under noisy label conditions.[10] The integration of SHAP Tree Explainer delivered theoretically grounded Shapley values with additive consistency across all features, fulfilling the interpretability standards that regulatory bodies increasingly demand of healthcare AI.[11,12] Doshi-Velez and Kim argue that interpretability is not a supplementary feature but a scientific necessity in high-stakes domains, because it enables practitioners to audit model behaviour, identify unintended patterns, and build justified confidence in outputs.[17] ADR•X operationalises this principle by presenting SHAP explanations as a first-class deliverable alongside the predicted signal score.

Several limitations must be acknowledged. Validation was conducted only on a held-out portion of the FAERS sertraline dataset; no independent external pharmacovigilance database was available, so

generalisation beyond this distribution cannot be confirmed.[15] The multi-omics proxy features approximate biological processes from curated literature rather than direct patient-level measurements, introducing a layer of interpretive uncertainty. The cross-sectional structure of FAERS precludes time-to-event analysis or characterisation of delayed adverse reactions. Future iterations will prioritise integration of longitudinal electronic health record data, validated patient-level genomic and transcriptomic information, and extension to additional therapeutic agents to evaluate the framework's broader applicability.

CONCLUSION

In conclusion, ADR•X establishes that biologically plausible, interpretable ADR signal detection from FAERS data is achievable within a fully leakage-aware modelling framework. Conservative AUC values are the expected and honest outcome when working with voluntary reporting databases and should be read as evidence of methodological rigor rather than model failure. The framework is offered as a research and educational resource that prioritises transparency, reproducibility, and responsible application of artificial intelligence in pharmacovigilance.

ACKNOWLEDGEMENTS

The authors thank the institutional leadership of Guru Nanak Khalsa College for their support. The publicly available FAERS database maintained by the U.S. FDA is gratefully acknowledged.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet. 2000;356(9237):1255-9.
2. WHO. The Importance of Pharmacovigilance: Safety Monitoring of Medicinal Products. Geneva: WHO Press. 2022. Available at: https://www.who.int/publications/i/item/10665-42493?. Accessed on 03 March 2026.
3. U.S. Food and Drug Administration. FDA Adverse Event Reporting System (FAERS) Public Dashboard. Silver Spring (MD): FDA. 2023. Available at: https://www.fda.gov/drugs/fda-adverse-event-reporting-system-faers. Accessed on 03 March 2026.
4. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther. 2012;91(6):1010.
5. Hauben M, Bate A. Decision support methods for the detection of adverse events in post-marketing data. Drug Discov Today. 2009;14(8):343-57.
6. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25:1337-40.
7. Stahl SM. Stahl's Essential Psychopharmacology: Neuroscientific Basis and Practical Applications. 4th ed. Cambridge: Cambridge University Press. 2013.
8. Kirchheiner J, Brosen K, Dahl ML, Gram LF, Kasper S, Roots I, et al. CYP2D6 and CYP2C19 genotype-based dose recommendations for antidepressants. Acta Psychiatr Scand. 2001;104(3):173-92.
9. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalised medicine. Clin Pharmacol Ther. 2012;92(4):414-7.
10. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146-54.
11. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-74.
12. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56-67.
13. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18(6):427-39.
14. Van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11(1):3-10.
15. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions: a systematic review. Drug Saf. 2006;29(5):385-96.
16. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med. 2012;31(30):4401-15.
17. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. 2017.

Cite this article as: Dubey AD, Parab RM, Nadar S, Uppal GK. ADR•X: an interpretable, leakage-aware machine learning framework for sertraline adverse drug reaction signal detection using FAERS pharmacovigilance data. Int J Basic Clin Pharmacol 2026;15:775-80.