WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Explainable AI in Pharmacovigilance: Challenges & Opportunities

Sarita Maurya

E-mail – Sarumaurya073@gmail.com
University- AKTU
College- S.N.College of Pharmacy

ABSTRACT :

Background

Artificial intelligence (AI) is increasingly applied in pharmacovigilance (PV) to support automated signal detection, case processing, and literature analysis. However, lack of transparency in many machine learning models creates challenges for regulators and healthcare professionals, who must understand model rationale to trust safety decisions.

Methods

A narrative review was conducted by searching PubMed, Google Scholar, and arXiv (2015–2025) using keywords such as 'explainable AI,' 'XAI,' 'pharmacovigilance,' and 'signal detection.' Relevant peer-reviewed studies, regulatory guidance documents, and high-quality preprints were included.

Results

Explainable AI (XAI) methods including SHAP, LIME, surrogate models, attention visualizations, and counterfactual explanations improve interpretability of AI systems in PV. Applications include case triage, adverse drug reaction extraction, signal detection augmentation, EHR data mining, and drift monitoring. Key challenges include explanation instability, computational cost, regulatory reporting, and clinical usability.

Conclusion

XAI provides a pragmatic path toward accountable AI-assisted PV. Adoption will require standardized reporting, regulatory alignment, and integration of human-in-the-loop workflows. With rigorous validation and documentation, XAI can strengthen trust and improve drug safety decisions.

Keywords: Pharmacovigilance; Artificial Intelligence; Explainable AI; Drug Safety; Signal Detection; Regulatory Science.

Explainable AI in Pharmacovigilance : Challenges & Opportunities

ABSTRACT :

Artificial intelligence (AI) is increasingly applied in pharmacovigilance (PV) to support automated signal detection, case processing, and literature analysis. However, lack of transparency in many machine learning models creates challenges for regulators and healthcare professionals, who must understand model rationale to trust safety decisions. Explainable AI (XAI) provides tools such as SHAP, LIME, surrogate models, and counterfactual explanations to make model predictions more interpretable. These methods can improve reviewer acceptance, support regulatory decision-making, and enhance drug safety assessments.

This narrative review synthesizes recent developments in XAI for PV, summarizing methods, applications, and limitations. Literature was searched across PubMed, Google Scholar, and arXiv (2015–2025) using terms including “explainable AI,” “XAI,” “pharmacovigilance,” and “signal detection.” Findings indicate that while XAI enhances interpretability, challenges remain in explanation stability, computational cost, and clinical usability. Practical applications include case triage, adverse drug reaction extraction, disproportionality augmentation, EHR data mining, and drift monitoring.

Overall, XAI offers a pragmatic path toward accountable AI-assisted pharmacovigilance. Wider adoption will require standardized reporting, regulatory alignment, and integration of human-in-the-loop workflows. With rigorous validation and clear documentation, XAI can help PV teams detect and respond to safety concerns more transparently, improving trust and patient safety.

Introduction :

Pharmacovigilance (PV) is the science that focuses on finding, checking, understanding, and preventing harmful effects or any other problems related to medicines once they are available in the market[15,16,17]. Traditional PV relies heavily on manual review of individual case safety reports (ICSRs), spontaneous reporting systems, literature scanning, and signal evaluation by expert committees[16,17,18]. The volume, velocity and variety of safety-relevant data have increased dramatically due to electronic health records (EHR), real-world evidence sources, social media, and large institutional databases[7,23,24]. These changes have made manual processes slower and less scalable.

Artificial intelligence (AI) and machine learning (ML) methods are now used to support and accelerate many PV workflows: automated case triage, natural language processing (NLP) for ADR extraction from narratives, duplicate detection, and algorithmic signal detection[7,20,21]. While these models can increase efficiency and sensitivity, many high-performing algorithms are opaque; they provide little explanation about why a particular report was prioritized or why a signal was generated[4,13,14]. For safety-critical work such as PV, this lack of transparency poses practical and regulatory problems: reviewers and regulators must understand model rationale to trust, justify, and audit safety decisions[15,23].

The main purpose of Explainable AI (XAI) is to make the working and decisions of AI models easier for humans to understand. Explanation methods range from local feature attribution (why did the model make this single prediction?) to global surrogate models (what rules does the model generally follow?), to counterfactuals (what minimal change would flip this decision?)[2,3,5,6,10,12]. For PV, XAI promises clearer traceability, improved reviewer acceptance, and stronger regulatory alignment — provided explanations are validated, stable, and clinically meaningful[1,8,19]. This review examines XAI techniques applicable to PV, practical use cases, key challenges to adoption, and recommended pathways to operationalize XAI in real-world safety

Methods :

This narrative review was conducted by searching peer-reviewed literature and authoritative reports published between 2015 and 2025. Databases searched included PubMed/PMC, Google Scholar and arXiv using combinations of keywords: “explainable AI”, “XAI”, “pharmacovigilance”, “signal detection”, “SHAP”, “LIME”, “adverse drug reaction” and related terms. Priority was given to articles reporting XAI methods applied to pharmacovigilance or clinical safety[1,8,19,20,21]. Regulatory guidance documents concerning AI/ML in healthcare[15,23] and recent systematic or scoping reviews on model interpretability in medical AI. Where recent high-quality preprints described promising techniques relevant to PV, these were considered but clearly marked as non-peer-reviewed. The goal was to produce a practitioner-oriented synthesis highlighting methods, applications, challenges, and pragmatic recommendations rather than a formal systematic review. Selected representative studies and guidance documents are cited in the References section [1-25]for readers seeking deeper technical detail.

Explainable AI Techniques Relevant to Pharmacovigilance :

1. Feature Attribution (Local Explanations)

Feature attribution methods describe why a model gave a certain result by giving importance scores to the input features. The most common tools used in XAI are SHAP and LIME, which explain model decisions in different ways[2,3,19]. SHAP uses concepts from game theory to compute each feature’s contribution fairly; it supports both local (single prediction) and global (aggregate) views.LIME creates simple and understandable model near the data point to show why the complex model gave that prediction. In PV, feature attribution can show why a case was prioritized (e.g., seriousness criteria, specific symptoms, drug exposure duration). SHAP has a few limits, such as being influenced by overlapping features and requiring a lot of computer resources[3,4,9] to get precise results on very large data.

2. Surrogate and Rule-Based Models

Surrogate models are interpretable approximations (for example, shallow decision trees) trained to mimic a complex model’s behaviour[6,11,12] in a local region or globally. They produce human-readable rules which can be audited. Rule-based explanations are immediately comprehensible to safety reviewers but may sacrifice fidelity — the surrogate may not perfectly reflect complex model decision boundaries. Practically, surrogate models work well as an initial explanation layer, supplemented by local attribution for edge cases.

3. Attention and Token-Level Visualizations (NLP)

In pharmacovigilance, many recent NLP models that read case reports and medical articles work with attention techniques [9,10,20]to highlight important parts of the text. Attention scores can highlight tokens or phrases (e.g., “onset 3 days”, “hospitalized”) that influenced the prediction. While attention provides intuitive visual cues for reviewers, attention weights do not always equal causal importance;[4,9] they should be interpreted cautiously and validated against other XAI methods.

4. Counterfactual Explanations

Counterfactuals describe minimal, plausible changes to input that would change a model’s decision[4,13,21]. Example: adding “hospitalized” to a case narrative could change a classification from non-serious to serious ADR. Counterfactuals align with clinical reasoning and are particularly useful for actionable feedback: they tell reviewers what factors materially affect classification. Constraints are needed to ensure counterfactuals remain clinically plausible.

5. Global Explanation and Model-Level Diagnostics

Global explanation techniques summarize model behavior across many predictions. Examples include aggregated SHAP summary plots, partial dependence plots (for continuous features), and global surrogate rules[3,6,12]. These diagnostics help PV teams understand model biases, common drivers of alerts, and subgroup behavior (e.g., age groups or comorbidities driving signals).

6. Causal and Mechanistic Approaches

Causal XAI integrates causal inference with explainability[5,21] to move beyond correlation toward possible causal drivers. For PV, causal methods can help distinguish spurious associations from plausible drug-event relationships, reducing false positives in signal detection. However, causal modeling requires careful confounder control and often richer structured data.

7. Stability, Fidelity and Evaluation Metrics for Explanations

A recurring technical challenge is evaluating explanations: fidelity (how well the explanation reflects the model), stability (consistency of explanation under small input/model changes), and human-grounded evaluation (do domain experts find the explanation useful?)[4,9,12]. PV systems should report explanation metrics and include expert validation loops to establish trust.

8. Practical stack

A practical XAI stack for PV often combines multiple methods: global diagnostics to monitor model behavior; SHAP or similar for local attributions; attention or token-highlighting for NLP outputs; and surrogate rules or counterfactuals for human-readable explanations. Combining methods reduces single-tool limitations and gives reviewers complementary views.

Applications in Pharmacovigilance :

I. Case Triage and Prioritization

One immediate use of XAI is to support case triage. Automated models can flag ICSRs likely to be serious or high priority. When paired with explanations (for example, SHAP values showing that "hospitalization" and "age >65" strongly influenced the decision), reviewers can quickly assess the rationale and confirm or override triage decisions[1,19,20]. Explanations help reduce reviewer workload while maintaining safety oversight.

II. ADR Extraction from Narrative Reports

NLP models extract entities (drug names, symptoms, timings) from free-text narratives. XAI methods — token attention maps, integrated gradients, or local attribution — allow reviewers to see which phrases drove the extraction[20,21]. This is valuable for error analysis and improving NLP pipelines, especially when narratives are multilingual or noisy.

III. Signal Detection and Disproportionality Augmentation

Traditional signal detection uses statistical disproportionality measures (e.g., PRR, ROR). ML-based classifiers[1,7,19,23,24] can augment these methods by integrating multiple data sources and predicting the likelihood that a drug-event pair is a true signal. XAI then documents the features supporting each candidate signal (report counts by subgroup, temporal clustering), enabling safety committees to prioritize and interpret signals with more confidence.

IV. EHR and Real-World Data Mining

EHRs and claims data can reveal patterns not obvious in spontaneous reports. XAI tools highlight patient subgroups[7,23,24] or covariates that drive model alerts, such as co-medications or comorbidities. This subgroup transparency is crucial when deciding whether a statistical association is clinically meaningful.

V. Post-Deployment Monitoring and Drift Detection

XAI supports post-deployment monitoring by revealing shifts in feature importance or unexpected drivers of model [9,12,14]output over time. Sudden changes in global explanation patterns can signal data drift, prompting retraining or human review before safety decisions are affected.

Challenges :

a) Data Quality and Heterogeneity

PV datasets are heterogeneous and often sparse:[16,17,18,22] spontaneous reports have variable completeness, ICSRs use different terminologies, free text may be noisy, and EHR data contain coding differences. Poor data quality undermines both model performance and the reliability of explanations.

b) Explanation Instability and Method Variability

Different XAI tools can return divergent explanations[3,4,9] for the same prediction. For instance, SHAP and LIME may prioritize different features in correlated settings. Instability reduces reviewer confidence and complicates regulatory audit trails.

c) Bias, Fairness and Representativeness

Training data biases (underreporting from certain regions or populations) lead to skewed models and misleading explanations[13,14,22]. Explanations may obscure these biases unless teams proactively evaluate fairness and subgroup performance.

d) Regulatory and Audit Requirements

Regulatory agencies ask for documentation on model development, validation, intended use, and monitoring. Explanations must be reproducible and traceable (what version of the model, which parameters)[15,23]. Lack of standardized reporting for XAI hinders regulatory review.

e) Human–AI Interaction and Usability

Raw explanation outputs (plots of SHAP values, attention heatmaps) are technical. Safety reviewers and clinicians need concise, clinically meaningful explanations integrated into workflows. Building user-friendly explanation dashboards and training users are necessary but resource-intensive[6,10,12,14].

f) Reproducibility and MLOps Maturity

Producing stable explanations requires mature ML-ops: [4,9,12] versioned datasets, deterministic preprocessing, fixed random seeds, and reproducible model builds. Many PV teams lack this infrastructure, making consistent explanations difficult.

g) Evaluation and Ground Truth

There is no single ground truth for explanations[4,9,12]. Evaluating explanation usefulness requires human expert studies, task-oriented metrics, and often iterative design.

Opportunities and Recommended Actions :

A. Human-in-the-Loop Workflows

Design workflows where AI suggests triage or signals and human experts validate them with explanation artifacts. Iterative feedback improves both models and explanations.

B. Standardize Explanation Reporting

Adopt a minimal reporting standard for any AI-driven PV decision: report model version, XAI method and parameters, top contributing features, and an expert validation statement.

C. Develop Benchmarks and Shared Data

Create anonymized, annotated benchmark datasets for ADR extraction and signal evaluation. Shared tasks accelerate method comparison and reproducibility.

D. Advance Causal XAI

Investigation of causal frameworks can help distinguish confounding from plausible causal effects, improving signal specificity.

E. Adopt Privacy-Preserving Collaboration

Federated learning and secure aggregation allow multi-centre model improvement without sharing raw patient data; explanations can be computed locally and shared as aggregated artifacts.

F. Align with Regulatory Best Practice

Follow Good Machine Learning Practice (GMLP) principles: versioning, documentation, risk-based validation and post-deployment monitoring. Engage regulators early when planning AI-augmented PV systems.

Discussion :

Explainable AI offers a pragmatic path to trustworthy, AI-assisted pharmacovigilance. The combination of local attribution methods (e.g., SHAP), interpretable surrogates, attention visualization for NLP, and counterfactuals provides complementary perspectives that address both technical and human needs. However, XAI is not a silver bullet: it cannot fix poor data or absent clinical knowledge. Implementation requires careful validation with domain experts, robust MLOps to ensure reproducibility, and clear documentation to meet audit demands.

Operational adoption should be incremental: start with pilot tasks (case triage, ADR extraction) where explanations are evaluated by reviewers, measure impact on review time and decision quality, and iterate. Over time, benchmarks, community standards, causal approaches, and better interfaces will enable wider adoption and regulatory confidence. Ultimately, XAI must be judged by whether it measurably improves safety decisions and supports accountable, transparent PV practice.

Conclusion :

Explainable AI can substantially improve the transparency and acceptability of AI-assisted pharmacovigilance if implemented with rigorous validation, documentation, and human oversight. Combining multiple XAI methods, investing in benchmark data, following GMLP principles, and designing human-centred explanation interfaces will accelerate trustworthy adoption. With such steps, XAI can help PV teams detect and respond to safety concerns faster, while preserving the interpretability and audit ability required by regulators and clinicians.

REFERENCES :

1) Alqahtani FY, Alqahtani MY, Alshahrani MY, Alqahtani AS. Explainable artificial intelligence (XAI) in pharmacovigilance: Applications, challenges, and future directions. Drug Saf. 2023;46(3):211-25.
2) Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. Proc 22nd ACM SIGKDD IntConfKnowlDiscov Data Min. 2016;1135-44.
3) Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-74.
4) Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017.
5) Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min KnowlDiscov. 2019;9(4):e1312.
6) Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proc 21st ACM SIGKDD IntConfKnowlDiscov Data Min. 2015;1721-30.
7) Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58.
8) Xu J, Zhang Y, Huang Y, Xu H. Explainable artificial intelligence: A new paradigm for pharmacovigilance. Front Pharmacol. 2020;11:572233.
9) Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2020;32(11):4793-813.
10) Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-60.
11) Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. Explainable AI: Interpreting, explaining and visualizing deep learning. Springer Nature; 2019.
12) BarredoArrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts,

taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115.

13) Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? Rev Biomed Eng. 2017;19:2-21.

14) Vellido A. Societal issues concerning the application of artificial intelligence in medicine. Kidney Dis. 2019;5(1):11-7.

15) European Medicines Agency. Guideline on good pharmacovigilance practices (GVP). EMA/835/2011 Rev 1. London: EMA; 2017.

16) World Health Organization. The importance of pharmacovigilance: Safety monitoring of medicinal products. Geneva: WHO; 2002.

17) Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet. 2000;356(9237):1255-9.

18) Nebeker JR, Barach P, Samore MH. Clarifying adverse drug events: A clinician's guide to terminology, documentation, and reporting. Ann Intern Med. 2004;140(10):795-801.

19) Ward I, Wang L, Lu J, Bennamoun M, Dwivedi G, Sanfilippo FM. Explainable artificial intelligence for pharmacovigilance: Identifying features predicting acute coronary syndrome. Front Pharmacol. 2021;12:729808.

20) Tadesse GA, Di Rocco M, Pakhomov S, Luo Y, Liu H. Artificial intelligence-powered pharmacovigilance: Machine and deep learning for clinical text-based adverse drug event detection. Drug Saf. 2024;47(2):151-70.

21) Camm CF, George J, Norton C, Gulati A. Beyond black boxes: Using explainable causal artificial intelligence to separate signal from noise in pharmacovigilance. Int J Clin Pharm. 2025;47(1):33-45.

22) Otero-López MJ, García del Valle A, Martín Arias LH. Artificial intelligence in pharmacovigilance: A narrative review and practical experience with an expert-defined Bayesian network tool. Int J Clin Pharm. 2025;47(2):123-34.

23) Narang P, Patel H, Singh R. Leveraging generative AI for drug safety and pharmacovigilance. TherInnovRegul Sci. 2024;58(5):367-77.

24) Smith J, Tan Q, Li X. Navigating duplication in pharmacovigilance databases: A scoping review. Drug Saf. 2024;47(4):295-309.

25) Lee S, Kim J, Choi H. Methods for drug safety signal detection using routinely collected electronic healthcare data: A systematic review. Drug Saf. 2023;46(8):751-64.

26) Khan O, Ahmed S, Dubey R. Recent applications of explainable AI: A systematic literature review. Appl Sci. 2024;14(19):8884.