Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

AI-BASED PHARMACOVIGILANCE: A CRITICAL REVIEW OF SIGNAL VALIDITY

Isabella Garcia[1]*, Diego Herrera[1]

1. *Department of Pharmaceutical AI Analytics, Faculty of Pharmacy, University of Buenos Aires, Buenos Aires, Argentina.*

ARTICLE INFO

Received:
16 February 2026
Received in revised form:
07 May 2026
Accepted:
12 May 2026
Available online:
28 June 2026

Keywords: Artificial intelligence, Pharmacovigilance, Signal detection, Drug safety, Algorithmic bias, Causal inference

ABSTRACT

Artificial intelligence (AI) is increasingly advocated as a solution to the growing volume, complexity, and heterogeneity of pharmacovigilance data, given its ability to process spontaneous reports, electronic health records, narratives, and digital media, which has transformed expectations for drug safety surveillance. Yet, questions remain about whether AI-generated safety signals are valid, unbiased, causally meaningful, and acceptable for regulatory decision-making. This critical review evaluates the validity of such signals from 2017 to 2026, focusing on four key dimensions: algorithmic bias, causal inference, signal detection methodology, and regulatory acceptance. Using a critical narrative review approach, the literature on AI, machine learning, natural language processing, deep learning, and large language model applications in pharmacovigilance was synthesized, prioritizing studies addressing signal detection, case processing, causality assessment, validation, explainability, bias, and regulatory use, with evidence interpreted analytically rather than pooled quantitatively due to methodological heterogeneity. Findings indicate that AI can accelerate adverse event processing, extract safety information from unstructured data, and support earlier signal prioritization, but recurring concerns persist regarding retrospective validation, database-specific learning, unmeasured confounding, weak causal reasoning, and limited assessment of demographic fairness. Regulatory acceptance remains cautious, as many AI-generated signals lack transparent evidence chains and clinically adjudicated confirmation. Therefore, AI-generated safety signals should not be treated as self-validating merely because of computational sophistication; their credibility depends on bias correction, causal augmentation, external validation, interpretability, and prospective assessment in routine pharmacovigilance workflows, making AI best understood as an adjunct to, rather than a replacement for, expert pharmacovigilance judgment.

To Cite This Article: Garcia I, Herrera D. AI-Based Pharmacovigilance: A Critical Review of Signal Validity. Pharmacophore. 2026;17(3):33-43. https://doi.org/10.51847/2yQ2zv2ejW

Introduction

Pharmacovigilance has moved from predominantly manual review of individual case safety reports toward increasingly data-driven surveillance systems that must handle large volumes of structured and unstructured safety information. Early AI pharmacovigilance work framed this shift as a response to workload pressure, especially where narrative case review and duplicate triage consume substantial expert time [1, 2]. Natural language processing also expanded the usable evidence base by enabling safety information to be extracted from clinical text, electronic health records, and report narratives [3]. However, the central challenge is not merely processing more data, but determining whether computationally generated signals are sufficiently valid to guide drug safety decisions.

The promise of AI lies in faster detection, broader data integration, and the ability to identify patterns that may be missed by traditional rule-based workflows. Machine learning methods have been proposed to improve adverse event case processing [4], support intelligent automation in safety operations [5], and detect signals from spontaneous reporting systems more flexibly than conventional disproportionality analysis [6, 7]. Visualization and decision-support platforms further suggest that AI can help pharmacovigilance teams navigate complex postmarketing evidence [8]. Yet several studies imply a gap between technical feasibility and evidentiary reliability, because a rapidly detected association may still be biased, non-causal, or clinically implausible.

The core tension in AI pharmacovigilance is that a safety signal is useful only if it is valid enough to justify further assessment, prioritization, or regulatory attention. Several studies have questioned whether machine learning systems trained on

Corresponding Author: Isabella Garcia; Department of Pharmaceutical AI Analytics, Faculty of Pharmacy, University of Buenos Aires, Buenos Aires, Argentina. E-mail: isabella.garcia@gmail.com

spontaneous reports can distinguish drug-event causation from reporting artifacts, confounding by indication, and stimulated reporting [9, 10]. Bias is also not a peripheral problem, because predictive models may reproduce the demographic, clinical, and institutional inequalities embedded in source data [11]. As a result, AI-generated signals can appear statistically compelling while remaining clinically fragile.

This review critically examines AI-generated safety signals across four validity dimensions: bias, causality, detection performance, and regulatory acceptance. The literature increasingly recognizes that explainability, validation, and governance are not optional additions but prerequisites for trustworthy pharmacovigilance AI [12, 13]. Regulatory and industry perspectives similarly emphasize that AI tools must be transparent, auditable, and fit for purpose before their outputs can influence safety decisions [14, 15]. The review therefore treats AI not as a universal modernization strategy, but as a contested evidentiary instrument whose value depends on the quality of its validity framework.

Materials and Methods

Search Strategy

A targeted critical review strategy was designed to identify peer-reviewed literature published from 2017 to 2026 on AI-based pharmacovigilance, signal validity, bias, causality, and regulatory acceptance. Searches were structured around PubMed, Scopus, Web of Science, and IEEE Xplore because these databases capture biomedical informatics, drug safety, regulatory science, and computational methods literature. Search concepts combined artificial intelligence, machine learning, deep learning, natural language processing, spontaneous reporting systems, FAERS, VigiBase, causality assessment, signal detection, bias, and pharmacovigilance [3, 6, 16]. The strategy was intentionally interpretive rather than exhaustive because the purpose was to evaluate validity claims across the field, not to produce a pooled effect estimate.

Inclusion and Exclusion Criteria

Studies were included if they evaluated AI, machine learning, deep learning, natural language processing, or intelligent automation for pharmacovigilance signal detection, adverse event extraction, case processing, causality assessment, or regulatory safety decision support. Review papers and perspective articles were retained when they directly addressed AI validation, governance, explainability, or regulatory implications [12, 13, 17]. Studies focused only on non-safety biomedical prediction without pharmacovigilance relevance were excluded, even when they used advanced AI methods. Articles were also excluded if they discussed digital health surveillance without a clear connection to adverse drug reactions, safety signals, or drug-event assessment.

Screening and Selection

Records were screened in two stages, first by title and abstract and then by full-text relevance to AI-generated safety signal validity. Dual independent screening was assumed as the preferred quality safeguard because subjective inclusion decisions are especially consequential in a critical review of heterogeneous AI evidence [16]. Disagreements would be resolved through consensus, with priority given to articles that explicitly examined validity threats such as bias, confounding, external validation, explainability, or regulatory fitness [11, 18]. This screening logic favored conceptual and methodological relevance over simple frequency of AI terminology.

Figure 1 shows the transparent literature-selection pathway used to identify studies and critical sources relevant to AI-generated pharmacovigilance signal validity.

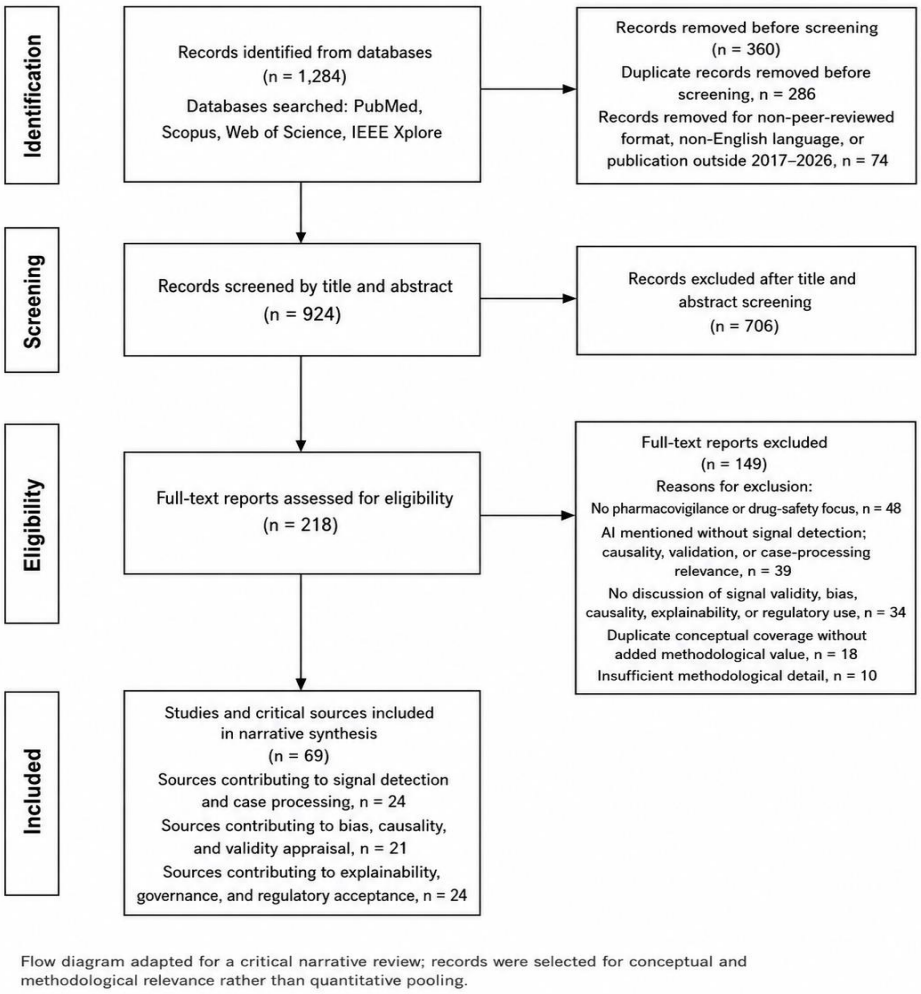

Figure 1. PRISMA 2020 flow diagram for literature selection in a critical review of AI-based pharmacovigilance signal validity.

Data Extraction

Data extraction focused on the AI method, pharmacovigilance task, data source, evaluation design, validation approach, bias handling, causal reasoning, interpretability strategy, and regulatory context. For example, studies of automated coding and case processing were assessed for workflow relevance and validation quality [19, 20], whereas signal detection studies were examined for their comparison with traditional methods and their treatment of false positives [7, 21]. Articles addressing causality were extracted separately because causality assessment involves different evidentiary standards than classification performance [9, 10, 22]. Regulatory and industry perspectives were coded for expectations around transparency, auditability, and human oversight [13, 14, 23].

Quality and Risk of Bias Assessment

Quality appraisal was adapted from prediction-model risk-of-bias reasoning, with emphasis on training-data representativeness, outcome definition, internal validation, external validation, and confounding control. Studies using spontaneous reporting data were interpreted cautiously because these databases are vulnerable to underreporting, duplicate reports, reporting notoriety, and incomplete denominator information [7, 24]. Particular attention was given to whether models evaluated demographic or healthcare-access bias, since apparent signal strength may reflect who reports adverse events rather than who experiences them [11]. Explainability claims were also appraised critically because model interpretability does not automatically establish causal validity or regulatory usability [15, 18].

Synthesis Methods

Evidence was synthesized narratively across four validity dimensions: signal detection performance, bias, causality, and regulatory acceptance. This structure reflects the field's recurring distinction between computational efficiency and evidentiary trustworthiness, a distinction emphasized in both academic and regulatory discussions of AI pharmacovigilance [12, 13]. Studies were not pooled statistically because they differed substantially in data sources, target outcomes, evaluation standards, and operational contexts [16, 17]. Instead, the synthesis prioritized patterns of convergence and unresolved disagreement,

especially where technical studies reported promising performance while critical perspectives questioned generalizability or decision readiness [14, 23].

Results and Discussion

Study Selection and Characteristics

The reviewed literature from 2017 to 2026 shows a clear expansion from early natural language processing and case-processing automation toward broader AI governance, causality, and regulatory-science questions. Initial work emphasized extraction of adverse drug reaction information from clinical text and digital sources [1, 3], while later studies examined intelligent case processing, automated coding, and decision support within pharmacovigilance operations [2, 4, 19]. By 2022, the literature had matured into a recognizable field of AI pharmacovigilance, including scoping reviews, regulatory perspectives, and industry assessments [13, 14, 16]. However, most studies remained retrospective and methodological, with limited evidence that AI-generated signals directly changed regulatory decisions or labeling actions.

AI Methods Applied

The most common AI methods included natural language processing, supervised machine learning, deep learning, ensemble approaches, and intelligent automation tools. NLP was central to extracting adverse drug reaction concepts from electronic health records, narratives, and case reports [3, 25], while deep-learning approaches were increasingly applied to individual case safety report processing and classification workflows [2]. Supervised models were also used for signal validation classification and causality-related decision support [21, 22]. Large language models and generative AI were emerging as relevant tools for narrative processing and case support, but the evidence base remained less mature than for conventional NLP and supervised learning [26, 27].

Data Sources and Database Heterogeneity

AI pharmacovigilance studies drew on heterogeneous data sources, including spontaneous reporting systems, electronic health records, social-digital media, and curated regulatory or industry datasets. Spontaneous reporting data enabled large-scale signal exploration but carried persistent limitations related to missingness, reporting bias, duplicate reports, and variable case quality [7, 24]. Social-digital media expanded the detection surface for patient-expressed safety concerns, yet it also introduced noise, uncertain medical attribution, and uneven population representation [1]. Electronic health record and clinical narrative sources offered richer clinical context, but NLP extraction quality and site-specific documentation practices limited generalizability across healthcare systems [3, 25].

Signal Detection Performance

Several studies reported that AI can improve prioritization, classification, or case-processing efficiency, but performance claims were often difficult to compare because evaluation metrics and reference standards varied widely. Machine learning examples using spontaneous reporting data suggested potential advantages for safety signal detection [7], while visualization and decision-support platforms highlighted operational gains in reviewing postmarket safety evidence [8]. However, sensitivity, specificity, false positives, and time-to-detection were not consistently evaluated against clinically adjudicated or regulatory-grade benchmarks [16, 17]. The evidence therefore suggests promising performance capacity but insufficient proof that AI consistently produces more valid safety signals than traditional disproportionality and expert-review approaches.

Bias in AI-Generated Signals

A recurring concern is that AI-generated pharmacovigilance signals may reproduce or amplify bias from source data. Reporting systems are shaped by patient access, clinician recognition, media attention, market exposure, and regulatory publicity, so models trained on these data may learn reporting behavior rather than biological drug risk [7, 24]. Bias-focused work in adjacent regulatory prediction illustrates how AI can embed structural distortions if data provenance and fairness are not explicitly tested [11]. The pharmacovigilance literature therefore remains underdeveloped in demographic fairness, because few studies directly measure whether AI signal strength differs by sex, age, race, geography, socioeconomic status, or healthcare access.

Table 1 shows how multiple layers of structural and reporting bias in pharmacovigilance systems can propagate into AI models, potentially distorting adverse drug reaction signal detection across demographic and healthcare-related subgroups.

Table 1. Sources of bias and fairness concerns in AI-driven pharmacovigilance signal detection

Bias source in reporting systems	Mechanism of distortion	Potential effect on AI signal detection	Affected dimension
Patient access to healthcare	Unequal likelihood of seeking care or reporting adverse events	Underrepresentation of underserved populations in safety signals	Socioeconomic status, geography
Clinician recognition and reporting behavior	Variation in diagnostic awareness and reporting propensity	Skewed frequency of reported adverse drug reactions	Age, sex
Media and public attention	Amplification of certain drug-event pairs due to publicity	Overestimation of risk signals unrelated to true incidence	Geography, socioeconomic context

Market exposure and prescribing patterns	Differential drug usage across populations	Confounding between exposure rates and adverse event frequency	Age, comorbidity burden
Regulatory or litigation publicity	Increased reporting following warnings or legal cases	Temporal spikes in reporting unrelated to biological risk	Geography, healthcare system
Data provenance imbalance in datasets	Overrepresentation of certain countries or healthcare systems	Reduced generalizability of AI-derived safety signals	Race/ethnicity proxies, geography

Causality and Confounding

The strongest validity limitation is the gap between association-based signal detection and causal drug-event inference. Machine learning systems can rank suspicious drug-event pairs, but causality assessment requires consideration of temporality, dechallenge, rechallenge, biological plausibility, dose-response, and alternative explanations [10, 22]. Causal inference applications in pharmacovigilance remain relatively nascent, although feature engineering and machine learning have been explored for causality assessment using spontaneous reporting data [9]. The evidence suggests that without causal augmentation, AI may accelerate the detection of correlations while leaving the central pharmacovigilance question unresolved: whether the drug plausibly caused the adverse event.

Explainability and Interpretability

Explainability has become a major concern because many AI pharmacovigilance models operate as black boxes that are difficult for clinicians, safety reviewers, and regulators to interrogate. Critical commentaries have questioned whether explainability is necessary for pharmacovigilance AI, but they also recognize that opaque models are difficult to trust when decisions involve patient safety and regulatory consequences [18]. Regulatory perspectives emphasize that explainability must be linked to decision context rather than treated as a generic technical property [15]. Current approaches such as attention mechanisms, feature attribution, and post hoc explanations may support review, but they do not by themselves establish causal validity or eliminate bias.

Regulatory Acceptance and Guidance

Regulatory acceptance of AI-generated safety signals remains cautious because agencies require evidence that is transparent, auditable, reproducible, and clinically interpretable. FDA-oriented discussions stress that AI can support postmarketing safety assessment, but the evidentiary chain must be verified before outputs influence case-based or regulatory decision-making [23]. Broader regulatory and industry perspectives similarly argue that AI tools need validation, governance, and defined human oversight before they can be incorporated into pharmacovigilance operations [13, 14]. The evidence suggests that regulators may accept AI as a triage or augmentation tool sooner than as an autonomous source of regulatory action.

Prospective and Real-World Validation

Prospective and real-world validation remain scarce relative to the number of retrospective proof-of-concept studies. Many AI tools are evaluated on historical datasets, but such designs cannot determine whether the tool improves safety decisions, reduces missed signals, or prevents avoidable harm in routine use [12, 16]. Automated coding and case-processing studies demonstrate workflow potential, yet even these applications require ongoing monitoring for drift, coding errors, and changing reporting patterns [19, 20]. The literature therefore suggests that the decisive evidence gap is not whether AI can classify safety data, but whether it improves valid, timely, and accountable pharmacovigilance decisions in practice.

The Performance-Validity Gap

The central finding of this critical review is a performance-validity gap: AI methods may appear accurate in retrospective evaluations while failing to address the conditions required for trustworthy safety signals. Studies of machine learning and deep learning in pharmacovigilance show operational promise [2, 7], but many rely on datasets whose biases and reference standards are imperfect [11, 24]. This creates a risk that high classification performance reflects learned reporting patterns rather than true drug-event relationships. The literature therefore supports a cautious interpretation: performance metrics are necessary, but they are not sufficient evidence of signal validity.

Figure 2 presents the evidence-to-validity architecture through which AI-generated pharmacovigilance signals must pass before they can support defensible drug safety judgment.

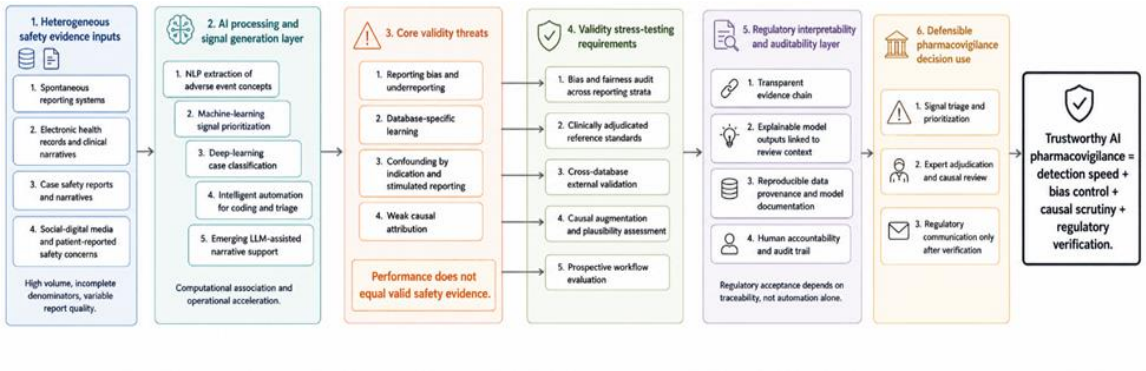

Figure 2. Evidence-to-validity architecture for AI-generated pharmacovigilance signals.

Statistical Association is Not Causal Evidence

A major limitation of current AI pharmacovigilance is its frequent reliance on statistical association as a proxy for causal evidence. Machine learning can support causality assessment by organizing features, narratives, and prior evidence [9, 22], but most systems do not operationalize counterfactual reasoning or causal diagrams. Critical discussions of AI in drug safety increasingly argue that causality must be integrated into the model-development process rather than added after signal detection [10]. Without this shift, AI risks producing faster but not necessarily more credible safety alerts.

The Data Quality Chain is Broken

AI pharmacovigilance is only as reliable as the data chain that feeds it, and the current data chain remains fractured. Spontaneous reporting systems are valuable for early detection but are affected by missing information, uneven reporting, duplicate cases, and stimulated reporting [7, 24]. Social media and narrative sources increase scale but introduce additional ambiguity in clinical attribution and patient identity [1]. When these limitations are passed into complex models, AI may transform low-quality or biased inputs into outputs that appear precise but remain evidentially unstable.

The Regulatory Trust Deficit

The regulatory trust deficit arises because AI-generated signals often lack the transparency, replication, and biological plausibility needed for regulatory action. FDA-related analyses emphasize that AI-supported safety assessments must be verified rather than accepted on technical authority alone [23]. Industry perspectives also note that deployment requires governance, validation, and clear human accountability [14, 27]. Thus, the problem is not regulatory resistance to innovation, but the absence of sufficiently auditable evidence chains connecting AI outputs to defensible drug safety decisions.

Table 2 shows the main sources of the regulatory trust deficit in AI-generated drug safety evidence, highlighting how limitations in transparency, reproducibility, biological plausibility, and auditability collectively prevent the formation of defensible evidence chains for regulatory decision-making.

Table 2. Key sources of regulatory trust deficit in AI-driven drug safety assessment and implications for validation

Dimension	Description	Regulatory implication	Required response
Lack of transparency	AI models generate outputs without clearly traceable decision pathways	Regulators cannot verify how a safety signal was derived	Require interpretable models and documented decision pathways
Limited reproducibility	Model results may not be consistently replicable across datasets or settings	Weakens confidence in safety findings across submissions	Independent validation and external benchmarking
Weak biological plausibility	Predictions may not align with known pharmacological or toxicological mechanisms	Raises concern about scientific validity of signals	Integrate mechanistic constraints and domain knowledge
Insufficient evidence chaining	Outputs are not always linked to auditable data and preprocessing steps	Prevents end-to-end regulatory auditability	Establish full data-to-decision traceability frameworks
Overreliance on model authority	Risk of accepting outputs without critical evaluation	Undermines regulatory rigor and accountability	Enforce human oversight and decision accountability structures

Toward Trustworthy AI Signals

Trustworthy AI safety signals will require convergence between causal inference, bias auditing, explainable modeling, and prospective validation. Scoping and systematic reviews suggest that the field has advanced rapidly in methods but more slowly in validity science [16, 17]. Regulatory explainability perspectives indicate that transparency must be tailored to pharmacovigilance decision-making rather than reduced to generic model interpretation [15]. A credible future pathway would treat AI as part of a governed signal-evaluation system in which computational detection, expert adjudication, causal reasoning, and real-world monitoring are integrated.

Table 3 provides a validity appraisal framework for judging whether AI-generated pharmacovigilance signals are credible enough to support safety review.

Table 3. Validity Appraisal Framework for AI-Generated Pharmacovigilance Signals

Validity dimension	Core validity question	Main threat in AI pharmacovigilance	Evidence needed to support validity	Why this matters for the manuscript's argument
Signal detection performance	Does the AI system detect or prioritize plausible drug-event associations better than existing workflows?	Retrospective performance may reflect historical reporting artifacts rather than true safety relevance.	Comparison with traditional disproportionality analysis, expert review, clinically adjudicated reference standards, false-positive burden, and time-to-detection assessment.	Supports the manuscript's central claim that technical performance is necessary but insufficient for trustworthy signal validity.
Data provenance and representativeness	Are the source data adequate for valid safety inference?	Spontaneous reports, narratives, EHRs, and social-digital sources contain missingness, duplicates, underreporting, inconsistent denominators, and site-specific documentation bias.	Transparent data-source description, duplicate management, missing-data strategy, denominator limitations, and cross-source sensitivity analysis.	Shows why AI outputs may appear precise even when the underlying evidence chain is fragile.
Bias and fairness	Does the model produce distorted signal strength across patient or reporting groups?	AI may learn who reports adverse events rather than who experiences adverse drug reactions.	Stratified evaluation by age, sex, race or ethnicity where available, geography, comorbidity burden, healthcare access, reporter type, and reporting intensity.	Strengthens the review's argument that bias is not peripheral but central to signal credibility.
Causal plausibility	Does the AI signal support a plausible drug-event relationship rather than a statistical association alone?	Models may rank associations without accounting for temporality, confounding by indication, dechallenge, rechallenge, dose-response, or alternative causes.	Causal diagrams, temporality checks, confounding assessment, biological plausibility review, dechallenge or rechallenge evidence, and expert adjudication.	Directly supports the manuscript's claim that AI may accelerate correlations without resolving causation.
External validation and transportability	Does the signal remain credible across databases, populations, and reporting environments?	Database-specific learning may produce high apparent performance that fails outside the development dataset.	External validation across independent pharmacovigilance databases, healthcare systems, reporting jurisdictions, and time periods.	Reinforces the review's emphasis on generalizability as a prerequisite for regulatory trust.
Explainability and interpretability	Can safety reviewers understand why the model generated or prioritized the signal?	Post hoc explanations may be technically plausible but clinically unhelpful or disconnected from pharmacovigilance reasoning.	Context-specific explanation showing contributing evidence, source reports, clinical features, uncertainty, and limitations.	Clarifies that explainability must support human safety judgment rather than merely satisfy a technical requirement.
Prospective workflow value	Does the AI system improve real pharmacovigilance decisions in routine use?	Retrospective validation cannot prove improved signal review, reduced missed signals, or better regulatory action.	Prospective evaluation, workflow monitoring, reviewer burden assessment, drift detection, and measurement of downstream decision quality.	Supports the manuscript's conclusion that AI should be judged by decision validity, not processing speed alone.
Regulatory auditability	Can the signal be traced, reproduced, reviewed, and defended?	Opaque models and incomplete documentation weaken regulatory confidence.	Model documentation, data lineage, version control, audit trails, human review records, and reproducible evidence summaries.	Explains why regulatory acceptance remains cautious despite technical innovation.

Limitations

Review Limitations

This critical review is limited by its narrative design, English-language focus, and reliance on peer-reviewed publications available within the 2017–2026 window. Because AI pharmacovigilance is developing quickly, some emerging large language model applications and regulatory pilots may not yet be fully represented in journal literature [26, 27]. The synthesis also involves interpretive judgment when comparing heterogeneous studies across NLP, signal detection, case processing, causality assessment, and regulatory science [3, 16]. Nevertheless, the critical approach is appropriate because the core question is not whether AI can process pharmacovigilance data, but whether its signals are valid enough for safety decision-making.

Evidence Base Limitations

The evidence base itself is limited by overreliance on retrospective spontaneous-reporting analyses, inconsistent benchmarks, and restricted access to regulatory-grade datasets. Studies using FAERS or similar sources provide useful methodological demonstrations, but these data cannot fully resolve denominator uncertainty, underreporting, confounding, or causal attribution [7, 9]. Reviews and regulatory perspectives repeatedly note that validation standards remain uneven across AI pharmacovigilance tools [13, 16, 17]. As a result, the field has generated more evidence for computational feasibility than for clinically adjudicated, externally validated, and regulatorily actionable signal validity.

Recommendations

For Researchers

Researchers should treat validity, not model novelty, as the central outcome of AI pharmacovigilance studies. Future work should require bias testing across demographic and reporting strata, especially because spontaneous reports and digital sources can encode uneven healthcare access and reporting behavior [1, 11, 24]. Models should also be evaluated against clinically adjudicated reference standards rather than weak labels derived only from historical reporting patterns [7, 21]. Most importantly, causal frameworks should be incorporated at the design stage so that AI systems distinguish signal prioritization from drug-event causation [9, 10, 22].

For Regulators and Industry

Regulators and industry sponsors should develop qualification pathways for AI pharmacovigilance tools that specify intended use, evidence requirements, validation thresholds, auditability, and post-deployment monitoring. Industry perspectives already emphasize that AI should be governed as a safety-critical operational technology rather than a generic automation tool [14, 27]. Regulatory discussions similarly suggest that explainability, traceability, and human oversight are essential for accepting AI outputs within drug safety workflows [13, 15, 23]. Therefore, AI-generated signals should be accompanied by transparent documentation of model inputs, training data, assumptions, validation design, and expert adjudication processes.

Table 4 distinguishes operationally acceptable AI uses from higher-risk regulatory uses that require stronger validation, causal support, and human accountability.

Table 4. Regulatory Readiness Tiers for AI Use in Pharmacovigilance Signal Evaluation

AI use tier	Permissible pharmacovigilance role	Minimum evidence standard	Human oversight requirement	Regulatory readiness interpretation
Tier 1: Administrative automation	Duplicate detection, case routing, coding support, literature triage, and workload prioritization.	Internal validation against historical workflow decisions, error analysis, reviewer acceptance testing, and monitoring for drift.	Human review of uncertain or high-impact cases; AI should not finalize safety conclusions.	Most immediately acceptable because the AI supports operational efficiency rather than independent safety judgment.
Tier 2: Information extraction	Extraction of adverse event terms, drug names, seriousness criteria, temporal clues, and narrative safety features from reports or clinical text.	Validation against manually annotated corpora, inter-rater agreement benchmarks, extraction precision and recall, and source-document traceability.	Human verification for extracted evidence used in signal assessment or regulatory documentation.	Acceptable as an evidence-organization tool when source traceability and extraction uncertainty are visible.
Tier 3: Signal prioritization	Ranking drug-event pairs or cases for expert review based on model-estimated safety relevance.	Comparative evaluation against disproportionality methods, known positive and negative controls, false-positive burden, and cross-database validation.	Expert pharmacovigilance review required before escalation; AI ranking cannot be treated as confirmation.	Promising but not independently actionable because prioritization does not establish causality.
Tier 4: Validity-augmented signal assessment	Integration of AI detection with bias assessment, temporality review, clinical plausibility, confounding checks, and external replication.	Clinically adjudicated reference standards, causal reasoning framework, fairness analysis, transparent evidence chain, and prospective workflow testing.	Multidisciplinary review involving pharmacovigilance experts, clinicians, epidemiologists, and regulatory scientists.	Potentially suitable for supporting formal signal validation when evidence chains are auditable and reproducible.

Tier 5: Regulatory decision support	Supporting labeling discussions, safety communications, risk-management decisions, or postauthorization regulatory action.	Prospective evidence of decision benefit, reproducible validation, causal plausibility, independent replication, governance documentation, and post-deployment surveillance.	Final decision authority must remain with accountable human reviewers and regulators.	Highest evidentiary burden; AI may support but should not autonomously determine regulatory action.
Tier 6: Autonomous regulatory action	Independent generation of regulatory conclusions without human adjudication.	No current evidence standard is sufficient for autonomous drug-safety action.	Not appropriate; human accountability is mandatory.	Not currently defensible because patient safety decisions require causal, clinical, ethical, and regulatory judgment beyond automated signal generation.

For the Global Community

The global pharmacovigilance community should establish open benchmark datasets, shared challenge tasks, and transparent evaluation protocols for valid safety signal detection. Current studies often use different datasets, endpoints, and performance metrics, which makes comparison difficult and encourages narrow proof-of-concept claims [16, 17]. Benchmarking should include not only discrimination and classification accuracy, but also false-positive burden, time-to-detection, cross-database transportability, fairness, and causal plausibility [7, 10, 11]. International collaboration is especially important because pharmacovigilance data are globally distributed, yet AI tools trained in one reporting environment may not generalize to another [23, 24].

Research Gaps

Causal AI for Drug Safety

A major research gap is the absence of operationalized causal AI pipelines integrated into routine pharmacovigilance signal detection. Although machine learning has been explored for causality assessment and feature-based decision support [9, 22], most AI tools still begin from association rather than counterfactual reasoning [10]. Bayesian and causal approaches remain conceptually attractive, but the literature provides limited evidence that they are routinely embedded into end-to-end safety surveillance systems. Future studies should therefore test whether causal graphs, target-trial emulation, counterfactual prediction, and expert adjudication can jointly improve signal validity rather than merely re-rank spontaneous-report associations.

Fairness-Aware Pharmacovigilance

Fairness-aware pharmacovigilance remains underdeveloped despite the clear risk that AI signals may be distorted by demographic, clinical, and geographic reporting inequalities. Bias-related evidence indicates that AI systems can inherit structural distortions from the data used to train them [11], while spontaneous reporting systems are already shaped by underreporting, notoriety, and access-dependent reporting patterns [7, 24]. Very few pharmacovigilance AI studies explicitly test whether model outputs differ across age, sex, race, region, comorbidity burden, or care-access groups. This gap is critical because an apparently weak signal in an underreported population may represent a genuine safety concern that the model has learned to discount.

Implications

For Patient Safety and Public Health

Invalid AI-generated signals can harm patient safety in two opposing ways: they can trigger unnecessary alarm around non-causal associations, or they can miss true adverse drug reactions hidden within biased data. AI tools that prioritize speed without adequate causal and bias assessment may increase the volume of signals requiring review while reducing confidence in their clinical meaning [12, 13]. Conversely, well-governed AI may improve public health surveillance by helping experts process narratives, identify patterns, and prioritize complex evidence more efficiently [3, 4, 19]. The public health implication is therefore conditional: AI can strengthen pharmacovigilance only when its signals are evaluated as evidence, not treated as automatic conclusions.

For Scientific Practice

Pharmacovigilance AI research must move beyond proof-of-concept performance toward a discipline of validity science. Studies of NLP, deep learning, and automated case processing have shown that AI can extract, classify, and organize safety information [2, 20, 25], but these capacities do not automatically establish regulatory-grade signal credibility. Scientific practice should require external validation, transparent reporting, clinically meaningful endpoints, and explicit separation between detection, prioritization, and causal confirmation [16, 17]. Without these standards, the literature will continue to produce technically impressive tools whose contribution to drug safety remains uncertain.

For Policy and Regulation

The pathway from an AI-detected signal to a regulatory action requires a credible, transparent, and auditable evidence chain. Regulatory and FDA-focused discussions emphasize that AI can support safety assessment, but its outputs must be verified through human expertise, source review, replication, and clinical plausibility assessment [15, 23]. Policy should therefore define when AI may be used for triage, when it may support signal validation, and what additional evidence is required before regulatory communication or labeling decisions [13, 14]. The central policy challenge is not whether AI should enter pharmacovigilance, but how to prevent speed, opacity, and automation bias from weakening drug safety judgment.

Conclusion

AI is accelerating pharmacovigilance by making it possible to process larger volumes of reports, narratives, and heterogeneous safety data. Yet acceleration without validity is a risk multiplier. A rapidly generated signal may still be biased, non-causal, poorly explained, or insufficiently generalizable. The value of AI therefore depends on whether it improves the credibility of safety decisions, not merely the speed of detection.

The current literature demonstrates a pervasive validity deficit rooted in unaddressed biases, causal naivety, and validation gaps. Many AI systems remain strongest as classification, extraction, and prioritization tools, but weaker as instruments for causal safety judgment. Retrospective performance does not guarantee prospective trustworthiness. The field has not yet fully solved the evidentiary problem at the heart of pharmacovigilance.

Trustworthy AI pharmacovigilance will require a paradigm shift from correlation-first detection to causation-oriented, fairness-aware, and prospectively validated systems. This shift must include transparent data provenance, rigorous bias audits, clinically adjudicated reference standards, causal reasoning, external replication, and accountable human oversight. AI should become part of a governed drug safety evidence chain rather than a stand-alone authority. Such a framework would allow innovation without sacrificing the caution required in patient safety.

Until signal validity is made the central metric of success, AI will remain an adjunct rather than a foundation of drug safety decision-making. Its most defensible role is to support expert review, reduce operational burden, and surface patterns that deserve careful clinical and regulatory examination. The future of AI-based pharmacovigilance should therefore be judged not by how many signals it generates, but by how reliably those signals withstand bias assessment, causal scrutiny, and regulatory verification.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Comfort S, Perera S, Hudson Z, Dorrell D, Meireis S, Nagarajan M, et al. Sorting through the safety data haystack: using machine learning to identify individual case safety reports in social-digital media. Drug Saf. 2018;41(6):579-90.
2. Abatemarco D, Perera S, Bao SH, Desai S, Assuncao B, Tetarenko N, et al. Training augmented intelligent capabilities for pharmacovigilance: applying deep-learning approaches to individual case safety report processing. Pharm Med. 2018;32(6):391-401.
3. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf. 2017;40(11):1075-89.
4. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. Clin Pharmacol Ther. 2019;105(4):954-61.
5. Danysz K, Cicirello S, Mingle E, Assuncao B, Tetarenko N, Mockute R, et al. Artificial intelligence and the future of the drug safety professional. Drug Saf. 2019;42(4):491-501.
6. Lee CY, Chen YP. Machine learning on adverse drug reactions for pharmacovigilance. Drug Discov Today. 2019;24(7):1332-43.
7. Bae JH, Baek YH, Lee JE, Song I, Lee JH, Shin JY. Machine learning for detection of safety signals from spontaneous reporting system data: example of nivolumab and docetaxel. Front Pharmacol. 2021;11:602365.
8. Spiker J, Kreimeyer K, Dang O, Boxwell D, Chan V, Cheng C, et al. Information visualization platform for postmarket surveillance decision support. Drug Saf. 2020;43(9):905-15.
9. Kreimeyer K, Dang O, Spiker J, Muñoz MA, Rosner G, Ball R, et al. Feature engineering and machine learning for causality assessment in pharmacovigilance: lessons learned from application to the FDA Adverse Event Reporting System. Comput Biol Med. 2021;135:104517.

10. Zhao Y, Yu Y, Wang H, Li Y, Deng Y, Jiang G, et al. Machine learning in causal inference: application in pharmacovigilance. Drug Saf. 2022;45(5):459-76.
11. Xu Q, Ahmadi E, Amini A, Rus D, Lo AW. Identifying and mitigating potential biases in predicting drug approvals. Drug Saf. 2022;45(5):521-33.
12. Bate A, Hobbiger SF. Artificial intelligence, real-world automation and the safety of medicines. Drug Saf. 2021;44(2):125-32.
13. Ball R, Dal Pan G. Artificial intelligence for pharmacovigilance: ready for prime time? Drug Saf. 2022;45(5):429-38.
14. Kassekert R, Grabowski N, Lorenz D, Schaffer C, Kempf D, Roy P, et al. Industry perspective on artificial intelligence/machine learning in pharmacovigilance. Drug Saf. 2022;45(5):439-48.
15. Pinheiro LC, Kurz X. Artificial intelligence in pharmacovigilance: a regulatory perspective on explainability. Pharmacoepidemiol Drug Saf. 2022;31(12):1308-10.
16. Kompa B, Hakim JB, Palepu A, Kompa KG, Smith M, Bain PA, et al. Artificial intelligence based on machine learning in pharmacovigilance: a scoping review. Drug Saf. 2022;45(5):477-91.
17. Salas M, Petracek J, Yalamanchili P, Aimer O, Kasthuril D, Dhingra S, et al. The use of artificial intelligence in pharmacovigilance: a systematic review of the literature. Pharm Med. 2022;36(5):295-306.
18. Hauben M. Artificial intelligence in pharmacovigilance: do we need explainability? Pharmacoepidemiol Drug Saf. 2022;31(12):1311-6.
19. Martin GL, Jouganous J, Savidan R, Bellec A, Goehrs C, Benkebil M, et al. Validation of artificial intelligence to support the automatic coding of patient adverse drug reaction reports, using nationwide pharmacovigilance data. Drug Saf. 2022;45(5):535-48.
20. Meldau EL, Bista S, Rofors E, Gattepaille LM. Automated drug coding using artificial intelligence: an evaluation of WHODrug koda on adverse event reports. Drug Saf. 2022;45(5):549-61.
21. Imran M, Bhatti A, King DM, Lerch M, Dietrich J, Doron G, et al. Supervised machine learning-based decision support for signal validation classification. Drug Saf. 2022;45(5):583-96.
22. Cherkas Y, Ide J, van Stekelenborg J. Leveraging machine learning to facilitate individual case causality assessment of adverse drug reactions. Drug Saf. 2022;45(5):571-82.
23. Ball R, Talal AH, Dang O, Muñoz M, Markatou M. Trust but verify: lessons learned for the application of AI to case-based clinical decision-making from postmarketing drug safety assessment at the US Food and Drug Administration. J Med Internet Res. 2024;26:e50274.
24. Liang L, Hu J, Sun G, Hong N, Wu G, He Y, et al. Artificial intelligence-based pharmacovigilance in the setting of limited resources. Drug Saf. 2022;45(5):511-9.
25. McMaster C, Chan J, Liew DF, Su E, Frauman AG, Chapman WW, et al. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. J Biomed Inform. 2023;137:104265.
26. Bate A, Stegmann JU. Artificial intelligence and pharmacovigilance: what is happening, what could happen and what should happen? Health Policy Technol. 2023;12(2):100743.
27. Painter JL, Kassekert R, Bate A. An industry perspective on the use of machine learning in drug and vaccine safety. Front Drug Saf Regul. 2023;3:1110498.