INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: https://www.ijpsjournal.com

OPEN ACCESS

Review Article

Artificial Intelligence in Pharmacovigilance: A Future Perspectives

Niraj Patil, Chetana Patil, Priyanka Patil, Lokesh Patil, Lalitkumar Patil*

SES's Arunamai College Of Pharmacy, Mamurabad, Jalgaon, Maharashtra, India-425002.

ARTICLE INFO

Published: 29 Nov 2026

Keywords:
artificial intelligence, pharmacovigilance, machine learning, natural language processing, adverse drug reactions, individual case safety reports, signal detection, large language models

DOI:
10.5281/zenodo.24645394

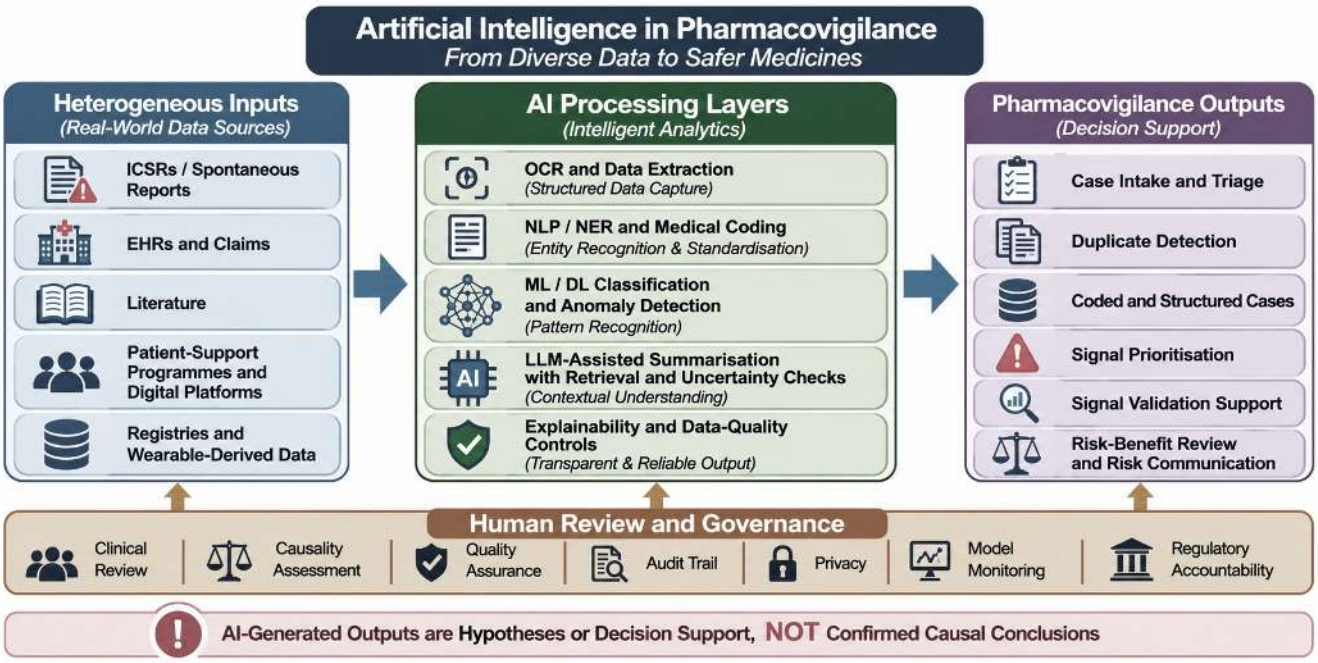
ABSTRACT

Pharmacovigilance (PV) protects patients by detecting, assessing, understanding, and preventing adverse effects and other medicine-related problems throughout the product lifecycle. Growth in spontaneous reports, electronic health records, scientific publications, patient-support programmes, and digital-platform data has increased the volume and heterogeneity of safety information. Artificial intelligence (AI)—including machine learning (ML), deep learning (DL), natural language processing (NLP), and large language models (LLMs)—can support extraction, classification, prioritisation, and analysis of this information. This narrative review examines AI applications across individual case safety report (ICSR) processing, adverse drug event (ADE) detection, signal detection and assessment, literature surveillance, and real-world data integration. Published studies demonstrate task-specific utility: a deep neural network applied to 299,028 hospital safety reports achieved an area under the receiver-operating-characteristic curve of 0.979 and identified 24.2% more confirmed allergic-reaction reports while reducing manual review by 63.8%; text-mining methods for medication-error detection achieved recall of 78.5–90.4%, although precision remained 1.6–3.4%. ML models have also shown promise in classifying ICSR assessability and predicting ADE risk, while recent LLM studies reveal important limitations in algorithm adherence and agreement with human assessors. Across applications, performance is constrained by incomplete reporting, confounding, selection bias, class imbalance, poor transportability, privacy risks, and limited explainability. AI should therefore augment—not replace—qualified PV judgment. Context-specific validation, human oversight, auditability, and regulatory alignment are prerequisites for safe implementation.

Graphical Abstract: proposed content and layout

***Corresponding Author:** Lalitkumar Patil
Address: SES's Arunamai College Of Pharmacy, Mamurabad, Jalgaon, Maharashtra, India-425002
Email ✉: Patillalitkumar.151@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



Highlights

- AI can accelerate extraction and prioritisation of safety information from structured and unstructured sources.
- Published performance is task-dependent: high recall may coexist with low precision and substantial false-positive review burden.29 Nov 2026
- LLMs can assist with case summarisation and structured reasoning, but current evidence shows incomplete algorithm adherence in ICSR causality assessment.
- External validation, data provenance, explainability, privacy safeguards, and human oversight are essential for deployment.
- ICH E2D (R1) (effective in the EU from 18 March 2026) explicitly addresses newer post-approval safety-data sources, including digital platforms.

Abbreviations

ADE, adverse drug event; ADR, adverse drug reaction; AE, adverse event; AI, artificial intelligence; AEFI, adverse event following immunisation; AUROC, area under the receiver-operating-characteristic curve; AUPRC, area under the precision-recall curve; BCPNN, Bayesian confidence propagation neural network; DL, deep learning; DILI, drug-induced liver injury; DNN, deep neural network; DDI, drug–drug interaction; EHR, electronic health record; EMA, European Medicines Agency; FAERS, FDA Adverse Event Reporting System; FDA, US Food and Drug Administration; GVP, Good Pharmacovigilance Practices; HITL, human-in-the-loop; ICSR, individual case safety report; ICH, International Council for Harmonisation; LLM, large language model; MedDRA, Medical Dictionary for Regulatory Activities; ML, machine learning; NER, named-entity recognition; NLP, natural language processing; ODCS, organised data collection system; PRR, proportional reporting ratio; PV, pharmacovigilance; ROR, reporting odds ratio; RWD, real-world data; RWE, real-world evidence; SHAP, Shapley additive explanations; XAI, explainable artificial intelligence.

INTRODUCTION

Pharmacovigilance is the science and set of activities concerned with detecting, assessing, understanding, and preventing adverse effects or other medicine-related problems. Its remit extends beyond spontaneous reporting to clinical-trial safety, post-authorisation surveillance, medication errors, product-quality issues, lack of efficacy in relevant settings, and the evaluation of risks in populations that may be underrepresented in pre-approval trials. Spontaneous reporting systems remain indispensable for detecting unexpected or rare

events, but reports are incomplete, heterogeneous, subject to stimulated and under-reporting, and generally do not provide reliable exposure denominators. Disproportionality statistics consequently identify reporting patterns rather than establish causality [1,2].

The expansion of EHRs, administrative claims, registries, biomedical literature, patient-support programmes, and digital platforms creates opportunities to triangulate safety evidence. It also introduces difficulties: incompatible terminologies, duplicated records, missing clinical context, confounding by indication, channeling bias, and variable data access. AI methods may help process these sources at scale, particularly where safety information is embedded in narrative text. However, model outputs are only as reliable as their labels, input data, validation design, and intended use [3–5].

The literature now includes both research prototypes and operational automation. Evidence is strongest for bounded tasks such as text classification, information extraction, and case prioritisation. Evidence is less mature for autonomous causality assessment, signal confirmation, and regulatory decision-making. A scoping review of 393 publications found that most studies still relied on traditional statistical methods and that only a minority reflected contemporary good machine-learning practice [3]. Industry surveys likewise describe increasing automation in ICSR workflows but identify training-data availability and regulatory harmonisation as persistent barriers [4].

This review critically evaluates AI across the PV lifecycle, distinguishes demonstrated results from proposed applications, and identifies the conditions required for responsible implementation. Particular attention is given to original studies with transparent datasets and quantitative performance reporting, alongside relevant reviews and current regulatory guidance.

2. Review methodology and literature search strategy

This manuscript was developed as a targeted narrative review, not a registered systematic review or meta-analysis. The supplied 35-page draft was examined in full, including its embedded diagrams and reference sections. Its relevant subject matter and citations were retained as the starting point; duplicated passages, unrelated general-healthcare material, and the telecommunications/radio-frequency discussion were excluded because they did not directly inform pharmaceutical safety surveillance.

A targeted literature search was conducted on 25 November 2026 using PubMed/PubMed Central-indexed records, publisher pages, and official regulatory or intergovernmental sources. Search concepts included combinations of: artificial intelligence, machine learning, deep learning, pharmacovigilance, adverse drug event, adverse drug reaction, ICSR, causality assessment, signal detection, text mining, NLP, large language model, explainable AI, EHR, FAERS, social media, and real-world data. Reference-list screening was used to identify relevant landmark and methodological papers.

Priority was given to peer-reviewed original research from 2020–2026, with earlier landmark work retained where needed for signal-detection principles, NLP, and social-media monitoring. Review articles were used to map the field, not as substitutes for original-study evidence. Official ICH, FDA, EMA, and WHO materials were used for regulatory and governance context. Studies were selected for direct relevance to PV tasks, clarity of methods, availability of performance measures, and relevance to implementation.

Because this was a targeted narrative search, no claim is made that all eligible publications were identified. No PRISMA flow diagram or pooled effect estimate is presented. The search date, databases/sources, and search concepts should be updated and adapted to the target journal before submission; a formal systematic review would require reproducible database-specific search strings, dual independent screening, a documented selection flow, and risk-of-bias assessment.

3. Pharmacovigilance foundations and data sources

3.1 Core PV activities and ICSR lifecycle

An ICSR is a structured report of one or more suspected adverse events in an individual. Case validity requires the minimum report elements specified by applicable standards: an identifiable patient, an identifiable reporter, a suspect medicinal product, and a suspected

adverse event/reaction. These elements establish reportability; they do not prove that the product caused the event. The lifecycle commonly includes receipt and intake, validity assessment, duplicate search, data entry and medical coding, narrative preparation, medical review, causality assessment, expedited or periodic reporting as applicable, follow-up, aggregate analysis, signal management, and risk-minimisation activity.

Medical coding commonly uses MedDRA for medical concepts and event terms, with product dictionaries such as WHODrug used where applicable. Automated coding must preserve the original verbatim term and provide traceability to the selected code, including uncertainty and reviewer changes. Causality tools such as the WHO–UMC system and Naranjo algorithm can structure assessment, but they do not eliminate the need to evaluate chronology, alternative causes, dechallenge/rechallenge, biological plausibility, concomitant medicines, and the completeness of the case.

3.2 Main data sources

- Spontaneous reports: national systems such as FAERS and international repositories such as VigiBase; useful for signal generation but affected by under-reporting, stimulated reporting, duplicates, and lack of exposure denominators.
- Clinical records and claims: provide longitudinal diagnoses, prescriptions, laboratory results, procedures, and outcomes; require careful phenotype definitions, linkage, and control of confounding.
- Clinical trials and registries: offer protocol-defined follow-up and more structured data, but may be limited by sample size, duration, eligibility criteria, and generalisability.
- Biomedical literature: includes case reports, observational studies, trial reports, and reviews; screening and extraction can be automated, but relevance and evidence quality require human appraisal.
- Digital platforms and patient-support programmes: may provide early patient-reported information, but identity, product exposure, event validity, consent, and reportability must be assessed under applicable requirements.
- Mobile health and wearables: can contribute repeated physiological measurements and patient-reported outcomes; device validity, adherence, missingness, and signal attribution remain important.

4. AI methods and scientific principles

4.1 Machine learning

ML algorithms infer patterns from examples and may perform classification, regression, ranking, clustering, or anomaly detection. Supervised models require labelled examples; unsupervised methods identify structure without explicit labels; semi-supervised and weakly supervised approaches use limited expert annotation alongside larger unlabelled datasets. In PV, ML can prioritise reports, classify event narratives, identify duplicates, estimate patient-level risk, or support the assessment of report completeness. Model selection should follow the intended task and data structure rather than a presumption that a more complex algorithm is inherently superior.

4.2 Deep learning

DL uses multilayer neural networks to learn representations from high-dimensional inputs. Convolutional networks, recurrent architectures such as long short-term memory networks, attention mechanisms, and transformer encoders have been applied to clinical text and safety narratives. DL can capture context and word order beyond exact

keyword matching, but often requires larger datasets, computational resources, careful regularisation, and external validation. Performance should be reported with metrics suited to class imbalance, including precision, recall, F1 score, AUROC, AUPRC, calibration, and workload-related measures.

4.3 Natural language processing

NLP converts unstructured language into computable representations. Typical PV tasks include named-entity recognition for drugs, events, doses, dates, and outcomes; relation extraction to connect a suspect product with an event; negation and temporality detection; abbreviation resolution; spelling normalisation; language identification and translation; document

classification; and mapping verbatim terms to controlled vocabularies. Context is essential: 'no rash', 'history of rash', and 'rash after treatment' should not be treated as equivalent. Clinical NLP systems must also distinguish an event from an indication, a pre-existing condition, a family history, or an event explicitly judged unrelated.

4.4 Large language models

LLMs can summarise long narratives, extract candidate fields, translate or standardise text, generate structured drafts, and assist reviewers in navigating evidence. Their fluent output can obscure factual errors, omissions, unsupported inference, or inconsistent application of a structured algorithm. A safe workflow should constrain the model to supplied source material, require field-level provenance, preserve the source narrative, return 'unknown' when information is absent, apply deterministic validation rules, and route uncertain or high-impact outputs to a qualified reviewer. LLMs should not autonomously create reportable facts, assign definitive causality, or submit regulatory reports without validated controls.

4.5 Explainable AI

XAI aims to make model behaviour more understandable. Intrinsically interpretable models include sparse regression, decision trees, and rule-based systems. Post-hoc approaches include local surrogate explanations (LIME), Shapley-value feature attribution (SHAP), feature permutation, and attention visualisation. Explanations can help detect spurious correlations and support case review, but a plausible explanation does not prove that a model is correct or causally valid. Explanations should be evaluated for stability, faithfulness, and usefulness to the intended reviewer.

4.6 Bayesian networks and statistical signal detection

Bayesian networks represent conditional dependencies among variables using a directed acyclic graph and can combine prior knowledge with observed data. They may support risk modelling where relationships are uncertain or involve multiple interacting factors. In spontaneous-report databases, established statistical screening methods include the reporting odds ratio, proportional reporting ratio, information component, and empirical Bayes geometric mean. These methods remain important comparators and should not be described as replaced by AI. Disproportionality is a screening signal, not a causal estimate; confounding, notoriety, reporting intensity, masking, and multiple testing require follow-up assessment [1,2].

5. Applications of AI across the pharmacovigilance lifecycle

5.1 ICSR intake, triage, and data extraction

Optical character recognition can convert scanned forms and PDFs into text, after which NLP or LLM-assisted extraction can identify the four minimum report elements and additional fields such as age, sex, suspect and concomitant products, dose, route, event onset, seriousness criteria, outcome, and medical history. Automated triage can route potentially serious, time-critical, incomplete, or non-valid cases to appropriate queues. These systems should retain source-page or source-span references so reviewers can verify each extracted value.

Kassekert and colleagues reported an industry survey of TransCelerate member companies across 2019–2021. The survey described increasing implementation of intelligent automation in ICSR processes, particularly rule-based automation, robotic process automation, lookups, and workflow tools. It also documented interest in ML and NLP, while identifying suitable training data and harmonised regulatory guidance as important barriers [4]. This is an industry implementation survey rather than a controlled efficacy trial; its findings should not be interpreted as a universal estimate of automation accuracy.

5.2 Medical coding and duplicate detection

Contextual NLP can map reporter verbatims to MedDRA terms and identify potential product/event relationships. Duplicate detection can combine demographic similarity, event terms, dates, product details, reporter information, and narrative embeddings. A similarity score should trigger review rather than automatic deletion because genuinely distinct follow-up reports may resemble one another. The

system should retain linkage between the original and suspected duplicate records and preserve the audit trail.

5.3 Adverse-event detection in clinical narratives

Luo and colleagues' structured review of EHR-based PV described the progression from rule-based and statistical NLP toward ML-based extraction and highlighted unresolved needs in context characterisation, on-label versus off-label use, polypharmacy, data integration, and shared benchmark corpora [5]. A key scientific requirement is to distinguish temporally associated symptoms from suspected drug-related events and to capture uncertainty rather than overstate causality.

5.4 Signal detection and prioritisation

AI can rank drug-event pairs, detect temporal or subgroup patterns, and combine narrative and coded information. It may improve prioritisation where the number of candidate associations is large. However, spontaneous-report databases have no reliable denominator for exposed patients and are vulnerable to reporting biases. Model outputs therefore require triangulation with clinical knowledge, product information, epidemiology, mechanistic plausibility, and independent data sources. A signal is information suggesting a new potentially causal association or a new aspect of a known association that warrants verification; it is not proof of causation.

5.5 Signal validation, evaluation, and causality assessment

AI may help assess whether a case contains sufficient information for expert evaluation, identify relevant chronology and alternative causes, retrieve known evidence, and prioritise signals for validation. It is particularly important to separate the relatively bounded task of identifying assessable reports from the much harder task of determining causality. Kreimeyer and colleagues analysed 326 redacted FAERS reports annotated using modified WHO–UMC criteria. Their best models achieved recall and F1 scores above 0.80 for identifying assessable reports and above 0.75 for identifying reports meeting at least a 'Possible' causality threshold. They emphasised that report text alone was insufficient and that clinical and temporal information, including external context, mattered [8].

5.6 Literature monitoring

Automated literature surveillance can retrieve candidate publications, classify safety relevance, extract product-event mentions, and prioritise full-text review. A robust workflow should distinguish original clinical evidence from reviews, editorials, preprints, conference abstracts, and non-peer-reviewed material. It should also detect duplicate publication, identify the population and exposure, extract the event definition and timing, and preserve a direct link to the source passage. LLM-generated summaries must be checked against the article because omission of caveats or misattribution of results can distort safety evaluation.

5.7 Social media and digital-platform surveillance

Social-media data can provide patient-reported experiences, but posts frequently lack verified identity, product exposure, dose, chronology, diagnosis, and outcome. Pierce and colleagues retrospectively examined 935,246 Facebook and Twitter posts related to ten FDA postmarketing safety signals and six negative controls. Automated classification identified 98,252 posts resembling adverse events; only 13 posts were selected for product-event causality assessment. The study illustrates the gap between high-volume text detection and clinically interpretable, valid case evidence [14].

The regulatory context has evolved. ICH E2D (R1), finalised in 2025 and implemented in the EU from 18 March 2026, updates definitions and reporting standards for post-approval safety data and addresses newer sources such as digital platforms, patient-support programmes, and market-research programmes [20]. This does not mean every online mention automatically constitutes a valid ICSR; source context, report minimum criteria, and regional requirements remain decisive.

5.8 Real-world evidence, distributed networks, and personalised safety

Linking EHRs, claims, registries, and other RWD can support active surveillance, longitudinal outcome ascertainment, subgroup analysis, and evaluation of rare

or delayed outcomes. Distributed data networks and federated approaches may allow analysis across institutions without centralising all patient-level data, but harmonisation of data models, phenotype definitions, and analytic protocols is essential. Wong and colleagues describe opportunities and methodological considerations for applying ML in distributed pharmacoepidemiologic and PV networks [15], while Lavertu and colleagues discuss the growing role of real-world data and digital monitoring in PV [18]. Patient-level prediction models may help identify vulnerable groups, but predictions must be calibrated and validated in the population and clinical setting in which they will be used.

5.9 Wearables and patient-generated health data

Wearables and mobile devices may contribute repeated heart-rate, activity, sleep, glucose, or other measurements. Such data can provide temporal resolution that conventional reports lack, but sensor error, device heterogeneity, adherence, missing-not-at-random patterns, and confounding by illness or activity can generate misleading alerts. Their role in PV is promising but should be considered investigational unless the specific device-data pipeline and intended use have been validated.

6. Recent research developments and evidence from original studies

The selected studies below illustrate different PV tasks and should not be compared as if they measured one common outcome. They differ in data source, target label, prevalence, validation design, and intended use.

Table 1: Selected Original Studies of AI and Text Mining Relevant to Pharmacovigilance

Study	Objective and methods	Main quantitative findings	Limitations and interpretation
Eskildsen et al., 2020 [6]	Compared two text-mining systems and an expert standard method for identifying a specific insulin medication-error pattern in 154,209 ICSRs; 2,533 reports were manually annotated.	Recall: CPR text mining 90.4%, SSA 88.1%, I2E 78.5%. Precision: 1.9%, 3.4%, and 1.6%, respectively.	High recall but very low precision means substantial false-positive review. One specific medication-error use case; results should not be generalised to all signal types.
Yang et al., 2020 [7]	Developed a four-layer attention-based DNN using free-text safety reports from two academic hospitals; 299,028 reports involving 172,854 patients.	AUROC 0.979 (95% CI 0.973–0.985); AUPRC 0.809 (0.773–0.845). Compared with keyword search, 24.2% more confirmed allergic reactions and 63.8% fewer reports for manual review.	Strong external/time-period evaluation, but test-set labelling constraints and keyword-enriched training may introduce selection bias. Hospital safety reports are not identical to spontaneous regulatory ICSRs.
Kreimeyer et al., 2021 [8]	Used supervised ML, engineered report features, NLP, and external information to classify 326 redacted FAERS reports for assessability and causality-related categories.	Best models: recall and F1 >0.80 for assessable reports; >0.75 for reports meeting at least Possible causality threshold.	Small annotated dataset; causal reasoning depends on external clinical context and complex logic. Most defensible use is workflow triage, not autonomous causality determination.
Mohsen et al., 2021 [9]	Combined rat liver gene-expression profiles from Open TG–GATEs with ADR associations from FAERS; trained 14 deep neural-network models for 14 ADR phenotypes.	Mean validation accuracy 89.94% (range 71.42–100%).	Preclinical toxicogenomics plus spontaneous-report labels; small numbers of positive/negative compounds per ADR and limited external clinical validation. Supports early discovery prioritisation, not patient-level prediction.

Study	Objective and methods	Main quantitative findings	Limitations and interpretation
Hu et al., 2022 [10]	Compared ML models for predicting ADEs in older inpatients; evaluated multiple algorithms and risk factors.	AdaBoost performed best: accuracy 88.06%, precision 68.57%, recall 48.21%, F1 52.75%, AUC 0.91.	Single-centre/setting-specific retrospective prediction; moderate recall and potential transportability limitations. Risk factors included medication burden, age, length of stay, and comorbidity-related variables.
Lee and Yoo, 2024 [11]	Developed InterDILI, combining molecular fingerprints/physicochemical descriptors and interpretable feature-importance/attention methods for DILI prediction.	Reported hold-out performance included AUROC 0.97, AUPR 0.95, and accuracy 0.90; ten-fold cross-validation performance was lower (AUROC 0.87, accuracy 0.78, AUPR 0.87).	Drug-level computational prediction rather than post-marketing signal detection; dataset composition and external validation determine generalisability. Explanations identify predictive features, not proven mechanisms.
Abate et al., 2025 [12]	Evaluated ChatGPT (GPT-3.5-turbo-16k) and Gemini 1.5 Pro for WHO-algorithm causality assessment in 150 VAERS myocarditis/pericarditis reports; compared with human expert assessment.	ChatGPT adherence 34% and median agreement 71%; Gemini adherence 7% and median agreement 53%. Inconsistency: 8.0% for ChatGPT and 46.7% for Gemini.	Focused on a specific vaccine-event subset and early LLM versions. Findings demonstrate feasibility but inadequate reliability for autonomous case causality assessment.
Heckmann et al., 2026 [13]	Compared biomedical LLMs, prompting strategies, and causality algorithms against two human evaluators in 150 ICSRs (140 FAERS; 10 VAERS).	Best combination (Medicine LLaMA-3 8B with Naranjo and chain-of-thought prompting) achieved 64% agreement for final causality classification; critical item-level agreement remained low.	Small sample and selected model/prompt combinations. Irrelevant responses and weak performance on temporality, alternative causes, and objective evidence limit clinical use.

Abbreviations: AUC/AUROC, area under the receiver-operating-characteristic curve; AUPR/AUPRC, area under the precision-recall curve; CPR, Center for Protein Research; DNN, deep neural network; FAERS, FDA Adverse Event Reporting System; ICSR, individual case safety report; SSA, safety surveillance advisor. Metrics are reproduced from the cited publications; they are not directly comparable across studies.

7. Mechanistic interpretation and workflow integration

AI contributes to PV through a sequence of computational operations: (i) acquisition and normalisation of source data; (ii) extraction of entities, events, temporal expressions, and contextual qualifiers; (iii) representation of records as structured variables or embeddings; (iv) classification, ranking, clustering, or prediction; (v) explanation and uncertainty estimation; and (vi) expert review and regulatory action. Each stage can introduce error. OCR errors propagate into entity extraction; incorrect event coding distorts signal counts; label noise biases supervised learning; and distribution shift can invalidate a model after changes in reporting practice, product mix, language, or clinical documentation.

A useful architecture is therefore modular. Data ingestion should retain source provenance and original text. Extraction services should return candidate values with confidence and source spans. A rules layer should validate mandatory fields, units, dates, and internal consistency. Statistical and ML services should prioritise cases or associations, with explicit thresholds tied to

the consequences of false positives and false negatives. A reviewer interface should display evidence, uncertainty, and model rationale. Final medical judgement, causality assessment, signal confirmation, and regulatory submission should remain governed by accountable PV personnel.

Table 2: AI Methods Mapped to PV Tasks and Principal Controls

PV task	Suitable AI methods	Expected contribution	Critical control
ICSR intake and validity	OCR, NER, NLP, constrained LLM extraction, rules	Candidate extraction of patient, reporter, product, event, and case details	Field-level provenance; four minimum criteria checked; source retained
Medical coding	Contextual NLP, transformer encoders, terminology matching	Verbatim-to-MedDRA/ WHODrug candidate mapping	Human verification of ambiguous codes; retain verbatim and coding rationale
Duplicate management	Fuzzy matching, embeddings, Siamese networks, graph linkage	Prioritise potentially related reports	Do not auto-delete; reviewer confirmation and audit trail
Literature screening	Information retrieval, classifiers, LLM-assisted extraction	Prioritise relevant articles and extract product-event evidence	Source verification; publication type and evidence quality labelled
Signal detection	Disproportionality statistics, ML ranking, temporal models, graph methods	Rank candidate drug-event associations	Do not equate ranking with causality; assess bias and background reporting
Signal validation and causality support	NLP, structured decision support, retrieval-augmented LLMs	Identify missing data, relevant evidence, chronology, and alternative causes	Human causality assessment; evidence traceability; abstention on insufficient data
Patient-level risk prediction	Supervised ML, survival models, multimodal models	Identify higher-risk groups for monitoring	External validation, calibration, subgroup fairness, clinical utility assessment

8. Critical comparative analysis

The evidence base supports a nuanced conclusion. Models can improve retrieval and prioritisation, but a high AUROC or recall does not by itself establish clinical utility. In a low-prevalence setting, even a model with apparently good discrimination may generate many false positives. Eskildsen et al. demonstrate this directly: recall was high, but precision was below 4% [6]. Conversely, Yang et al. reported strong discrimination and workload reduction for allergic-reaction detection in hospital safety reports, but the target task and data source were different [7]. These findings should not be collapsed into a single claim that AI improves PV accuracy by a fixed percentage.

A second distinction is between operational automation and scientific inference. Extracting a product name, detecting a duplicate, or identifying a report with sufficient information is a constrained task. Determining whether a medicine caused an event requires evaluation of time sequence, alternative explanations, dose-response, dechallenge/rechallenge, prior knowledge, and external evidence. Kreimeyer et al. showed that text engineering alone did not capture all components of causality assessment [8]. LLM studies in 2025 and 2026 reinforce this distinction: models can execute a structured workflow in some cases, yet adherence and agreement remain inadequate for autonomous decisions [12,13].

A third distinction is between model performance and transportability. Internal validation may overestimate performance when training and test data share institutions, time periods, coding practices, or patient characteristics. External validation across hospitals, countries, languages, drug classes, and calendar periods is essential. Prospective silent-mode evaluation should precede deployment, and performance should be monitored after implementation for data drift, alert fatigue, and subgroup disparities.

Finally, interpretability is necessary but not sufficient. Feature attribution may help identify influential variables, but does not establish causal mechanisms. A deployable PV model requires a clear context of use,

defined decision thresholds, documented data lineage, version control, change management, reproducible validation, and a process for investigating errors.

9. Pharmaceutical and clinical applications

- Pharmaceutical companies: case intake, data extraction, duplicate prioritisation, literature triage, coding support, narrative drafting, and quality-control checks.
- Regulatory authorities: prioritisation of high-volume reports, exploration of signal clusters, structured review support, and analysis across distributed data sources.
- Hospitals and health systems: detection of medication-related harm in incident reports and EHR narratives; identification of allergy and medication-error patterns.
- Clinical development and drug discovery: integration of toxicogenomics and clinical safety databases to prioritise compounds or potential toxicity phenotypes for further testing.
- Population and precision safety: risk stratification, subgroup surveillance, and longitudinal monitoring using linked clinical and patient-generated data, subject to robust validation.

These applications are complementary. Preclinical prediction can inform candidate selection, clinical-trial surveillance can identify emerging tolerability issues, and postmarketing PV can monitor broader and longer-term populations. However, the evidentiary standards differ across these settings; a drug-discovery model should not be presented as proof of effectiveness in routine postmarketing PV.

10. Challenges and limitations

10.1 Data quality, missingness, and bias

Spontaneous reports are selectively submitted and often incomplete. Missingness may be informative: serious, publicised, or legally salient events may be reported more frequently. Duplicate cases, inconsistent coding, language variation, and changing reporting practices can distort model inputs. Training labels may reflect reviewer disagreement or historical coding conventions rather than an objective ground truth.

10.2 Confounding and causal inference

Drug-event co-occurrence does not establish a causal relationship. Indication, disease severity, concomitant medicines, age, comorbidity, and healthcare utilisation can influence both drug exposure and event occurrence. Models trained on spontaneous reports may learn reporting behaviour, product notoriety, or clinical workflow patterns rather than pharmacological mechanisms.

10.3 Class imbalance and performance reporting

Serious or novel events are often rare relative to all reports. Accuracy can therefore be misleading. Studies should report sensitivity/recall, specificity, precision/positive predictive value, F1, AUROC, AUPRC, calibration, confidence intervals, and the number of cases requiring manual review. Threshold selection should reflect the relative costs of missed cases and false alerts.

10.4 Privacy, security, and data governance

EHRs, patient narratives, and digital platforms may contain identifiable or sensitive information. Data minimisation, lawful processing, access controls, encryption, retention limits, de-identification where appropriate, vendor controls, and secure model-development environments are required. Federated learning can reduce central data movement but does not eliminate privacy risks or the need for governance.

10.5 Explainability, reproducibility, and model drift

Opaque models can make it difficult to investigate why a report was prioritised or why a code was assigned. Reproducibility may be affected by model updates, prompt changes, stochastic generation, external retrieval changes, or vendor-side modifications. Versioning, locked configurations, validation datasets, audit logs, monitoring, and change-control procedures are essential.

10.6 Regulatory and operational constraints

AI-enabled PV systems operate within existing obligations for case validity, reporting timelines, quality systems, record retention, and accountability. FDA's January 2025 draft guidance proposes a risk-based credibility framework for AI used to support regulatory decision-making on drug and biological products; it is

draft guidance and should not be misrepresented as a final PV-specific rule [22]. ICH E2D (R1) updates post-approval safety-data definitions and management, including digital platforms, but does not endorse unsupervised AI decision-making [20]. WHO guidance emphasises ethics, human rights, accountability, transparency, and governance for AI in health [21].

Table 3: Key Challenges, Consequences, and Mitigation Strategies

Challenge	Potential consequence	Recommended mitigation
Incomplete, noisy, and heterogeneous data	Missed events, false alerts, unreliable extraction	Data-quality profiling; standard vocabularies; source provenance; uncertainty flags
Under-reporting and stimulated reporting	Distorted drug-event associations	Triangulate spontaneous reports with EHR, claims, literature, and epidemiology
Confounding and selection bias	Spurious associations and poor causal inference	Causal design principles; negative controls; stratification; expert review
Rare outcomes and class imbalance	Misleading accuracy; low positive predictive value	AUPRC, sensitivity/precision, calibration, threshold and workload analyses
Poor transportability and drift	Performance degradation across settings or time	External and temporal validation; prospective silent deployment; continuous monitoring
Opaque or generative outputs	Unreviewable decisions; fabricated details	Source-grounded extraction; explanations; constrained outputs; mandatory reviewer approval
Privacy and cybersecurity	Disclosure or misuse of patient data	Data minimisation, access control, encryption, secure processing, vendor oversight
Unclear accountability	Regulatory non-compliance and unsafe decisions	Defined intended use, SOPs, audit trails, validation, change control, named accountable owners

11. Future perspectives

11.1 Multimodal and federated safety analytics

Future systems may combine coded reports, clinical narratives, laboratory trajectories, imaging-derived features, genomics, and patient-generated data. Multimodal integration could improve event characterisation, but only if the data are temporally aligned, clinically meaningful, and validated across sites. Federated approaches may enable multicentre learning while retaining local data custody, although harmonisation and privacy protections remain necessary.

11.2 Retrieval-grounded and domain-adapted LLMs

LLMs should be evaluated on PV-specific tasks using curated benchmark datasets, blinded expert reference standards, and transparent error taxonomies. Retrieval-augmented systems should cite the source document and exact evidence span for every extracted claim. Structured output schemas, constrained vocabularies, calibrated abstention, and deterministic post-processing can reduce—but not eliminate—hallucination risk. Evaluation should include factual consistency, omission rates, temporal reasoning, multilingual performance, and reviewer time.

11.3 Prospective validation and clinical utility

Research should progress from retrospective model development to temporal and external validation, prospective silent-mode testing, and controlled implementation studies. Outcomes should include not only model metrics but also time to case processing, reviewer workload, signal yield, false-alert burden, time to escalation, and effects on patient-safety decisions. Cost-effectiveness and workflow integration should be assessed transparently.

11.4 Shared benchmarks and reporting standards

The field needs de-identified, multilingual, representative datasets with transparent annotation protocols, adjudication procedures, and standardised task definitions. Benchmarking should prevent leakage between patient, product, and time partitions. Reporting should document data provenance, intended use, model architecture, hyperparameters, thresholds, calibration, fairness, external validation, and post-deployment monitoring.

11.5 Human-AI collaboration and governance

The most defensible near-term model is augmented intelligence: AI handles repetitive retrieval and candidate prioritisation, while trained PV professionals interpret evidence, resolve uncertainty, and make accountable decisions. Governance should define permitted and prohibited uses, escalation thresholds, human override, auditability, incident reporting, retraining controls, and vendor responsibilities. WHO's 2024/2025 guidance on large multimodal models further underscores the need for governance proportional to the capabilities and risks of generative systems [21].

CONCLUSION

AI has credible potential to improve the efficiency of pharmacovigilance by extracting safety information from unstructured data, prioritising reports, supporting coding and duplicate review, and integrating evidence across data sources. Original studies show meaningful gains for specific tasks, including allergic-reaction detection and medication-error text mining, but also demonstrate the limitations of low precision, restricted datasets, and incomplete causal reasoning. Recent LLM evaluations are particularly clear that fluent output does not equal reliable causality assessment. AI-derived associations should therefore be treated as candidates for expert evaluation rather than confirmed adverse drug reactions or regulatory conclusions. Responsible implementation requires fit-for-purpose validation, source traceability, transparent performance reporting, privacy safeguards, continuous monitoring, and human accountability. The field's central opportunity is not the replacement of pharmacovigilance expertise, but the creation of more scalable and evidence-integrated workflows that allow experts to focus on clinically meaningful safety questions.

13.Tables and scientific figure suggestions

The three tables above are designed for direct inclusion. The following additional table is recommended if the target journal permits a fifth table.

Table 4: Comparison of Established Signal-Detection Approaches and AI-supported Approaches

Approach	Input and output	Strength	Key limitation
PRR/ROR	Counts in spontaneous-report contingency tables; disproportionality score	Transparent, relatively simple, established screening baseline	Sensitive to reporting patterns, confounding, sparse counts, and multiple testing
BCPNN/IC	Bayesian shrinkage estimate of disproportionality	Stabilises estimates for sparse data and supports signal screening	Still detects reporting disproportionality, not causality
Supervised ML	Annotated reports or drug-event pairs; predicted class/rank	Can combine structured and text-derived features	Requires reliable labels and external validation; may reproduce annotation bias
DL/NLP	Narratives and high-dimensional clinical text; entities, relations, classifications	Captures context and linguistic variation beyond exact matching	Computationally intensive; can be opaque and domain-specific
LLMs	Narratives, case fields, and retrieved literature; summaries/extractions/reasoning support	Flexible summarisation, translation, and structured extraction	Hallucination, inconsistent reasoning, prompt sensitivity, privacy and reproducibility risks

Approach	Input and output	Strength	Key limitation
Bayesian networks	Probabilistic variables and dependencies; conditional risk estimates	Can incorporate prior knowledge and represent uncertainty	Structure specification and causal interpretation require care

Recommended figures:

Figure 1: End-to-end AI-enabled Pharmacovigilance Architecture. Show Data Sources, Preprocessing, NLP/ml/LLM Modules, PV Outputs, and a Human-Governance Layer. Place After Section 5.

Figure 2: ICSR Processing Workflow with Human-In-The-Loop Checkpoints. Show Intake, Minimum-Criteria Validation, Duplicate Search, Extraction, MedDRA Coding, Seriousness Triage, Narrative Drafting, Medical Review, Submission, and Audit Trail. Place After Section 3.1.

Figure 3: Signal Management Funnel. Distinguish Signal Generation, Prioritisation, Validation, Assessment, Recommendation, and Risk Minimisation; Indicate That AI Supports but Does Not Replace Causal Assessment. Place After Section 5.4.

Figure 4: LLM Safety Architecture for PV. Show Source-Grounded Retrieval, Structured Extraction, Provenance Spans, Rule-Based Validation, Uncertainty/Abstention, Human Review, and Versioned Audit Logging. Place After Section 5.5 Or 10.5.

REFERENCES

1. Jeetu G, Anusha G. Pharmacovigilance: a worldwide master key for drug safety monitoring. J Young Pharm. 2010;2 (3):315-320. doi:10.4103/0975-1483.66802.
2. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18 (6):427-436. doi:10.1002/pds.1742.
3. Kompa B, Hakim JB, Palepu A, Kompa KG, Smith M, Bain PA, et al. Artificial intelligence based on machine learning in pharmacovigilance: a scoping review. Drug Saf. 2022;45 (5):477-491. doi:10.1007/s40264-022-01176-1.
4. 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):439-448. doi:10.1007/s40264-022-01164-5.
5. 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):1075-1089. doi:10.1007/s40264-017-0558-6.
6. Eskildsen NK, Eriksson R, Christensen SB, Aghassipour TS, Bygsø MJ, Brunak S, et al. Implementation and comparison of two text mining methods with a standard pharmacovigilance method for signal detection of medication errors. BMC Med Inform Decis Mak. 2020;20:94. doi:10.1186/s12911-020-1097-0.
7. Yang J, Wang L, Phadke NA, Wickner PG, Mancini CM, Blumenthal KG, et al. Development and validation of a deep learning model for detection of allergic reactions using safety event reports across hospitals. JAMA Netw Open. 2020;3 (11):e2022836. doi:10.1001/jamanetworkopen.2020.22836.
8. 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. doi:10.1016/j.compbiomed.2021.104517.
9. Mohsen AM, Tripathi LP, Mizuguchi K. Deep learning prediction of adverse drug reactions in drug discovery using Open TG-GATEs and FAERS databases. Front Drug Discov. 2021;1:768792. doi:10.3389/fddsv.2021.768792.
10. Hu Q, Wu B, Wu J, Xu T. Predicting adverse drug events in older inpatients: a machine learning study. Int J Clin Pharm. 2022;44 (6):1304-1311. doi:10.1007/s11096-022-01468-7.

11. Lee S, Yoo S. InterDILI: interpretable prediction of drug-induced liver injury through permutation feature importance and attention mechanism. J Cheminform. 2024;16:1. doi:10.1186/s13321-023-00796-8.
12. Abate A, Poncato E, Barbieri MA, Powell G, Rossi A, Peker S, et al. Off-the-shelf large language models for causality assessment of individual case safety reports: a proof-of-concept with COVID-19 vaccines. Drug Saf. 2025;48:805-820. doi:10.1007/s40264-025-01531-y.
13. Heckmann NS, Papoutsi DG, Barbieri MA, Battini V, Mølgaard SN, Schmidt SØ, et al. Biomedical large language models and prompt engineering for causality assessment of individual case safety reports in pharmacovigilance. Pharm Res. 2026;43 (8):2761-2774. doi:10.1007/s11095-026-04121-4.
14. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, Van Le H, et al. Evaluation of Facebook and Twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug Saf. 2017;40 (4):317-331. doi:10.1007/s40264-016-0491-0.
15. Wong J, Prieto-Alhambra D, Rijnbeek PR, et al. Applying machine learning in distributed data networks for pharmacoepidemiologic and pharmacovigilance studies: opportunities, challenges, and considerations. Drug Saf. 2022;45 (5):493-510. doi:10.1007/s40264-022-01158-3.
16. 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. doi:10.3389/fdsfr.2023.1110498.
17. Ball R, Dal Pan G. Artificial intelligence for pharmacovigilance: ready for prime time? Drug Saf. 2022;45 (4):429-438. doi:10.1007/s40264-022-01157-4.
18. Lavertu A, Vora P, Giacomini KM, et al. A new era in pharmacovigilance: toward real-world data and digital monitoring. Clin Pharmacol Ther. 2021;109 (5):1197-1202. doi:10.1002/cpt.2172.
19. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. ISBN:978-92-4-002920-0.
20. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH harmonised guideline E2D (R1): post-approval safety data: definitions and standards for management and reporting of individual case safety reports. Geneva: ICH; 2025.
21. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2025. ISBN:978-92-4-008475-9.
22. US Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: draft guidance for industry and other interested parties. Silver Spring (MD): FDA; 2025.

HOW TO CITE: Niraj Patil, Chetana Patil, Priyanka Patil, Lokesh Patil, Lalitkumar Patil, Artificial Intelligence in Pharmacovigilance: A Future Perspectives, Int. J. of Pharm. Sci., 2026, Vol 4, Issue 11, 5949-5962. https://doi.org/10.5281/zenodo.24645394