Review Article

What Happens to the Safety Case When Artificial Intelligence Enters Pharmacovigilance? An Umbrella Review from Intake to Signal Action

David Thompson[1*], Sarah Mitchell[1], Rachel Adams[2], James Brown[3], Michael Lee[4]

[1]Department of AI and Pharmacovigilance Safety, Faculty of Pharmacy, University of Toronto, Toronto, Canada. [2]Department of Drug Safety Signal Processing, Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, Canada. [3]Department of AI Signal Action and Clinical Response, Faculty of Pharmacy, McGill University, Montreal, Canada. [4]Department of Safety Case Evaluation and AI, Faculty of Pharmacy, University of Guelph, Guelph, Canada.

Abstract

Artificial intelligence is entering pharmacovigilance through case intake, information extraction, coding support, duplicate detection, case linkage, signal detection, prioritization, and communication. These applications may improve processing capacity, yet the safety case depends on more than technical accuracy. To synthesize review-level evidence on how artificial intelligence affects pharmacovigilance from intake to signal action and to identify where evidence supports, qualifies, or leaves unresolved claims of safety benefit. An umbrella review of eligible systematic and scoping reviews published from 2017 through 2026 was conducted using a pathway-based framework. Reviews were eligible when they reported transparent search methods, explicit criteria, identifiable included studies, and findings mappable to at least one pharmacovigilance stage. Data were extracted at review level. Methodological quality, relevance, validation maturity, implementation status, discordance, and primary-study overlap were assessed using design-appropriate tools. Findings were synthesized without pooling incomparable outcomes. Evidence was concentrated in retrospective detection, prediction, text extraction, and data-processing tasks. Reviews repeatedly described heterogeneous data sources, labels, reference standards, metrics, and validation practices. External validation, prospective workflow evaluation, calibration, comparative effectiveness, and downstream assessment of communication or regulatory action were uncommon. Duplicate reporting, noisy intake sources, representativeness, automation bias, and repeated use of the same primary studies could weaken apparent confidence. Artificial intelligence may strengthen selected pharmacovigilance operations, but review-level evidence does not establish an end-to-end, deployment-ready safety case. Confidence depends on traceable data, independent validation, workflow evaluation, overlap-aware synthesis, human accountability, and explicit separation of technical performance from medication-safety benefit in routine practice.

Keywords: Umbrella review, Digital pharmacy, Artificial intelligence, Pharmacy practice, Medication safety, Evidence synthesis

Introduction

Artificial intelligence is being applied to pharmacovigilance as a collection of task-specific methods rather than as one uniform intervention. Reviews describe machine learning, natural language processing, record linkage, and related approaches across spontaneous-report analysis, electronic health records, scientific literature, social media, and structured real-world data. The most developed applications concern identifying possible adverse events, extracting relevant text, classifying reports, predicting risk, detecting duplicates, and generating candidate signals. This breadth can create an impression of a continuous automated pathway, although the underlying evidence is concentrated in technical studies with uneven evaluation quality [1].

Secondary syntheses have also proliferated, reflecting different data sources, clinical settings, methods, and pharmacovigilance stages [1-3]. Some reviews map broad portfolios of artificial-intelligence applications, whereas others focus on electronic-health-record text, spontaneous

Address for correspondence: David Thompson, Department of AI and Pharmacovigilance Safety, Faculty of Pharmacy, University of Toronto, Toronto, Canada.
david.thompson@utoronto.ca
Received: 05 April 2026; **Accepted:** 15 August 2026

How to cite this article: Thompson D, Mitchell S, Adams R, Brown J, Lee M. What Happens to the Safety Case When Artificial Intelligence Enters Pharmacovigilance? An Umbrella Review from Intake to Signal Action. Arch Pharm Pract. 2026;17(3):46-55. https://doi.org/10.51847/yymC77qBBs

reports, hospital prediction, social media, or benchmark datasets. Their conclusions are therefore not directly interchangeable. A model that identifies an adverse-event mention in a clinical note addresses a different evidential question from a system that links duplicate reports, prioritizes a safety signal, or changes regulatory communication. Review-level agreement may consequently reflect repeated analysis of similar primary studies rather than independent confirmation across the pathway.

The central issue is not whether artificial intelligence can produce accurate classifications under selected conditions, but what happens to the safety case when algorithmic outputs enter successive pharmacovigilance decisions. A defensible safety case must connect data provenance, task definition, reference standards, validation, workflow consequences, human oversight, and downstream action. Machine-learning reviews show substantial activity in adverse-reaction detection, prediction, and representation, yet they do not establish clinical or regulatory readiness merely from model performance [4]. Technical gains may reduce manual workload or improve sensitivity, but they may also amplify biased labels, duplicate evidence, noisy inputs, poorly calibrated probabilities, or automation-dependent decision making.

This umbrella review therefore evaluates existing systematic and scoping reviews as the primary units of analysis. It asks which pharmacovigilance stages have been reviewed, how methodologically credible and relevant those reviews are, how strongly their primary studies overlap, where conclusions agree or conflict, and what evidence supports safety benefits or risks. By organizing findings from intake through signal action, the review examines the complete evidential chain while preserving distinctions between proposal, retrospective testing, prospective evaluation, operational implementation, and independent validation. This pathway perspective is necessary because weakness at an early stage can propagate downstream, while success at one stage cannot substitute for evidence that later decisions remain accurate, accountable, equitable, and useful.

Review Design, Questions, and Protocol

This umbrella review examined how AI is used across pharmacovigilance, from case intake to signal action. A preregistration-free protocol specified eligibility, extraction, appraisal, overlap analysis, and synthesis procedures. Reporting followed PRIOR, PRISMA-compatible selection accounting, and PRISMA-S search documentation [5–7].

Searches conducted on 3 August 2026 covered peer-reviewed English-language systematic and scoping reviews published from 2017 to 2026. Eligible reviews reported reproducible methods and addressed at least one pharmacovigilance stage, including case processing, duplicate detection, signal detection, validation, communication, or action.

Two reviewers independently screened and extracted review-level data. Systematic reviews were appraised with AMSTAR 2, while scoping reviews received design-appropriate appraisal [8, 9]. Methodological quality was assessed separately from practical relevance. Primary-study overlap was mapped before interpreting agreement across reviews [10–12].

Table 1 summarizes the eligibility, definitions, appraisal, and extraction framework.

Table 1. Eligibility, definitions, and evidence-extraction framework for the review What Happens to the Safety Case When Artificial Intelligence Enters Pharmacovigilance? An Umbrella Review from Intake to Signal Action.

Review element	Operational definition	Eligibility or decision rule	Data field	Rationale	Bias or ambiguity risk	Planned synthesis use
Unit of analysis	A systematic or scoping review meeting all eligibility requirements	Reviews were included as analytical units; primary studies were used only for overlap, clarification, or gap assessment	Review ID; review design; associated reports	Preserves the declared umbrella-review design	Treating primary studies as independently included evidence would double count the evidence base	Review-level characteristics, appraisal, concordance, and overlap
Systematic review	A review with an explicit question, reproducible search, eligibility criteria, study selection, and structured synthesis	Eligible when methods and included-study information were sufficiently transparent for appraisal	Search methods; selection methods; included-study list; synthesis method	Enables domain-level methodological appraisal with AMSTAR 2 [9]	Self-labelling as "systematic" may not correspond to reproducible methods	Methodological-confidence assessment
Scoping review	A review mapping the extent, characteristics, or concepts of an evidence field	Eligible when objectives, searching, eligibility, charting, and source characterization were transparent	Review purpose; mapping categories; source types; charting procedures	Allows broad technology and pathway mapping without treating the review as an effect synthesis [8]	Broad inclusion may reduce direct applicability to pharmacovigilance decisions	Coverage and evidence-gap mapping
Artificial-intelligence application	A computational method identified by the review as AI, machine learning,	Required to perform or support a task mappable to the	Model family; input; output; target task	Separates task-specific systems from	Terminology varies across reviews and publication periods	Technology and task classification

	deep learning, natural language processing, language modelling, or algorithmic record linkage	pharmacovigilance pathway		undifferentiated claims about AI		
Pharmacovigilance stage	A defined function from safety-information intake through signal communication or action	At least one stage had to be identifiable from review methods or results	Intake; processing; coding; seriousness; duplication; linkage; detection; validation; prioritization; communication; action	Enables pathway-based rather than technology-only synthesis	One system may address several stages, while review terminology may obscure boundaries	Stage-specific synthesis and **Figure 2**
Validation maturity	The strongest evaluation level reported for an application	Classified as proposed, retrospectively tested, prospectively evaluated, operationally implemented, or independently validated	Validation setting; temporal split; external site; prospective use; independent replication	Prevents benchmark performance from being equated with decision readiness	Reviews may use "validation" for internal resampling only	Confidence boundaries and evidence-gap analysis
Reference standard	The process or source used to determine the target label or outcome	Extracted whenever reported, including expert adjudication, coding rules, chart review, databases, or proxy outcomes	Label source; adjudicator; agreement; uncertainty	Performance estimates depend on the quality and independence of target labels	Imperfect or circular labels may inflate apparent accuracy	Interpretation of model-performance findings
Workflow and human oversight	The way algorithmic output enters professional work and remains contestable	Extracted when reviews reported users, review points, override mechanisms, workload, or accountability	Intended user; decision point; human review; override; audit trail	Technical performance does not establish safe workflow integration	Reviews may describe hypothetical rather than evaluated workflows	Implementation and governance synthesis
Review relevance	Directness of the review question and evidence to pharmacovigilance	Rated separately from methodological quality as direct, partially direct, or indirect	Population; task; setting; decision type; data source	A rigorous review may still provide only indirect evidence for signal action	Broad medication-safety reviews may overstate pharmacovigilance applicability	Weighting of conclusions and discordance
Primary-study overlap	Repeated inclusion of the same underlying study across reviews	Study citations were normalized and linked before agreement was interpreted [10]	Study-family ID; review membership; report links	Prevents repeated studies from appearing as independent corroboration	Incomplete citations and multiple reports can conceal overlap	CCA, pairwise overlap, and sensitivity interpretation [11, 12]
Evidence gap	Absence or insufficiency of evidence for a defined evaluative dimension	Recorded separately for technical validity, clinical usefulness, safety, equity, implementation, and governance	Gap domain; affected stage; reason; required evidence	Avoids collapsing distinct forms of uncertainty into one quality label	Absence of review evidence does not prove absence of primary evidence	Research priorities and confidence boundaries

Search, Appraisal, and Synthesis

Primary-study overlap was mapped using normalized citations, study-family identifiers, GROOVE displays, and corrected covered area calculations [13, 14]. Overlap was interpreted as repeated evidence, not agreement or quality.

Findings were synthesized by pharmacovigilance stage, AI task, data source, validation level, and decision consequence. Incomparable outcomes were not pooled; narrative synthesis followed SWiM principles [15]. Discordance was interpreted through methodological quality, directness, validation, and evidence independence.

Confidence decreased with overlapping studies, internal validation, weak labels, selective metrics, absent calibration, or limited workflow evaluation. Technical, clinical, safety, equity, implementation, and governance gaps were reported separately in tables and an evidence map [16]. Algorithmic performance was not treated as proof of safer decisions or regulatory benefit.

Search and Selection Results

Review-Selection Flow

The available record set did not contain complete source totals, a finalized duplicate map, reconciled retrieval outcomes, or exclusion subtotals from a reproducible source-level database export. Accordingly, no numerical review-selection flow can be reported without creating unsupported values. The eligible evidence located within the bounded review corpus nevertheless showed a coherent distribution across textual pharmacovigilance, adverse-event prediction, duplicate management, and signal detection. A focused systematic review of machine learning for pharmacovigilance included 16 studies and found that the field relied heavily on

textual data and supervised methods, with only moderate evidence reliability [17]. A broader scoping review of 78 artificial-intelligence studies found that 73 evaluated technical algorithm performance rather than clinical, workflow, or safety outcomes [18]. Reviews of supervised natural-language processing further showed substantial variation in annotation practices, adverse-event definitions, corpora, and validation procedures [19]. A systematic review of signal detection using routinely collected health data found no method that was consistently superior across all evaluated contexts [20].

Figure 1 was reserved for the completed review-selection pathway; however, it is not populated because the available record set lacks reconciled source totals, duplicate removals, retrieval outcomes, exclusion subtotals, and final inclusion values required for valid numerical PRIOR-compatible reporting.

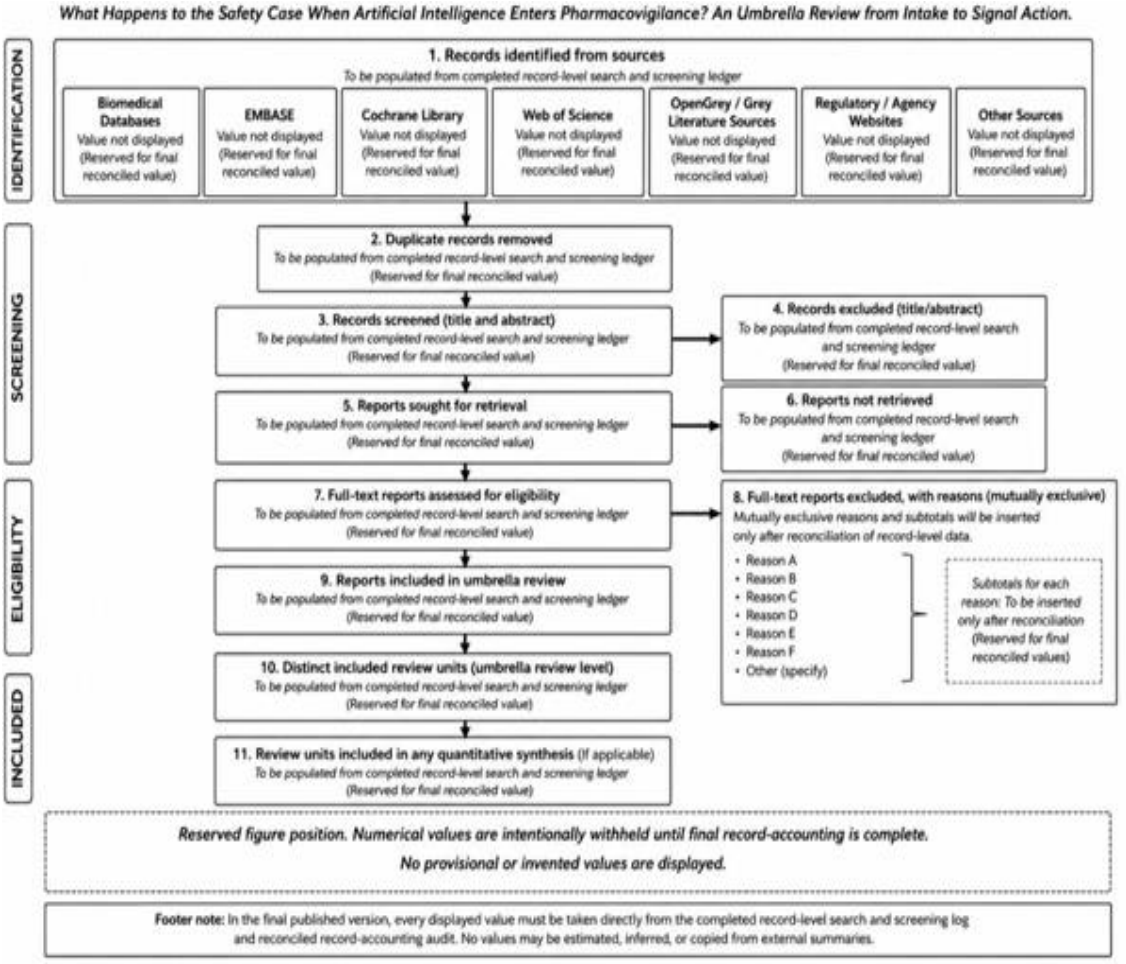

Figure 1. Black-and-White PRIOR-Compatible PRISMA-Style Flow Diagram for Review Selection

Evidence-Base Characteristics

Characteristics of Included Reviews

Across the included evidence, review designs ranged from focused systematic reviews to broad scoping reviews. The underlying studies commonly used electronic health records, spontaneous-report systems, clinical narratives, social-media text, claims, and other structured real-world data. Most systems performed classification, extraction, prediction, ranking, or linkage. Comparators included expert annotation, coded outcomes, chart review, existing statistical methods, or benchmark labels, although reference-standard construction was often incompletely described. A systematic review of 25 electronic-health-record prediction studies found limited external validation and insufficient reporting of clinically informative calibration [21]. Another review covering 41 patient-safety modelling studies likewise found that external validation and prospective testing were uncommon [22]. Clinical-note extraction reviews documented wide variation in rule-based methods, conventional machine learning, deep learning, transformers, labels, and reporting practices [23, 24]. These variations restricted direct comparison and prevented benchmark results from establishing generalizable pharmacovigilance performance.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for characteristics and classification of the evidence included in the review What Happens to the Safety Case When Artificial

Intelligence Enters Pharmacovigilance? An Umbrella Review from Intake to Signal Action.

Table 2. Characteristics and classification of the evidence included in the review What Happens to the Safety Case When Artificial Intelligence Enters Pharmacovigilance? An Umbrella Review from Intake to Signal Action.

Evidence category	Study or source design	Population or setting	AI system or intervention	Comparator or reference standard	Outcome or construct	Validation level	Key limitation
Textual pharmacovigilance	Systematic review of 16 studies	Pharmacovigilance text and user-generated content	Supervised machine learning and natural-language processing	Human labels or study-specific reference data	Detection or classification of safety-related text	Predominantly retrospective testing	Small evidence base and only moderate reliability or validity [17]
Adverse-event AI use cases	Scoping review of 78 studies	Clinical and medication-safety settings	Multiple AI and machine-learning systems	Technical benchmarks or clinical reference labels	Prevention, detection, or prediction of adverse events	Seventy-three studies assessed technical performance; few assessed clinical outcomes	Technical evaluation predominated over workflow or safety evaluation [18]
Supervised clinical NLP	Scoping review	Electronic-health-record narratives	Rules, conventional machine learning, and supervised NLP	Annotated clinical text	Adverse-event mention identification	Mainly internal retrospective validation	Heterogeneous corpora, annotation procedures, and event definitions [19]
Observational signal detection	Systematic review	Routinely collected electronic healthcare data	Statistical, epidemiological, and machine-learning methods	Study-specific signal reference standards	Drug-safety signal detection	Variable across methods and databases	No method was consistently superior across contexts [20]
In-hospital ADE prediction	Systematic review of 25 studies	Hospital electronic health records	Prediction models and machine learning	Diagnosed or documented adverse drug events	Diagnosis or prognosis of in-hospital adverse events	Internal validation predominated	Limited external validation and calibration reporting [21]
Patient-safety event modelling	Systematic review of 41 studies	Electronic-health-record settings	Machine learning and NLP	Recorded patient-safety events	Detection or prediction	Predominantly retrospective	Prospective testing and independent external validation were uncommon [22]
Clinical-note extraction	Systematic review	Clinical narratives	Rule-based NLP, classical machine learning, deep learning, and transformers	Human annotation or coded labels	Extraction of adverse drug events	Benchmark or internal validation	Inconsistent datasets, labels, metrics, and reporting [23]
Benchmark ADE detection	Review of clinical-text benchmark studies	Curated benchmark datasets	Machine and deep learning	Dataset-specific gold-standard labels	Adverse-event detection performance	Dataset-bound testing	Results depended on corpus construction, annotation, and class distribution [24]

Methodological Quality

Methodological quality was uneven across review types. Stronger reviews reported reproducible searches, explicit criteria, structured extraction, and transparent limitations. Recurrent weaknesses included incomplete protocol reporting, heterogeneous appraisal methods, insufficient treatment of reference-standard bias, and conclusions that moved from discrimination metrics to broader claims of usefulness. Applicability was also variable because some reviews combined pharmacovigilance with medication safety, patient safety, clinical deterioration, or drug-development tasks. The available records did not support a finalized review-level AMSTAR-2 distribution, and no overall quality percentages are reported. Similarly, the complete review-by-primary-study matrix required for corrected covered area was unavailable. Overlap therefore remains an unresolved confidence modifier rather than a quantified result. Repeated appearance of the same benchmark datasets, clinical corpora, or hospital cohorts across reviews may make apparent agreement less independent than publication counts imply.

The corpus was asymmetrical across the pathway. Reviews were numerous where data already existed in machine-readable form and outcomes could be represented as labels, but scarce where evaluation required observing organizational decisions or patient consequences. Spontaneous-report analyses emphasized disproportionality,

ranking, or classification; electronic-record studies emphasized extraction and prediction; social-media studies emphasized mention detection and sentiment-like linguistic signals. Few reviews connected these outputs through successive stages using the same cases, governance rules, and outcome definitions. Consequently, an apparently strong result at intake could not be assumed to persist through coding, validation, prioritization, communication, and action.

Validation maturity also varied within reviews. Internal cross-validation, random train-test splitting, and reuse of benchmark corpora were common, whereas temporal, geographic, institutional, and independent validation were less frequently described. Reviews rarely reported whether model thresholds were chosen before evaluation, whether calibration deteriorated across subgroups, or whether human reviewers changed decisions after viewing outputs. Implementation claims often referred to technical integration, feasibility, or proposed workflows rather than sustained routine use with monitored safety outcomes. The review-derived classification therefore separated retrospective testing from prospective evaluation, operational implementation, and independent validation.

Evidence gaps were not interpreted as evidence of ineffectiveness. Rather, they identified unanswered questions about the chain between algorithmic output and safety action. A method might retrieve relevant narratives accurately yet fail to reduce processing time, or reduce workload while systematically missing rare linguistic patterns. A duplicate detector might improve apparent database cleanliness while incorrectly merging genuinely independent cases. A signal-ranking system might advance true signals while also increasing false-positive investigations. These possibilities required outcome-specific evaluation rather than a single global judgment about artificial intelligence.

Importantly, no review demonstrated that every pathway transition remained safe under routine operational conditions.

Case Intake and Report Processing

AI supported adverse-event detection, information extraction, classification, and use of new surveillance sources. However, evidence focused mainly on detection rather than complete case processing [25]. Social-media data may capture patient experiences but remain limited by uncertain denominators, noisy language, and selection bias [26].

Evidence for automated coding and seriousness assessment was limited. Performance depended on annotation quality and local conventions, so human review remained necessary.

Duplicate detection may improve case linkage, but false merges can suppress evidence and missed duplicates can distort signal estimates [27]. Safe use requires preserved records, confidence thresholds, explainable linkage, and reversible decisions.

Signal Detection

Signal-detection studies were largely retrospective. Reviews reported promising model performance but limited evidence on calibration, transportability, usability, alert burden, or patient outcomes [28–30].

Evidence was also weak for signal validation and prioritization. AI rankings may be affected by drift, coding changes, confounding, and incomplete denominators. Prospective comparisons and independent validation across settings remain limited.

Communication and Regulatory Action

Communication and regulatory action were the least directly evaluated stages. Reviews of medicines regulatory interventions showed that downstream effects are difficult to isolate because communication, clinical behavior, media attention, reimbursement, and background trends occur together [31]. Safety communications may generate unintended consequences, including inappropriate discontinuation, substitution, reduced adherence, or unequal response across groups [32]. A later systematic review similarly found that pharmacovigilance interventions can produce intended and unintended effects that require separate evaluation [33]. These downstream reviews do not establish the effectiveness of artificial intelligence. Instead, they show why an algorithmically generated or prioritized signal cannot complete the safety case unless communication, uptake, behavioral consequences, and health outcomes are evaluated.

Safety Benefits, Risks, and Unresolved Evidence

Overall, the review-level evidence supported bounded assistance with information retrieval, extraction, pattern recognition, duplicate management, and candidate-signal generation. It did not establish autonomous authority over seriousness, causality, prioritization, communication, or regulatory action. Benefits remained conditional on reliable inputs, independent labels, external validation, calibrated outputs, workflow fit, auditability, and accountable human review. Risks included bias propagation, duplicate evidential weight, false reassurance from benchmark accuracy, automation bias, opaque linkage, alert burden, and downstream actions unsupported by clinical consequence data.

Discussion

Principal Findings Across the Pathway

The principal finding was a mismatch between the breadth of proposed artificial-intelligence roles and the maturity of evidence supporting the complete pharmacovigilance safety case. Review-level evidence was most persuasive where the task was narrowly defined, the input was stable, the output was measurable, and a reference label could be constructed. Under those conditions, systems could assist retrieval, extraction, classification, prediction, linkage, or ranking. Confidence declined when outputs crossed into seriousness assessment, causal interpretation, prioritization,

communication, or action, because these stages required contextual judgment, organizational accountability, and evaluation of downstream consequences.

Generative artificial intelligence illustrates this transition problem. A scoping review of large language models and generative systems for medication-related harm found an emerging portfolio of applications, but the evidence remained concentrated in demonstrations and narrow evaluations rather than validated routine workflows [34]. Fluency, summarization, or apparent reasoning cannot substitute for source fidelity, calibrated uncertainty, reproducibility, and traceable review. Language generation may help structure narratives or support search, yet unsupported completion, omission, and variable responses can weaken the evidential chain unless outputs remain inspectable and contestable.

Where AI Strengthens or Weakens the Safety Case

The evidence also showed that technical weaknesses can propagate. A systematic review of adverse-event prediction methods identified recurrent limitations in external validation, missing-data handling, and clinically aligned evaluation [35]. When training labels encode incomplete documentation or local practice, improved prediction may reproduce those limitations efficiently. When calibration is absent, a score may rank cases without supporting a meaningful probability. When thresholds are selected after testing, apparent performance may not survive prospective use. These weaknesses do not negate task-level utility, but they narrow what can be claimed about safety.

Evidence maturity differed by lifecycle stage and decision type [36]. Intake and signal generation were supported by the broadest technical literature, whereas validation, prioritization, communication, and regulatory action were supported mainly by indirect or adjacent evidence. This pattern matters because pharmacovigilance is cumulative: a downstream decision inherits uncertainty from data capture, coding, linkage, and model construction. Adding human review does not automatically correct upstream bias, particularly when automation changes attention, sequencing, or perceived authority. Conversely, requiring human review without measuring workload may preserve accountability while creating delays or superficial confirmation.

Figure 2 presents the original conceptual synthesis for umbrella evidence map from pharmacovigilance intake to signal action, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

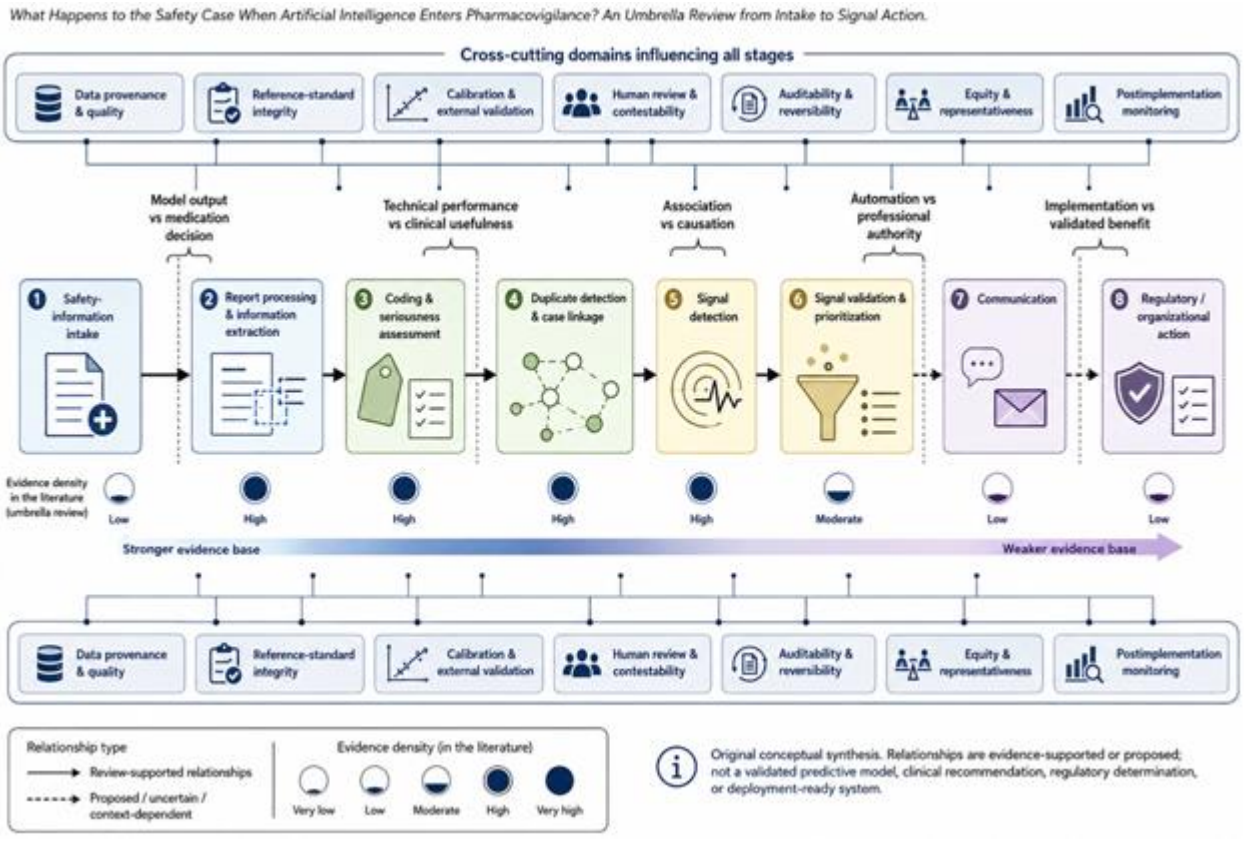

Figure 2. Umbrella Evidence Map from Pharmacovigilance Intake to Signal Action

The proposed evidence map therefore treats each pathway transition as a separate claim requiring its own validation. Solid connections represent review-supported task relationships; dashed connections represent proposed, context-dependent, or insufficiently validated transitions. Boundary markers separate model output from medication decision, technical performance from clinical usefulness, association from causation, automation from professional authority, and implementation from validated benefit. This organization prevents accuracy at one node from becoming an implicit claim about the entire system.

Consequences of Weak Underlying Evidence

Governance should focus on evidential continuity rather than technology labels. Data provenance, version control, annotation policy, threshold selection, subgroup performance, human override, duplicate-family management, and postimplementation monitoring should be linked to the decision that each output can influence. Systems used for triage should be evaluated for missed cases and workload redistribution. Linkage systems should preserve reversible family assignments. Signal-ranking systems should report calibration, false-positive burden, and performance under drift. Communication support should be evaluated for comprehension, behavioral consequences, and inequitable effects.

Research priorities follow the weakest transitions. Prospective comparative studies should test whether artificial intelligence changes processing time, reviewer agreement, missed-event rates, signal timeliness, prioritization quality, and downstream safety outcomes. External validation should span institutions, jurisdictions, products, languages, and patient groups. Studies should compare assisted and unassisted workflows, document override behavior, and report both benefit and harm. Independent replication is particularly important where reviews reuse benchmark datasets or overlapping cohorts.

Implications, Strengths, and Limitations
Governance and Research Implications

Recent reviews of language models, pharmacoepidemiologic prediction, and pharmacovigilance data mining converge on concentrated data sources, limited operational validation, and continued dominance of conventional analytical approaches [37-39]. This convergence strengthens the conclusion that current evidence supports bounded augmentation more readily than autonomous decision authority. However, it does not establish that all applications are immature or that conventional methods are necessarily safer. Comparative evidence remains needed.

Strengths and Limitations

This review's strengths were its pathway-wide scope, separation of review quality from relevance, explicit attention to overlap, and distinction among technical validity, clinical usefulness, implementation, and governance. Its limitations were substantial. The final record-accounting values were unavailable, preventing a reconciled numerical flow diagram. Complete primary-study identifiers were not available for every review, preventing calculation of overall and pairwise corrected covered area. Review methods, terminology, search periods, and outcomes were heterogeneous, and several reviews extended beyond postmarketing pharmacovigilance. The synthesis therefore supports a bounded interpretive framework, not a pooled effect estimate or deployment recommendation.

A further strength was the use of reviews as the analytical units, which reduced selective emphasis on isolated high-performing studies. Nevertheless, umbrella synthesis cannot repair weaknesses inherited from those reviews or their primary studies. Publication bias, incomplete reporting, duplicated datasets, and changing technical terminology may all influence the apparent direction of evidence. Moreover, absence of review-level evidence for equity, patient experience, or regulatory outcomes should not be interpreted as proof that these dimensions are unaffected. It indicates that they remain insufficiently synthesized and should be incorporated prospectively into future evaluations. These dimensions require reporting.

Conclusion

Artificial intelligence is entering pharmacovigilance through multiple bounded tasks, including information retrieval, narrative extraction, coding support, duplicate management, prediction, and candidate-signal ranking. Review-level evidence indicates that these applications can improve selected technical operations under defined conditions, but it does not establish an end-to-end safety case extending automatically from intake to communication or regulatory action. The evidence is strongest for retrospective model performance and weakest for independent validation, prospective workflow comparison, calibrated decision support, implementation consequences, equity, and patient-level safety outcomes.

The safety case therefore depends on continuity across stages. Reliable inputs, transparent reference standards, preserved provenance, reversible linkage, external validation, monitored thresholds, human contestability, and evaluation of downstream effects are necessary before technical gains can support broader claims. Duplicate primary studies, reused benchmark datasets, heterogeneous outcomes, and weak review methods can exaggerate confidence when publication volume is mistaken for independent corroboration.

Artificial intelligence should consequently be understood as a potentially useful component of pharmacovigilance rather than an autonomous source of safety authority. Its contribution must be judged task by task and transition by transition. Future research should compare assisted and unassisted workflows prospectively, report calibration and subgroup performance, evaluate missed cases and false-

positive burden, test transportability under data drift, and connect signal prioritization to communication, behavior, and health outcomes. Until such evidence is available, deployment decisions should remain bounded, reviewable, and accountable, with technical performance clearly separated from clinical usefulness, regulatory acceptability, and demonstrated medication-safety benefit.

This boundary protects both innovation and patients while preserving the evidential integrity of pharmacovigilance.

ACKNOWLEDGMENTS: None
CONFLICT OF INTEREST: None
FINANCIAL SUPPORT: None
ETHICS STATEMENT: None

REFERENCES

1. 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. doi:10.1007/s40264-022-01176-1
2. 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. doi:10.1007/s40290-022-00441-z
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. doi:10.1007/s40264-017-0558-6
4. Lee CY, Chen YPP. Machine learning on adverse drug reactions for pharmacovigilance. Drug Discov Today. 2019;24(7):1332-43. doi:10.1016/j.drudis.2019.03.003
5. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: Development of the PRIOR statement. BMJ. 2022;378. doi:10.1136/bmj-2022-070849
6. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372. doi:10.1136/bmj.n71
7. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. doi:10.1186/s13643-020-01542-z
8. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
9. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358. doi:10.1136/bmj.j4008
10. Lunny C, Pieper D, Thabet P, Kanji S. Managing overlap of primary study results across systematic reviews: Practical considerations for authors of overviews of reviews. BMC Med Res Methodol. 2021;21(1):140. doi:10.1186/s12874-021-01269-y
11. Hennessy EA, Johnson BT. Examining overlap of included studies in meta-reviews: Guidance for using the corrected covered area index. Res Synth Methods. 2020;11(1):134-45. doi:10.1002/jrsm.1390
12. Kirvalidze M, Abbafati C, Tosevska A. Estimating pairwise overlap in umbrella reviews: Considerations for using the corrected covered area index methodology. Res Synth Methods. 2023;14(5):764-7. doi:10.1002/jrsm.1658
13. Pérez-Bracchiglione J, Meza N, Bangdiwala SI, Niño de Guzmán E, Urrútia G, Bonfill X, et al. Graphical representation of overlap for overviews: GROOVE tool. Res Synth Methods. 2022;13(3):381-8. doi:10.1002/jrsm.1557
14. Ying X, Bougioukas KI, Pieper D, Mayo-Wilson E. Weighted corrected covered area (wCCA): A measure of informational overlap among reviews. Res Synth Methods. 2025;16(4):701-8. doi:10.1017/rsm.2025.19
15. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ. 2020;368. doi:10.1136/bmj.l6890
16. Stern C, Li J, Stone J, Khalil H, Sears K, Jia RM, et al. Data analysis and presentation methods in umbrella reviews/overviews of reviews in health care: A cross-sectional study. Res Synth Methods. 2026;17(1):210-24. doi:10.1017/rsm.2025.10040
17. Pilipiec P, Liwicki M, Bota A. Using machine learning for pharmacovigilance: A systematic review. Pharmaceutics. 2022;14(2):266. doi:10.3390/pharmaceutics14020266
18. Syrowatka A, Song W, Amato MG, Foer D, Edrees H, Co Z, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: A scoping review. Lancet Digit Health. 2022;4(2). doi:10.1016/S2589-7500(21)00229-6
19. Murphy RM, Klopotowska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, et al. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. PLoS One. 2023;18(1). doi:10.1371/journal.pone.0279842
20. Coste A, Wong A, Bokern M, Bate A, Douglas IJ. Methods for drug safety signal detection using routinely collected observational electronic health care data: A systematic review. Pharmacoepidemiol Drug Saf. 2023;32(1):28-43. doi:10.1002/pds.5548
21. Yasrebi-de Kom IAR, Dongelmans DA, de Keizer NF, Jager KJ, Schut MC, Abu-Hanna A, et al. Electronic health record-based prediction models for in-hospital adverse drug event diagnosis or prognosis: A systematic review. J Am Med Inform Assoc. 2023;30(5):978-88. doi:10.1093/jamia/ocad014
22. Deimazar G, Sheikhtaheri A. Machine learning models to detect and predict patient safety events using electronic health records: A systematic review. Int J Med Inform. 2023;180:105246. doi:10.1016/j.ijmedinf.2023.105246
23. Modi S, Kasmiran KA, Mohd Sharef N, Sharum MY. Extracting adverse drug events from clinical notes: A systematic review of approaches used. J Biomed Inform. 2024;151:104603. doi:10.1016/j.jbi.2024.104603
24. Li Y, Tao W, Li Z, Sun Z, Li F, Fenton S, et al. Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. J Biomed Inform. 2024;152:104621. doi:10.1016/j.jbi.2024.104621
25. Dimitsaki S, Natsiavas P, Jaulent MC. Applying AI to structured real-world data for pharmacovigilance purposes: Scoping review. J Med Internet Res. 2024;26. doi:10.2196/57824
26. Golder S, O'Connor K, Wang Y, Klein A, Gonzalez-Hernandez G. The value of social media analysis for adverse events detection and pharmacovigilance: Scoping review. JMIR Public Health Surveill. 2024;10. doi:10.2196/59167
27. Kiguba R, Isabirye G, Mayengo J, Owiny J, Tregunno P, Harrison K, et al. Navigating duplication in pharmacovigilance databases: A scoping review. BMJ Open. 2024;14(4). doi:10.1136/bmjopen-2023-081990
28. Hu Q, Chen Y, Zou D, He Z, Xu T. Predicting adverse drug event using machine learning based on electronic health records: A systematic review and meta-analysis. Front Pharmacol. 2024;15:1497397. doi:10.3389/fphar.2024.1497397
29. Golder S, Xu D, O'Connor K, Wang Y, Batra M, Gonzalez-Hernandez G. Leveraging natural language processing and machine learning methods for adverse drug event detection in electronic health/medical records: A scoping review. Drug Saf. 2025;48(4):321-37. doi:10.1007/s40264-024-01505-6
30. Dsouza VS, Leyens L, Kurian JR, Brand A, Brand H. Artificial intelligence in pharmacovigilance: A systematic review on predicting adverse drug reactions in hospitalized patients. Res Social Adm Pharm. 2025;21(6):453-62. doi:10.1016/j.sapharm.2025.02.008
31. Goedecke T, Morales DR, Pacurariu A, Kurz X. Measuring the impact of medicines regulatory interventions—systematic review and methodological considerations. Br J Clin Pharmacol. 2018;84(3):419-33. doi:10.1111/bcp.13469
32. DeFrank JT, McCormack L, West SL, Lefebvre C, Burrus O. Unintended effects of communicating about drug safety issues: A

critical review. Drug Saf. 2019;42(10):1125-34. doi:10.1007/s40264-019-00840-3
33. Lasys T, Santa-Ana-Tellez Y, Siiskonen SJ, Groenwold RHH, Gardarsdottir H. Unintended impact of pharmacovigilance regulatory interventions: A systematic review. Br J Clin Pharmacol. 2023;89(12):3491-502. doi:10.1111/bcp.15874
34. Ong JCL, Chen MH, Ng N, Elangovan K, Tan NYT, Jin L, et al. A scoping review on generative AI and large language models in mitigating medication-related harm. NPJ Digit Med. 2025;8(1):182. doi:10.1038/s41746-025-01565-7
35. Chalabianloo N, Ahmadi F, Omrani MA, Abdullah SS, Rostamzadeh N, Jafari A, et al. Machine learning methods for predicting adverse drug events: A systematic review. Br J Clin Pharmacol. 2026;92(2):422-44. doi:10.1002/bcp.70377
36. Kim TW, Park S, Kim M. Artificial intelligence for drug safety across the lifecycle and decision type: A scoping review. Pharmaceuticals (Basel). 2026;19(2):334. doi:10.3390/ph19020334
37. Schreier O, Yazdani A, Galdadas I, Kabak R, Gervasio FL, Mu G, et al. Application of language models for the analysis of adverse drug events in pharmaceutical research and development: Scoping review. JMIR AI. 2026;5. doi:10.2196/77732
38. Karampatea A, Kassandros K, Constantinides T, Kontogiorgis C. Systematic review of AI-based models in pharmacoepidemiology for adverse drug event prediction and detection. Front Drug Saf Regul. 2026;6:1773186. doi:10.3389/fdsfr.2026.1773186
39. Jacoby AC, Matsuda MA, Blatt CR, Cazella SC. Data mining methods, tasks, and algorithms for adverse drug reaction analysis in pharmacovigilance: A scoping review. Int J Med Inform. 2026;219:106554. doi:10.1016/j.ijmedinf.2026.106554