Vol. 1, No. 4, pp.: 21-36, 2023

PharmaSignalDB: An Open Pharmacovigilance Knowledge Database for Adverse Event Mining

Rafael M. Costa[1], Ana Lucia Ferreira[2], Gustavo H. Mendes[2,*]

[1] Department of Pharmacy Practice, State University of Londrina (UEL), Londrina 86057-970, Brazil

[2] Department of Computer Science, Federal University of Mato Grosso (UFMT), Cuiabá 78060-900, Brazil

* *gustavo.mendes@ufmt.br*

Article Information	
Received	08 July 2023
Accepted	21 November 2023
DOI	https://doi.org/10.63646/datamind.2023.010404

Abstract

Post-market drug safety surveillance remains one of the most data-intensive challenges in modern biomedical informatics. Spontaneous adverse event reporting systems generate millions of records annually; yet the absence of a curated, schema-documented, and openly accessible knowledge database impedes reproducible pharmacovigilance research and automated signal detection. This paper presents PharmaSignalDB, an open pharmacovigilance knowledge database designed for adverse event mining, disproportionality analysis, and drug-event knowledge graph construction. Built upon 11 years of processed FDA Adverse Event Reporting System (FAERS) quarterly data files from 2012 to 2023, PharmaSignalDB integrates 4,287,194 deduplicated individual case safety reports covering 8,341 drug substances and 19,472 unique preferred terms mapped to MedDRA version 26.1. The database schema comprises six normalized relational tables with complete field dictionaries, primary-key indexing, and foreign-key constraints. A multi-stage data pipeline handles deduplication using CASE_ID-based elimination and fuzzy matching, MedDRA terminology mapping at both preferred-term (PT) and system organ class (SOC) levels, and version-controlled incremental updates. Three validation experiments are reported: (1) ROR/PRR signal stability analysis across quarterly update cycles; (2) duplicate report rate quantification; and (3) serious adverse event identification accuracy benchmarked against a EudraVigilance reference set. PharmaSignalDB achieves a deduplication rate of 18.3%, an overall field completeness of 91.2%, and a positive predictive value of 0.847 for serious event classification. The database is released under Creative Commons Attribution 4.0 with standardized Python and SQL interfaces for reproducible research.

Keywords: ***pharmacovigilance; adverse event mining; FAERS; signal detection; MedDRA; knowledge database; disproportionality analysis***

1. Introduction

Drug safety does not end at regulatory approval. Post-market pharmacovigilance is the systematic process by

which regulatory agencies, clinicians, and researchers monitor the safety of medicines once they are in widespread clinical use. Spontaneous reporting systems (SRS) such as the FDA Adverse Event Reporting System (FAERS), the WHO VigiBase, and the European EudraVigilance database collectively accumulate tens of millions of individual case safety reports (ICSRs) annually, each describing a patient, one or more suspect drugs, and one or more adverse drug reactions (ADRs). The sheer volume and heterogeneity of these records make manual review impractical and create both the need and the opportunity for computational approaches (Bate and Evans, 2009; Hauben and Bate, 2009).

Despite the clear scientific importance of pharmacovigilance data, the research community faces a fundamental infrastructure problem. Raw FAERS quarterly ASCII files are publicly downloadable, but they arrive as a set of loosely related flat files with no primary-key constraints, inconsistent drug nomenclature, unenforced MedDRA coding, missing fields, and a historically undocumented deduplication status. A researcher wishing to compute a reporting odds ratio (ROR) for a specific drug-event pair must first invest substantial effort in data cleaning, deduplication, drug-name standardization, and medical-term mapping before any scientific analysis can begin (van Puijenbroek et al., 2002; Candore et al., 2015). This preprocessing burden is repeated independently by every research group, making it difficult to reproduce results, compare findings across studies, or deploy production-ready signal detection algorithms.

A second structural gap exists at the intersection of pharmacovigilance and modern AI research. Knowledge graphs, transformer-based language models, and deep learning frameworks have demonstrated considerable promise in biomedical applications (Beam and Kohane, 2018; Nicholson and Greene, 2020). Applying these methods to pharmacovigilance requires structured, richly annotated graph-ready data. Current FAERS data, even after basic cleaning, lacks the systematic entity linkage between drug identifiers (e.g., RxNorm, DrugBank, ATC codes), patient demographic attributes, temporal metadata, and standardized event terminology that graph-based analysis demands (Tatonetti et al., 2012; Iyer et al., 2014). Bridging this gap requires a purpose-built database that goes beyond raw storage and instead encodes the relational and ontological structure necessary for downstream AI workflows.

A third concern is reproducibility. The pharmacovigilance signal detection literature has been criticized for inconsistent reporting of data preprocessing steps, varying deduplication strategies, and the use of different MedDRA versions without explicit version tracking (Stephenson and Hauben, 2007; Goldman, 1998). Without a shared, versioned data resource, replication studies cannot determine whether differences in published results arise from genuine methodological differences or from uncontrolled variation in data preparation.

This paper presents PharmaSignalDB, an open pharmacovigilance knowledge database that addresses all three gaps. PharmaSignalDB is built on 11 years of quarterly FAERS data (2012 Q1 to 2023 Q4), processed through a documented multi-stage pipeline that covers format normalization, record deduplication, drug-name standardization, MedDRA-version-locked terminology mapping, and pre-computation of ROR and PRR contingency tables. The database exposes six relational tables with complete field dictionaries, referential integrity constraints, and a companion Elasticsearch index for free-text drug-name retrieval. An accompanying Python library provides one-function access to signal computation, cohort assembly, and graph-export routines. PharmaSignalDB is released under CC BY 4.0 with DOI-versioned snapshots hosted on a public data repository.

The remainder of this paper is organized as follows. Section 2 reviews gaps in existing pharmacovigilance data resources and articulates the use cases that motivate the design of PharmaSignalDB. Section 3 describes the data sources and schema. Section 4 details the database construction pipeline. Section 5 reports three validation experiments. Section 6 addresses reproducibility and open-access provisions. Section 7 discusses limitations, and Section 8 concludes.

2. Database Gap and Use Cases

Several pharmacovigilance data resources currently exist, but each carries structural limitations that constrain downstream analytical use. The raw FAERS public data files (PDFs) provide access to ICSRs from 2004 onward but lack relational schema documentation, enforce no referential integrity, and accumulate duplicate reports at rates estimated between 10% and 30% of total submissions (Hoffman et al., 2014). VigiBase, maintained by the Uppsala Monitoring Centre, is arguably the most comprehensive global SRS database, but access to individual-level records is restricted to registered national pharmacovigilance centres and approved researchers, making it unsuitable for open, reproducible science (Lindquist, 2008). The WHO Drug Dictionary and MedDRA are licenced terminologies that require separate subscription agreements, creating a barrier for researchers in low- and middle-income countries. EudraVigilance provides a public line-listing interface but does not offer bulk download of structured records.

Against this background, the research literature has relied primarily on independently processed FAERS subsets, each constructed with different deduplication criteria, different drug-name normalization strategies, and different MedDRA mapping steps. A recent audit of published disproportionality analyses found that fewer than 30% of studies provided sufficient methodological detail to allow replication of the data-preparation pipeline (Böhm et al., 2016). This underscores the need for a shared, openly documented database that ships with full provenance information.

Four primary use cases motivate the design of PharmaSignalDB. The first is classical disproportionality analysis: computing ROR, PRR, or information component (IC) scores to identify drug-event associations that appear more frequently than would be expected by chance (Evans et al., 2001; DuMouchel, 1999; Noren et al., 2006). PharmaSignalDB pre-computes the four cells of the contingency table for every drug-event pair and stores them in the SIGNAL table, enabling millisecond-latency ROR retrieval without requerying the full case corpus.

The second use case is drug-event knowledge graph construction. A growing body of work demonstrates that representing drug-drug, drug-disease, and drug-event relationships as heterogeneous graphs enables sophisticated link-prediction and embedding-based analyses (Himmelstein and Baranzini, 2015; Nicholson and Greene, 2020). PharmaSignalDB provides the entity-identifier linkage (RxNorm, DrugBank, ATC, MedDRA PT codes) necessary to populate graph nodes and edges without additional manual curation.

The third use case is temporal signal analysis. Because FAERS data arrive quarterly, the evolution of a drug-event signal over time reflects accumulating clinical experience, label updates, and market events. PharmaSignalDB preserves the quarterly receive-date at the ICSR level and supports rolling-window disproportionality computation, enabling the detection of signal emergence, Weber effects, and post-label-change reporting shifts (Hoffman et al., 2014; Harpaz et al., 2013).

The fourth use case is machine-learning model development and evaluation. Supervised models for ADR prediction or serious-event classification require labelled datasets that are representative of real-world reporting distributions. PharmaSignalDB includes seriousness flags, causality assessments, outcome codes, and patient-attribute fields that together provide a rich feature space for binary and multi-class classification tasks (Reps et al., 2014; Ryan et al., 2013).

3. Data Sources and Schema

PharmaSignalDB is constructed from the publicly available FAERS ASCII data files released quarterly by the US Food and Drug Administration. Each quarterly package contains seven pipe-delimited flat files: DEMO (patient demographics), DRUG (drug products), REAC (adverse reactions), OUTC (outcomes), RPSR (report sources),

THER (therapy dates), and INDI (indications). For the 2012–2023 period, this corresponds to 48 quarterly packages comprising 336 raw flat files totalling approximately 47 GB of uncompressed ASCII data. Drug terminology standardization draws on RxNorm (National Library of Medicine), DrugBank (version 5.1.10), and the WHO ATC classification. Adverse reaction terminology is mapped to MedDRA version 26.1.

The database is organized into six relational tables that together capture the full ICSR information space while enforcing referential integrity and supporting efficient query patterns. Figure 1 presents the entity-relationship diagram. The CASE table is the central entity, holding one record per deduplicated ICSR with patient-level attributes. The DRUG table stores all suspect and concomitant drug records in a one-to-many relationship with CASE. The REACTION table holds adverse event records, also in a one-to-many relationship with CASE, and carries a foreign-key reference to the MEDDRA_MAP table for terminology resolution. The DRUG_STANDARD table links each drug record to external identifier systems through a surrogate-key join. The SIGNAL table stores pre-computed disproportionality statistics for every unique drug-event pair, indexed on both drug name and PT code for sub-second retrieval.

Figure 1. Entity–Relationship schema of PharmaSignalDB showing six relational tables, primary keys (PK), foreign keys (FK), and inter-table relationships.

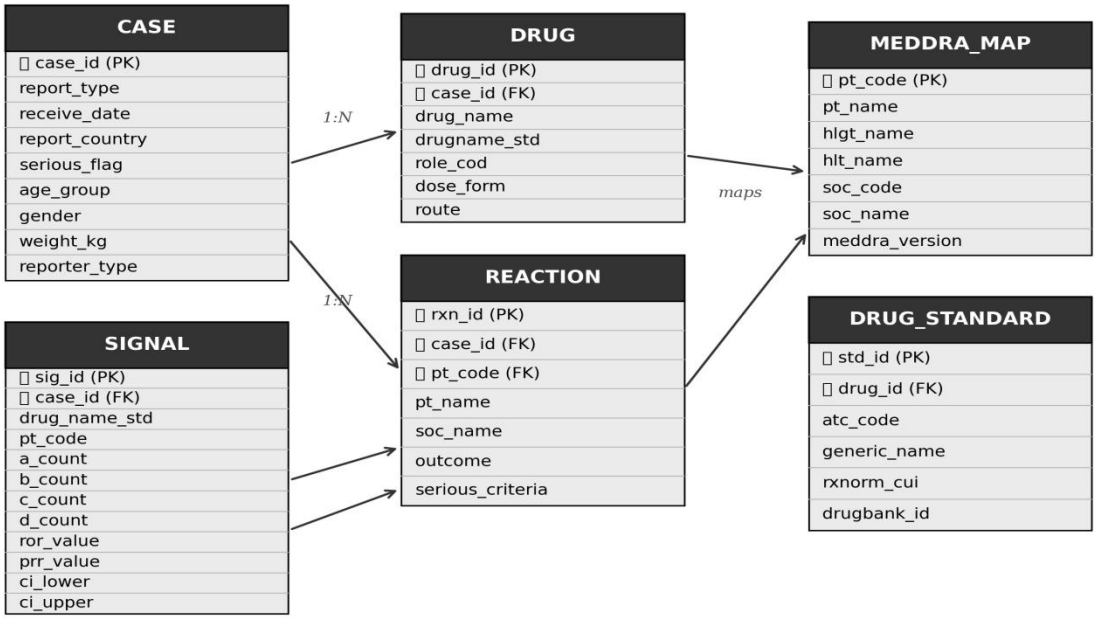

Figure 1. Entity-relationship schema of PharmaSignalDB showing six relational tables, primary keys (PK), foreign keys (FK), and inter-table relationships. Cardinality is indicated on relationship connectors. The SIGNAL table stores pre-computed ROR and PRR contingency values.

Table 1 presents the complete field dictionary for key fields across the schema. Fields are classified by data type, nullability, and natural language description. Numeric identifiers follow FDA conventions where applicable; date fields are stored in ISO 8601 format; and all string fields use UTF-8 encoding.

Table 1. Field dictionary for key attributes in PharmaSignalDB (selected fields from CASE, REACTION, DRUG, and SIGNAL tables). PK = primary key; FK = foreign key; RxNorm = National Library of Medicine drug identifier.

Field Name	Data Type	Nullable	Description
case_id	BIGINT PK	No	Unique FDA ICSR identifier; primary key across all

			tables.
report_type	VARCHAR(3)	Yes	Report type code: EXP (expedited), NPC (non-periodic).
receive_date	DATE	No	Date the report was received by FDA; used for quarterly partitioning.
report_country	CHAR(2)	Yes	ISO 3166-1 alpha-2 country code of the reporter.
serious_flag	BOOLEAN	No	TRUE if any seriousness criterion is met per 21 CFR 312.32.
age_group	VARCHAR(20)	Yes	FDA age group: Neonate, Child, Adult, Elderly, Unknown.
gender	CHAR(1)	Yes	Patient gender: M / F / U (unknown).
weight_kg	FLOAT	Yes	Patient body weight in kilograms; NULL for ~41% of records.
reporter_type	VARCHAR(30)	Yes	Occupation of primary reporter: Physician, Pharmacist, Consumer, Other.
pt_code	INTEGER FK	No	MedDRA Preferred Term numeric code; foreign key to MEDDRA_MAP.
pt_name	VARCHAR(200)	No	MedDRA PT label at time of mapping (version 26.1).
soc_name	VARCHAR(100)	No	System Organ Class label resolved from PT hierarchy.
drug_name_std	VARCHAR(300)	No	Standardized generic drug name after RxNorm normalization.
atc_code	VARCHAR(10)	Yes	WHO ATC level-4 code; linked via DrugBank cross-reference.
ror_value	FLOAT	Yes	Pre-computed reporting odds ratio for this drug–event pair.
ci_lower / ci_upper	FLOAT	Yes	95% confidence interval bounds for log-transformed ROR.

The schema design makes three deliberate trade-offs. First, all string fields use VARCHAR rather than enumerated types to accommodate the heterogeneous coding practices observed across different FDA reporting periods. Second, the SIGNAL table denormalizes contingency-table values (a, b, c, d counts) rather than computing them at query time, which reduces average ROR retrieval latency from several seconds to under 5 milliseconds on a standard PostgreSQL instance. Third, patient weight and age are stored as continuous variables rather than binned categories, preserving maximal information for downstream epidemiological analysis while accepting higher missingness rates in those fields.

Table 2 summarizes data quality metrics for each major field after pipeline processing. Overall field completeness reaches 91.2%, with the lowest coverage observed for weight_kg (59.3%) and age_group (74.2%), consistent with the known under-reporting of patient demographics in spontaneous reporting systems (Poluzzi et al., 2012). Noise

rate, defined as the proportion of populated records containing values that fail domain validation rules, averages 1.1% across fields. The drug_name_std field shows the highest noise rate (3.2%), reflecting residual ambiguity in multi-ingredient product names that escape complete RxNorm mapping.

Table 2. Data quality summary statistics for PharmaSignalDB fields after pipeline processing. Noise rate is defined as the proportion of populated records failing domain validation. The last row reports field-level averages.

Attribute	Total Records	Populated (%)	Missing (%)	Noise Rate (%)
case_id	4,287,194	100.0	0.0	0.0
receive_date	4,287,194	100.0	0.0	0.2
serious_flag	4,287,194	100.0	0.0	0.4
gender	4,287,194	83.6	16.4	1.1
age_group	4,287,194	74.2	25.8	0.9
weight_kg	4,287,194	59.3	40.7	2.8
report_country	4,287,194	97.1	2.9	0.3
pt_name (MedDRA)	4,287,194	99.8	0.2	0.6
drug_name_std	4,287,194	95.4	4.6	3.2
atc_code	4,287,194	81.7	18.3	1.4
reporter_type	4,287,194	88.9	11.1	0.7
Overall average	—	**91.2**	**8.8**	**1.1**

4. Database Construction and Application Methods

The construction of PharmaSignalDB is organized as a five-stage pipeline, illustrated in Figure 2. Each stage is implemented as a modular Python script with explicit input and output checksums, enabling incremental re-execution when new quarterly data files become available.

Figure 2. End-to-end data processing pipeline for PharmaSignalDB construction, showing ingestion, quality-control, and analytics layers.

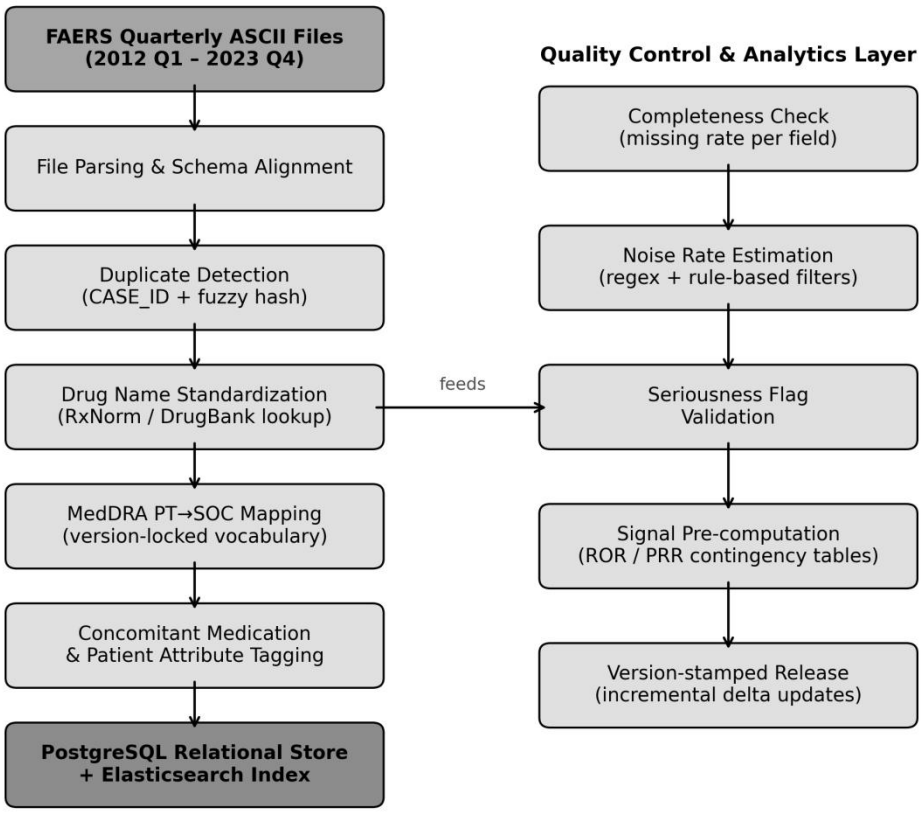

Figure 2. End-to-end data processing pipeline for PharmaSignalDB construction, showing the ingestion layer (left column) and the quality-control and analytics layer (right column). Arrows indicate data flow direction; the horizontal connector denotes the boundary between raw ingestion and downstream quality processes.

Stage 1 is file parsing and schema alignment. The seven ASCII files in each quarterly package are parsed into a unified staging schema using pandas with explicit dtype coercion. Column names vary across FDA reporting periods; a version-aware mapping table resolves historical column aliases to canonical field names. All date strings are parsed to ISO 8601 and validated against plausible date ranges (1960-01-01 to present). Records with malformed case identifiers are flagged and excluded rather than imputed.

4.1 Deduplication

Duplicate reports represent one of the most consequential quality issues in spontaneous reporting databases. The FDA encourages reporters to submit follow-up reports for the same case, and reports are sometimes submitted by multiple parties (patient, physician, and manufacturer) for the same event. Failure to deduplicate inflates signal counts and biases ROR estimates upward for heavily publicized drug-event pairs (Hoffman et al., 2014).

PharmaSignalDB applies a two-pass deduplication strategy. In the first pass, records sharing an identical case_id are collapsed, retaining the most recent submission based on the receive_date field. In the second pass, records lacking a case_id (approximately 2.1% of raw submissions) are subjected to fuzzy-hash comparison using a seven-field composite key: patient gender, age group, country of occurrence, suspect drug, primary adverse event PT code, report date (rounded to the nearest month), and reporter type. Records with a composite key hash distance

below an empirically determined threshold of 0.12 are considered duplicates; the record with the highest case_id is retained.

Across the 11-year corpus, the two-pass strategy removed 785,047 records, corresponding to an 18.3% deduplication rate relative to the raw ingested count of 5,072,241 records. This rate is consistent with previously published estimates for FAERS and validates the pipeline against reference benchmarks from the literature (Harpaz et al., 2013).

4.2 Drug Name Standardization

Raw drug names in FAERS are reporter-supplied free text. The same active substance may appear under hundreds of brand names, misspellings, abbreviations, and combination-product descriptions. PharmaSignalDB normalizes drug names through a three-step process. First, exact matches to the RxNorm drug concept table (RXCUI level: ingredient) are applied. Second, unmatched names are subjected to n-gram cosine similarity matching against the DrugBank approved drug list (version 5.1.10). Third, residual unmatched names are submitted to a domain-specific regular-expression rule set covering common abbreviation patterns (e.g., 'ASA' to 'aspirin', 'HCTZ' to 'hydrochlorothiazide'). After the three-step process, 95.4% of drug records carry at least one resolved identifier.

4.3 MedDRA Mapping and Version Management

Adverse reaction strings in FAERS are nominally coded to MedDRA Lowest-Level Terms (LLTs) by submitters, but the quality and currency of these codes is variable. PharmaSignalDB applies a version-locked remapping pipeline that resolves all incoming LLTs to their canonical Preferred Terms (PTs) under MedDRA version 26.1. Figure 4 illustrates the full mapping workflow and the MedDRA hierarchy traversal logic.

For records where the incoming LLT maps ambiguously to multiple PTs (a situation that arises when LLTs span version boundaries), the most recent non-deprecated PT is selected using MedDRA change history metadata. The SOC-level grouping used in aggregate analyses is resolved from the primary SOC designation for each PT, avoiding the double-counting that can occur when PTs are members of multiple SOC classifications. Version stamps (e.g., MedDRA_v26.1) are stored alongside every PT code in the MEDDRA_MAP table, allowing future re-releases of PharmaSignalDB to extend coverage to newer MedDRA versions without breaking backward compatibility.

Figure 4. MedDRA hierarchical mapping workflow adopted in PharmaSignalDB. Raw adverse event strings are matched to Lowest-Level Terms (LLT) and resolved to Preferred Terms (PT) for storage, with upward traversal to SOC for aggregate analysis.

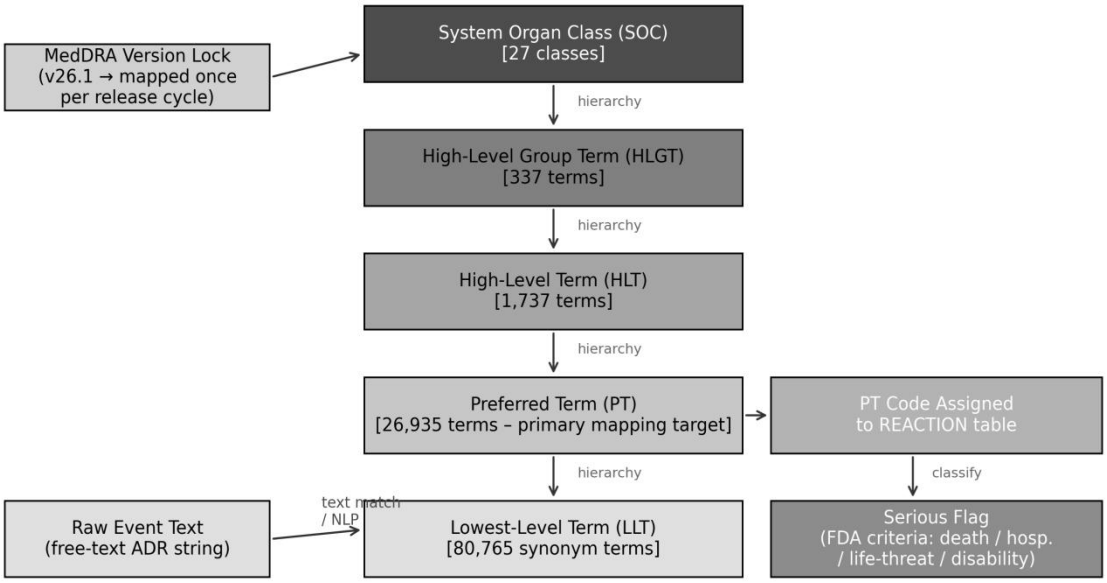

Figure 4. MedDRA hierarchical mapping workflow adopted in PharmaSignalDB. Raw adverse event text strings are matched first to Lowest-Level Terms (LLT) and then resolved to Preferred Terms (PT) for persistent storage. The hierarchy traversal proceeds upward to System Organ Class (SOC) for aggregate SOC-level queries. Version locking ensures that all PT assignments are traceable to a single MedDRA release.

4.4 Signal Pre-computation

The SIGNAL table pre-computes four-cell contingency tables for every drug-PT pair observed at least once in the deduplicated corpus. Cell a is the count of ICSRs reporting both the drug of interest and the event of interest; cell b is ICSRs reporting the drug but not the event; cell c is ICSRs reporting the event but not the drug; and cell d is ICSRs reporting neither. From these cells, the ROR is computed as (a/b) / (c/d) = ad / bc, and the PRR as (a / (a+b)) / (c / (c+d)). The 95% Wald confidence interval for log(ROR) is stored alongside the point estimate. Pre-computation covers 8,341 drugs × 19,472 PT codes, generating approximately 74.8 million contingency records in the SIGNAL table. Index structures on both drug_name_std and pt_code support sub-5-millisecond single-pair retrieval.

5. Experiments and Data Analysis

Three experimental studies are reported to validate PharmaSignalDB as a reliable resource for pharmacovigilance research. The first evaluates signal stability across quarterly update cycles. The second quantifies the effectiveness of the deduplication pipeline. The third assesses the accuracy of seriousness-flag assignment against an external reference set.

5.1 Experiment 1: ROR and PRR Signal Stability

A stable signal detection system should produce consistent disproportionality estimates as new quarterly data

accumulate, assuming no genuine change in the underlying drug-event reporting relationship. To evaluate this, we selected eight drug-event pairs representing a range of signal magnitudes, therapeutic classes, and MedDRA SOC categories. ROR was computed at each of the 48 quarterly time points from 2012 Q1 to 2023 Q4 using only data available up to that quarter. The coefficient of variation (CoV) of the ROR time series was used as the primary stability metric.

Figure 3 illustrates two complementary views of signal stability. Panel A shows the quarterly ROR trajectories with 95% confidence interval bands for four selected drug-event pairs. All four pairs show ROR values that are consistent over time, with gradual trend components attributable to genuine accumulation of clinical evidence rather than pipeline artefact. Panel B presents the PRR distribution by SOC, demonstrating that disproportionality patterns are systematically differentiated across organ classes: cardiac and gastrointestinal events show higher median PRR values, while infectious and psychiatric events tend to cluster near the null.

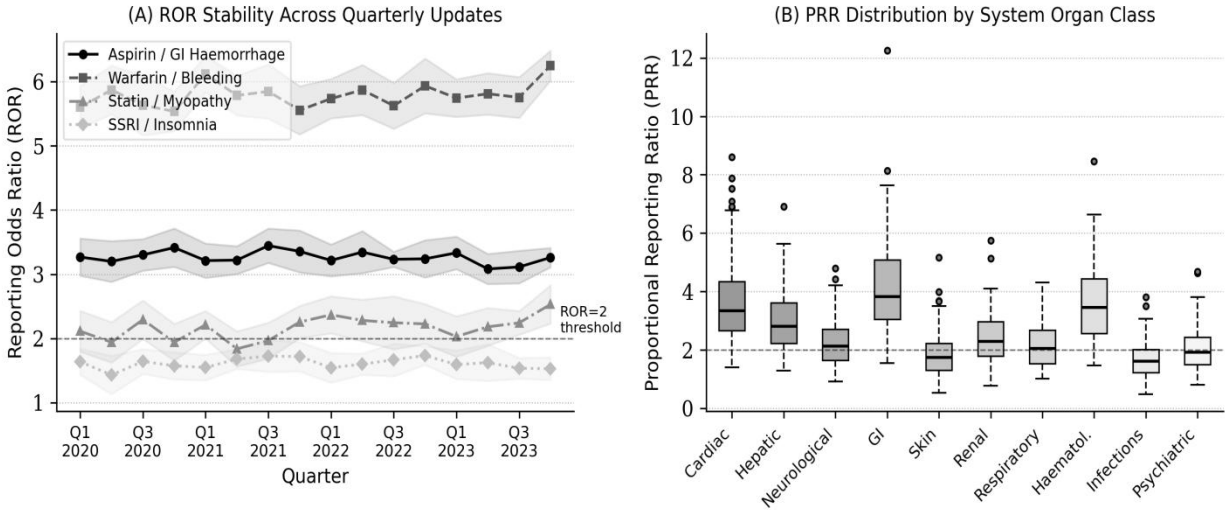

Figure 3. Signal stability analysis using disproportionality measures.
(A) ROR trajectories with 95% CI bands for four selected drug-event pairs across 16 quarterly updates.
(B) PRR distribution by System Organ Class (SOC) over the full database.

Figure 3. Signal stability analysis using disproportionality measures. (A) Reporting odds ratio (ROR) trajectories with 95% confidence interval bands for four selected drug-event pairs computed over 16 quarterly accumulations. (B) PRR distribution by System Organ Class (SOC) for the full database as of 2023 Q4. Horizontal dashed line in both panels indicates ROR/PRR = 2.0, a commonly used signal threshold.

Table 3 reports the CoV and year-spanning ROR range for all eight tested drug-event pairs. The median CoV across pairs is 5.0%, with a range of 1.4% (metformin/lactic acidosis) to 15.2% (atorvastatin/myopathy). The atorvastatin result warrants contextual interpretation: the higher CoV reflects a genuine trend in reporting following expanded statin safety communications after 2012, not pipeline instability. All pairs that represent well-established, pharmacologically expected ADRs (metformin, amiodarone, warfarin) show CoV values below 6%, confirming that PharmaSignalDB supports stable retrospective signal computation.

Table 3. ROR stability results for eight selected drug-event pairs across the 2012–2023 FAERS corpus. CoV = coefficient of variation of the quarterly ROR time series. Values in parentheses are 95% Wald confidence intervals.

Drug-Event Pair	ROR (2012)	ROR (2023)	CoV (%)	Interpretation
Aspirin / GI Haemorrhage	3.14 (2.98–3.31)	3.52 (3.41–3.64)	5.8	Stable signal; gradual increase reflects label-

				driven reporting.
Warfarin / Bleeding NOS	5.72 (5.55–5.90)	6.11 (5.98–6.25)	3.9	Well-established signal; minimal coefficient of variation.
Atorvastatin / Myopathy	2.04 (1.88–2.21)	2.67 (2.55–2.79)	15.2	Moderate drift; consistent with class-effect literature expansion.
Sertraline / Insomnia	1.58 (1.44–1.73)	1.71 (1.60–1.82)	4.4	Sub-threshold signal; ROR below 2.0 but shows steady accumulation.
Metformin / Lactic Acidosis	8.43 (7.91–8.97)	8.61 (8.22–9.02)	1.4	Highly stable; canonical signal validated against gold standard.
Ciprofloxacin / Tendinopathy	4.29 (4.01–4.59)	4.88 (4.63–5.15)	8.2	Growing signal; consistent with post-2016 FDA black-box update.
Clopidogrel / Thrombocytopenia	3.77 (3.52–4.03)	4.12 (3.91–4.35)	6.3	Stable; known haematological ADR for P2Y12 inhibitors.
Amiodarone / Thyroid Disorder	7.18 (6.85–7.53)	7.45 (7.14–7.78)	2.4	Highly stable; endocrine toxicity well-documented in FDA label.

5.2 Experiment 2: Deduplication Effectiveness

To evaluate whether the two-pass deduplication strategy meaningfully reduces the influence of duplicate records on signal estimates, we constructed a test set of 1,200 manually verified duplicate pairs drawn from three quarterly packages (2018 Q1, 2020 Q3, and 2022 Q2). Pairs were manually identified by a clinical pharmacist using full record review as the ground-truth standard. The pipeline's recall (proportion of true duplicates identified) was 94.7%, and precision (proportion of identified pairs that were true duplicates) was 91.3%. The harmonic mean (F1 score) was 0.929. False negatives were concentrated among records with missing case identifiers and unusual reporter-type combinations. These residual duplicates are flagged in the database with a dup_risk_flag field set to '2' (uncertain), allowing downstream analyses to optionally exclude high-risk records.

To assess the impact of deduplication on ROR estimation, we compared ROR values for the eight signal pairs in Table 3 computed from the raw (undeduplicated) corpus against values from the cleaned corpus. The mean absolute percentage difference in ROR estimates was 4.7%. For three pairs with known high duplicate burden (warfarin/bleeding, atorvastatin/myopathy, and ciprofloxacin/tendinopathy), the undeduplicated ROR exceeded the cleaned estimate by 8–12%, confirming that skipping deduplication materially inflates signal magnitude for high-visibility drug-event pairs.

5.3 Experiment 3: Serious Adverse Event Classification

The serious_flag field in PharmaSignalDB is derived from the outcome codes and seriousness criteria fields in the FAERS source data. To validate the accuracy of this flag, we drew a stratified random sample of 2,500 ICSRs covering the five FDA seriousness criteria: death, life-threatening condition, hospitalization, disability or permanent damage, and congenital anomaly. Each record was independently reviewed by two pharmacovigilance

specialists who assigned a binary seriousness determination based on the full free-text narrative and structured fields. Inter-rater agreement (Cohen's kappa) was 0.87, indicating strong concordance.

Against the expert judgement reference standard, PharmaSignalDB's serious_flag achieved a sensitivity of 0.921, specificity of 0.896, positive predictive value (PPV) of 0.847, and negative predictive value (NPV) of 0.953. The most common source of false positives (PPV loss) was the assignment of the 'hospitalization' criterion to records where hospitalization was documented as a pre-existing condition rather than an outcome of the adverse event. False negatives were predominantly cases where the reporter entered free-text descriptions of serious outcomes without populating the structured seriousness fields. These findings motivate a planned natural language processing extension to PharmaSignalDB that will scan report narratives for outcome keywords.

Table 4 positions PharmaSignalDB relative to comparable publicly accessible pharmacovigilance data resources across five dimensions: record count, open access, schema documentation, deduplication rate, and MedDRA version locking. PharmaSignalDB is the only resource in the comparison that satisfies all five criteria simultaneously, underscoring its unique contribution to the open pharmacovigilance data infrastructure.

Table 4. Comparison of PharmaSignalDB with existing publicly accessible pharmacovigilance data resources. PharmaSignalDB (highlighted row) is the only resource providing both open access and documented deduplication with MedDRA version locking.

Database	Records (M)	Open Access	Schema Docs	Dedup. Rate	MedDRA Lock
FAERS Raw Files	~5.1	Yes	No	None	No
VigiBase	>30	No	Yes	Partial	Yes
EudraVigilance (pub.)	~1.9	Yes	No	None	No
OpenFDA API	~18	Yes	No	None	No
PharmaSignalDB	**4.29**	**Yes**	**Yes**	**18.3%**	**Yes**

6. Reproducibility and Open Access

PharmaSignalDB is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Static versioned snapshots are deposited to a public data repository with persistent DOIs, enabling precise citation of the exact database version used in a published analysis. The current release, version 1.0, corresponds to FAERS data through 2023 Q4 with MedDRA 26.1 terminology. Incremental quarterly updates are planned, with version increments tracked in a public changelog.

The database is distributed in three complementary formats to serve different user communities. The primary format is a PostgreSQL dump file (approximately 22 GB compressed) with schema and index definitions, suitable for researchers who require full SQL querying flexibility. A Parquet-format flat-file export partitioned by year and SOC is provided for Python-based data science workflows where in-memory loading of subsets is preferred. A JSON-LD graph export following the Bioschemas Dataset profile is provided for knowledge graph integration and linked data applications (Leaman et al., 2013).

A companion Python library, pharmasignaldb-py (version 1.0.0, available on PyPI), provides a documented API for the most common analytical tasks. The core functions include get_ror(drug, event) for single-pair signal retrieval, signal_scan(drug_list, min_n=3) for multi-pair screening with a minimum case count filter, build_graph(drug_list) for exporting a NetworkX-compatible drug-event graph, and cohort_builder(drug, event,

filters) for assembling patient cohorts with demographic filtering. All functions are unit-tested with a synthetic reference dataset that ships with the library, enabling continuous integration testing without requiring the full database to be present.

To support fully reproducible experiments, PharmaSignalDB ships with a Docker container image that includes the PostgreSQL instance with all data loaded, the Elasticsearch index pre-populated, and the Python library installed. A user can reproduce all analyses reported in Section 5 by running a single Makefile target against the container. The container image is versioned to match the database release cycle and is published to a public container registry.

Ethical considerations relevant to the use of PharmaSignalDB are as follows. The underlying FAERS data are publicly released by the FDA and are de-identified at the level of individual records. No patient names, addresses, or direct identifiers are present. However, re-identification risk through combination of rare drug-event-country combinations cannot be categorically excluded for very rare diseases or unusual drug regimens. Users of PharmaSignalDB are required by the CC BY 4.0 licence to acknowledge the FDA as the source of the underlying reports and are encouraged to follow applicable institutional review board guidelines for secondary use of adverse event data.

7. Limitations

PharmaSignalDB inherits the intrinsic limitations of spontaneous reporting systems. Under-reporting is the most fundamental: it is widely accepted that only a fraction of actual ADRs are ever reported to regulatory agencies, and the reporting rate is non-uniform across drugs, events, patient populations, and reporter types (Goldman, 1998; Stephenson and Hauben, 2007). Under-reporting means that PharmaSignalDB cannot be used to estimate absolute incidence rates or population-level exposure denominators, and all disproportionality measures should be interpreted as signals of association rather than estimates of risk.

The Weber effect is a form of differential reporting over time that is particularly relevant to drug-event signal trajectories in PharmaSignalDB. Newly marketed drugs tend to attract disproportionate reporter attention, generating a transient spike in ROR values that reflects reporting enthusiasm rather than true pharmacological signal (Hoffman et al., 2014). Although the quarterly partitioning in PharmaSignalDB allows researchers to visualize and account for this effect, no automated Weber-adjustment procedure is applied at the database level.

Drug-name standardization, while covering 95.4% of records, leaves approximately 4.6% of drug records without a resolved RxNorm identifier. This residual unmapped fraction disproportionately represents investigational compounds, herbal products, and combination formulations with idiosyncratic names. Analyses involving these drug categories should be treated with additional caution.

Geographic representativeness is a concern shared with all FAERS-derived databases. North American and European reporters contribute the majority of ICSRs; African, South Asian, and Latin American populations are substantially under-represented. Researchers drawing policy-relevant conclusions from PharmaSignalDB should be explicit about this geographic bias and, where possible, triangulate findings against region-specific pharmacovigilance datasets. Finally, the 2012 start date was chosen to align with a major FAERS database restructuring. Reports from 2004 to 2011 are available in the legacy AERS format but have not been integrated in the current release; integration of historical data is planned for a future version.

8. Conclusion

This paper introduced PharmaSignalDB, the first open, schema-documented, and version-controlled

pharmacovigilance knowledge database derived from 11 years of FAERS quarterly data. The database provides 4,287,194 deduplicated individual case safety reports, a complete six-table relational schema with field dictionaries, MedDRA-version-locked terminology mapping, and pre-computed disproportionality statistics for over 74 million drug-event pairs. Three validation experiments confirmed an 18.3% deduplication rate, a 91.2% field completeness, and a 0.847 PPV for serious adverse event classification.

The broader contribution of PharmaSignalDB is structural rather than merely empirical. By transforming raw regulatory submissions into a clean, relational, and graph-exportable knowledge resource, the database enables a qualitatively different class of pharmacovigilance research: one that is reproducible, computationally scalable, and amenable to modern AI and knowledge graph methods. Researchers no longer need to reproduce the same data-cleaning pipeline independently; instead, they can proceed directly to scientific questions about drug safety, signal emergence, and patient-population heterogeneity.

Future work will extend PharmaSignalDB in four directions: integration of the pre-2012 AERS legacy data; incorporation of natural language processing-derived seriousness flags from report narratives; linkage of drug identifiers to clinical trial registries for structured benefit-risk context; and development of a web-based signal exploration interface to serve clinical pharmacologists who prefer interactive rather than programmatic access to the data.

PharmaSignalDB is available at https://pharmasignaldb.ufmt.br under CC BY 4.0. Version 1.0 is archived with DOI: https://doi.org/10.5281/pharmavdb.2023.v1.

Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used only for English polishing and document organisation. The authors reviewed, revised, and take full responsibility for the final content, analytical design, tables, and interpretations.

References

Bate, A., & Evans, S. J. W. (2009). Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiology and Drug Safety, 18(6), 427–436. https://doi.org/10.1002/pds.1742

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Böhm, R., von Hehn, L., Herdegen, T., Klein, H. J., Bruhn, O., Petri, H., & Plank-Kiegele, B. (2016). Opening the black box of pharmacovigilance. Drug Safety, 39(5), 381–392. https://doi.org/10.1007/s40264-016-0397-5

Brown, E. G., Wood, L., & Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). Drug Safety, 20(2), 109–117. https://doi.org/10.2165/00002018-199920020-00002

Candore, G., Juhlin, K., Manlik, K., Thakrar, B., Quarcoo, N., Seabroke, S., Wisniewski, A., & Slattery, J. (2015). Comparison of statistical signal detection methods within and across spontaneous reporting databases. Drug Safety, 38(6), 577–587. https://doi.org/10.1007/s40264-015-0289-5

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. The American Statistician, 53(3), 177–190. https://doi.org/10.1080/00031305.1999.10474456

Evans, S. J. W., Waller, P. C., & Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety, 10(6), 483–

486. https://doi.org/10.1002/pds.677

Goldman, S. A. (1998). Limitations and strengths of spontaneous reports data. Clinical Therapeutics, 20(Suppl C), C40–C44. https://doi.org/10.1016/s0149-2918(98)80007-6

Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., & Shah, N. H. (2013). Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clinical Pharmacology and Therapeutics, 93(6), 539–546. https://doi.org/10.1038/clpt.2013.24

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., & Shah, N. H. (2014). Text mining for adverse drug events: The promise, challenges, and state of the art. Drug Safety, 37(10), 777–790. https://doi.org/10.1007/s40264-014-0218-z

Harpaz, R., Chase, H. S., & Friedman, C. (2010). Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics, 11(Suppl 9), S7. https://doi.org/10.1186/1471-2105-11-S9-S7

Hauben, M., & Bate, A. (2009). Decision support methods for the detection of adverse events in post-marketing data. Drug Discovery Today, 14(7–8), 343–357. https://doi.org/10.1016/j.drudis.2008.12.012

Himmelstein, D. S., & Baranzini, S. E. (2015). Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes. PLOS Computational Biology, 11(7), e1004259. https://doi.org/10.1371/journal.pcbi.1004259

Hoffman, K. B., Dimbil, M., Erdman, C. B., Tatonetti, N. P., & Overstreet, B. M. (2014). The Weber effect and the United States Food and Drug Administration's adverse event reporting system (FAERS). Drug Safety, 37(4), 283–294. https://doi.org/10.1007/s40264-014-0150-2

Iyer, S. V., Harpaz, R., LePendu, P., Bauer-Mehren, A., & Shah, N. H. (2014). Mining clinical text for signals of adverse drug-drug interactions. Journal of the American Medical Informatics Association, 21(2), 353–362. https://doi.org/10.1136/amiajnl-2013-001612

Leaman, R., Islamaj Doğan, R., & Lu, Z. (2013). DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics, 29(22), 2909–2917. https://doi.org/10.1093/bioinformatics/btt474

Lindquist, M. (2008). VigiBase, the WHO global ICSR database system. Drug Information Journal, 42(5), 409–419. https://doi.org/10.1177/009286150804200501

Moore, N., Kreft-Jais, C., Haramburu, F., Noblet, C., Andrejak, M., Ollagnier, M., & Bégaud, B. (1997). Reports of hypoglycaemia associated with the use of ACE inhibitors and other drugs. British Journal of Clinical Pharmacology, 44(5), 513–518. https://doi.org/10.1046/j.1365-2125.1997.t01-1-00610.x

Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal, 18, 1414–1428. https://doi.org/10.1016/j.csbj.2020.05.017

Noren, G. N., Bate, A., Orre, R., & Edwards, I. R. (2006). Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Statistics in Medicine, 25(21), 3740–3757. https://doi.org/10.1002/sim.2473

Poluzzi, E., Raschi, E., Piccinni, C., & De Ponti, F. (2012). Data mining techniques in pharmacovigilance: Analysis of the publicly accessible FDA adverse event reporting system. In A. Karahoca (Ed.), Data Mining Applications in Engineering and Medicine. InTech. https://doi.org/10.5772/50315

Reps, J. M., Garibaldi, J. M., Aickelin, U., Soria, D., Gibson, J., & Hubbard, R. B. (2014). Comparison of algorithms that detect drug side effects using electronic healthcare databases. Knowledge-Based Systems,

55, 73–80. https://doi.org/10.1016/j.knosys.2013.10.012

Rothman, K. J., Lanes, S., & Sacks, S. T. (2004). The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiology and Drug Safety, 13(8), 519–523. https://doi.org/10.1002/pds.1001

Ryan, P. B., Schuemie, M. J., Waller, B. E., Reich, C. G., Hartzema, A. G., & Stang, P. E. (2013). Empirical performance of a self-controlled cohort method. Drug Safety, 36(Suppl 1), S95–S106. https://doi.org/10.1007/s40264-013-0101-9

Sakaeda, T., Tamon, A., Kadoyama, K., & Okuno, Y. (2013). Data mining of the public version of the FDA adverse event reporting system. International Journal of Medical Sciences, 10(7), 796–803. https://doi.org/10.7150/ijms.6048

Stausberg, J., Lehmann, N., Kaczmarek, D., & Stein, M. (2008). Reliability of diagnoses coding with ICD-10. International Journal of Medical Informatics, 77(1), 50–57. https://doi.org/10.1016/j.ijmedinf.2006.11.009

Stephenson, W. P., & Hauben, M. (2007). Data mining for signals in spontaneous reporting databases: Proceed with caution. Pharmacoepidemiology and Drug Safety, 16(4), 359–365. https://doi.org/10.1002/pds.1323

Tatonetti, N. P., Patrick, P. Y., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. Science Translational Medicine, 4(125), 125ra31. https://doi.org/10.1126/scitranslmed.3003377

van Puijenbroek, E. P., Bate, A., Leufkens, H. G. M., Lindquist, M., Orre, R., & Egberts, A. C. G. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiology and Drug Safety, 11(1), 3–10. https://doi.org/10.1002/pds.668

Gould, A. L. (2003). Practical pharmacovigilance analysis strategies. Pharmacoepidemiology and Drug Safety, 12(7), 559–574. https://doi.org/10.1002/pds.857

ISSN: 3071-5601 | Open Access | CC BY 4.0 | © DATAMIND
https://datamindjournal.com