Data Mining Applications in Pharmacovigilance Databases: A Scoping Review

Ana Carolina Jacoby [a,b], Mel Amisa Matsuda [a,] Sílvio César Cazella [a],Carine Raquel Blatt.[a]

ORCID IDs:
Ana Carolina Jacoby [https://orcid.org/0009-0006-9045-5018]
Mel Amisa Matsuda [https://orcid.org/0009-0005-0244-6852]
Sílvio César Cazella [https://orcid.org/0000-0003-2343-893X]
Carine Raquel Blatt [https://orcid.org/0000-0001-5935-1196]

Emails: ana.jacoby@ufcspa.edu.br, mell.matsuda@ufcspa.edu.br, silvioc@ufcspa.edu.br, carineblatt@ufcspa.edu.br.

Affiliations: [a] Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre/RS, Brazil.

Corresponding Author: [b] Corresponding author: Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) Rua Sarmento Leite, 245 - Centro Histórico, CEP 90050-170, Porto Alegre/RS, Brazil E-mail: ana.jacoby@ufcspa.edu.br

1. Background

Adverse drug reactions (ADRs), defined as harmful and unintended responses to drugs administered in appropriate doses for the prevention, diagnosis, treatment of diseases, or modification of physiological functions, represent one of the leading causes of emergency hospital admissions (1).

It is estimated that the median prevalence of hospitalizations associated with ADRs is 6.3% in developed countries and 5.5% in developing countries, with a considerable proportion of these cases being preventable (2).

In this context, pharmacovigilance, as defined by the World Health Organization, is the science and set of activities related to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problems, and is an essential component of patient safety (3). Although medicines undergo rigorous regulatory protocols before commercialization, many adverse effects only become evident after widespread use,

outside the controlled conditions of clinical trials (4). Pharmacovigilance therefore enables early risk identification, promotes the rational use of medicines, and enhances public health policies (5)(6).

Beyond the clinical impacts, ADRs generate substantial costs for health systems (7). A study conducted in Japan found that 11% of patients evaluated required hospitalization due to adverse drug events. Direct costs per patient ranged from approximately USD 144 to 153 for outpatient care and from USD 5,769 to 5,914 for hospitalizations. Annual costs related to preventable adverse events in older adults exceeded USD 1.74 billion (8), reinforcing the urgency of effective monitoring strategies.

With technological advances and the exponential growth of available health data, it is increasingly necessary to incorporate advanced computational tools into the pharmacovigilance process (9). Data mining and machine learning have emerged as promising approaches to address this complexity, enabling the identification of hidden patterns, early detection of warning signals, and more accurate prediction of adverse events (10). These techniques have been applied to large official pharmacovigilance databases, such as the FDA Adverse Event Reporting System (FAERS) (11) and the European Medicines Agency's EudraVigilance (EMA) (12), among others. Descriptive techniques, such as exploratory data analysis (EDA), and predictive methods based on supervised and unsupervised learning algorithms have been successfully employed in the analysis of these databases (13). The integration between pharmacovigilance and data science is a strategic area for advancing research and innovation in public health (14).

Considering this scenario, the present scoping review aims to map the data mining algorithms applied to pharmacovigilance databases, identifying the most used approaches, their objectives — such as detection, classification, or prediction of adverse events — and their respective contexts of application. Additionally, this mapping is expected to contribute to the scientific literature by synthesizing existing knowledge and highlighting gaps, guiding future investigations in the field.

2. Methods

2.1. Type of Study

This study is a scoping review conducted according to the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) (15).

2.2. Review Question

The research question was developed based on the PCC (Population, Concept, Context) framework:

P – Problem: Cases of adverse drug reactions

C – Concept: Data mining algorithms applied to pharmacovigilance

C – Context: Scientific studies using data mining in the field of pharmacovigilance

Thus, the guiding question of this study is: “Which data mining algorithms have been described for exploring adverse drug events in pharmacovigilance studies?” In addition, this research also seeks to explore the following aspects: which data mining tasks have been applied in these studies (such as classification, clustering, association rule mining, or signal detection), whether machine learning or deep learning algorithms have been employed and, if so, which ones, and whether the datasets used are available for further research, thereby fostering Open Science.

2.3. Eligibility Criteria

The eligibility criteria include works published between 2015 and 2025, in English, that are complete original studies addressing the application of data mining techniques using real pharmacovigilance database data, and articles available in full in academic portals accessible through the university.

Exclusion criteria: review studies (systematic, narrative, scoping), editorials, expert opinions, books, book chapters, dissertations, theses, non-peer-reviewed preprints, and conference abstracts. Studies that do not use real pharmacovigilance data or do not apply data mining will also be excluded.

2.4. Search Strategy

The search strategy was built using a combination of controlled descriptors (e.g., MeSH and Emtree) and free terms, with Boolean operators AND and OR to structure the search logic. The search will be conducted in PubMed, Embase, Scopus, and Web of Science databases. The terms were selected to retrieve studies describing the application of data mining for the identification, analysis, prediction, or detection of adverse drug reactions in pharmacovigilance databases.

The combination of terms will be adapted to each database's syntax. Filters will be applied for publication year (2015–2025), language (English), and publication type (full open-access article). Retrieved records will be managed in Rayyan, where duplicates will be removed, and screening by title, abstract, and full text will be conducted. Selection will follow the PRISMA-ScR flow, ensuring transparency, reproducibility, and methodological traceability.

MeSH terms
• **D060735 Pharmacovigilance** • **D016903 Drug Monitoring** *Monitoring, Drug* *Therapeutic Drug Monitoring* *Drug Monitoring, Therapeutic* *Monitoring, Therapeutic Drug* • **D064420 Drug-Related Side Effects and Adverse Reactions** *Drug Related Side Effects and Adverse Reactions* *Side Effects of Drugs* *Drug-Related Side Effects and Adverse Reaction* *Drug Related Side Effects and Adverse Reaction* *Adverse Drug Reaction* *Adverse Drug Reactions* *Drug Reaction, Adverse* *Drug Reactions, Adverse*

MeSH terms
Reactions, Adverse Drug *Adverse Drug Event* *Adverse Drug Events* *Drug Event, Adverse* *Drug Events, Adverse* *Drug Side Effects* *Drug Side Effect* *Effects, Drug Side* *Side Effect, Drug* *Side Effects, Drug* *Drug Toxicity* *Toxicity, Drug* *Drug Toxicities* *Toxicities, Drug* ● **D057225 Data Mining** *Mining, Data* *Text Mining* *Mining, Text*

The base string used to guide the database-specific adaptations was as follows:

("pharmacovigilance" OR "drug monitoring" OR "therapeutic drug monitoring" OR "monitoring, drug" OR "drug monitoring, therapeutic" OR "monitoring, therapeutic drug" OR "drug related side effects and adverse reactions" OR "side effects of drugs" OR "drug-related side effects and adverse reaction" OR "adverse drug reaction" OR "adverse drug reactions" OR "drug reaction, adverse" OR "drug reactions, adverse" OR "reactions, adverse drug" OR "adverse drug event" OR "adverse drug events" OR "drug event, adverse" OR "drug events, adverse" OR "drug side effects" OR "drug side effect" OR "side

effect, drug" OR "side effects, drug" OR "drug toxicity" OR "toxicity, drug" OR "drug toxicities" OR "toxicities, drug")

AND

("data mining" OR "text mining" OR "mining, data" OR "mining, text")

2.5. Study Selection

Selection will be carried out by two independent reviewers who will screen titles and abstracts, followed by full-text reading. Disagreements will be resolved by consensus or with the assistance of a third reviewer. Rayyan will be used to manage the screening and data extraction process.

2.6. Data Extraction

Data extraction will be performed using a table containing the following fields: reference identification (author and year), journal, year of publication, country, language, type of study, and application objective. Information about the database used (number of notifications or records analyzed, data scope, pre-processing or enrichment steps) will also be included. Regarding data mining techniques, the applied methodology, analysis objectives, and use of cross-validation or other validation methods will be recorded. Finally, the main results reported by the authors, as well as methodological quality assessment and other relevant information, will be described.

2.7. Data Analysis and Synthesis

Data synthesis will be conducted through a descriptive approach and thematic categorization, including tables and graphical visualizations to present algorithms, data types, and objectives. Trends, gaps, and potential future opportunities will be highlighted.

3.0 AI Use Disclosure

OpenAI’s ChatGPT (version GPT-5) was used exclusively for the purpose of translating this review protocol from Portuguese to English. The authors have carefully reviewed, edited, and

verified the translated content to ensure accuracy and fidelity to the original text. No other content generation or modification was performed by the AI tool.

4. References

1. Hodel KVS, Fiuza BSD, Conceição RS, Aleluia ACM, Pitanga TN, Fonseca LMDS, et al. Pharmacovigilance in Vaccines: Importance, Main Aspects, Perspectives, and Challenges-A Narrative Review. Pharm Basel Switz. 19 de junho de 2024;17(6):807.
2. Komagamine J. Prevalence of urgent hospitalizations caused by adverse drug reactions: a cross-sectional study. Sci Rep. 13 de março de 2024;14(1):6058.
3. Khan MAA, Sara T, Babar ZUD. Pharmacovigilance: the evolution of drug safety monitoring. J Pharm Policy Pract. 2024;17(1):2417399.
4. Defining and assessing adverse events and harmful effects in psychotherapy study protocols: A systematic review - PubMed [Internet]. [citado 24 de julho de 2025]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/35049321/
5. Abiri OT, Johnson WCN. Pharmacovigilance systems in resource-limited settings: an evaluative case study of Sierra Leone. J Pharm Policy Pract. 2019;12:13.
6. Cui X, Yang DQ, Xie ZN, Li YY, Wang ZF, Liu H, et al. [Pharmacovigilance guidelines for clinical application of Chinese patent medicines for external use]. Zhongguo Zhong Yao Za Zhi Zhongguo Zhongyao Zazhi China J Chin Mater Medica. agosto de 2024;49(16):4285–90.
7. Robinson EG, Hedna K, Hakkarainen KM, Gyllensten H. Healthcare costs of adverse drug reactions and potentially inappropriate prescribing in older adults: a population-based study. BMJ Open. 23 de setembro de 2022;12(9):e062589.
8. Evaluation of the Direct Costs of Managing Adverse Drug Events in all Ages and of Avoidable Adverse Drug Events in Older Adults in Japan - PubMed [Internet]. [citado 23 de julho de 2025]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34867381/
9. Alomar M, Palaian S, Al-Tabakha MM. Pharmacovigilance in perspective: drug withdrawals, data mining and policy implications. F1000Research. 2019;8:2109.
10. Magana J, Gavojdian D, Menahem Y, Lazebnik T, Zamansky A, Adams-Progar A. Machine learning approaches to predict and detect early-onset of digital dermatitis in dairy cows using sensor data. Front Vet Sci. 2023;10:1295430.

11. Potter E, Reyes M, Naples J, Dal Pan G. FDA Adverse Event Reporting System (FAERS) Essentials: A Guide to Understanding, Applying, and Interpreting Adverse Event Data Reported to FAERS. Clin Pharmacol Ther. 19 de maio de 2025;
12. Chiappini S, Vickers-Smith R, Guirguis A, Corkery JM, Martinotti G, Harris DR, et al. Pharmacovigilance Signals of the Opioid Epidemic over 10 Years: Data Mining Methods in the Analysis of Pharmacovigilance Datasets Collecting Adverse Drug Reactions (ADRs) Reported to EudraVigilance (EV) and the FDA Adverse Event Reporting System (FAERS). Pharm Basel Switz. 27 de maio de 2022;15(6):675.
13. Advancements within Modern Machine Learning Methodology: Impacts and Prospects in Biomarker Discovery - PubMed [Internet]. [citado 24 de julho de 2025]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/33557728/
14. Ben Abacha A, Chowdhury MFM, Karanasiou A, Mrabet Y, Lavelli A, Zweigenbaum P. Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. J Biomed Inform. dezembro de 2015;58:122–32.
15. 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. 2 de outubro de 2018;169(7):467–73.