Data mining applications in pharmacovigilance databases: A scoping review

Ana Carolina Jacoby [a,b,] Mel Amisa Matsuda [a,], Sílvio César Cazella [a,], and 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

Abstract: 300 palavras

Background: +- 3

Methods: +-8

Results:+-8

Conclusion:+-5

Keywords: Pharmacovigilance, Data Mining, Adverse Drug Reactions, Signal Detection.

1. Introduction

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)

Beyond the clinical impacts, ADRs generate substantial costs for health systems. (6) 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 (7) , 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.(8) 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.(9) These techniques have been applied to large official pharmacovigilance databases, such as the FDA Adverse Event Reporting System (FAERS) (10) and the European Medicines Agency's EudraVigilance (EMA) (11), 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. (12) The integration between pharmacovigilance and data science is a strategic area for advancing research and innovation in public health.(13)

However, the scope and variety of data mining algorithms used for adverse event analysis in pharmacovigilance have not yet been systematically mapped.

Considering this scenario, the present scoping review aims to fill this gap, guided by the following research question, based on the PCC framework: "What data mining algorithms have been described for exploring adverse drug events in pharmacovigilance studies?"

Specifically, this review seeks to map the data mining tasks applied (such as signal detection, classification, clustering, or association rules), identify the prevalence and types of

algorithms employed, and determine the state of research transparency by assessing the availability of datasets (Open Science) to promote replicability.

It is expected that this mapping will contribute to the scientific literature by synthesizing existing knowledge, identifying current gaps, and thereby guiding future investigations in this field.

2. Methods

This scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines (14), and its protocol was registered on the Open Science Framework (OSF) on April 28, 2025 (DOI: 10.17605/OSF.IO/KZJDT). The review was structured using the PCC (Population, Concept, and Context) framework, considering pharmacovigilance databases as the population, the application of data mining algorithms as the concept, and pharmacovigilance studies as the context. The research question guiding the review was: "What data mining algorithms have been described for exploring adverse drug events in pharmacovigilance studies?" In addition, this review aims to explore 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 publicly available, thereby promoting Open Science. This review included studies published in the last 10 years in English that addressed the application of data mining algorithms to pharmacovigilance databases. Studies comprising reviews, editorials, books, book chapters, guidelines, expert opinions, dissertations, theses, and conference abstracts were excluded. Additionally, studies not involving pharmacovigilance research, those not applying data mining techniques to adverse drug events, or those not providing sufficient methodological details were also not

considered.

2.1 Search approach

A comprehensive search strategy was developed and executed to identify relevant studies on the application of data mining algorithms in pharmacovigilance. The search was conducted on July 16, 2025, in the PubMed, Scopus, Embase, and Web of Science databases. The strategy utilized Medical Subject Headings (MeSH) and keywords in English, as listed in Table 1, combining terms related to pharmacovigilance and data mining with the Boolean operators AND and OR. The search syntax was adapted to the requirements of each database, and the full strategy is detailed in Figure 1. Additionally, a manual search of the reference lists of included articles was performed to identify other potentially eligible studies. The search was restricted to articles published in English between January 2015 and July 2025, as established by the eligibility criteria.

Table 1- Medical subject headings (MeSH) used in the search strategy

CONCEPT 1: PHARMACOVIGILANCE / ADVERSE REACTIONS
PHARMACOVIGILANCE **DRUG MONITORING** ● Monitoring, Drug ● Therapeutic Drug Monitoring ● Drug Monitoring, Therapeutic ● Monitoring, Therapeutic Drug

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
- Adverse Drug Reaction
- Adverse Drug Reactions
- Drug Reaction, Adverse
- Drug Reactions, Adverse
- 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

CONCEPT 2: DATA MINING

DATA MINING

- Mining, Data
- Text Mining
- Mining, Text

Figure 1 - Full formatted search strategy.

("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")

2.2. Study selection procedure

A total of 1,468 articles were found (PubMed: 161; Scopus: 543; Web of Science: 339; Embase: 425) in the search. The articles were extracted from the databases in Research Information Systems (RIS) format and imported into the reference manager Zotero to standardize the metadata and remove duplicates, resulting in the exclusion of 978 duplicate articles. The remaining 490 articles were then exported to Rayyan, the software used to

manage the review. The selection process was carried out by two authors, and disagreements were reviewed by a third author. In the first phase, the articles were screened by reading titles and abstracts, where 281 articles were excluded for not meeting the inclusion criteria established by the authors. In the second phase, the 209 articles were read in full to confirm their inclusion criteria. Thirty-two articles were excluded for not being fully available for reading, and 14 articles were excluded at this stage for being outside the established eligibility criteria (articles comparing the quality of data mining algorithms). Finally, during the inclusion phase, the reference lists of the screened studies were reviewed to identify additional relevant studies not initially retrieved, but no new articles were included. After all phases, 163 articles were included in this review.

2.3. Data extraction

A data extraction form was developed using Google Sheets to collect relevant information from the included articles. The extraction process was conducted in an iterative manner to refine the variables, with any discrepancies resolved by consensus among the researchers. The data were grouped into four main thematic areas. The first, Context and Data Source, included the year of publication (year_publication), the source database (database), and the pharmaceutical class or drugs in focus (categorized according to the ATC code) (drug_class). The second area, focused on Methodology and Collection, addressed the study period (start_data_collection, end_data_collection), as well as the availability of raw data (open_data) and the processed dataset (dataset_made_available). Subsequently, the third area, Algorithms and Analysis, detailed the disproportionality techniques and/or algorithms (disproportionality_techniques), the additional data mining algorithms (data_mining_algorithms_techniques), and other analytical techniques (other_techniques). Finally, the fourth area, Study Outcomes, was included to characterize the articles by

summarizing the main objective and the conclusions presented by the authors.

2.4. Data analysis

The extracted data were independently verified and validated by the researchers, with any divergences resolved by consensus. Synthesis of the findings was performed through a narrative and thematic approach, focused on mapping the literature in response to the guiding question of the review. All data analyses and result generation were conducted using the Python programming language via the Google Colaboratory (Colab) environment. The complete source code and scripts were developed by the first two authors and cross-checked by the team of specialized professors in Data Mining and Pharmacovigilance, aiming for maximum transparency and reproducibility. This material is publicly available in a GitHub repository for readers' consultation. The main analysis consisted of identifying, grouping, and describing the different data mining algorithms, in addition to establishing patterns of application and identifying methodological gaps in the literature.

3. Results

3.1. General Characteristics of the Included Studies

Contexto Básico: Apresentação inicial do universo de 163 artigos. O que entra aqui: • Tendência Temporal: Distribuição dos artigos por ano de publicação. • Origem Geográfica e/ou Institucional (se mapeado). • Objetivos e Conclusões: Breve resumo dos focos dos artigos (para caracterização, como você mencionou).

3.2. Drug Classes and Data Sources

Contexto Clínico/Fonte: Onde a pesquisa de farmacovigilância está concentrada. O que entra aqui: • Classes Farmacológicas (ATC): As classes mais frequentemente estudadas. •

Fontes de Dados: Quais bancos de dados de farmacovigilância foram mais usados (FAERS, VigiBase, etc.).

3.3 Data Mining Algorithms and Techniques

SEU FOCO PRINCIPAL. O que entra aqui: • Agrupamento e Descrição: Lista e descrição clara de TODOS os algoritmos encontrados. • Frequência: Qual algoritmo é o mais utilizado. • Classificação: Agrupar por função (Desproporcionalidade: ROR/PRR/BCPNN vs. Aprendizagem de Máquina: SVM/RF/DL).

3.4. Open Science and Reproducibility

Seu tópico de Ciência Aberta. O que entra aqui: • Disponibilidade de Código/Scripts: Quantos artigos disponibilizaram o código-fonte (GitHub, etc.). • Disponibilidade de Dados: Quantos estudos usaram dados abertos (open_data) e quantos disponibilizaram seus datasets tratados. • Avaliação: Percentual de estudos que permitem a reprodutibilidade.

4. Discussion

Síntese e Mapeamento dos Algoritmos:

Qual é o Panorama? Comece sintetizando as principais descobertas: Qual é o algoritmo mais usado? Os algoritmos de desproporcionalidade ainda dominam, ou há uma migração para Aprendizagem de Máquina/DL?

Contexto vs. Ferramenta: Discuta a relação entre os algoritmos escolhidos e as classes farmacológicas/bases de dados mais comuns.

Implicações para a Informática Médica (Foco da Revista):

Avanço Metodológico: O que o mapeamento desses algoritmos significa para o avanço da Farmacovigilância baseada em dados e para a tomada de decisão clínica.

Desafios do Big Data: Aborde como a diversidade de algoritmos reflete a complexidade de lidar com bancos de dados grandes e ruidosos de eventos adversos.

Análise de Transparência e Reprodutibilidade (Seu Ponto Forte):

Crítica à Ciência Aberta: Discuta o baixo ou alto nível de adesão à Ciência Aberta (seus achados no 3.4). Se poucos artigos compartilham códigos e dados, use isso para criticar a falta de reprodutibilidade na área.
Recomendação de Políticas: Use sua análise de transparência para sugerir que revistas (como a IJMI) exijam o compartilhamento de código (link para o GitHub/Colab).

Lacunas e Direções Futuras:
O que está faltando? Use o mapeamento para identificar classes de medicamentos negligenciadas ou algoritmos pouco explorados (por exemplo, uso de NLP para dados de texto não estruturados em notificações).
Sugestão de Pesquisa: Apresente propostas concretas para futuras pesquisas baseadas nas lacunas encontradas.

5. Conclusion

Esta seção deve ser curta e focada. Não repita resultados ou discussões.

Tópicos que sua Conclusão deve abordar:

Resposta Direta à Pergunta: Uma frase clara que resume os principais algoritmos encontrados (o "quais" da sua pergunta de pesquisa).

Impacto Primário: Reforce o principal insight do estudo (ex: o domínio dos algoritmos tradicionais, a baixa reprodutibilidade, etc.).

Contribuição para a Área: Afirme, em uma frase final e poderosa, a contribuição do seu mapeamento para a informática em saúde.

CRediT authorship contribution statement

Ana Carolina Jacoby: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Writing – original draft, Project administration.

Mel Matsuda: Investigation, Writing – review & editing.

Silvio Cesar Cazella: Supervision, Resources, Writing – review & editing.

Carine Raquel Blatt: Supervision, Resources, Writing – review & editing.

Declaration of competing interest

The authors affirm they have no competing financial interests or personal relationships that could have influenced the research reported in this manuscript.

References