Data Mining based Pharmacovigilance System for Pharma Drug Safety

Nidhi Patel | Moumita Baidya

1. Problem Statement:

The FDA's Adverse Event Reporting System contains over 4 million reports of drug side effects, but current manual review methods fail to detect complex drug interactions and rare safety patterns in time to prevent patient harm. This delay in identifying dangerous drug combinations costs thousands of lives annually and billions in healthcare expenses.

2. Objective:

To develop an unsupervised machine learning framework using association mining, clustering, and dimensionality reduction techniques to automatically discover hidden drug safety patterns, identify high-risk drug combinations, and stratify vulnerable patient populations from three years of FAERS data, enabling earlier detection of adverse events and improving patient safety.

3. GitHub Repository: https://github.com/orgs/2025-F-CS6220/teams/pharmacovigilance-system

4. Dataset Description:

- **Dataset:** FDA Adverse Event Reporting System (FAERS)
- **Source:** https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html
- **Access:** Public, free to download
- **Total Size:** ~12–15 GB uncompressed
- **Update Frequency:** Quarterly

We are currently working from Q1 2024 to the latest quarter of 2025, covering 7 quarterly releases. The data is provided in ASCII text format and is delimited using dollar signs ($). This amount of data is large enough to capture emerging safety patterns, yet manageable for local analysis.

5. Dataset Overview:

FAERS contains voluntary and mandatory adverse drug event reports submitted by healthcare professionals, manufacturers, and consumers. Each quarterly release includes seven core relational files:

- DEMO: Patient demographics & report metadata
- DRUG: Drugs used, active ingredients, dose & route
- REAC: Reported adverse reactions
- OUTC: Patient outcomes (hospitalization, death, etc.)
- INDI: Indications (reason for drug use)
- THER: Therapy timelines
- RPSR: Report sources

High-Level Dataset Statistics (7 quarters):

- ~2.2–2.7 million adverse event reports
- 10,000+ unique drugs
- 15,000+ unique adverse events
- 200K–350K demographic entries per quarter
- 800K–1.2M drug entries per quarter

Using 7 quarters allows me to observe seasonal patterns, emerging reactions.

6. Sample Records

(i) DEMO File

primaryid	caseid	age	age_cod	sex	wt	occr_country	event_dt
205431871	20543187	67	YR	F	70.4	US	20240412
205431872	20543187	50	YR	M	88.1	US	20240414
205431873	20543188	73	YR	F	60.2	GB	20240502

(ii) DRUG File

primaryid	drug_seq	drugname	prod_ai	route	dose_amt	dose_unit
205431871	1	OZEMPIC	SEMAGLUTIDE	ORAL	1	MG
205431871	2	METFORMIN	METFORMIN HCL	ORAL	500	MG
205431872	1	APIXABAN	APIXABAN	ORAL	5	MG

(iii) REAC (Reactions)

primaryid	pt
205431871	HYPOGLYCAEMIA
205431871	NAUSEA
205431872	GASTROINTESTINAL HAEMORRHAGE
205431873	ATRIAL FIBRILLATION

(iv) OUTC (Outcomes)

primaryid	outc_code
205431871	HO # Hospitalization
205431872	DE # Death
205431873	LT # Life-threatening

7. Data Mining Problem Definition:

Three Interconnected Unsupervised Learning Problems:

(a). Association Rule Mining

Objective: Discover patterns: "IF [Drug A + Drug B] AND [Age > 65] THEN [Adverse Event X]"

- Drug combinations causing severe events
- Demographic factors linked to side effects
- Temporal sequences: Drug A → Drug B → Event C

Success Metrics:

- Support ≥ 0.001, Confidence ≥ 0.60, Lift ≥ 2.0
- 15-20 clinically significant rules
- 10+ unknown drug-drug interactions
- Precision@20 ≥ 0.60 against FDA warnings

(b). Clustering Analysis

Targets: Drugs with similar safety profiles, patients with similar susceptibility, co-occurring adverse events

Success Metrics:

- Silhouette ≥ 0.35, Davies-Bouldin ≤ 1.0, Calinski-Harabasz > baseline
- 10-15 drug clusters, 5-7 patient risk groups
- Within-cluster similarity > 0.70

(c) . Dimensionality Reduction

Applications: 2D/3D drug safety visualization, patient risk trajectories, temporal pattern evolution

Success Metrics:

- Explained variance ≥ 80% (PCA)
- Trustworthiness ≥ 0.85, Continuity ≥ 0.85
- Clear drug class separation, smooth temporal transitions

6.Validation Strategy

Association Rules

- Medical Literature: Compare with DrugBank, FDA black box warnings (2023-2025)
- Temporal: Train on 2023-2024, test on 2025; track quarterly evolution
- Clinical Review: Top 20 rules categorized as Known/Novel/Spurious

Clustering

- Manual Inspection: 10 drug pairs per cluster for profile verification
- External: ATC classification comparison, withdrawn drug groupings
- Stability: Bootstrap (100 iterations), cross-quarter consistency

Dimensionality Reduction

- Neighborhood: Similar drugs remain proximate, recalled drugs as outliers
- Temporal: Track quarterly positions, ensure gradual signal emergence

Cross-Validation Dataset Split

- Historical Set (2023 Q1-Q2): Known FDA warnings
- Training (2023 Q3-2024 Q2): Pattern discovery
- Validation (2024 Q3-Q4): Parameter tuning
- Test (2025 Q1-Q3): Final evaluation
- Stratified Testing by: Age groups, therapeutic areas, report sources, outcome severity

7. Success Criteria

- 30% patterns confirmed in literature
- 5+ novel findings for investigation
- Expert interpretability score ≥ 4/5
- Early detection of 2+ drugs requiring FDA warnings

8. Expected Impact

- Early warning system for drug safety signals
- Risk stratification for personalized medicine
- 10-20 actionable safety insights

This comprehensive validation ensures unsupervised discoveries translate into clinically relevant findings for improved patient safety and regulatory decisions.