**Milestone 3 – White Paper**

Akshay Sharma

Department of Data Science, Bellevue University

DSC680-T302: Applied Data Science

Amirfarrokh Iranitalab

October 27, 2025

Detecting Prescription Fraud and Drug Diversion Using Data Analytics

## **Detecting Prescription Fraud and Drug Diversion Using Data Analytics**

This project focuses on using data analytics to detect fraudulent prescribing and dispensing patterns that may indicate prescription fraud or drug diversion. By identifying anomalies early, pharmacies and regulatory bodies can reduce financial losses and improve patient safety.

**Business Problem**

Prescription fraud and drug diversion continue to challenge the U.S. healthcare system, contributing to escalating costs and the ongoing opioid crisis. Fraudulent behaviors—such as forged prescriptions, “doctor shopping,” and overprescribing of controlled substances—are often difficult to detect using traditional rule-based systems. These methods rely on static thresholds and manual review, which cannot adapt to emerging fraud schemes.  
This project proposes a data-driven anomaly detection framework to identify suspicious prescribing and dispensing activities in real time. By leveraging unsupervised machine learning, pharmacies, insurers, and regulators can proactively detect unusual patterns and mitigate misuse of controlled substances.

**Background / History**

Historically, prescription oversight was fragmented across agencies and states. The **DEA’s Automated Reports and Consolidated Orders System (ARCOS)** and state-based **Prescription Drug Monitoring Programs (PDMPs)** were created to track controlled substances through the supply chain. However, data remained siloed, delaying detection and response.  
In the 2010s, as the opioid epidemic intensified, the need for integrated data analytics became clear. Advances in cloud computing and machine learning now enable real-time analysis of millions of transactions to identify patterns indicative of diversion or fraud.

**Figure 1. Project Overview Diagram**  
*(A conceptual flowchart illustrating data sources such as DEA ARCOS, CMS, and PDMP feeding into Data Preparation → Anomaly Detection → Flagged Entities → Regulatory Actions.)*

A diagram with text on it

AI-generated content may be incorrect.

**Data Explanation**

**Data Sources**

* **DEA ARCOS Dataset:** Tracks the flow of controlled substances from manufacturers to pharmacies.
* **CMS Medicare Part D Prescriber Data:** Contains prescription claims by provider, drug, and location.
* **Synthetic PDMP Dataset:** Represents anonymized records of prescriber–pharmacy–patient interactions.

**Data Preparation**

Data are cleaned, standardized, and integrated using prescriber and pharmacy identifiers. Continuous variables such as dosage and quantity are normalized. Missing data are imputed where appropriate. Categorical features like payment type and prescriber specialty are encoded.

**Figure 2. Data Preparation Workflow**  
*(A horizontal diagram showing sequential steps: Data Cleaning → Normalization → Feature Engineering → Integration → Model Input.)*

A diagram of a data processing workflow

AI-generated content may be incorrect.

**Sample Data Dictionary**

| **Variable** | **Description** | **Data Type** |
| --- | --- | --- |
| prescriber\_id | Unique ID for prescribing provider | String |
| drug\_name | Name of controlled substance | String |
| quantity | Total units dispensed | Numeric |
| dosage\_mg | Average dosage strength | Numeric |
| payment\_type | Cash, insurance, or government plan | Categorical |
| state | Prescriber’s state | String |
| prescription\_date | Date of prescription issue | Date |

**Methods**

Due to limited labeled data, the study uses **unsupervised anomaly detection** methods to identify abnormal prescribing behavior.

* **Exploratory Data Analysis (EDA):** Identifies distribution patterns and potential outliers.
* **Feature Engineering:** Derived features such as prescription frequency, dosage ratios, and payment proportions.
* **Anomaly Detection Models:**
  + *Isolation Forest* – isolates rare, irregular patterns.
  + *Local Outlier Factor (LOF)* – detects density-based anomalies.
  + *DBSCAN clustering* – identifies unusual clusters of prescribers or pharmacies.
* **Geospatial and Temporal Analysis:** Detects geographic hotspots and sudden volume spikes.
* **Validation:** Compares flagged entities to public enforcement or policy reports.

**Figure 3. Anomaly Detection Architecture Diagram**  
*(A conceptual block diagram depicting: Input Data → EDA & Feature Engineering → Anomaly Detection Algorithms (Isolation Forest, LOF, DBSCAN) → Flagged Outliers → Analyst Review Dashboard.)*

**A close-up of a diagram

AI-generated content may be incorrect.**

**Analysis**

Exploratory analysis of synthetic data revealed that most prescribers fall within expected dosage and prescription ranges, while a small subset exhibited significantly higher metrics.

**Figure 4. Scatterplot of Prescriber Patterns Highlighting Outliers**  
*(Displays total prescriptions vs. average dosage; normal prescribers form a cluster while red markers indicate outliers.)*

**A graph with blue dots and white text

AI-generated content may be incorrect.**

**Figure 5. Synthetic U.S. Heatmap Showing Prescription Hotspots**

*(A U.S. map illustrating higher prescription intensity in Florida, Ohio, and West Virginia.)*

A map of the united states

AI-generated content may be incorrect.

These anomalies align with known geographic trends of opioid misuse. Time-series analysis further revealed spikes in prescription volume during specific months, suggesting potential diversion patterns tied to policy or supply events.

**Figure 6. Prescription Volume Distribution**

A graph of a number of blue and black bars

AI-generated content may be incorrect.

A graph of a prescription volume distribution

AI-generated content may be incorrect.

**Figure 7. Correlation Heatmap of Key Prescription Features**

A screenshot of a diagram

AI-generated content may be incorrect.

**Conclusion**

The analysis indicates that unsupervised machine learning, particularly Isolation Forest and LOF, can successfully flag abnormal prescribing behaviors. When coupled with geospatial visualization, these methods can enhance the ability of regulators and pharmacies to act proactively, reducing fraud risk and improving patient safety.

**Assumptions**

* The data accurately represents prescription behavior.
* Fraudulent activity manifests as measurable statistical anomalies.
* Synthetic data are representative of real-world patterns.

**Limitations**

* Lack of true fraud labels limits direct accuracy validation.
* Aggregated datasets reduce granularity at the patient level.
* Risk of false positives requires manual verification.

**Challenges**

* **Data Imbalance:** True fraud cases are rare compared to normal prescribing, making model training and validation difficult.
* **Data Quality & Availability:** Public datasets may be aggregated or delayed, requiring careful preprocessing and possible use of synthetic augmentation.
* **Evolving Fraud Schemes:** Fraud patterns change rapidly; models must adapt to new behaviors over time.
* **Ethical & Legal Sensitivity:** Extra caution is needed when analyzing data that involves real prescribers and controlled substances.

**Future Uses / Additional Applications**

* Expansion to telehealth and mail-order pharmacy analysis.
* Integration with **NLP** tools to analyze prescription notes.
* Adoption of **blockchain** for end-to-end supply transparency.
* Real-time dashboards for policymakers and healthcare organizations.

**Recommendations**

1. Create a centralized anomaly detection platform combining DEA, CMS, and PDMP data.
2. Implement explainable AI modules to enhance interpretability.
3. Conduct quarterly retraining of models with updated data.
4. Establish cross-agency collaboration for data consistency and sharing.

**Implementation Plan**

| **Phase** | **Objective** | **Deliverables** |
| --- | --- | --- |
| **1. Data Integration** | Merge ARCOS, CMS, PDMP data | Unified data repository |
| **2. Model Development** | Train anomaly detection algorithms | Isolation Forest & LOF results |
| **3. Validation** | Evaluate with known enforcement cases | Performance report |
| **4. Deployment** | Integrate dashboards into compliance workflows | Operational monitoring system |

**Ethical Assessment**

* **Privacy & HIPAA:** Patient and prescriber data must be anonymized; any real data must comply with federal and state privacy laws.
* **False Positives:** Incorrectly flagging a prescriber or pharmacy could harm reputations and patient care — models must be interpretable and human-reviewed.
* **Bias & Fairness:** Avoid disproportionate targeting of specific regions, specialties, or patient populations.
* **Transparency:** Findings should support explainable decision-making, not automated punitive actions.

**FAQ’s**

**1. How does the model distinguish between legitimate high-volume prescribers and fraudulent ones?**

The model evaluates prescribers based on multidimensional statistical profiles—such as dosage per prescription, refill frequency, payment method mix, and patient volume—relative to peers within the same specialty and region. Legitimate high-volume prescribers typically have balanced metrics aligned with patient load, whereas fraudulent prescribers show disproportionate spikes in dosage, refill frequency, or cash payments.

**2. What safeguards ensure HIPAA compliance and data anonymity?**

All datasets are de-identified before analysis, removing personal identifiers such as patient names, addresses, and SSNs. Aggregation at the prescriber or pharmacy level ensures that individuals cannot be re-identified. Additionally, any real-world implementation would require compliance with HIPAA’s *minimum necessary rule* and institutional data-use agreements.

**3. Why were unsupervised methods chosen over supervised learning?**

Because verified labels of fraud cases are rare and often unavailable, unsupervised methods like Isolation Forest, Local Outlier Factor (LOF), and DBSCAN are ideal for discovering unknown or evolving anomalies without prior examples. They help detect both new and subtle fraud patterns that rule-based or supervised models might miss.

**4. How will the model adapt to newly emerging fraud schemes?**

Model retraining is scheduled quarterly using recent data. Drift detection metrics track when prescription behavior changes significantly, prompting adaptive threshold updates. Integration of streaming data pipelines allows near real-time recalibration, while human analysts validate newly detected anomalies to reinforce learning.

**5. Can this system integrate with existing pharmacy or insurance workflows?**

Yes. The model’s outputs—risk scores, anomaly flags, and prescriber summaries—can feed directly into existing pharmacy claim systems, PDMP dashboards, and insurer fraud-detection platforms through API integrations. This minimizes workflow disruption while enhancing decision support for claims and compliance teams.

**6. What performance metrics (precision, recall, F1) will define success?**

Success will be measured by **precision** (proportion of correctly flagged frauds), **recall** (proportion of actual frauds identified), and **F1-score** (balance of both). Since fraud datasets are imbalanced, **precision@k** (top-ranked anomalies) and **false positive rate** per review cycle will be critical operational KPIs.

**7. How are false positives managed to protect reputations?**

Each flagged prescriber is subject to a two-stage review: (1) automated scoring with explainable AI (SHAP-based feature importance) and (2) manual verification by compliance analysts. Only validated anomalies trigger further investigation, ensuring fairness and minimizing reputational harm.

**8. Which U.S. regions show the highest anomalies, and what policy actions are recommended?**

Synthetic analysis identified higher anomaly densities in **Florida, Ohio, and West Virginia**, aligning with historical opioid crisis hotspots. Recommended policy actions include reinforcing PDMP data sharing, increasing prescriber audits in high-risk counties, and offering prescriber education on controlled-substance protocols.

**9. What collaborations are needed between DEA, CMS, and state PDMPs for scaling?**

A unified data governance framework is essential—linking ARCOS, CMS, and PDMP systems via standardized prescriber identifiers. Joint oversight committees could coordinate anomaly verification, share cross-agency intelligence, and standardize alert formats for nationwide consistency.

**10. How could future enhancements (e.g., NLP, blockchain) improve fraud detection?**

NLP models could analyze prescription notes and patient records for semantic cues of misuse, while blockchain could ensure immutable audit trails for controlled-substance transactions, improving transparency, accountability, and cross-agency trust.

**References (APA Style)**

Centers for Medicare & Medicaid Services. (2023). *Medicare Part D Prescriber Data*. Retrieved from <https://data.cms.gov>

U.S. Drug Enforcement Administration. (2023). *ARCOS Data Release*. Retrieved from <https://www.deadiversion.usdoj.gov/arcos>

U.S. Department of Health and Human Services. (2021). *Combating Prescription Drug Fraud, Waste, and Abuse.*

Van Hout, M. C., & Norman, I. J. (2016). Misuse of prescription drugs: Epidemiology, policy, and program responses. *International Journal of Drug Policy*, 38, 1–6.

Wang, H., & Xu, J. (2018). A data-driven approach for detecting prescription fraud. *IEEE Transactions on Information Forensics and Security*, 13(11), 2775–2788.