API-Driven: Drug Data Management and Integration System

Data Management Project Report

Academic Year: 2025-2026

Submission Date:

Presented By

Abdul Sami Student Id: 939542	Davina David Student Id: 946851	Shiraz Ahmed Student Id: 948290

Presented To

Prof. Andrea Maurino

Abstract

This report investigates the accuracy of RxCUI identifier extraction from the RxNorm REST API when integrated with drug label information from the OpenFDA API. Reliable drug identifiers are crucial for effective healthcare data integration and interoperability across clinical systems. Using automated Python-based data acquisition and integration workflows, the study evaluates whether RxNorm reliably provides RxCUI identifiers and whether key drug-related fields remain complete and consistent in the combined dataset. The results indicate that RxNorm consistently delivers accurate RxCUI identifiers, and integration with OpenFDA data enhances the completeness and regulatory context of the dataset. These findings support the use of RxNorm in conjunction with OpenFDA as a dependable source for drug information in healthcare data management applications.

Keywords

RxNorm, Rxcui, Accuracy, Completeness

Project Work Distribution

Component	Team Member	Responsibility
Data Acquisition	Sheraz Ahmed	To Gather and Evaluate relevant Web Apis
Data Integration	Abdul Sami	To merge RxNorm and OpenFDA data into a consistent, unified dataset, linking identifiers
Data Pre-Processing	Sheraz Ahmed and Abdul Sami	To clean and standardize the two API dataset
Data Storage	Abdul Sami	To storing Datasets in Sql and apply Filteration
Data Exploratory Analysis	Davina David	Assess Data Quality through parameters
Reporting	Davina David	To compile Report and Slides with references

Table of Contents

1. Introduction

The increasing digitization of healthcare has led to the generation of vast amounts of data, offering significant potential to improve patient care, support informed decisions, and streamline healthcare operations. However, effectively managing, analyzing, and deriving insights from this data remains a considerable challenge.

Healthcare professionals often face challenges when managing large datasets, especially when integrating information from multiple sources. Ensuring the accuracy and completeness of drug-related data is essential for informed decision-making, safe drug prescription, and effective patient care.

This project focuses on acquiring, integrating, and analyzing drug-related data using various APIs, with a specific emphasis on the RxNorm API. Our report aims to evaluate "How accurate is the RxNorm API extraction in providing RxCUI identifiers when combined with OpenFDA data? By addressing this question, we aim to evaluate the reliability of RxNorm as a standardized source of drug information.

The study involves sourcing data from multiple platforms to ensure a comprehensive view, covering aspects such as drug characteristics, dosages, classifications, and interactions. These datasets are integrated into a unified structure to facilitate efficient data usage, including extraction, cleaning, and merging processes. Rigorous validation and quality checks are applied to ensure that the information is accurate, complete, and suitable for healthcare applications.

The project addresses the challenge of managing and integrating large volumes of healthcare data. Accurate and complete drug information is critical for informed decision-making and patient safety. The team's goal is to evaluate the reliability of the RxNorm API for Rxcui extraction and overall data completeness.

2. Research Questions

Q: How accurate is the RxNorm API extraction in providing RxCUI identifiers when combined with OpenFDA data?

3. Data Source

This study utilizes two publicly available web APIs to collect and analyze drug-related data: the **RxNorm API** and the **openFDA Drug API**. The RxNorm API serves as the primary data source for standardized drug identifiers and nomenclature, while the openFDA Drug API is used as a reference source for regulatory and safety-related information. Using these two complementary sources enables the evaluation of the accuracy and completeness of RxNorm data by comparing it against authoritative regulatory records.

3.1 RxNorm API

The RxNorm API, maintained by the U.S. National Library of Medicine, provides a standardized drug nomenclature system designed to unify drug information across healthcare applications. It assigns unique identifiers known as **RxNorm Concept Unique Identifiers (Rxcui)** to drugs, enabling consistent representation of drug names across different systems and databases. The API provides detailed drug-related information, including drug names, dosages, and classifications, in JSON format. Due to its authoritative maintenance and widespread adoption, RxNorm is a reliable source for standardized drug data.

3.2 openFDA Drug API

The openFDA Drug API, maintained by the U.S. Food and Drug Administration (FDA), provides comprehensive regulatory information on drugs approved for use in the United States. It includes data on FDA approvals, safety warnings, adverse events, and drug labeling, delivered in JSON format. As the information is sourced directly from the FDA, the openFDA API serves as a trusted benchmark for validating regulatory status and safety-related attributes of drugs referenced in the RxNorm dataset.

4. Data Acquisition

The data acquisition phase involved collecting drug-related information from two publicly available web APIs: the RxNorm API and the openFDA Drug API. The RxNorm API was used to obtain standardized drug identifiers and nomenclature, while the openFDA Drug API provided regulatory and safety-related information.

Data collection was automated using Python scripts, which issued HTTP GET requests to the APIs with drug names as input parameters. The APIs returned structured responses in JSON format, which were programmatically retrieved to ensure consistency and reproducibility of the acquisition process. The acquired data were then parsed to extract key attributes such as drug names, RxNorm Concept Unique Identifiers (Rxcui), drug purpose, and safety warnings for subsequent preprocessing, integration, and analysis.

5. Data Integration

Data integration was performed by programmatically combining drug-related information obtained from the RxNorm and openFDA APIs into a unified data structure. The RxNorm Concept Unique Identifier (Rxcui) was used as a standardized identifier to represent drugs consistently across sources. Integration was implemented using Python scripts that extracted relevant attributes from each API response and consolidated them into a single record for each drug. Where information from the openFDA API was unavailable, default placeholder values were assigned to maintain structural consistency. The integrated dataset was then stored in a relational SQLite database, enabling unified access to standardized identifiers, safety warnings, and drug purpose information for subsequent analysis.

6. Data Pre-Processing

Before storing the acquired data, the JSON responses from the RxNorm and openFDA APIs were parsed to extract key attributes such as drug names, RxNorm Concept Unique Identifiers (Rxcui), drug purpose, and safety warnings. Missing or unavailable information from the APIs ensuring that all records were consistent and aligned with the database schema. This initial preprocessing step prepared the data for reliable insertion into the SQLite database and maintained structural consistency across all entries.

7. Data Storage

The integrated drug data obtained from the RxNorm and openFDA APIs is stored locally using a relational database implemented with SQLite. A structured table named drug_data is created to persist key attributes, including the drug name, RxNorm Concept Unique Identifier (Rxcui), safety warnings, and therapeutic purpose. SQLite is accessed programmatically through Python's built-in sqlite3 library, enabling seamless insertion, querying, and management of records. To support data quality and scalability for analysis, the database stores multiple entries and allows bulk insertion of records. Additional validation steps are performed after storage, including checks for missing values and duplicate records, with duplicates removed directly at the database level to maintain data integrity. This approach provides a lightweight, reliable, and reproducible storage solution suitable for structured healthcare data.

8. Data Quality Check

After storing the data in the SQLite database, the dataset underwent validation to ensure data integrity and quality. Checks were performed for missing values and duplicate entries. Duplicate records were removed while retaining the first occurrence, ensuring the uniqueness of each entry. These validation steps produced a clean, reliable dataset ready for further analysis and integration.

9. Data Exploratory Analysis

The analysis showed that the RxNorm API consistently returned valid Rxcui for commonly queried drug names, demonstrating high accuracy in identifier extraction. The majority of API responses included all essential identification fields, indicating strong data completeness. However, some responses lacked extended descriptive attributes beyond basic identification. While this does not affect identifier accuracy, it suggests that RxNorm is primarily optimized for standardization rather than comprehensive clinical detail.

9.1 Data Quality Assessment:

After completing the data integration and analysis phases, several findings emerged that highlight both the strengths and areas for improvement in the dataset. The following summarizes the key results of the analysis:

- **Data Accuracy**

The accuracy of **our** dataset is largely dependent on the **source APIs**: RxNorm and openFDA, both of which are reputable, authoritative sources for drug-related information. RxNorm provides **standardized identifiers (**Rxcui**)** and nomenclature, ensuring that drug names are uniquely and correctly represented. openFDA supplies regulatory information, including drug purpose and safety warnings, which are accurate as per FDA records. However, some limitation exist:

Random data was fetched from the two source apis for testing purpose.

- **Data Consistency**

Consistency is ensured through multiple mechanisms in our workflow:

- Standardized identifiers (Rxcui) link data across RxNorm and openFDA.
- Python scripts enforce uniform data types and field structures before storage.
- Post-storage preprocessing removes duplicate entries and aligns missing values with placeholders.

These steps maintain **internal consistency**, ensuring that every record conforms to the same schema.

10. Software Architecture

The data management pipeline was implemented using a combination of **Python programming**, **web APIs**, and a **relational database** to support structured storage, integration, and analysis of drug-related information.

Tools and Technologies Used:

- **Python:** The primary programming language used for scripting the data acquisition, integration, and preprocessing workflows. The requests library was utilized for making HTTP GET requests to web APIs, while json handled parsing and structured storage of API responses. The pandas library was used for post-storage data validation and analysis.
- **RxNorm API:** A publicly available API providing standardized drug identifiers and nomenclature, used for retrieving Rxcui and related drug data.
- **openFDA Drug Label API:** Provides regulatory, safety, and purpose-related information for drugs, complementing the standardized RxNorm data.
- **SQLite:** A lightweight relational database used to store integrated drug data. Python's sqlite3 library was employed for creating tables, inserting records, and executing queries. This allowed for efficient data storage, retrieval, and post-storage validation.

The overall architecture follows a **modular ETL-like approach**, where data is **extracted** from external APIs, **transformed** through preprocessing and integration, and **loaded** into a structured SQLite database. This setup ensures reproducibility, maintainability, and scalability, making it suitable for future extensions such as analytics or machine learning applications.

11. Conclusion and Future Work

This technical evaluation confirms that the RxNorm API is a reliable source for accurate Rxcui extraction and basic drug identification. Its high accuracy and strong completeness for core fields make it suitable for healthcare data integration and interoperability tasks. Future work may involve integrating additional APIs to enrich drug profiles, performing cross-source validation to further assess accuracy, and applying more advanced data quality metrics such as consistency and timeliness.

12. References

1. RxNorm API Documentation

URL: https://rxnav.nlm.nih.gov/RxNormAPIs.html

2. openFDA Drug API Documentation

URL: https://open.fda.gov/apis/drug/

3. SQLite Documentation

URL: https://sqlite.org/docs.html