Safety and Adverse Event Characterization of Type 2 Diabetes Treatments Using R

Final Project Proposal

Team Members: Jiayi Chu, Yixin Xue, Shuya Guo, Runjiu Chen

2025-11-14

1. Overview

Project Title: Safety and Adverse Event Characterization of Type 2 Diabetes Treatments Using R

Team Name: Jet2Holiday

Team Members:

- Jiayi Chu - jchu47
- Yixin Xue - yxue39
- Shuya Guo - sguo66
- Runjiu Chen - rchen137

Submitted: November 14, 2025

2. Background and Importance

Type 2 diabetes mellitus (T2DM) is a highly prevalent chronic disease that requires long-term pharmacologic management across multiple mechanistic classes of medications. As newer therapies such as GLP-1 receptor agonists, dual GIP/GLP-1 agonists, and SGLT2 inhibitors are increasingly prescribed alongside older agents like metformin, DPP-4 inhibitors, and basal insulin, the overall safety landscape has grown more complex. Each drug class carries distinct adverse event profiles that may influence treatment decisions, patient adherence, and clinical outcomes. However, real-world safety information is vast, difficult to navigate, and often not presented in a way that allows meaningful comparisons across medications. Understanding how adverse events differ between drug classes is therefore essential not only for clinicians and patients, but also for researchers, safety analysts, and individuals involved in drug development.

3.Project Description

This project aims to build an R-based analytical tool that systematically characterizes and compares adverse event patterns across commonly used T2DM medications. Using a fully reproducible R workflow, the project integrates name standardization, event classification, descriptive statistics, and signal detection to generate clear safety summaries for each drug and drug class. The final

product will be an interactive dashboard that allows users to explore adverse event frequencies, serious event proportions, class-level comparisons, and temporal patterns. While not intended to replace the specialized pharmacovigilance systems used by regulatory agencies or pharmaceutical companies, this tool provides a transparent and accessible way to examine the safety profiles of multiple T2DM treatments in one place. It offers clinicians, students, and early-career researchers a practical starting point for understanding real-world medication risks and for conducting comparative safety evaluations using modern analytical techniques.

4.Existing Work

Post-marketing safety studies using FAERS have examined individual T2DM drug classes, such as GLP-1 receptor agonists, SGLT2 inhibitors, and DPP-4 inhibitors, and have identified known risks including gastrointestinal events, diabetic ketoacidosis, and pancreatitis. However, most published FAERS studies analyze only one drug or one class at a time, and there are very few cross-class comparative analyses covering all major modern T2DM therapies within a unified framework. This leaves a gap for systematic comparisons across mechanisms.

5. Data Collection and Database Integration

We collect data by programmatically retrieving all quarterly ASCII submissions from the official adverse event reporting system using an automated R workflow. This process extracts download links directly from the source webpage, downloads all available quarterly ZIP archives for the specified period, and imports the included ASCII datasets into R. The raw data consist of multiple standardized tables, covering drug exposures, reported adverse events, demographic variables, outcomes, indications, and therapy information, which together provide the foundational structure required for comprehensive adverse event analysis.

After downloading, we follow a structured multi-step data-processing pipeline to prepare the dataset for analysis. This includes cleaning and standardizing field formats, removing duplicate case versions, filtering implausible values, and harmonizing drug names through both fuzzy matching and the RxNorm API to ensure consistent identification of our target medications. We also map adverse event preferred terms to system-organ classes using MedDRA, retrieve WHO ATC codes to classify drugs into mechanistic categories, and finally integrate all component tables by CASEID to form a complete analysis-ready dataset. This reproducible workflow ensures that our data collection and preprocessing are both reliable and fully automated within R.

To support efficient handling of the multi-million record FAERS dataset, we store all standardized tables in a relational SQLite database using the DBI and RSQLite packages. After preprocessing each quarterly file, the DRUG, REAC, DEMO, OUTC, INDI, and THER tables are written into the database as separate relations indexed by CASEID. This enables fast and structured querying during analysis. Throughout the project, subsets of interest, such as reports for a specific drug, a particular mechanistic class, or a selected time period, are extracted directly from the database using SQL queries within R. This database component ensures scalable data management and demonstrates the required database programming paradigm for the project.

6. Selected Drugs & Mechanistic Classes

We selected **15 FDA-approved T2DM medications** across **5 mechanistic classes**:

Mechanistic Class	Generic Name	ATC Code
Biguanide	Metformin	A10BA02
GLP-1 receptor agonist	Semaglutide	A10BJ06
	Liraglutide	A10BJ02
	Dulaglutide	A10BJ05
	Exenatide	A10BJ01
SGLT2 inhibitor	Empagliflozin	A10BK03
	Dapagliflozin	A10BK01
	Canagliflozin	A10BK02
	Ertugliflozin	A10BK04
DPP-4 inhibitor	Sitagliptin	A10BH01
	Saxagliptin	A10BH03
	Linagliptin	A10BH05
	Alogliptin	A10BH04
Basal insulin	Insulin glargine	A10AE04
	Insulin degludec	A10AE06

7. Programming Paradigms

In this project, we integrate several programming paradigms within an R-based analytical framework to support automated data ingestion, structured processing, reproducible analysis, and interactive visualization.

- **Command-line programming:**

We use command-line scripts to automate the batch downloading and extraction of all quarterly FAERS archives and to trigger R preprocessing scripts through Rscript, enabling reproducible and efficient data ingestion.

- **Functional programming:**

Functional tools such as dplyr, purrr, and custom helper functions are used to create reusable pipelines for tasks including filtering case records, joining multi-table data, mapping drugs to mechanistic classes, and computing summary metrics.

- **Object-oriented programming:**

We implement simple S3 or R6 structures to encapsulate repeated analytical components, such as drug-level profiles or forecasting modules, allowing drug-specific summaries, plots, and analyses to be generated through associated methods.

- **Machine learning paradigms:**

Time-series forecasting models (e.g., ARIMA or Prophet-style approaches) are applied to analyze longitudinal reporting patterns and identify potential emerging safety signals within the adverse event data.

8. Software, Packages and Tools

● **Software:**

We will use the R programming language for data processing, analysis, and dashboard development. Our work will be conducted using two IDEs: **RStudio** and **Visual Studio Code (VS Code)**.

● **Core R Packages for Data Processing and Functional Programming:**

○ **tidyverse** (dplyr, tidyr, purrr, stringr) for data wrangling, functional pipelines, multi-table integration, and modularized processing functions.

○ **data.table** for efficient handling of large FAERS datasets.

○ **readr** for standardized import of ASCII files.

● **Packages for Drug Name Standardization and Classification:**

○ **httr** and **jsonlite** for accessing the RxNorm API to normalize drug names.

○ **rvest** for web scraping ATC classifications and related drug information.

● **Database and SQL Integration:**

○ **DBI** and **RSQLite** for constructing relational database structures and running SQL queries to join tables, filter cases, and aggregate adverse event records efficiently.

● **Adverse Event Mapping and Signal Detection:**

○ **epitools** or custom functions for computing PRR and ROR.

○ **forcats** for factor handling in MedDRA PT-to-SOC mapping workflows.

● **Visualization Packages:**

○ **ggplot2** for creating core statistical and comparative visualizations.

○ **plotly** for interactive graphics and exploratory data displays.

○ **DT** for searchable interactive tables within the dashboard.

● **Machine Learning and Time-Series Forecasting:**

○ **forecast** or **fable** for ARIMA-based time-series modeling.

○ **prophet** for machine learning–oriented forecasting.

○ **tsibble** for managing time-indexed adverse event data.

● **Dashboard and Reactive Programming:**

○ **Shiny** and **shinydashboard** for building the interactive dashboard.

○ **shinyWidgets** and **bslib** for enhanced UI components and modern theming.

9. Data Analytic Product

Our final data product will be an interactive, R-based dashboard built using Shiny. The dashboard will be organized into six pages, each corresponding to a different analytical perspective on ad-

verse events reported for type 2 diabetes treatments. These pages together will cover class-level comparisons, individual drug summaries, temporal patterns, and methodological documentation.

Page 1 — Home / Overview

This page will introduce the project context, research motivation, main research questions, and a brief summary of the analytical workflow. It will also describe the included medications and provide basic instructions for navigating the dashboard.

Page 2 — Global Trends

This page will display aggregated reporting trends for all selected medications from 2019–2021. Planned visualizations include annual reporting counts, changes over time, distributions of serious versus non-serious outcomes, and high-level patterns across major mechanistic classes (e.g., GLP-1 receptor agonists, SGLT2 inhibitors, DPP-4 inhibitors, biguanides, and basal insulins).

Page 3 — Mechanism Comparison

This section will focus on comparisons between drug classes. It will include summaries of system-organ class (SOC) distributions, class-level disproportionality metrics such as PRR/ROR, and ranked lists of the most frequently reported preferred terms (PTs) within each mechanistic class.

Page 4 — Individual Drug Profiles

On this page, users will be able to select any of the 15 included medications (for example, semaglutide, tirzepatide, or empagliflozin) and view drug-specific summaries. Outputs will include reporting counts over time, top adverse events, SOC distributions, and signal-detection statistics. Tables and plots will update reactively based on the selected drug.

Page 5 — Temporal & Emerging Signals (with Forecasting)

This page will examine longitudinal trends and potential emerging signals. It will present time-series plots of adverse event reporting for selected drugs or classes and apply time-series forecasting methods (e.g., ARIMA or Prophet-style models) to estimate expected reporting patterns. Deviations from these forecasts will be used to explore potential emerging safety patterns.

Page 6 — Methods & Downloads

The final page will document the data processing pipeline, definitions of key analytic measures (such as PRR and ROR), and the approach used for drug classification. It will also provide access to downloadable summary tables and reports, as well as a link to the project's GitHub repository and a brief description of team roles.

10. Timeline

● **Nov 11–14:** Draft and submit the final project proposal (research questions, data plan, programming paradigms, and dashboard design).

● **Nov 17–21:** Meet with the instructor/TAs to discuss feasibility; refine the scope, finalize the list of 15 drugs and mechanistic classes, and lock in the analysis plan.

● **Nov 22–30:** Implement data acquisition and preprocessing pipeline in R (automated FAERS download, table import, de-duplication, drug name standardization, MedDRA SOC mapping, ATC classification, and case-level integration).

● **Dec 1–7:** Develop core analytical modules (descriptive summaries, class-level comparisons, PRR/ROR calculations) and build initial Shiny dashboard structure with the six main pages.

● **Dec 8–12:** Add temporal and forecasting components (time-series trend plots and ARIMA/Prophet-style models), refine visualizations and interactivity, and prepare final project presentation slides.

● **Dec 13–18:** Finalize the Shiny app and GitHub repository, complete the written project report for GitHub Classroom, perform internal testing, and submit the group participation self- and peer-evaluations.

11. Task Allocation

● **Shuya Guo:** Develop **Page 2 – Global Trends**, including overall reporting trends, serious vs. non-serious event distributions, and cross-class comparisons.

● **Jiayi Chu:** Develop **Page 3 – Mechanism Comparison**, focusing on SOC-level heatmaps, class-level PRR/ROR summaries, and ranked adverse event patterns across mechanisms.

● **Yixin Xue:** Develop **Page 4 – Individual Drug Profiles**, implementing drug-level reporting trends, top adverse events, SOC breakdowns, and signal detection tables with reactive updates.

● **Runjiu Chen:** Develop **Page 5 – Temporal & Emerging Signals**, including time-series visualization and machine learning forecasting models (ARIMA/Prophet-style).

● **Team Shared Responsibilities:** Build Page 1 (Home/Overview) and Page 6 (Methods & Downloads) collaboratively; implement data collection and preprocessing pipelines (FAERS import, standardization, MedDRA/ATC mapping); set up the Shiny framework; and ensure overall integration and quality control across all six pages.