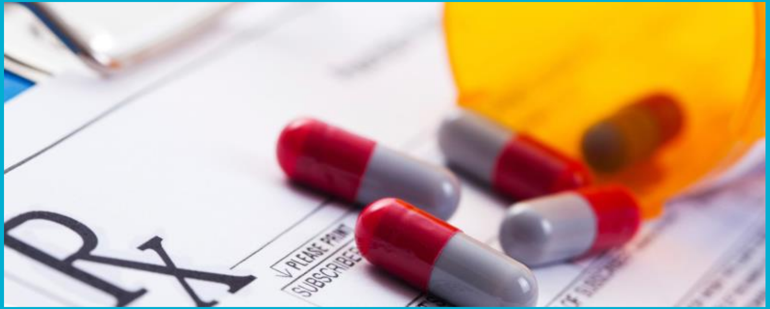
Predicting adverse drug reactions based on patient demographic information and other factors



James Lagrotteria, Anthony Olvera, and Annie Lydens

Project Description

For our semester project we intend to analyse the FDA's drug adverse events data collection using data mining techniques in hopes of obtaining interesting findings. We're particularly interested in looking at how outcomes correlate with a variety of influencing factors from demographic information (such as sex, age and location) to drug-specific information (such as type, manufacturer, dosage and administration method). Ideally, this will lead us to interesting predictive analyses and/or knowledge discovery. Examples include *"If you are age x, the likelihood of product abc has a higher chance of causing a significant medical outcome"* and *"patients in the UK suffered from the most adverse reactions to drug xyz ".*

Potential Interesting Questions:

- What demographic information is correlated most strongly with the most adverse reactions, or the most significant adverse reactions?
- Given a patient's demographic information (age, sex, weight, reported location), what is the likelihood of a specific class of drug causing a significant medical outcome?
- What set of most frequent patient characteristics tend to lead to a specific adverse drug reactions?

Prior Work

Maciejewski, M. et al. Reverse translation of adverse event reports paves the way for de-risking preclinical off-targets. eLife 6, (2017).

https://elifesciences.org/articles/25818

This study classified drugs by their active ingredients in order to correlate ingredients and outcomes.

A curated and standardized adverse drug event resource to accelerate drug safety research

https://www.nature.com/articles/sdata201626

This paper discusses some de-duping efforts, as well as value imputation across cases, which is something we should consider.

A method for estimating the probability of adverse drug reactions

https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1038/clpt.1981.154

Predicting the probability of adverse drug reactions based on clinical judgment and statistical analysis

Datasets

Dataset Summary

- List of datasets to use
    - The datasets utilized will be the FDA’s Adverse Event Reporting System (FAERS) Quarterly Extract datasets for the period of 2016Q1-2018Q1
- Where found (URL and who is supplying the data, e.g., NASA)
    - The FAERS quarterly data can be found on the FDA’s website and is publically available at the following location: https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html
    - Data Dictionary: TBD
- Whether it you have it downloaded (on who’s machine)
    - We have the data loaded into a Postgres relational database.
    - We also have the data set downloaded locally to Anthony’s 1.5 TB external drive as backup.

Proposed Work

There are known issues with the FAERS dataset, including duplicate reports of the same adverse event, duplicate cases being reported under different case numbers, symptoms of the disease being treated being reported as an adverse event, and multiple different names for drugs containing the same active ingredients.

Cleaning:

- Missing value imputation before removing duplicates
- Remove duplicate adverse event cases and reports
- Standardize free-text fields
    - Align misspelled drug and substance names whenever possible

Preprocessing:

- Data reduction by attribute subset selection (eliminate additional attributes not used in analysis)
- Data reduction by numerosity reduction (group by class of drug or by active substance rather than by official drug name)
- Bucket age ranges for easier analysis

Integration:

- Put FAERS data into multiple tables in a relational database, and properly associate matching records by case ID, and/or any other matching information

Intended Tools

- Python
- Python JSON library
- Pandas
- Numpy
- SciPy
- Sklearn
- Seabourn
- Patsy
- Matplotlib
- MySQL

Evaluation

How will the results be evaluated?

Our results can be evaluated using statistical analysis.

- For prediction we can use various parameters such as P-value, R-Squared and F statistic to evaluate how well we model the data.
- For classification we will use other methods such as confusion matrices, linear discriminant analysis and K nearest neighbors classification.