Predicting Adverse Drug Reactions Based on Patient Demographics

CSCI 4502 Semester Project

Anthony Olvera
Computer Science
University of Colorado
Boulder, Colorado, USA
anthony.olvera@colorado.edu

James Lagrotteria
Computer Science
University of Colorado
Boulder, Colorado, USA
james.lagrotteria@colorado.edu

Annie Lydens
Computer Science
University of Colorado
Boulder, Colorado, USA
anne.lydens@colorado.edu

PROBLEM STATEMENT

Our project will analyze the FDA FAERS data set, specifically examining links between various demographic attributes or drug characteristics and subsequent adverse outcomes. We hope to be able to find interesting correlations and relationships that will lead us to be able to predict the likelihood of a certain groups of people having adverse reactions to specific drugs. This research could be hugely beneficial for medical staff and pharmaceutical companies in allowing them to develop a more nuanced understanding of previously unidentified factors that may contribute to adverse reactions in patients. It could also help identify groups of patients who may have previously been considered low-risk for an adverse reaction to a drug, and correctly reclassify them as higher risk.

LITERATURE SURVEY

Because the FAERS database is readily available and large, it is a good candidate for data mining, and therefore there are good examples of prior work examining this data set.

In many cases, these prior studies have selected a single drug or drug ingredient, and looked at the risk of a specific outcome or set of outcomes occurring across demographic groups. For example, Fadini et al. [2017] examined the link between SGLT2 inhibitors and diabetic ketoacidosis using the FAERS dataset. They also examined outcomes across groups of people, and concluded that diabetic ketoacidosis associated with SGLT2 inhibitor use does not seem linked to specific demographic attributes.

Umetsu et al. [2015] used the FAERS dataset to determine the relationship between selective serotonin reuptake inhibitor therapy and suicidality. They specifically looked at this relationship across patients grouped by age, and discovered a stronger relationship between SSRIs and suicidality in groups age 18 and above using a logistic regression model.

Feng et. al [2013] took a similar approach to their research question, and mined the FAERS dataset to look at the pancreatic cancer risk associated with the use of dipeptidyl peptidase 4 inhibitors. They found a significant risk of pancreatic cancer associated with the use of dipeptidyl peptidase 4 inhibitors, and were able to identify that the combination of an additional drug, metformin, actually correlated with a decreased risk of pancreatic cancer.

Aside from studies linking drug or demographic attributes to specific adverse outcomes, there is also relevant work on the dataset itself. Banda et al. [2016] tried to address common data cleaning issues with the FAERS data set by providing a standardized version of the database with duplicates removed, missing values imputed, and drug names properly mapped. Though we are not using this standardized version of the database for our own

research, referencing techniques used by this team could be invaluable as we undertake data cleaning ourselves. Likewise, Maciejewski et al. [2017] also sought to standardize elements in the FAERS database, choosing to focus specifically on mapping drugs to their ingredients in order to provide a more accurate way to examine drug outcomes.

At a higher level, Sakaeda et al. [2013] published a paper entitled, “Data Mining of the Public Version of the FDA Adverse Event Reporting System” which contains an overview of some data mining algorithms selected for use on FAERS and the algorithms’ ability to detect “signals” in the data, where signals are “a statistical association between a drug and an adverse event or a drug-associated adverse event”. This study also touching on possible shortcomings of this type of analysis, which may help us avoid pitfalls or erroneous conclusions in our own investigation.

PROPOSED WORK

While the studies covered in the literature survey tended to focus either on the mechanics of cleaning the data or on a specific drug-outcome relationship, we want to focus with more granularity on the demographic attributes available in the dataset. Because of the size of the dataset, we may need to focus our efforts on a specific drug, but since the FAERS dataset includes a number of interesting demographic attributes, we have an opportunity to uncover how adverse outcomes correlate with a variety of influencing factors like sex, age, weight, and location. Ideally, this will lead us to interesting predictive analyses.

Data Collection

For this project we will be working with the FAERS dataset in two different forms: firstly through a Postgresql relational database, and secondly by parsing JSON files using Python.

For the first method, FAERS data from 2016 Q1 through Q3 was downloaded using shell scripts, and loaded into seven different tables in a Postgresql relational database. This will allow us to use standard SQL queries to begin to explore the data.

For the second method, data over the same period was downloaded as JSON files, 788 in all. This was achieved using a Chrome batch download extension. The files are named by their numbering in a given set for that quarter, so after initial download, there are multiple files with the same name. Each file will need to be given a unique name to prevent overwrites when inflating the zip archives, which will be done using a shell script.

Having data in two formats will allow us to explore the dataset in different ways, and ideally give us the most flexibility possible as we begin analysis.

Preprocessing

For most of our data manipulation, Python will be the appropriate tool. We will first load the files into a Pandas dataframe in order to use Python for our data preprocessing.

The first step in the preprocessing stage will be data cleaning. We will need to handle null values if they exist, and check for and remove any outliers. As our data set is quite large (on the order of several million data points) we plan to simply delete all rows of data that are missing a value for one or more of the needed attributes as long that this method does not sacrifice the data integrity.

The next step in the cleaning process is to identify any outliers that may be present. For each numerical attribute we will calculate the first and third quartiles as well as the interquartile range and eliminate all data that fall outside of $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.

The next step in preprocessing will be data integration. At this point we only have one

source of data, so there will be minimal integration necessary. When querying the relational database, cases are uniquely identified across tables by a caseid attribute, so we can use that identifier as a way to join data tables, but more intensive data integration efforts are unneeded.

For data reduction we will simply disregard the attributes that are unnecessary and do not add value to our investigation. This will make the mining process easier as there will be no redundant data to deal with. Data transformation will be done by normalizing all numeric attributes using min-max normalization, so they can be analyzed statistically.

DATA SET

The data set can be located in different file formats (JSON, XML and txt) at the following URLs:

- JSON: https://open.fda.gov/tools/downloads/
- XML/.txt: https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html

More information about the Postgresql database we are using for data analysis is available at the following URL:

- https://github.com/CSCI-4502-DataMining Project/Drug_Reactions

We have downloaded a subset of the data available and placed it into tables mirroring tables in the FAERS database. These tables and their row counts are as follows:

- indi (indications) - 7,901,564
- outc (outcomes) - 2,122,914
- reac (reactions/adverse events) - 9,007,602
- rpsr (report sources) - 113,657
- ther (drug therapy start/end dates) - 4,200,330
- demo (patient demographic information)- 3,088,728
- drug (drug/biologic information) - 11,312,890

The next form of the data set exists as JSON files. For each file there are two subsections of the data, meta and results. Meta contains metadata about the query, including a disclaimer, link to data license, last-updated date, and total matching records, if applicable. Results contains an array of matching results, dependent on which endpoint was queried. The data set provides all necessary information about the patient such as age, gender, condition and outcome as well as detailed information about the drug used for treatment including but not limited to drug name, drug substance, dosage and administration method.

EVALUATION METHODS

For data evaluation it may first be useful to match probability distributions to each of the retained attributes in order to learn more about the data. We will attempt to answer questions such as: Do most attributes fit a normal distribution? Is the skew in certain attributes? If so what could be causing it? Next, we will calculate correlation coefficients between each pair of numerical attributes in order to identify significant correlations that may exist in hopes of exploring what demographic attributes may be highly correlated with negative outcomes.

Once we've discovered interesting correlations we can hopefully construct decision trees which support our claims.

Another method of evaluation will be to use the Apriori algorithm to discover sets of factors and reactions in patients which frequently appear together. If we find a significant amount of these conditions, we may be able to conclude with a certain probability that factor x will cause adverse reaction y.

A last method of evaluation will be implementing a classification algorithm. After data cleaning and preprocessing we can construct a training and test set in which we can use methods such as naive Bayesian and rule-based classification which will ideally lead to some interesting knowledge discovery. After all evaluation is finished it will be important to check the significance of our conclusion using metrics such as P-value, R squared, confusion matrices and F-score.

Ideally, our work will result in predictive analyses that will allow us to forecast adverse drug reactions for certain demographic subgroups. We hope to find interesting and unexpected indications that a subgroup may be more at-risk of an adverse reaction than otherwise anticipated.

TOOLS

We plan to use the following tools to aid in the analysis and evaluation detailed above:

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

MILESTONES

Our expected milestones are as follows:

- Friday, October 26: Finish any data collection for the JSON files. Continue data cleaning and preprocessing efforts. Select a single drug or drugs to examine.
- Friday, November 2: Finish data preprocessing efforts, and have completed initial investigation of the dataset. Match attributes to probability distributions and identify existing correlations.
- Sunday, November 4: Submit progress report.
- Friday, November 9: Construct predictive models.
- Friday, November 16: Implement Apriori in python and run on data set.
- Friday, November 23: Construct decision trees based on results.
- Friday November 30: Validate that our findings are accurate using appropriate methods.
- Friday December 7: Finish project code and descriptions and project final report.
- Friday December 14: Finish project presentation.
- Sunday, December 16: Submit Final presentation.

REFERENCES

Feng X, Cai A, Dong K, Chaing W, Feng M, et al. 2013. Assessing Pancreatic Cancer Risk Associated with Dipeptidyl Peptidase 4 Inhibitors: Data Mining of FDA Adverse Event Reporting System (FAERS). Pharmacovigilance 1: 110. DOI:http://dx.doi.org/10.4172/2329-6887.1000110

Mateusz Maciejewski et al. 2017. SGLT2 inhibitors and diabetic ketoacidosis: data from the FDA Adverse Event Reporting System. eLIFE 6, e25818 (August 2017). DOI:http://dx.doi.org/10.7554/eLife.25818

Gian Paolo Fadini, Benedetta Maria Bonora, and Angelo Avogaro. 2017. SGLT2 inhibitors and diabetic ketoacidosis: Data from the FDA Adverse Event Reporting System. Diabetologia 60, 8 (May 2017), 1385–1389. DOI:http://dx.doi.org/10.1007/s00125-017-4301-8

Ryogo Umetsu et al. 2015. Association between Selective Serotonin Reuptake Inhibitor Therapy and Suicidality: Analysis of U.S. Food and Drug Administration Adverse Event Reporting System Data. Biological and Pharmaceutical Bulletin 38, 11 (2015), 1689–1699. DOI:http://dx.doi.org/10.1248/bpb.b15-00243

Juan M. Banda, Lee Evans, Rami S. Vanguri, Nicholas P. Tatonetti, Patrick B. Ryan, and Nigam H. Shah. 2016. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data3 (May 2016). DOI:http://dx.doi.org/10.1038/sdata.2016.26

Toshiyuki Sakaeda, Akiko Tamon, Kaori Kadoyama, and Yasushi Okuno. 2013. Data Mining of the Public Version of the FDA Adverse Event Reporting System. Data Mining of the Public Version of the FDA Adverse Event Reporting System 10, 7 (April 2013). DOI:http://dx.doi.org/10.7150/ijms.6048