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

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

ABSTRACT

This project analyzed the FDA FAERS data set, specifically examining links between various demographic attributes or drug characteristics and subsequent adverse outcomes. We set out to find interesting correlations and relationships between demographic characteristics and their likelihood of having an adverse drug reaction. We also sought to resolve whether there are pairs of frequently occurring sets of specific patient demographic attributes. In addition, we also paired sets of common multiple reactions with certain medications which tend to occur together in the treatment process.

In the course of our research, we found specific patient profiles that had a higher predicted probability of having a life-threatening reaction after taking a drug. Older men outside of the US tended to have a higher likelihood of serious, life-threatening adverse reactions. Younger women had the least risk of suffering a life-threatening reaction.

We also found that, among all adverse reactions, the most common is drug ineffectiveness. The most common patient demographic paired with drug reactions are drug ineffectiveness amongst females and drug ineffectiveness amongst individuals in the US. Lastly, the most frequent co-occurring drug and reaction are the drug Lunesta and drug ineffectiveness. The most common reactions found in a pair are drug ineffective and insomnia. In fact, those who suffer from insomnia have nearly a 50% chance of claiming a drug used for treatment is ineffective.

INTRODUCTION

Over the course of our research, we sought to find the most common set of patient characteristics and adverse drug reactions, as well as identify demographic attributes that might indicate a patient is more at-risk for a serious, life-threatening outcome.

Furthermore, we want to determine if there are sets of multiple reactions which tend to occur together in the treatment process. That is, regardless of patient physical characteristics, do certain medical reactions occur together frequently? And if so, what are they? An example of a hypothetical result for this line of inquiry could be that drowsiness and weight loss frequently occur as a pair, perhaps even more so than one or the other by themselves.

These questions are important, as they could give medical professionals insight as to what types of treatments are best tailored to certain individuals. This could eventually lead to a reduction in adverse drug reactions by more educated decisions in treatment options for individuals who fit a specific profile.

In addition, our research could provide better preparation for subsequent treatment selection given information that a person has already experienced a certain adverse reaction.

This way, medical professionals may have a better idea of what other reactions may follow, and be ready to handle them if they in fact do.

RELATED WORK

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. 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 touched on possible shortcomings of this type of analysis, shortcomings we kept in mind as we performed our own research.

DATASET

All data for this investigation came from a publicly available data source provided by the US food and drug administration called openFDA, specifically the FDA adverse event reporting system or FAERS. FAERS is an open source database that provides reports and information regarding adverse reactions and errors regarding drugs and medications. More information about openFDA and FAERS can be found here https://open.fda.gov/about/ https://open.fda.gov/data/faers/

The data sets we worked with 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 dataset initially existed as a collection of 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 relevant 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. Below is a complete list of all attributes included in the raw dataset before any preprocessing was done. A detailed description of these attributes can be found at the openFDA API fields descriptions page https://open.fda.gov/apis/drug/event/searchable-fields/

```
['companynumb',
 'fulfillexpeditecriteria',
 'patient.drug',
 'patient.patientdeath.patientdeathdate',
 'patient.patientdeath.patientdeathdateformat',
 'patient.patientonsetage',
 'patient.patientonsetageunit',
 'patient.patientsex',
 'patient.patientweight',
 'patient.reaction',
 'primarysource',
 'primarysource.qualification',
 'primarysource.reportercountry',
 'receiptdate',
 'receiptdateformat',
 'receivedate',
 'receivedateformat',
 'receiver',
 'safetyreportid',
 'sender.senderorganization',
 'serious',
 'seriousnesscongenitalanomali',
 'seriousnessdeath',
 'seriousnessdisabling',
 'seriousnesshospitalization',
 'seriousnesslifethreatening',
 'seriousnessother',
 'transmissiondate',
 'transmissiondateformat']
```

MAIN TECHNIQUES APPLIED

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 wanted to focus with more granularity on the demographic attributes available in the dataset. Since the FAERS dataset includes a number of interesting demographic attributes, we had an opportunity to uncover how adverse outcomes may change with different demographic influencing factors like sex, age, weight, and location.

Our overall data strategy involved the Postgresql database, which contained all the FDA FAERS data and acted as our data warehouse. We relied heavily on the data warehouse as a source of data for analysis.

Preprocessing and Integration

The first milestone we completed was data cleaning, merging and preprocessing. We performed data cleaning and preprocessing steps on both the JSON version of the dataset and the version of the dataset loaded into our Postgresql database. As mentioned earlier, having the data in two formats allowed us to have lots of flexibility in our analysis.

JSON Files

A total of 788 JSON files provided by the FDA at https://open.fda.gov/tools/downloads/ and containing all patient information, ranging from patient demographic information such as age, weight and sex, to drugs used for treatment and reactions to those drugs, were downloaded and inflated to an external drive. The data, in JSON format, contained a lot of information in nested dictionaries which needed to be flattened into a table format to allow for querying and numerical manipulation. Using a Python Jupyter notebook, which can be found here https://github.com/CSCI-4502-DataMiningProject/Drug_Reactions, the files were loaded into a pandas dataframe. The first nested dictionary structure of each file was flattened and its attributes listed.

Immediately a number of attributes were dropped as they are irrelevant to our project and will not add any value to our investigation. If desired, those attributes can be referenced in the Jupyter notebook. Out of the remaining attributes, two (drug and reaction) contained a further nested data structure needing to be unpacked. This was done by creating two new data frames for each file containing arrays of the drug(s) and reaction(s) associated with each subject. The columns were split by drug and reaction, then merged with the original data frame.

We wished to start our investigation small, so we limited our dataset to include only one drug per data object. We also limited the dataset to include only up to the first three reactions per data object, and dropped all others. Information on how that drug was selected, and why we chose to limit the reactions, is detailed below in the next section. However, we've kept all files stored in case we wish to expand our project.

As the current data were currently all represented as objects, we needed to convert the data to the appropriate data types so the could be compared using boolean operators in future exploration and mining. The following attributes were converted to the following types. Age and weight were converted to floating point values. Sex, transmission date, receive date, and seriousness were converted to integers. And country, drug_0, reaction_1, reaction_2 and reaction_3 were converted to strings.

The next step in the process was to eliminate any outliers. This was done on the numerical attributes using the following formula Q1-1.5×IQR and Q3+1.5×IQR, where Q1 and Q3 are the first and third quartiles, and IQR is the interquartile range. Next patient weight was converted from kilograms to pounds, all data objects with null demographic information, that is age, sex and weight were dropped from the data frame as well as data with 0 or unindicated sex.

The last step in the JSON preprocessing stage was to discretize the numerical attributes in order to categorize patients and match using the Apriori algorithm. Patient onset age and patient weight were cut into the following six equal sized bins. For patient onset age the sets were given the names ['child','adolescent','young_adult','adult','senior','elderly'] and patient weight ['light','medium_light','medium','medium_heavy','heavy','very_heavy'] this will make future mining more practical as there are a wide variety of ages and weights in the data.

Postgresql Data

The data preprocessing and cleaning steps on the Postgresql database began with us creating several different subsets of the dataset for exploration. We joined the demographic, reactions, outcomes, and drug tables in the database, matching each record by case ID. Case ID is the FDA unique identifier for a particular patient case. We dropped all attributes unhelpful in our analysis.

In order to focus our investigation, we chose to only analyze the “primary suspect” drug, which is the drug suspected of causing the adverse reaction. The FDA FAERS database denotes the role of a drug with an attribute called “Drug Role Code”. Patients, especially chronically ill ones, may be taking multiple drugs at a time, and the FDA FAERS database will have a record of each of these drugs as part of a case report. However, some drugs are not as likely to cause adverse reactions as others. The drug most likely to have caused the reaction has a Drug Role Code of “PS” (corresponding to “primary suspect”). In order to minimize noise, we chose to retain only these “primary suspect” drugs.

Additional data cleaning steps included normalizing date fields to a MM/DD/YYYY format, standardizing weight to kilograms, and removing duplicates.

The attributes retained in our data subsets are:

- Primary Suspect Drug Name
- Drug role code
- Case ID
- Age
- Age Group
- Sex
- Weight
- Country
- Event Occurrence Date
- Reaction

These data subsets include:

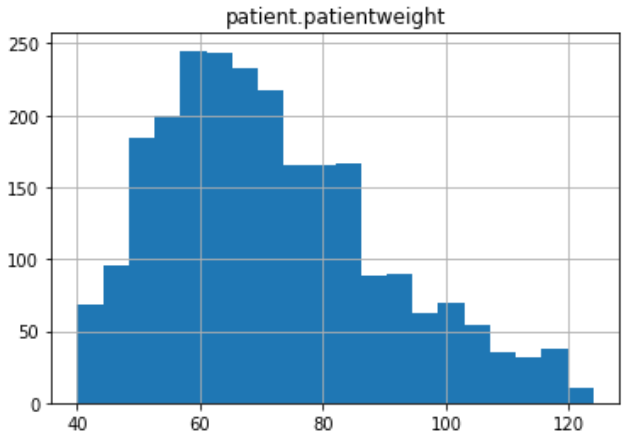
Subset A
Subset A contains 1.5 million records with the the attributes above. However, it only includes the primary (first) reaction per case. Reactions can be extremely specific and interrelated, so using a primary reaction as a kind of “descriptive” reaction will work well for higher-level analysis. For example, a patient whose primary reaction is listed as “Abdominal pain lower” may also have “Pelvic pain” and the more general “Fatigue” and “Pain” listed as reactions. Using the primary reaction will help us narrow our focus, and will streamline our investigation when we’re not interested in breaking analysis out by specific (and sometimes unhelpful) reactions.

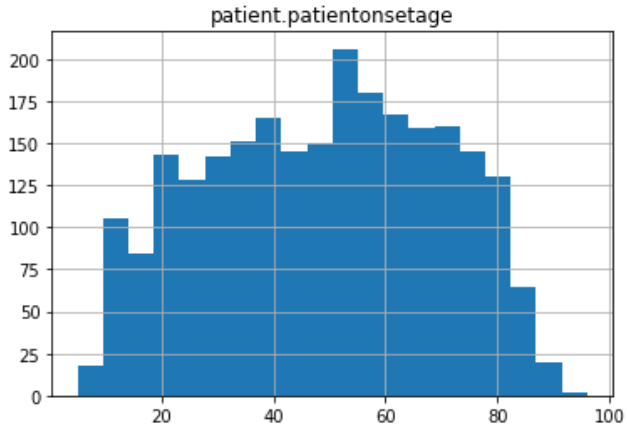
Subset B
Subset B contains all the attributes above aside from drug name, and includes all reactions, not just primary reactions, for finer-grained analysis. It contains 18 million records.

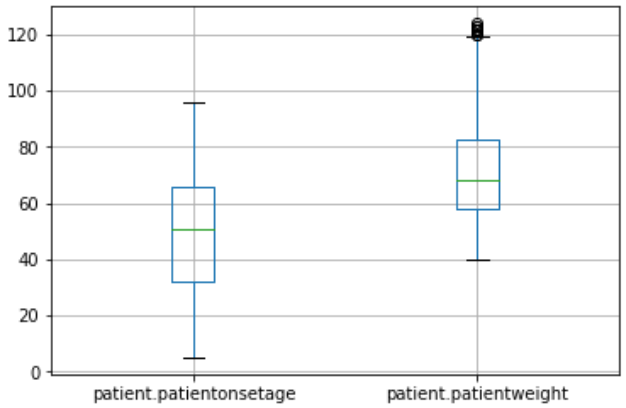
Data Exploration and Analysis

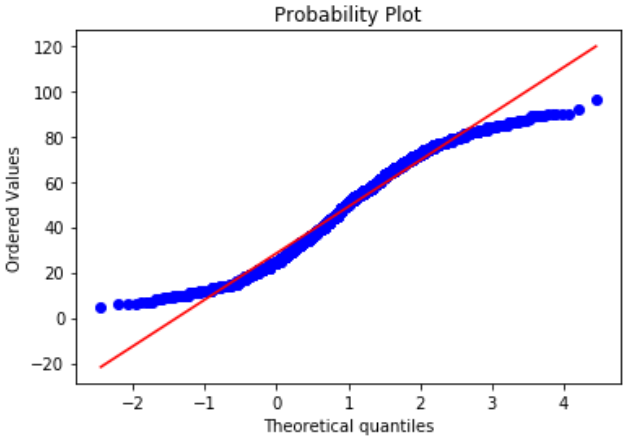
As part of our initial investigation of the data, we plotted histograms of patient demographics. Specifically, patient age, weight and sex were plotted for visual inspection in an attempt to identify any obvious trends and or skews in the data that may exist. Age appears to be evenly distributed, a normal distribution can be verified give the quantile-quantile plot below. There is a slight deviation near the outer ends, however, this can be expected as a result of outlier elimination. Interestingly, patient weight appears to be slightly skewed to the right. This may be an indicator that weight loss is present among a majority of patients in this data set. Using the Scipy statistical analysis tool for Python, the skewness was measured to roughly 0.628 meaning nearly 63% of the weight exists in the left tail of the distribution. In addition a skew test was ran which returned the following result.

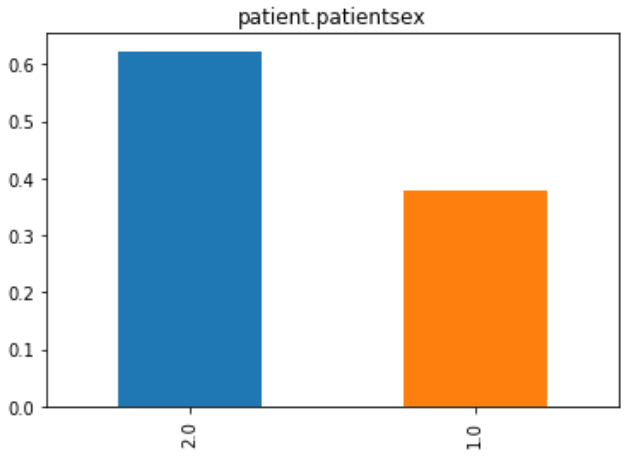
SkewtestResult(statistic=11.766181724799404, pvalue=5.83041859092924e-32). This will be analyzed in further detail in the near future. Lastly, and for reasons unknown, the number of female patients appears to be greater than the number of male patients in most of the files. A bar chart of overall percentage of males compared to females averaged over 5 files can be reviewed below, where males are given by the number 1 and females the number 2. All plots were also recaptured before any preprocessing was done to verify that these trends are indeed genuine and were not a result of the data cleaning and preprocessing.







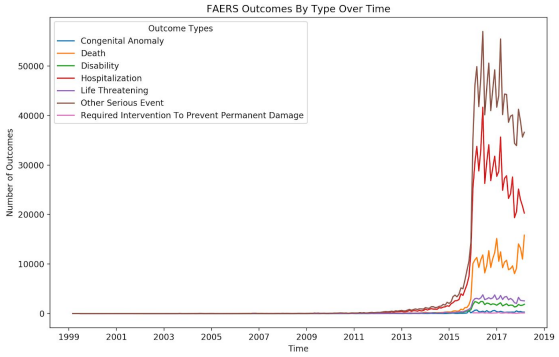




We also examined patient outcomes, which are broken out by the FAERS database into seven types: "Congenital Anomaly", "Death", "Disability", "Hospitalization","Life Threatening", "Other Serious Event", and "Required Intervention". A patient might have multiple

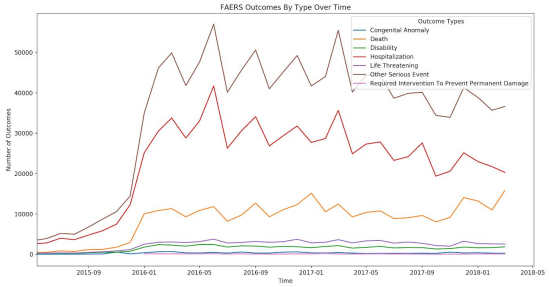
adverse reactions per case, but only one outcome.

To explore how the dataset might have changed over time, we created a temporal graph of all outcome reports.



Cumulatively, we can clearly see from the graph that there was a significant increase in reporting starting in 2013.

If we examine this period in greater detail, we can see the beginnings of a pattern, where reporting volume spikes approximately once a quarter, with larger spikes around the first few months of the year. It's possible that the quarterly spikes are due to reporting behaviors by medical staff; perhaps reports tend to get filed around the same time each quarter. It would be interesting to continue to observe outcomes reporting into 2019 and 2020 to see if these patterns continue.

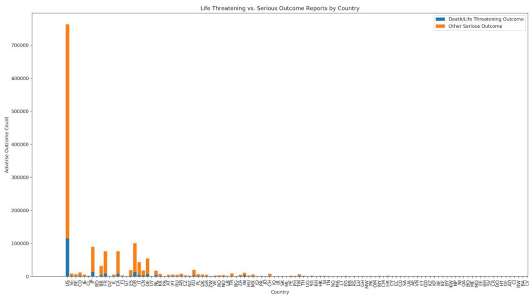


We can see that "Hospitalization" and "Other" are the outcomes categories most commonly used, with "Required Intervention" near the bottom. This is not unreasonable -- since cases only have one outcome, if a patient is hospitalized at any point, even during a case that "Required Intervention", their case may get categorized as a "Hospitalization". This could explain the heavy reliance on categorizing outcomes as hospitalizations.

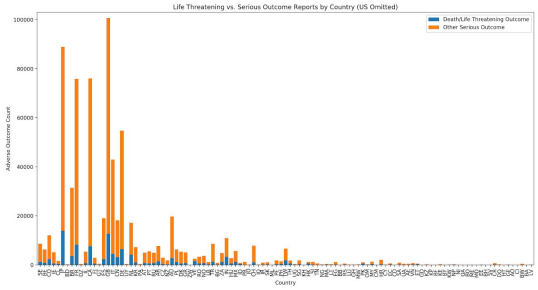
We also looked at outcomes broken out by country. In the FAERS database, "country" is where the adverse reaction occurred. In the graph below, we can see, rather unsurprisingly, that there is an outsize number of reports that occurred in the US.

To create the graph below, we divided outcomes into two categories, rather than the full seven used by the FAERS database. The two categories we used were "Life Threatening/Death" (hereafter referred to as LTR) and "Other Serious Outcome". The first category includes the FAERS cases that were marked as "Death" or "Life Threatening", and the second category includes all other outcomes.

We omitted all countries with less than 20 total outcome reports.



The graph below omits the US, giving us a better view of the other countries.



Japan, Canada, Great Britain, and Germany are all countries leading the way after the US in terms of reporting numbers. These are not particularly surprising candidates, but there are some potentially interesting questions one could examine based on this data. For example, Great Britain and Japan have roughly the same number of LTR reports, but Japan's total outcome reports are only about 90% of Great Britain's outcome reports. What might account for the relatively higher LTR reports in Japan?

Data Mining Methods Applied

We applied two main data mining techniques to our data. We ran the Apriori algorithm to find frequent co-occurring demographic characteristics and outcomes. We also analyzed the data using Naive Bayes classifiers.

Apriori was done in Python using the following algorithm:

C_k: Candidate item set of size k
L_k : Frequent item set of size k
L_1= {frequent items};
For (k=1; L_k!=Φ;k++) do begin
C_{k+1} = candidates generated from L_k;
For each transaction t in database do
Increment the count of all candidates in C_{k+1}
Those are contained in t
L_{k+1} = candidates in C_{k+1} with min_support
End
Return U_k L_k;

The first part of the process was to generate C1, the candidate item set of all distinct data values in the data. This was done by means of a function which loops over each row then each data value per row then checks if that value is in the set C1, if it is not it is added to the set. The function then converts each element of the list to a frozen set so it can be used as a key in the dictionary which will be created in the following step.

Once the C1 set is returned it is a run through another function which runs successive scans of the data, thus returning the frequent itemsets we are interested in given the original dataset, the candidate set generated and the minimum support value we supply. The function works by looping over each row of data and for each candidate in the candidate set checks if that candidate set is a subset of the given row. If so the algorithm proceeds to check if the set is in the dictionary then increments a count for that given set. A second loop eliminates all data in the dictionary that does not meet the minimum support provided. Once the entire set is processed the function returns all frequent itemsets which meet the criteria along with their support data.

To implement our Naive Bayes classifiers, we used the scikit-learn Python library. We classified patients by demographic attributes into two categories: those that we predict will experience a life-threatening reaction (LTR), and those we predict will not. We also looked at the probability estimates for particular patient test vectors.

We used both the Bernoulli and Gaussian Naive Bayes models provided by scikit-learn. We chose attributes and prepared the data differently for each model. This is because while the Bernoulli Naive Bayes model works only with binary attributes, the Gaussian model supports inputs where a normal (or near-normal) distribution is assumed.

For the Bernoulli Naive Bayes model, we used two attributes: sex and reporting country. We prepared our dataset for the model by coding patients listed as “male” as 1, and “female” as 0. Similarly, we bucketed reporting country values into “US” (1) and “non-US” (0). We discarded rows without sex or country values.

For the Gaussian Naive Bayes model, we examined four attributes: sex, country, age, and weight. Like in our preparation for the Bernoulli Naive Bayes model, we dropped rows where the sex or country values were missing, as we were not able to easily impute those values. We also chose to drop rows without age or weight values.

For both models, we split our data into two sets, one for training and one for testing, with 40% of the data being randomly selected for training. We used the scikit-learn Model Selection module to assist with generating the training and test sets.

Finally, we used the scikit-learn Metrics module to calculate the overall accuracy of our classifiers, and to assess their performance.

KEY RESULTS

The Apriori algorithm was ran repetitively on our data, our main interest being in candidate one and candidate two item sets. Each time the algorithm was ran, the minimum support threshold was incrementally decreased until interesting patterns began to emerge.

Item sets were generated for each of the following minimum support thresholds: 0.2, 0.1, 0.07, 0.05, 0.03 and 0.01. After each candidate set was generated, the given data values in that set were eliminated from the next sequential item set with a smaller minimum support as this set is guaranteed to be a subset of the next and redundant data need not be duplicated. As the minimum support decreases the itemsets generated tend to contain more non-interesting itemsets such as null values, unimportant dates and seriousness level integers. In future analysis these attributes should be dropped from the dataset. The scans with a minimum support of 0.2, 0.1 and 0.07 reveal no interesting itemsets however, all the results were maintained in a text file aprori_results.txt which can be found in the GitHub repository’s apriori directory. At a minimum support of 0.05 the scan revealed the most common set of a single drug reaction was “DRUG INEFFECTIVE ” occurring in at least 5% with a support of 0.05468 of the data.

The later scan at 0.03 reveals the most common drug in use in the data set is named “FORTEO” used to treat osteoporosis. Other common drugs which emerged in this scan were “BYETTA”, “CRESTOR” and “LUNESTA” used to treat diabetes, high cholesterol and insomnia respectively. Other common reactions which emerged in the 0.03 scan were “ASTHENIA” and “NAUSEA”. Interesting frequent two item sets also emerged in this scan the most frequent being “DRUG INEFFECTIVE” and “female” with a support of 0.03497. This finding is validated with the following association rules DRUG INEFFECTIVE => female(0.03497, 0.59910), female => DRUG INEFFECTIVE(0.03497, 0.05463). Another interesting two itemset which was found were “DRUG INEFFECTIVE” and “UNITED STATES” with support value of 0.03276, however this finding may be influenced by a large portion of the dataset being made up of reports from the US. The association rules are DRUG INEFFECTIVE => UNITED STATES(0.03276, 0.63964) and UNITED STATES => DRUG INEFFECTIVE(0.03276, 0.09219).

On the last scan with a minimum support threshold of 0.01, many frequent item sets were revealed most of which are not of interest to us. The frequent two itemsets we are most interested in are those which pair a common

drug and reaction and the two drug reactions the most frequently occur together. The first set happens to be the drug LUNESTA being ineffective below are the association rules DRUG INEFFECTIVE => LUNESTA(0.02126 ,0.38889), LUNESTA => DRUG INEFFECTIVE(0.02126, 0.47654), and the last interesting bit of information discovered is that DRUG INEFFECTIVE and INSOMNIA is the most common pair of patient reactions found together, the association rules are DRUG INEFFECTIVE => INSOMNIA(0.011656, 0.21321) and INSOMNIA => DRUG INEFFECTIVE(0.01166, 0.46942). These results are likely related to the fact that the drug LUNESTA is used to treat insomnia which we saw has a high chance of being ineffective (nearly 50%) in the previous association rules.

The Bernoulli Naive Bayes model allowed us to discover some interesting relationships and produced broad predictions for each high-level demographic group. While the overall likelihood of having a life-threatening reaction (LTR) is 11%, our model estimated the following probabilities of a LTR for each pair of demographic attributes examined by our classifier:

	US	Non-US
Male	10.936%	14.262%
Female	6.954%	9.194%

We can see that the group most likely to experience a LTR is non-US males by a large margin. Based on our exploration, women in general are less likely to suffer life-threatening reactions, with US females being the group least likely to experience LTRs.

We calculated an accuracy score of 89.93% for our Bernoulli Naive Bayes model using the metrics module.

Our Gaussian Naive Bayes model gave us many different patient profiles to explore and predict outcomes for, as each attribute could potentially change the predicted probability of experiencing a life threatening outcome. We included a few example patient profiles below to illustrate how tweaking just one attribute can significantly affect the predicted probability of a LTR:

	Male, 60, 80kg	Female, 60, 80kg	Male, 30, 80kg	Female, 30, 80kg
JP	13.36%	8.32%	10.60%	6.52%
GB	11.54%	7.14%	9.12%	5.58%
US	10.83%	6.67%	8.54%	5.21%
CH	13.90%	8.69%	11.05%	6.82%
CN	12.38%	7.68%	9.80%	6.02%
CA	12.85%	7.99%	10.19%	6.26%
DE	11.95%	7.40%	9.45%	5.79%
IT	15.13%	9.50%	12.06%	7.47%

The Gaussian Naive Bayes model allowed us to classify patient profiles with more granularity. Digging into the sex divide, we can see that for the countries listed above, women have a lower chance of a LTR across the board. As we might expect, a 30-year-old woman is predicted to have a lower probability of a LTR than an 80-year-old of the same weight. The same pattern holds true for men.

We calculated an accuracy score of 89.15% for our Gaussian Naive Bayes model using the metrics module.

Country-by-country comparisons are interesting as well, though direct comparisons might be less illuminating than they first appear.

A question we posed earlier in the report was why Japan might have a relatively higher rate of LTR reports. In digging into the country-by-country differences illuminated by the Gaussian Naive Bayes analysis, we formulated a hypothesis as to why Japan might have higher LTR reports, and why we in general might see such different results in different countries.

Simply put, each country maintains its own pharmacovigilance reporting systems and standards, as well as education around how to code adverse outcomes.

Cultural standards may also play a significant part in over- or under-reporting. In a study comparing spontaneous reporting systems in Korea, Japan, and Taiwan, Kimura et al. [2011] pointed out that reporter sources varied wildly between country for the reports they examined. Physicians comprised the majority of reporters (51%) in Korea, and pharmacists were the majority (65%) in Taiwan, but a stunningly-high 89% of reporters in Japan were drug manufacturers. It’s possible that manufacturers are motivated to report serious adverse outcomes for cultural or legal reasons in Japan, leading to a much higher reporting rate for LTRs.

Other country-specific patterns might be explained with more exploration into regional cultural or legal obligations on the part of adverse reaction reporters.

In our research so far, we have identified some patient profiles that may be more at-risk for life-threatening reactions. An interesting future investigation would be to dig further into some of the regional differences in reporting metrics.

APPLICATIONS

Our research could help reduce adverse drug reactions by bringing awareness to common adverse reactions that occur together. Knowing common pairs of reactions from specific drugs could help physicians and other medical professionals prepare patients for treatment, as well as make more educated treatment decisions.

Additionally, this research could be beneficial for medical staff and pharmaceutical companies in allowing them to develop a more nuanced understanding of previously unidentified demographic 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 a life-threatening outcome, and correctly reclassify them as higher risk.

An interesting software application of this research could be to build a web application that allows medical professionals to either enter patient demographic characteristics and view the patient's predicted probability of a life-threatening outcome, or input drugs or adverse reactions and see common co-occurring reactions.

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. 10, 7 (April 2013).

Kimura T, Matsushita Y, Yang YH, Choi NK, Park BJ. 2011. Pharmacovigilance systems and databases in Korea, Japan, and Taiwan. Pharmacoepidemiol Drug Saf. 20, 12 (Dec 2011).