Analyzing Adverse Events from Publicly Available Web Sources

Alexander Salamun
Marquette University

ANALYZING ADVERSE EVENTS FROM PUBLICLY AVAILABLE WEB SOURCES

by

Alex Salamun, M.S.

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Computing

Milwaukee, Wisconsin

December 2020

ABSTRACT
ANALYZING ADVERSE EVENTS FROM PUBLICLY AVAILABLE WEB SOURCES

Alex Salamun, M.S.

Marquette University, 2020

Abstract

Data mining for drug-reaction associations is a major topic in the pharmaceutical industry. Historically the focus has been on using privately owned and maintained datasets consisting of information that has been transformed via the FDA Adverse Event Reporting System (FAERS) and privatized reporting systems that house the data from clinical trials. Our focus will be on building a pipeline that demonstrates an open source solution for building a drug's safety profile from data collection through signal detection. In contrast this pipeline primarily uses the openFDA and social media data available through Reddit with all analysis being done in the R statistical programming language. The aim was to collect the information available in these public sources and apply popular data mining methodologies used to identify and predict the occurrence of adverse events. The results show the ability of the openFDA and social media sites to create real-time drug safety occurrence profiles by applying the same statistical methods applied in clinical trials. Social media will be shown to provide the best results when applied to prescribed daily use medications compared to common over-the-counter drugs or last line of defense medications. The information and results reported in this paper are not intended or implied to be a substitute for professional medical advice, diagnosis, or treatment. Do not delay seeking medical treatment or advice because of something you have read in this paper.

ACKNOWLEDGEMENTS

Alex Salamun, M.S.

I sincerely want to thank my partner Stephanie Duque for her continued support and motivation as well as her contributions to this work. I would also like to thank my parents for the love and guidance they have always given me.

I am extremely grateful to my advisor, Dr. Praveen Madiraju for his contributions and insight throughout the entirety of this process. Finally, I want to acknowledge Justin Ott and Steve Gao for helping shape the focus of this project and always being available to provide feedback.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

There is an ever-increasing amount of prescription drugs on the market with different options from brands to doses to chemical composition all for the purpose of treating a particular condition. This flux of available drugs makes it critically important to have a way to monitor their safety. The field of Pharmacovigilance is primarily focused on the monitoring of the effects of medical drugs after they have been made available and sold to the public with a focus on identifying potentially unreported or underreported reactions. The most robust profile of a drug's side effects can only be properly created after the drug is approved and used by the public in a non-clinical setting (Fram, Almenoff, & DuMouchel, 2003). The main reporting body for such adverse events is the FAERS (FDA Adverse Event Reporting System) database. Adverse events are defined as injuries related to medical management (in contrast to complications of the disease)

Mining the drug-adverse event pairs from the FAERS database has been well studied and analyzed (Harpaz et al., 2012). In 2014, the FDA released the openFDA API (Kass-Hout et al., 2016) that allows users to get direct access to the FAERS data without the cumbersome task of downloading and storing all of the quarterly FAERS data files. As an API or application programming interface it has the key feature of allowing the sharing of information between programs with a predefined setup. Prior to the creation of the API, researchers and organizations would have to ingest and maintain a relational database if they wanted to perform analysis of the spontaneous reporting adverse event data. This task not only requires large computing power but also a diverse technical background in order to create and maintain such a database which can be expensive for a company and almost impossible for the unequipped. The primary source for the openFDA is the FAERS database that is transformed with special fields indicated with the openfda tag that serve to increase the out-of-the-box effectiveness of the data. These openfda tags offer additional drug information and conditions that normally require industry knowledge to parse out. In an effort to study the validity of this API for public implementation this data was compared using the safetyreportid field and was shown to have an 11% higher row count than the FAERS database (Shin, 2014) which is the result of the pre-processing done on the text that caused some reports to cover multiple drug-adverse event pairs and result in duplicate entries. This testing of the data shows the validity of using the openFDA for signal detection analysis granted, but suggests that it should be viewed under a more conservative threshold.

One of the biggest problems associated with the FAERS data itself is that while it does allow reports to be submitted by anyone and at anytime it often is mostly reported either by a physician or manufacturing company and primarily during the initial clinical trials of a drug. This lack of consumer interaction in the reporting process has led to only half of newly discovered serious adverse drug reactions being detected within seven years after drug approval (Alatawi & Hansen, 2017). This under-utilization of the report system by consumers is due to its voluntary nature as opposed to the required reporting that exists in industry. This reporting bias from a system that is verified as the gold standard of a drug’s side effects profile creates a misrepresentation of the actual rate of occurrence. Initiative then has to be taken to identify potential other streams of adverse event reporting. As a result of the ubiquitous nature of technology and the internet the most accessible and equalizing source for identifying additional reports is from online social media. The goal then becomes to find social media that is poised for use as repository of drug-adverse event data. Research into the use of Reddit for detection of drug-adverse event pairs has shown to have promising results (Foufi, Timakum, Gaudet-Blavignac, Lovis, & Song, 2019). The success of this source is attributed to two main factors: (1) It has a public facing and open source API of all data. (2) The majority of content is in long-text format due to the conversational nature of the website.

This paper will examine the impact and potential inclusions of novel adverse reactions to the drug safety profile of three drugs that each represent a different level of patient/consumer interaction: Lisinopril, Tigecycline, and Aspirin. These reactions will be established through social media data in combination with the openFDA platform through signal detection analysis developed in the R statistical software. The primary focus of this research will be the creation of a data pipeline that will ingest and transform the available data for use with disproportionality analysis to determine signal detection that uses ideas originally developed for use in the market basket problem, but that have now been regularly used in drug-adverse reaction detection. The pipeline will be evaluated across the three drugs to identify the best data source for rapid post-market surveillance of drug-adverse event pairs.

Overview of Pipeline

The only inputs required for this pipeline to work are the desired drug to be investigated and a personal API key that can be obtained for free from the openFDA (discussed further in Chapter 4). Figure 1.1 below provides a high level overview of the steps taken by the pipeline to build the calculated drug safety profile.

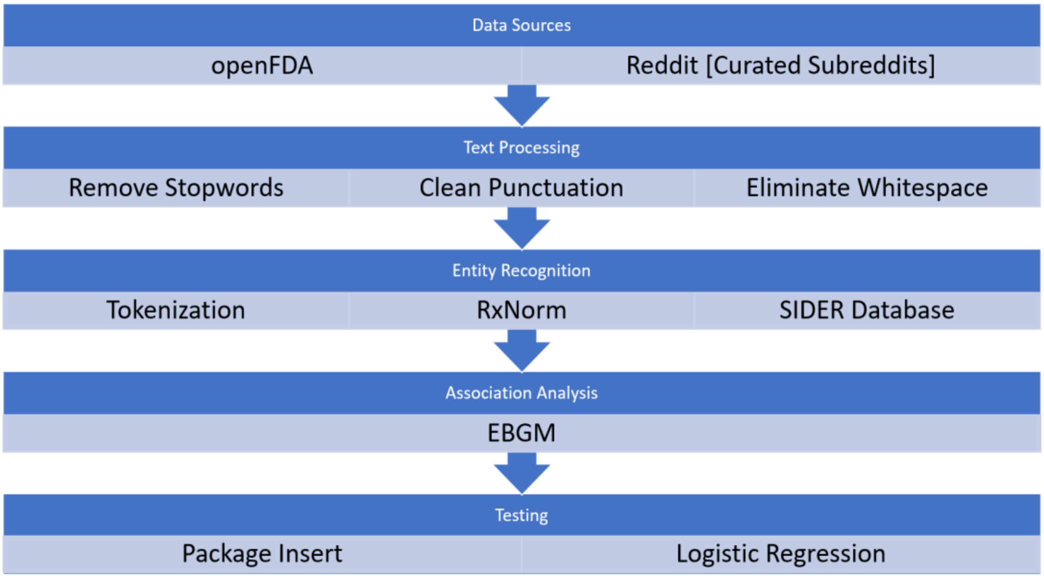


Figure 1.1: Pipeline Overview

CHAPTER 2: BACKGROUND

The Value of Open-Source Technologies

The largest sector that is pushing for increased adoption of open-source technologies is bioinformatics because it exists in a previously privatized realm of information. Now as technology has increased to a point where all the unstructured medical data can be analyzed their is further need to connect systems and make them publicly available. The main benefits of an open source health information system is the reduced development of costs, lack of vendor lock-in, and the increased potential for innovation and application usage (Janamanchi, Katsamakas, Raghupathi, & Gao, 2009). Open Source technologies rely on self-organizing communities to be able to come together and build aspects of a system from system databases to user interfaces and everything in between. The biggest struggle can come from a lack of guidance on a project by the main team that is driving the technology. Having a sponsored team at the base is one of the best ways to ensure open source technologies come to fruition because it gives resources and enthusiasm to the team and makes it more attractive for other groups to get engaged.

Role of Social Media in Pharmacovigilance

Social media is a fairly recent industry having only arisen in the past couple of decades and it really exploded with the advent of the internet. Originally the internet and personal computer users were quite rare and mostly constrained to hobbyists and computer scientists, but with the rising ubiquitous nature of the internet and high internet speeds, social media has come to reflect a diverse and complete view of humanity. The number one largest commodity in social media is user-generated content (Zeng, Chen, Lusch, & Li, 2010). This is true across most of the top social media sites like Facebook, Reddit, Twitter, and Second Life where the platform provides a format for communication and sharing of ideas and opinions and the actual content is then generated by the end user. This source of data has huge applications for businesses because it provides them with a consumer experience factor that previously would have required huge marketing budgets for them to generate and use for increasing sales and generating leads. In healthcare, this has a similar impact but instead of discussing their consumer experiences users are discussing their experiences as patients. This can help a pharmaceutical company identify drugs that have a negative connotation to patients online as potential places to focus their marketing resources or to identify diseases that don’t have an adequate treatment that may be potential pathways for break-through drugs. It is also a unique source of information and can provide regular users with potential drug treatment information and options that can be discussed with their doctor beyond what they see advertised (Yang, Kiang, & Shang, 2015) which is a great opportunity for the general public to stay up to date on medical advances by including them in a worldwide medical community. The FDA has identified several potential benefits that could be gained from using social media for adverse event detection. Social media makes it possible for faster signal detection and to get information from consumers that has previously been inaccessible. The FDA does also make a note of the potential pitfalls of using this information. There is a potential for duplicating reporting if the adverse events are serious enough that they led to hospitalization or a doctor’s visit. The terminology is not always consistent with medical reporting and requires a new vocabulary of colloquial medical terms to be added for further evaluation. However, the FDA does suggest that if all four of the requirements that are needed for a spontaneous report are available in a social media post then it should be treated as such. These four criteria are:

- Patient Identifier
- Product
- Adverse Event
- Reporter

Based on the format of most social media posts this information is readily available, given the assumption that the reporter and the patient are the same person. Identifying spontaneous reports in this format provides the inherent benefit that they are already anonymized. The requirement then becomes to be able to identify all potential mentions of a product and the reported adverse events associated with it.

Spontaneous Reports and Post-Marketing Surveillance

The FDA Office of Drug Safety is the primary reporting body that governs how adverse events should be reported and by whom. It started collecting these reports back in 1969 and has continued to grow since then. Reports are entered over the phone or through the MedWatch (Wysowski & Swartz, 2005) form both in paper and digitally. These reports are then organized and collected in the FAERS Database. This database provides evidence that certain approved drug products pose serious safety problems and in some cases has even led to those drugs being removed from the market. This is not to say that during clinical trials the drugs are not adequately studied or that the results are falsified. During a clinical trial, there are multiple inclusion and exclusion criteria that are needed to meet the efficacy requirements of a drug study but this is not the case in normal clinical practice and does not always give an accuracte example of real world application. Additionally, clinical trials are conducted under specific protocols and only last for about six or seven years before the drug goes to market (Takeuchi et al., 2008). Once a drug has gone to market it is the repsonsibility of the manufacturer to ensure that the drug is still studied under what is called post-marketing surveillance usually done through in depth review of individual spontaneus reports against a given drug.

The main purpose of collecting spontaneous reports of adverse events that are related to drugs either during a clinical trial or post-marketing is to increase the safety and usability of the drug to the patient population. This information provides physicians with the information so that they can determine whether or not a certain drug regiment is appropriate for their patients. (Freifeld et al., 2014). The majority of adverse events are reported in a hospital where it is reported to administration to allow them to carry out an investigation and determine root cause before reporting to an external body to aggregate the data. (Leape, 2002). In recent years, regulatory authorities have noted that a majority of adverse events went unreported because they were perceived as either not outside of reporting guidelines or were due to an individual not following protocol. Regulatory agencies have also noted that social media has become an increased source of medical information for general consumers as well as a potential place for reporting of issues or adverse events that are associated (Health & of Inspector General, 2012).

Signal Detection Methodologies

Signal detection is a vital tool in pharmacovigilance that is used to investigate the relative occurrence of drugs and their reactions. The overarching goal of signal detection is to be able to identify false alerts that would compromise the efficiency of a pharmacogivilance system. Traditionally this analysis has been done by an expert reviewing the reports that have been submitted. Over the last few decades with the rising increase in overall reports that have been submitted data mining techniques have been used to supplement the expert review of reports by creating mathematical models to determine whether a drug-reaction pair belongs on a safety profile (Puijenbroek et al., 2002). These methods are best validated by literature research and the existing drug product label to serve as indicators of positive or negative test results for determining whether they are accurate. The accuracy is most often tested based on their overall sensitivity and specificity scores.

CHAPTER 3: RELATED WORKS

openFDA for Adverse Event Signal Detection

In 2020, Alex Wright et al. published their findings on the use of the openFDA for identifying malfunctions and injuries following Balloon sinuplasty (BSP) surgery (Wright, Davis, Khan, & Chaaban, 2020). This study used the Medical Device endpoint to query all of the adverse events that involved BSP from January 2015 to December 2018. The Medical Device endpoint is connected to the openFDA thorught Manufacturer and User Facility Device Experience (MAUDE) ("MAUDE - Manufacturer and User Facility Device Experience," n.d.). This dataset contains a similar structure to that of the FAERS database being that the reports can be submitted by either a manufacturer, physician, or consumer and that the atomic grain of each individual report is the safetyreportid. In using the openFDA they were able to collect several data points about each adverse event including reporting type, sinuses involved, complication type, type of malfunction, source of event, procedure type, source of event, procedure type, and the submitted medical device encounter text. They were able to obtain 78 adverse event reports involving BSP for the management of Chronic Rhinosinusitis among other sinus conditions. Due to the nature of the data contained in the openFDA there was a limited patient history available to the research team, however the availability of possible adverse events does provide surgeons in discussing the potential complications of BSP with their patients in order to receive informed consent. The research conducted here takes an expert review of the reports generated from the openFDA rather than using any data mining methods to determine a signal.

Social Media for Adverse Events

In 2015, Yang et al. stated that drug safety currently depends heavily on post-marketing surveillance and is conducted through centralized volunteering reports, with a majority of adverse events never being reported (Yang et al., 2015). They grouped the data from targeted online communities focused on patient reaction reporting into “threads”, where each thread included the original post and the series of following comments focusing on the same discussion topic. In reviewing the lexicon of most social media topics they found that consumers do not use vocabularies and expressions which makes it difficult to match them to standard medical lexicons. Their research combined the typical text mining techniques used such as tokenizing, removing whitespace and punctuations, and removing stopwords. However, they did not perform word stemming. This study uses a lot of the same text processing techniques that are recorded here and will follow the example of not using word stems. It seems that the stems will cause a degree of confusion when using a domain vocabulary. They determined that a spontaneous report could meet the requirements by pulling the subject, username, timestamp, This research did not investigate further into the threads limiting the insight that these spontaneous reports can provide to a physician and keeping the information only high level for use with data mining. The social media sources used here are privatized and require different levels of user account status to gain access to the unique information. In terms of automation most do not provide a publicly available API and in some cases the site event directly prohibits download of the data (e.g. PatientsLikeMe.com)

Combining Social Media and the FAERS database

In 2020, Li et al. conducted a study to combine data available in the FAERS database with those that exist in Twitter and compare the performance of the combined system with signals generated by individual data sources (Zeng et al., 2010). They found that there can be a potential increased association of identifying adverse event signals by combining the two source systems, but suggested that they could not be combined directly and instead should be given different weights to reduce the influx of less serious reports that may come from social media. They suggest that Twitter's character limit makes it difficult to gather a full picture of the reporter's sentiment and suggested that future work also combine additional health websites and expand the scope of the reference standard. This research uses Reddit as its main social media source which often features content in a more discussion board format as compared to the one-hundred-and-forty character limit of Twitter.

Performance of Signal-Detection Algorithms

In 2013, Harpaz et al. conducted a study to determine which of the suggested signal detection methodologies were the most accurate and effective (Harpaz et al., 2013). The focus concerned the two main approaches of either Disproportionality Analysis or Multivariate Modeling Analysis. They provided performance guidelines for operating scenarios to inform the trade-off between sensitivity and specificity for different drug cases. These drug cases were classified based on what the overall drug indication was for such as acute myocardial infarction, acute renal failure, acute liver injury, and gastrointestinal bleeding. Their analysis is evaluated using the same accuracy of clinical diagnostic tests and is based on performance stats computed from Receiver Operating Characterstic (ROC) Curves and Area under the ROC curve (AUC). They found that across the different unique drug indications that the multivariate modeling approach through logistic regression had the highest AUC score compared to the other methods. In defining a positive test case of a drug-reaction pair for logistict regression they required that it meet at least one of three criteria:

- Event listed in Boxed Warning of active FDA strucured product label
- Drug listed as causative agent in Tidsale and Miller, 2010 “Drug induced Diseases”
- Literature review identified no powered studies showing evidence that refuted the drug effect.

A similar comparison of analysis is employed in this research in Chapter 6: Data Analysis, however instead of classifying performance across drugs based on indication this study compares performance across drugs using level of consumer interaction. The method for determining a positive test case for the logistic regression will only look at the information provided in the text of the FDA structured product labels. The literature review for accuracy testing is determined to be too wide of a scope to ensure reproducibility of the results.

CHAPTER 4: DATA SOURCES

A pipeline is created through identifying and ingesting multiple data sources. Each source serves a point of raw data that helps to transform the data in its raw form and convert it into a format that will be useful for performing signal detection analysis. The following data points constitute all of the pieces that are needed for analysis and in the order that they must be pulled for best effect:

- Cloud-based FDA Adverse Event Reporting System, openFDA
- Unified Medical Language System, RxNorm library
- Reddit
- SIDER Database of medical adverse events

All steps for ingesting and storing these data points are performed using the R programming language, but any software that can import CSV and JSON data files will be able to perform the same operations.

Accessing Data from the OpenFDA

The openFDA was created as a cloud based format to increase the accessibility of spontaneous drug reports. The cloud based platform makes it easy to identify the desired searchable fields along with the required APIquery structure and an interactive dashboard for quick testing of searches. For the analysis required here most drugs will require of an average of one thousand different API calls to the openFDA making it necessary to obtain a personal API key for reproducing the results. This will allow up to two-hundred-and-forty queries per minute and one-hundred-and-twenty thousand queries per day, per key. The openFDA API is a (Representation State Transfer) REST API that uses GET requests to deliver information directly from the source for the purpose of making the results reproducible and machine readable. With the API there is no need to perform any downloads of the data directly from the FAERS database or to maintain this information for performing the analysis on multiple drugs. Being a worldwide reporting organization means that the Spontaneous Reporting System is a mixed bag of nomenclature and protocols across different hospitals and different countries around the world. The diversity in the data around adverse events is why the FDA suffers critically from the need to standardize and clean the data (Hesha J. Duggirala (CVM), Joseph M. Tonning (CDER), Ella Smith (CFSAN), Roselie A.Bright (OITI), John D. Baker (CVM), Robert Ball (CBER), Carlos Bell (CDER), 2018). A major issue in collating this data is the implementation of text mining since the majority of reports are filed in free-text format. Hiring developers or purchasing products to perform text mining can greatly add to the cost of a project and this combined with the limited nature of information and the need to condense report information to no more than one word responses across multiple variables such as sex, date, product, etc. can become expensive and error prone. Additionally, without a clear list of labels being shared with all event reporters the drug names can be given in multiple formats as a generic name, brand name, abbreviation, or proper versus non-proper nouns. The openFDA API has met this challenge through creation of “openfda” fields that create a standardized version of these field variables. However, it doesn’t exist for all fields because for each drug name these different possibilities must all be programmed into the tool. The openFDA uses a multi-faceted url which can be further explored directly from the source to look at additional endpoints for Drug Product Labeling, Device Adverse Events, NDC Directory, etc.

The data from this API is available in three formats: (1) Data is available through API calls, allowing a search function in the form of a URL. This is a quick reference and has little to no barriers to access because it removes the need to understand the functionality of relational databases or how to build SQL queries. Aggregate data is readily available through API calls with a limit of one thousand rows of results. This type of access will be the primary focus of this report with the procedure requiring no more than one row of data per call. (2) Data is available through public download from https://api.fda.gov/download.json consisting of multiple JSON files of varying file size. This is an easier format than downloading the data directly from the FAERS database since the JSON file already delivers some of the relational design needed to connect drugs to their reactions. This method however would require an enormous amount of system memory power to extract meaning from these files and can be time consuming to perform. (3) Data is available through the openFDA S3 bucket hosted on AWS (Ryan Hood Vikram Anand & Rocamora, 2017) which allows for a cloud based connection to the raw data files. However, this method not only requires a background in using relational databases but also the AWS cloud infrastructure services. In addition this methodology will also be the most expensive option as each query through this method will be charged through the users AWS account. Across these different formats the complexity of use and the capacity for analysis increase proportionally making one no better than the other beyond the required scope of analysis. The scope of this project is to use the most easily accessible format which will be the first option to make calls directly to the API.

Unified Medical Language System and RxNorm

The UMLS integrates and distributes key terminology, classification, coding standards, and associated resources to promote the creation of more effective and interoperable biomedical information systems and services, including electronic health records. Drugs are manufactured by multiple different companies and are given different brand names each time. Consumers most often talk about the drug they are taking by brand name rather than by generic name so before doing the search through all subreddits the pipeline will need to connect to the RxNorm dictionary (Nelson, Zeng, Kilbourne, Powell, & Moore, 2011) to capture all associated drug names. This database of information is commonly used in verification for the analysis of the FAERS database as well so in creating a new data source such as social media it is important to use the same sources of information for verification of drugs.

User Thread Extraction from Reddit

There are millions of dissusion boards, called subreddits, within Reddit and performing a search through all of them for drug adverse events would result in an overly large and unfocused dataset. Instead it is more useful to only search through a small curated list of subreddits that deal with health and disease. The Reddit wiki hosts such a list of subreddits (‘https://www.reddit.com/r/ListOfSubreddits/wiki/health’) however the Reddit wiki format is not available for use via the API so it has to be gathered through a webscraper. This allows the search population to stay up-to-date as new subreddits are created and added to the wiki. This list is combined with an additional list of chronic disease based subreddits cultivated by other researchers (Foufi et al., 2019).This reduces the total list of subreddits down to only eighty three at the time of writing. This is a much more manageable and focused subset to search through, for the full list of subreddits used in this research see the Appendix. The data extraction will then loop through each of these subreddits and search for each of the potential drug names based on the information collected from RxNorm. The specific drug brand name that was used in each Reddit search is recorded against its available threads so that a drilled down analysis can be performed if needed. The scope of this research will focus on the larger subgroup of data based on the generic drug name.

The FDA requires that for something to be considered as a spontaneous report that it must contain four distinct criteria:

- Patient Identifier
- Product
- Event
- Reporter

Based on this criteria the data from Reddit had to be collected in a way that all of these fields could be gathered together. In this research these four requirements were met by taking the information and treating each identified thread as a spontaneous report. A thread in this sense is defined as the body post where a specific drug product was referenced and all of the associated comments associated with that post. The patient is identified based on the Reddit username and the reporter is always classified as a consumer in all cases because no credentials can be verified based on the data that is collected. The amount of associated threads that are collected will often vary based on which drug was searched, and the more common or older the drug the larger the amount of available threads.

SIDER Drug Database

The SIDER database contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts (Kuhn, Letunic, Jensen, & Bork, 2015). The available information that will be used for this study comes from three files “meddra_all_se.tsv.gz”, “meddra_all_indications.tsv.gz”, and “drug_names.tsv”. The SIDER database is currently in version 4.1 released as of October 21, 2015. The Side Effect names and Indications are copied from the Medical Dictionary for Regulatory Activities (MedDRA). MedDRA is the clinical terms and diagnoses that will be provided by a physician to a patient. These terms are available in both the lower level term and the preferred term for increased text matching potential. The SIDER database consists of adverse drug reactions and drug pairs that have been collected from biomedical literature and package inserts through natural language processing techniques. SIDER has been tagged with PubChem and MedDRA identifiers to quickly trace the drug side effects and indications. The drug names in the SIDER database only exist for the generic drug compound which in some cases consists of the chemical formula of the drug. If the desired drug being searched with the pipeline is not available in the SIDER database yet then the list of possible adverse reactions will use the list of potential reactions from the openFDA.

With the data now available in the system each post is parsed into substrings for entity matching with the list of adverse events. After parsing all words in this manner the list is cleaned down to only English words and MedDRA terms as a discrete vocabulary. This paper focuses on assigning discussions to either talking about drug-adverse event pairs or talking about drug-indications. The posts that are associated with the drug-adverse event pairs are then counted to be used in the disproportionality analysis outlined in Chapter 6.

CHAPTER 5: DATA PROCESSING

Chapter 4 focused on the upstream process of the pipeline to create a complete dataset. The data now needs to go through the downstream process to clean it up and transform it for performing analysis. The focus of this analysis is on signal detection so the data will need to be grouped into two subsets, one for drugs and one for reactions, which can be used for association rule mining,

Preprocessing the openFDA for Signal Detection

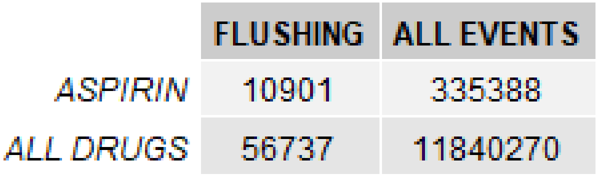
The key identifier for records from the openFDA is the safetyreportid, this is the field that can be used to find all of the information related to a single report and will provide more detailed information then is needed for the analysis performed here. The goal of signal detection analysis requires these safetyreportid's to be aggregated across different drugs and reactions in order to build a contingency table. In order to build a contingency table from the openFDA the information will be collected through the following GET requests:

- Count of all unique pair adverse events associated with the drug.
- Count of all instances of the drug across all reactions.
- Count of all instances of the reaction associated with any drug.
- Total count of all other drug-adverse event pairs for any other drug and reaction.

The additional processing operations needed will require use of a statistical analysis software. One of the drugs being studied in this paper is Aspirin because it is a common over-the-counter medication that will make it easy to understand and recognize the results. Before building the contingency table, it will be helpful to create an API call that tabulates all the reactions that have been seen with ASPIRIN to understand the variety of reactions that may be candidates for the drug safety profile. As an example the API call for identifying the occurrence of Flushing as a reaction to Aspririn is shown below.

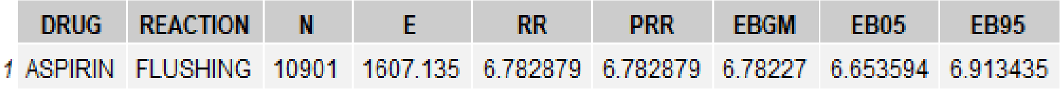
- The API call for all adverse events: https://api.fda.gov/drug/event.json?api_key=API_KEY &search=patient.drug.openfda.generic_name.exact:“ASPIRIN”&count=patient.reaction.reactionmeddrapt.exact&limit=1000
- The API call for ASPIRIN with FLUSHING: https://api.fda.gov/drug/event.json?api_key=API_KEY&search=patient.reaction.reactionmeddrapt.exact:“FLUSHING”+AND+patient.drug.openfda.generic_name.exact:“ASPIRIN”&limit=1
- The API call for ASPIRIN with all adverse events: https://api.fda.gov/drug/event.json?api_key=API_KEY&search=patient.drug.openfda.generic_name.exact:“ASPIRIN”&limit=1
- The API call for all drugs with FLUSHING: https://api.fda.gov/drug/event.json?api_key=API_KEY&search=_exists_:(patient.drug.openfda.generic_name.exact)+AND+patient.reaction.reactionmeddrapt.exact:“FLUSHING”&limit=1
- The API call for all drugs and all adverse events: https://api.fda.gov/drug/event.json?api_key=API_KEY&search=_exists_:(patient.reaction.reactionmeddrapt.exact)&limit=1

To create the contingency table there needs to be four different counts from the data. The unique pair of the drug with the adverse event, the drug with all other adverse events, the adverse event with all other drugs, and the count of all other drugs and other adverse events in the source data. The example contingency for the Aspirin-Flushing pair is shown below.



	FLUSHING	ALL EVENTS
ASPIRIN	10901	335388
ALL DRUGS	56737	11840270

Table 5.1. Aspiring-Flushing Contingency Table



	DRUG	REACTION	N	E	RR	PRR	EBGM	EB05	EB95
1	ASPIRIN	FLUSHING	10901	1607.135	6.782879	6.782879	6.78227	6.653594	6.913435

Table 5.2. Aspiring-Flushing EBGM Results

With a successfully created contingency table the user can now use these values to perform signal detection calculations and determine whether a specific reaction should be included in the drug safety profile. The formulas for calculating these association metrics are described in further detail in Chapter 6.

Text Mining of Reddit Threads

The most complex part of the data processing step is centered on taking the unstructured text found in the Reddit threads and converting it into structured data. The data source when completely pulled for a given drug consists of five columns: Post_ID, Product, Subreddit, Username, PostTitle, Thread. The main field to be examined here is the ‘Thread’ field which contains all the raw unstructured text. The processing of this text can be broken down into three main parts: cleaning, tokenizing, and entity recognition. The first step is cleaning where the text is scrubbed of any urls, punctuation and whitespace. A unique aspect of using Reddit for thread cleaning is that many subreddits features an automated bot. These bots are used to share basic information about the subreddit and its terms of use. To increase the value of the text from each thread it is important to remove these automated messages by deleting any comments that contain the phrase “I am a bot”. The second step tokenizing is done by splitting the threads at different window lengths. The different window lengths are from one to six words to accommodate for the longest adverse event name according to MedDRA. After the threads have been parsed into different window lengths the phrases are matched with stopwords to streamline the dataset. The R library qdapDictionaries provides an english dictionary that is also used to remove the different abbreviations and colloquial terms that are used across the different threads. Using the english dictionary to create a standard dictionary allows the text data to be as uniform as possible and decreases the chance that certain phrases may be used differently. The next step is to create a domain vocabulary by matching the SIDER database for indications and adverse events among the array of strings parsed out of the threads. To achieve the best results SIDER data is transformed through union between the lists of indications and side effects and then that list is filtered based on the drug being searched. There are occasions where a MedDRA term will be found as both an indication and a side effect, the associations for these are labeled as “Mix”. Due to the use of the MedDRA terminology, there are obscure spellings of several different conditions that need to be changed to match common spelling such as “diarrhoea” and “foetal” to match normal public use as “diarrhea” and “fetal” respectively. The nature of the data from SIDER is that it comes from multiple literature sources which will sometimes yield conflicting results where a certain MedDRA term may be listed as both an Indication and a Side Effect. In this case these results will be labeled as “Mix”, but they will treated as though they are Side Effects, this will allow them to be included in the signal detection tests.

CHAPTER 6: DATA ANALYSIS

The final step of the pipeline is to generate a usable product in the form of a recommended drug safety profile based on signal detection levels. These are signal detections of each drug with its adverse events. This chapter reviews the different types of analyses possible for determining a drug profile.

Signal Detection Analysis

The FDA has continuously strived to be transparent in the way they perform analytical methods so they have released the raw data in all of their databases plus they have also released their methodologies as well. The FDA has provided a white paper (Hesha J. Duggirala (CVM), Joseph M. Tonning (CDER), Ella Smith (CFSAN), Roselie A.Bright (OITI), John D. Baker (CVM), Robert Ball (CBER), Carlos Bell (CDER), 2018) on the strategies and types of data mining methodologies that they routinely apply in the surveillance of pharmacovigilance and drug safety. This makes it possible to evaluate the feasibility of performing these calculations for the general public and determining for themselves the risks associated with commercially available drugs. The FDA White Paper highlights several data mining techniques and ongoing projects involved in the assessment of safety reports. The focus of data mining efforts is to detect signals of drug associations and generate related hypotheses, but it is not used to establish causality that is the role of extensive clinical research. Extending this purpose to the openFDA and Reddit data populations we can gauge the level of efficacy that can be provided. Beyond this limitation the analysis of data is limited by two main factors (1) the lack of knowledge of the relative extent of use of the drugs involved, and (2) the overwhelming amount of various other alternative explanations for the observed association. Every statistical method applied to the data is focused on getting around and avoiding the impact of these two facts. The two main type of analysis used for signal detection are disproportionality analysis and logistic regression.

Disproportionality Analysis

The FDA routinely uses disproportionality analysis to create signal detection values that lead to associations between drugs and adverse events. One of the biggest difficulties with applying traditional data mining techniques to adverse events is that the database lacks a patient count identifier. This specific piece of data would allow users to see the number of patients who actually took the drug. As such, FDA safety reports rely heavily on proportionate differences rather than raw counts of associations among different stratifications to determine a relationship. There are a handful of statistical measures that can be used to identify these signals but the most powerful among these for the FDA is the Empirical Bayes Geometric Mean (EBGM) because it often leads to a reduced number of false-positive safety signals than its counterpart the Proportional Reporting Ratio (PRR). Several FDA centers rely on EBGM scores when performing surveillance activities. The EBGM is a Bayesian association rule mining method that was originally created with the market basket problem in mind and was developed for working with large databases where it is to hard to properly estimate values with a basic contingency table. The EBGM is a measure of association and is evaluated on the same scale as the Relative Risk (RR), so if the number is greater than one the strength of theassociation rises. The EBGM has three main inputs including the total count of occurrences (N), the baseline frequency of counts (E), and a vector, theta, of five optimized parameters. The estimated starting points for these parameters along with the calculations for EBGM are outlined by DuMochel (Dumouchel, 1999). As an input the EBGM requires all parts of a contingency table which can then be used to evaluate if a specific reaction should be included in the drug safety profile. The FDA takes a conservative approach to this analysis choosing to use the EB05 value which represents the lower end of the 95% confidence interval so any drug-reaction pair that has a value greater than the set cutoff value has a high confidence that a signal existsFor the code to see how this formulat was implemented on the available data see the Appendix. The next section outlines the mathematical functions of this formula.

Empirical Bayes Geometric Mean (Multi-Item Gamma Poisson Shrinker)

This formula defines the baseline frequencies using the following notations:

$N_{i,j} = \sum_k N_{i,j,k}$ as the number of unique pairs of drug *i* and event j, across all possible pairs *k*.

$$N_{i,k} = \sum_j N_{i,j,k}$$

$$N_{j,k} = \sum_i N_{i,j,k}$$

$$N_k = \sum_i \sum_j N_{i,j,k}$$

$$E_{i,j} = \sum_k \frac{N_{i,k} N_{j,k}}{N_k}$$

These are all the notations needed to create the Relative Risk.

$$RR_{i,j} = \frac{N_{i,j}}{E_{i,j}}$$

In order to perform the Empirical Bayes approach the following calculations assume that $N_{i,j} \approx Poisson(E_{i,j})$ with an unknown mean $\mu_{i,j}$ and that the associations are drawn from a common prior distribution $\lambda_{i,j}$. Then the relative risk measures can be adjusted for sampling variation and written like below.

$$\lambda_{i,j} = \frac{\mu_{i,j}}{E_{i,j}}$$

The calculation assumes that prior distribution, λ to be a mixture model with five free parameters.

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P) = Pg(\lambda; \alpha_1, \beta_1) + (1 - P)g(\lambda; \alpha_2, \beta_2)$$

These parameters are used to model populations, $\theta = (\alpha_1 = 0.2, \beta_1 = 0.1, \alpha_2 = 2, \beta_2 = 4, P = \frac{1}{3})$ by mixing them into two gamma distributions based on the α and β values. Assuming that θ and E are known then the distribution of N is

$$f(N; \alpha, \beta, E) = (1 + \frac{\beta}{E})^{-n}(1 + \frac{E}{\beta})^{-\alpha} \times \frac{\Gamma(\alpha + N)}{\Gamma(\alpha)N!}$$

Calculating these two gamma distributions will allow for the calculation of the posterior probability of λ. Here using P' to denote the posterior probability.

$$P' = \frac{Pf(N; \alpha_1, \beta_1, E)}{Pf(N; \alpha_1, \beta_1, E) + (1 - P)f(N; \alpha_2, \beta_2, E)}$$

With this posterior probability, the posterior distribution of λ will be $(\lambda|N) = \pi(\lambda; \alpha_1 + N, \beta_1 + E, \alpha_2 + N, \beta_2 + E, P')$

So calculating the posterior expectations of $log(\lambda)$ and using Ψ to denote the digamma function:

$$E[\log\lambda|N] = P'[\Psi(\alpha + N) - \log(\beta_1 + E)] + (1 - P')[\Psi(\alpha_2 + N) - \log(\beta_2 + E)]$$

Then the Empirical Bayes measure used to rank the cell counts, denoted as $EB\log2_{i,j}$ can be calculated as

$$EB\log2_{i,j} = \frac{E[\log\lambda|N_{i,j}]}{\log(2)}$$

and this value can be exponentiated onto the same scale as the relative risk for the final result. (Dumouchel, 1999)

$$EBGM_{ij} = 2^{EBlog2_{ij}}$$

Using a 95% confidence interval calculation allows the drug reaction pairs to take an even more conservative view of the data. With this consideration any drug-reaction pair with $EB05 = EBGM\exp(-2/\sqrt{N+1}) \geq 2$ is considered to be a known adverse event.

Logistic Regression Enhancement

The threshold value above is determined by best practices and previous research (Harpaz et al., 2013). This metric can be improved by using a test statistic to determine at which threshold value the signals detected best match the information available on the drug product label. To create a success measurement for the data the labels need to be ingested, these can be found at the FDA website in pdf format. For the purpose of this pipeline these PDFs were ingested individually and matched with the available adverse events reported from the openFDA and Reddit. These labels consist of a large amount of semantic and medical knowledge that are best when checked against by someone with a medical background. The pipeline here has limited semantic analysis capability but will attempt to use word stems to match adverse events to the labels. Prior to implementing the stem matching the overall list of adverse events found from the openFDA will be filtered based on two screening criteria. The first is a serious/death screen that consists of terms that the manufacturer would have included on a label to avoid potential lawsuits. The second screen will consist of terms based on the indications of a specific drug. For example the screen for Tigecycline will consist of the terms “infection”, “bacteria”, “pneumonia”, “staph”, and “klebsiella”. These are removed from the analysis because they can cause misleading classification of results.

After creating a success factor based on the drug product label the accuracy of the signal detection methodology can be calculated. This is based on the overall accuracy, sensitivity, and specificity of each adverse event across the different drug products and the different data sources. The goal of performing this threshold analysis is to determine at what value do the signals contirbute a significant signal to warrant investigation and whether or not for a specific drug the current threshold of $EB05 \geq 2$ should be raised or lowered. The nature of this analysis means that above a certain $EB05$ value if the label did not include the adverse event that these would automatically require further investigation and root cause analysis to understand the reason of their exclusion from the label. To maintain a focused approach the data is filtered to require that the unique pair meets at least a minimum support of (0.001) or $\frac{1}{1000}$ of the total number of reactions associated with the drug for each source.

To calculate the probability of a specific adverse event given a specific drug, the EBGM and EB05 values are used as the predicting variables and combined with the logit formula. The goal of performing this analysis is to find an EBGM and EB05 cutoff that can be used to achieve a minimum 90% specificity according to the available drug data.

$$\log \frac{P(AdverseEvent|Drug)}{1 - P(AdverseEvent|Drug)} = \beta_0 + \beta_1(EBGM) + \beta_2(EB05)$$

CHAPTER 7: RESULTS

The goal of this paper is to identify to what extent and with what level of accuracy a proper drug safety profile can be created during post-marketing surveillance through readily available web sources. This section of the paper will compare the accuracy of the signal detection analysis across the different data sources available for each drug. These drugs were chosen because they each represent a different level of consumer interaction with the medication. The goal of performing the comparison in this way is to examine the impact of consumer interactions on the availability of representational post-marketing surveillance. The drugs used in this study are:

- Lisinopril - Prescribed, regular daily use
- Aspirin - Over-the-counter, no prescription required
- Tigecycline - Last line of defense, hospital administered

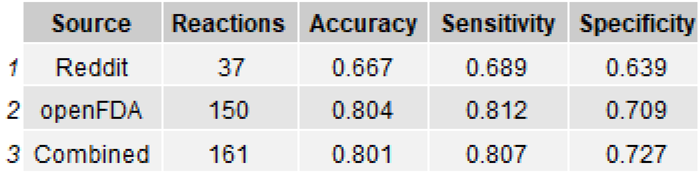
Each of these drugs are ran through the proposed pipeline to perform signal detection analysis and develop a proposed drug safety profile. The accuracy of each newly developed drug safety profile will be tested against the most recent available package insert label. The three data sources being used to generate a new drug safety profile are:

- Reddit - thread and comment text analysis
- openFDA - counts generated using the API version of FAERS
- The combination of the openFDA data with the Reddit data

Lisinopril

Lisinopril is helpful in lowering high blood pressure, preventing strokes, heart attacks, and kidney problems. Lisinopril belongs to the class of drugs known as angiotensin-converting enzyme (ACE) inhibitors. The mechanism of action for this class of drugs is to prevent the body from producing the enzyme angiotensin-II which causes the blood vessels of the body to narrow, when the blood vessels narrow it can cause high blood pressure and force the heart to work harder leading to an increase risk of heart attack. The drug was first approved in 1987, and has on average over one hundred million prescriptions per year for the past ten years likely due to the high prevalence of high blood pressure in the United States and its status as one of the first line treatments for certain groups of patients (Sean P. Kane, n.d.-b). The drug is administered orally on a once daily basis and requires prescription from a physician in order to get it from the pharmacy. According to GoodRx ("Lisinopril Prices, Coupons & Savings Tips," n.d.), the lowest available price for the tablets is estimated at $4.00 for 30 tablets.

The current drug package insert for Lisinopril lists 63 different possible adverse reactions (Merck, 2016). Based on the calculated EB05 with a cutoff of two or higher, the Reddit label has 38, the openFDA has 154, and the combined has 166 possible adverse reactions the full table of these adverse reactions is available in the Appendix. The next step was to evaluate the performance of the web-based data using a confusion matrix shown in Table 7.1. Both the openFDA and the Combined methods have a specificity of greater than 70% and even the Reddit data provides a specificity of 60%. These are meaningful results because it can show the potential ability of these sources to identify new adverse reaction pairs for post-marketing surveillance quicker than would be possible through the normal process.



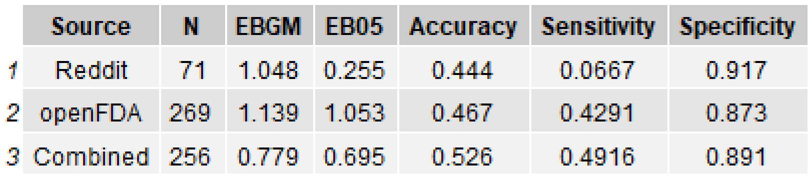
	Source	Reactions	Accuracy	Sensitivity	Specificity
1	Reddit	37	0.667	0.689	0.639
2	openFDA	150	0.804	0.812	0.709
3	Combined	161	0.801	0.807	0.727

Table 7.1   Lisinopril Signal Detection Accuracy.

This table shows that Lisinopril is consistent across data sources. This is likely due to the high number of prescriptions associated with this drug. It is the most common drug on the market (Sean P. Kane, n.d.-b) for high blood pressure and so will likely have the best cross representation between both social media reporting and reporting through the proper channels by manufacturers and physicians. In addition, to the higher level of consistency it also has the highest initial level of specificity with each data source giving a baseline value of between 60% and 70%. This provides even more value when considering that the power for the Reddit signal detection is less than 1% of the total power from the openFDA, and it demonstrates how well this small sample size represents the larger overall population. The results from Lisinopril are a great example case that using social media to determine signal detection for drugs is possible. Not every drug is quite so adaptable to social media analysis as the next two sections will show, there are certain criteria that Lisinopril meets which separate it from the other two drugs studied here:

1. It is easy to self-administer being taken orally rather than via intravenous fluids.
2. The drug is prescribed by a physician who understands the intended use of the drug.
3. It is a relatively cheap drug and easy to make part of a routine taking on a daily basis.

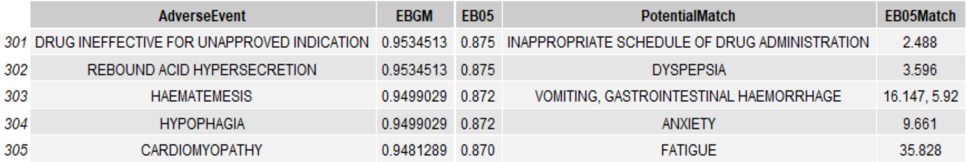
To maximize the specificity using web-based resources the EBGM calculations were enhanced with logistic regression using the EBGM and EB05 values as predicting variables. Based on these results in order to achieve this level of accuracy in the openFDA source data the values would have to be set at $EBGM = 1.14$ and $EB05 = 1.053$. These numbers are much lower than their current threshold and will greatly increase the total amount of the already large proposed drug safety profile from 154 drug-reaction pairs to 269 drug-reaction pairs. This is because the current package insert for Lisinopril makes mention of adverse events that are extremely rare based on the large power of adverse reactions reported. The details of the logistic regression enhanced profiles for each data source are listed in Table 7.2.



	Source	N	EBGM	EB05	Accuracy	Sensitivity	Specificity
1	Reddit	71	1.048	0.255	0.444	0.0667	0.917
2	openFDA	269	1.139	1.053	0.467	0.4291	0.873
3	Combined	256	0.779	0.695	0.526	0.4916	0.891

Table 7.2 LR Enhanced Lisinopril Signal Detection Accuracy.

This table shows that the logistic regression enhancement to choose the EB05 threshold requires a drastic decrease in the threshold in order to match all of the adverse events listed on the drug label. At this point the drug safety profile will be all inclusive of any potential side effects, but possibly too many which can lead to mistrust of the inclusion criteria. This logistic enhancement, while not viable for development of a new drug profile, does help to identify a key area that is lacking in the FAERS due to a non-standardized protocol for report submission. Where some physicians or manufacturers might submit each symptom individually whereas others only submit the overlaying disease which would explain why several adverse reactions should have higher EB05 values. Table 7.3 shows the bottom five adverse reactions associated with Lisinopril from the openFDA and while provide other potential options. This table makes it easy to see how each of the adverse reactions could have potentially been called something else that not only already had a signal, but by combing the multiple terminologies would give an even clearer signal of the potential risk.

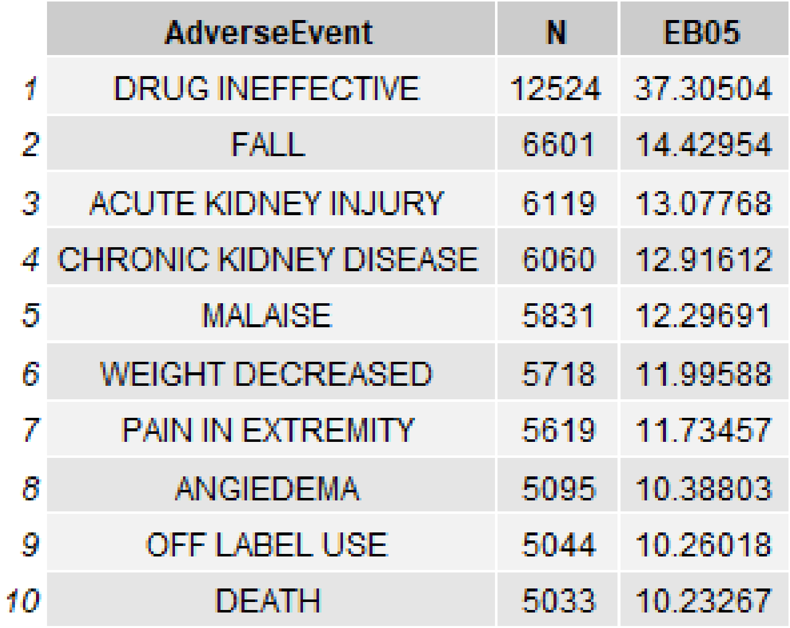


	AdverseEvent	EBGM	EB05	PotentialMatch	EB05Match
301	DRUG INEFFECTIVE FOR UNAPPROVED INDICATION	0.9534513	0.875	INAPPROPRIATE SCHEDULE OF DRUG ADMINISTRATION	2.488
302	REBOUND ACID HYPERSECRETION	0.9534513	0.875	DYSPEPSIA	3.596
303	HAEMATEMESIS	0.9499029	0.872	VOMITING, GASTROINTESTINAL HAEMORRHAGE	16.147, 5.92
304	HYPOPHAGIA	0.9499029	0.872	ANXIETY	9.661
305	CARDIOMYOPATHY	0.9481289	0.870	FATIGUE	35.828

Table 7.3 Lisinopril Reaction Matches.

Using this LR-enhanced EBGM cutoff value for the openFDA generated a signal for 269 adverse reactions. To highlight the most severe of these the suggested post-marketing surveillance profile of adverse events to be added to the drug safety profile for

Lisinopril that meet a $EB05 \geq 10$ and are not already included on the drug package insert are listed in table 7.4 below. These results seem to really highlight issues where Lisinopril was ineffective or that it resulted in death. This is surprising since this is a first line defense drug so there should be much more serious procedures given in order to prevent death due to high blood pressure. The kidney issues are also probably more related to the traditional drug indication since it is an ACE inhibitor making it an ideal candidate for treating diabetic kidney disease. Keeping these considerations in mind, the adverse reactions ranked 5, 6, 7 and 8 are all potential candidates to include for post-marketing surveillance of the drug. Weight decrease and malaise receive additional support based on the post-marketing reports currently available on Medscape ("Lisinopril," 2020) which identifies nutritional disorders which would be directly connected to these conditions. For the full list of potential drug safety profiles of Lisinopril across Reddit, openFDA, and the Combined data see the Appendix.

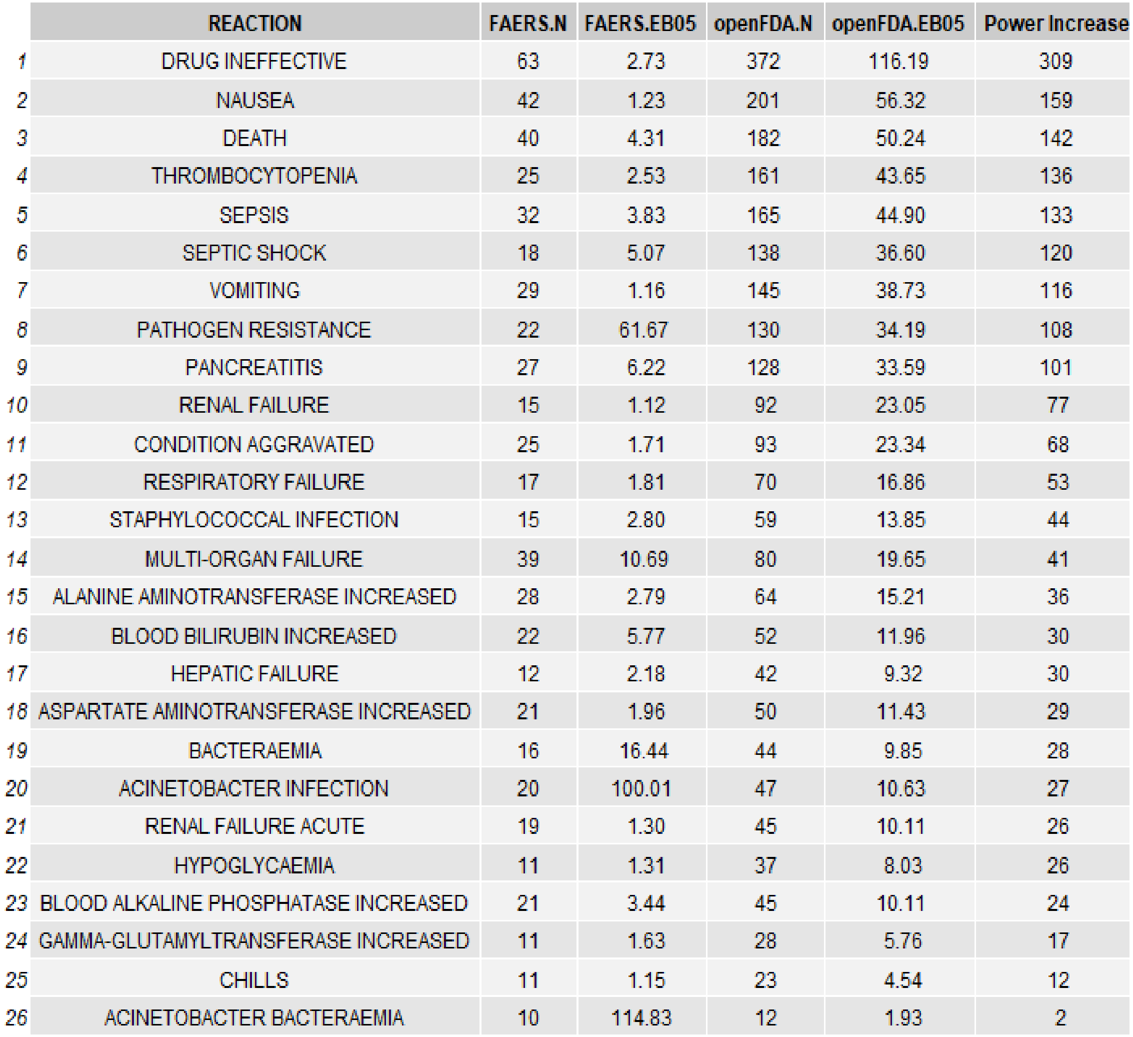


	AdverseEvent	N	EB05
1	DRUG INEFFECTIVE	12524	37.30504
2	FALL	6601	14.42954
3	ACUTE KIDNEY INJURY	6119	13.07768
4	CHRONIC KIDNEY DISEASE	6060	12.91612
5	MALAISE	5831	12.29691
6	WEIGHT DECREASED	5718	11.99588
7	PAIN IN EXTREMITY	5619	11.73457
8	ANGIEDEMA	5095	10.38803
9	OFF LABEL USE	5044	10.26018
10	DEATH	5033	10.23267

Table 7.4 EB05 > 10 Lisinopril Profile.

Tigecycline

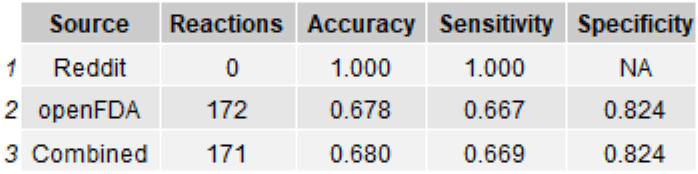
This section will review the drug safety profile for Tigecycline based on the reactions that meet the EB05 threshold value. Tigecycline doesn’t have readily available prescription because it is always hospital administered but it’s increase in use over time is compared with the results from a similar signal detection analysis performed using the FAERS database. This previous research consists of all the adverse events associated with Tigecycline during the period 2004-2009 (Kadoyama, Sakaeda, Tamon, & Okuno, 2012), that study has already calculated which drug-adverse event pairs created a signal based on the EB05 values. This information can be compared with the data from openFDA and from Reddit for Tigecycline extending the period from 2004 - 2020. There was a total of 3176 adverse events reported due to the use of Tigecycline throughout the period of 2004 - 2020 this is a marked increase from the original 1906 events found from 2004 - 2009. Table 7.5 shows how the counts of the original safety profile have increased over the past eleven years across the openFDA data source along with the EB05 signal level associated with each of the original adverse events. The table is sorted by largest increase in power based on the openFDA. Tigecycline was first introduced to the market in 2005, this comparison study can now show the increase of the drug prevalence from 2009 to 2020.



	REACTION	FAERS.N	FAERS.EB05	openFDA.N	openFDA.EB05	Power Increase
1	DRUG INEFFECTIVE	63	2.73	372	116.19	309
2	NAUSEA	42	1.23	201	56.32	159
3	DEATH	40	4.31	182	50.24	142
4	THROMBOCYTOPENIA	25	2.53	161	43.65	136
5	SEPSIS	32	3.83	165	44.90	133
6	SEPTIC SHOCK	18	5.07	138	36.60	120
7	VOMITING	29	1.16	145	38.73	116
8	PATHOGEN RESISTANCE	22	61.67	130	34.19	108
9	PANCREATITIS	27	6.22	128	33.59	101
10	RENAL FAILURE	15	1.12	92	23.05	77
11	CONDITION AGGRAVATED	25	1.71	93	23.34	68
12	RESPIRATORY FAILURE	17	1.81	70	16.86	53
13	STAPHYLOCOCCAL INFECTION	15	2.80	59	13.85	44
14	MULTI-ORGAN FAILURE	39	10.69	80	19.65	41
15	ALANINE AMINOTRANSFERASE INCREASED	28	2.79	64	15.21	36
16	BLOOD BILIRUBIN INCREASED	22	5.77	52	11.96	30
17	HEPATIC FAILURE	12	2.18	42	9.32	30
18	ASPARTATE AMINOTRANSFERASE INCREASED	21	1.96	50	11.43	29
19	BACTERAEMIA	16	16.44	44	9.85	28
20	ACINETOBACTER INFECTION	20	100.01	47	10.63	27
21	RENAL FAILURE ACUTE	19	1.30	45	10.11	26
22	HYPOGLYCAEMIA	11	1.31	37	8.03	26
23	BLOOD ALKALINE PHOSPHATASE INCREASED	21	3.44	45	10.11	24
24	GAMMA-GLUTAMYLTRANSFERASE INCREASED	11	1.63	28	5.76	17
25	CHILLS	11	1.15	23	4.54	12
26	ACINETOBACTER BACTERAEMIA	10	114.83	12	1.93	2

Table 7.5 Tigecycline Signal Growth for Adverse Reactions.

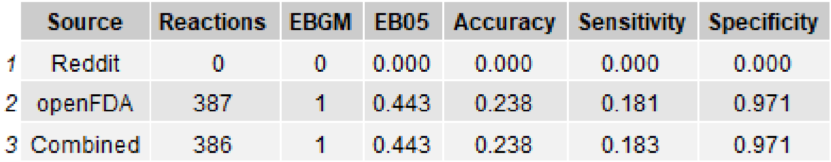
This information shows how important post-marketing surveillance can be for the overall safety profile of a drug. It shows a median increase of more than ten for the EB05 values of each adverse events. This indicates that for Tigecycline each of these reactions is now ten times more likely to occur when administered to a patient. Table 7.6 below provides the accuracy information of Tigecycline across the different data sources to estimate what new adverse reactions may be attributed to this drug.



	Source	Reactions	Accuracy	Sensitivity	Specificity
1	Reddit	0	1.000	1.000	NA
2	openFDA	172	0.678	0.667	0.824
3	Combined	171	0.680	0.669	0.824

Table 7.6 Tigecycline Signal Detection Accuracy.

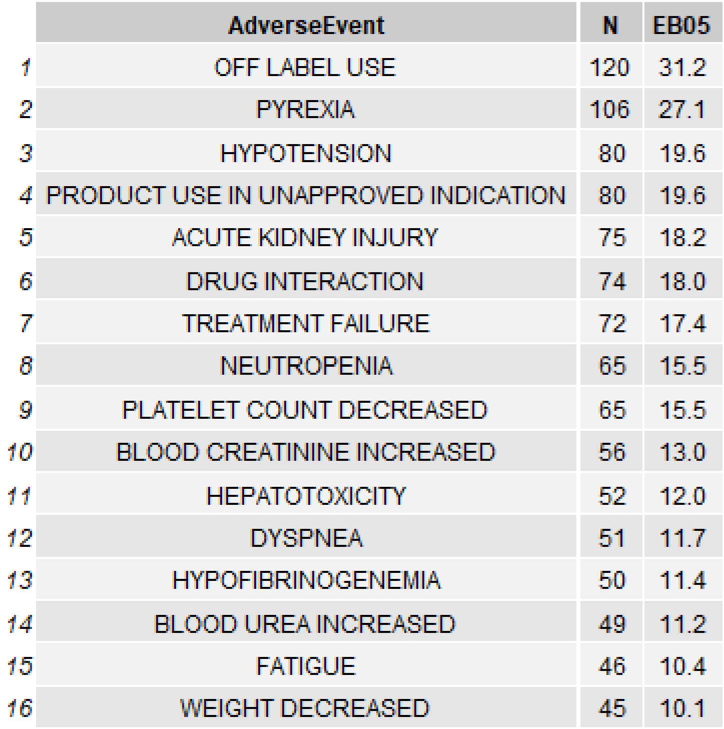
The accuracy measures for table 7.6 demonstrate one of the main limitations of using social media as a source for developing a signal detection and that is the lack of information about hospital administered drugs. In most cases, even if the patient did follow through to make the report via social media it would most likely be either a duplicate entry following a report previously made by a physician or it would not be as clinically significant since the choice to use a last line of defense medication is not always clear. The Reddit also has too small of a power to provide any valuable metrics, so no logistic regression analysis was able to be performed. Table 7.7 reviews the accuracy measures associated with the enhanced EBGM threshold to optimize specificity. The Reddit values have all been set to zero since no additional analysis was able to be performed. However, the Reddit data source still can contribute to the combined data source in terms of the overall count of reports as well as the number of reports available for each adverse event. This table reflects similar changes as identified with Table 7.2 for Lisinopril, because the lower EB05 threshold increases the number of reactions that potentially contribute a signal by over two-hundred additional reactions compared to the signal detection counts at the $EB05 \geq 2$ threshold. It is also interesting that even though the Reddit data doesn’t contribute any signal for the drug Tigecycline of its own, but the associated reactions are so commonly reported across



	Source	Reactions	EBGM	EB05	Accuracy	Sensitivity	Specificity
1	Reddit	0	0	0.000	0.000	0.000	0.000
2	openFDA	387	1	0.443	0.238	0.181	0.971
3	Combined	386	1	0.443	0.238	0.183	0.971

Table 7.7 LR Enhanced Tigecycline Signal Detection Accuracy.

There are many limitations and challenges that can arise when using these web-based sources for performing signal detection because of the need for secondary review by a physician, but the results can be very helpful in determining issues that may have been previously overlooked in a clinical setting. An example of this is, Hypofibrinogenemia, which indicates an abnormally low concentration of fibrinogen in the blood plasma. This specific finding is interesting because it has a strong signal detection of $EB05 = 11.43$, but is not located anywhere on the most recent package insert. Checking sources other than the package insert for this drug-reaction pair provides a validation of this result. The Medscape website, which is a publicly available physician’s reference that maintains drug data and is continually updated by staff pharmacists (“Tigecycline,” 2020), has Hypofibrinogenemia listed as part of the post-marketing surveillance results. Table 7.8 provides a list of additional adverse events for Tigecycline that don’t exist on the most current package insert but that have an $EB05 > 10$. The difficulty in assigning adverse events to a drug like Tigecycline is that it is usually administered to the person when they are already dying so it is hard to identify an association between reactions like death or organ failure as being associated with the drug itself. The full results of the suggested drug safety profiles for Tigecycline based on Reddit, openFDA, and the combined associations are available in the Appendix.

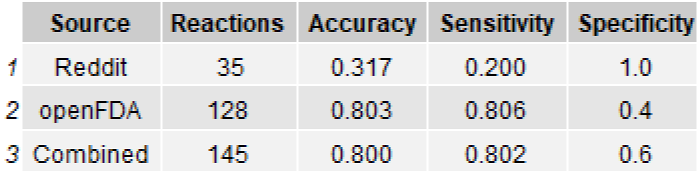


	AdverseEvent	N	EB05
1	OFF LABEL USE	120	31.2
2	PYREXIA	106	27.1
3	HYPOTENSION	80	19.6
4	PRODUCT USE IN UNAPPROVED INDICATION	80	19.6
5	ACUTE KIDNEY INJURY	75	18.2
6	DRUG INTERACTION	74	18.0
7	TREATMENT FAILURE	72	17.4
8	NEUTROPENIA	65	15.5
9	PLATELET COUNT DECREASED	65	15.5
10	BLOOD CREATININE INCREASED	56	13.0
11	HEPATOTOXICITY	52	12.0
12	DYSPNEA	51	11.7
13	HYPOFIBRINOGENEMIA	50	11.4
14	BLOOD UREA INCREASED	49	11.2
15	FATIGUE	46	10.4
16	WEIGHT DECREASED	45	10.1

Table 7.8 EB05 > 10 Tigecycline Profile.

Aspirin

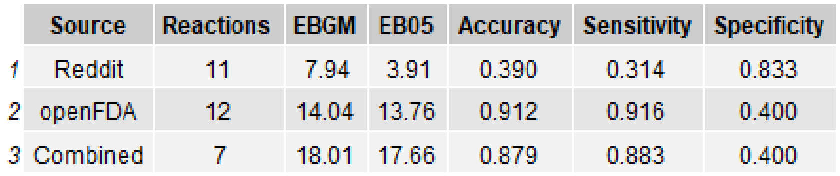
In this section the results of the signal detection analysis performed on aspirin are compared and evaluated. Aspirin is an OTC medication that can be purchased without a prescription. Additionally, it has also been prescribed an average of twenty million times per year over the last ten years (Sean P. Kane, n.d.-a). As an OTC medication, one of the largest risk factors of aspirin is due to consumers taking it to resolve non-indicated issues. The mechanism of action for aspirin varies based on the size of the dose. The main use of aspirin is for anti-platelet aggregation for the prevention of heart attacks. An accurate drug safety profile is estimated by using three different data sources here tested against the current package insert for 325 mg Aspirin (Bayer, 2013).



	Source	Reactions	Accuracy	Sensitivity	Specificity
1	Reddit	35	0.317	0.200	1.0
2	openFDA	128	0.803	0.806	0.4
3	Combined	145	0.800	0.802	0.6

Table 7.9. Aspirin Signal Detection Accuracy.

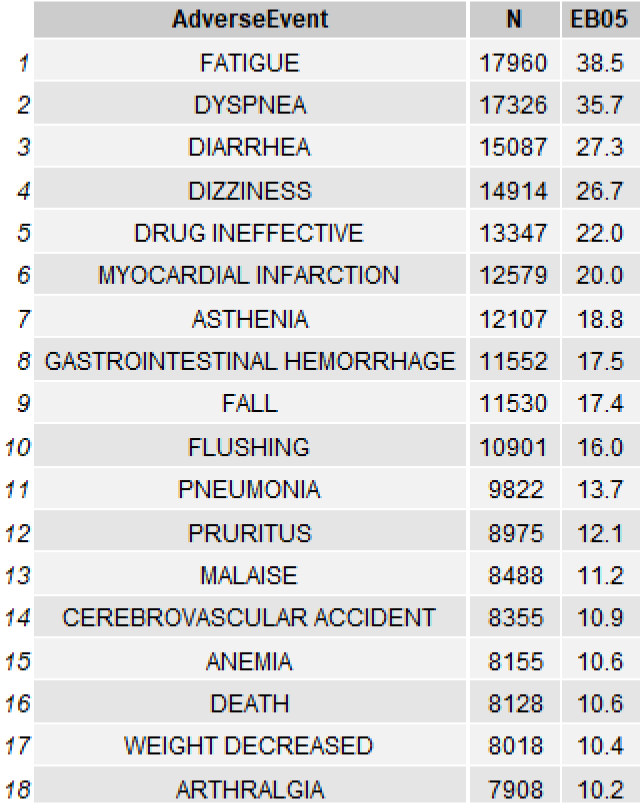
Table 7.9 above shows that across the three data sources Reddit data has the lowest overall accuracy, but also a 100% specificity. Specificity is the best determinant of performance for the signal detection method so this would indicate that Aspirin is best characterized from a full analysis with social media. This is also characterized because the current drug label for Aspirin only has six possible adverse reactions listed so Reddit was able to find a relatively large enough amount of posts related to each one. The openFDA has a much larger index of drugs and reactions which would explain why several of the drug-adverse event pairs did not meet the threshold requirements. The results for the openFDA were also not stratified by dosage which is especially important for drugs like aspirin that can commonly be administered at multiple different dosages based on the patient's indication. Because of the high accuracy of Reddit, it is assumed that most consumer reports are associated with the 325 mg Aspirin that matches the label.



	Source	Reactions	EBGM	EB05	Accuracy	Sensitivity	Specificity
1	Reddit	11	7.94	3.91	0.390	0.314	0.833
2	openFDA	12	14.04	13.76	0.912	0.916	0.400
3	Combined	7	18.01	17.66	0.879	0.883	0.400

Table 7.10 LR Enhanced Aspirin Signal Detection Accuracy.

Table 7.10 provides the logistic regression enhanced accuracy measures for Aspirin. Since the results from Reddit already achieved a 100% specificity the goal of this task was to increase the EBGM values to increase overall accuracy without sacrificing specificity. However, since there were so few adverse reactions indicated on the package label that even losing one additional reaction cost about 17% accuracy. This table also recommends an increased EB05 cutoff for the openFDA and Combined tables increasing to 13.76 and 17.66 respectively. To highlight the most likely of these the suggested post-marketing surveillance profile of adverse events to be added to the drug safety profile for Aspirin that meet a $EB05 \geq 10$ based on the openFDA and are not already listed on the package insert are included in table 7.11 below. Many of these adverse reactions are common and not usually serious such as fatigue, diarrhea, dizziness, or flushing. However, there are some with more severe health consequences such as gastrointestinal hemorrhage, this is supported by a study created to review the effects of long-term aspirin use (Huang, Strate, Ho, Lee, & Chan, 2011). There are several other reactions of note in this table as well but due to the use of Aspiring for so many different ailments based on dosage and the lack of a fully understood mechanism of action the others would require further analysis by a physician. For the full list of potential drug safety profiles of Aspirin see Appendix.



	AdverseEvent	N	EB05
1	FATIGUE	17960	38.5
2	DYSPNEA	17326	35.7
3	DIARRHEA	15087	27.3
4	DIZZINESS	14914	26.7
5	DRUG INEFFECTIVE	13347	22.0
6	MYOCARDIAL INFARCTION	12579	20.0
7	ASTHENIA	12107	18.8
8	GASTROINTESTINAL HEMORRHAGE	11552	17.5
9	FALL	11530	17.4
10	FLUSHING	10901	16.0
11	PNEUMONIA	9822	13.7
12	PRURITUS	8975	12.1
13	MALAISE	8488	11.2
14	CEREBROVASCULAR ACCIDENT	8355	10.9
15	ANEMIA	8155	10.6
16	DEATH	8128	10.6
17	WEIGHT DECREASED	8018	10.4
18	ARTHRALGIA	7908	10.2

Table 7.11 EB05 > 10 Aspirin Profile.

CHAPTER 8: CONCLUSION

Summary of Research

In this study, a pipeline was developed to determine the data mining applications of the openFDA and the social media site Reddit. The practicality of these data sources was assessed by comparing the results to pre-existing drug safety profiles developed directly from the manufacturer and printed on the current drug package insert. The purpose behind creating the data pipeline using the chosen sources was make it possible for any user regardless of association to the health care field to be able to quickly determine the likelihood that a drug-adverse event pair would occur. This was accomplished by creating contingency tables of each drug and its available reactions and then performing disproportionality analysis on these contingency tables to score and rank the likelihood of each association. To evaluate the performance of this ranking criteria a logistic regression analysis was performed to identify what an appropriate threshold cutoff should be to maximize the specificity compared to the test criteria. This online based approach eliminates the need to perform the multi-step process of creating a relational database because each data source ingestion piece can be performed through interactions with the available APIs that are explained and reproducible based on the API documentation of each source respectively. The full code for pipeline development is available in through the link in the Appendix.

The proposed pipeline was used to calculate a potential drug safety profile for three drugs: Tigecycline, Lisinopril, and Aspirin. These drugs are only a small sample of the available prescription drugs on the market, but they provide valuable information as to what aspects of the drug are most susceptible to being analyzed in this format. This research found that the pipeline was most effective at creating a proper drug safety profile for Lisinopril and least effective at creating the proper drug safety profile for Tigecycline. When an examination was performed against the drug profile of Tigecycline, the openFDA found over one-hundred additional drug-adverse event pairs that produced a signal, but there were few results found in social media. A drug such as Tigecycline which was developed in response to multidrug resistant bacterial organisms is sparingly used even in the hospital setting. As such the drug is tailored for a specific patient demographic which would in turn make the reporting of any adverse reactions not as plentiful. It is also difficult to determine the efficacy from last line of defense drugs due to lack of semantic reasoning behind the given adverse event which is a condition of the FAERS database as well because it requires an in-depth analysis from a physician to determine the true root cause. Lisinopril also out performed Aspirin according to the openFDA and Combined data sources because it is an OTC drug making it easier for the patient to misuse the drug, where Lisinopril is a once daily prescribed medication that is only given to patients with high blood pressure.

The most common limitations that analysts face in working with the FAERS database are the delayed releases of information and the misnomer of drug names due to different brands and dosage levels having their own names. This can make it difficult to correctly map adverse events to the same drug exposure. One of the biggest advantages of openFDA is that it functions as a Data-as-a-Service platform where the FDA team has already performed the data aggregation and preprocessing tasks. As with the nature of an API, this service works like a black box, but the FDA has made great efforts to increase transparency with public documentation. The openFDA team has made the data transformation scripts available on their Github. The efficacy of this implementation accompanied by the low barriers to access as well as the speed to receive information makes it a valuable source for both public and industry consumption. Adding a social media source such as Reddit adds another layer of robustness to this data and aids in overcoming the reporting bias. The reporting bias faced by FAERS and the openFDA alike can be overcome with the addition of this consumer facing tool. Reddit has a lot of potential compared to the other social media sites currently available because it is already anonymized and is written as a long thread or discussion-based format making it possible to conduct further investigation.

Future Work

The future implications of this pipeline are the ability to tailor the freely available information specifically to patient demographics. By applying stratifications such as age, gender, or location users can readily see the correlations and frequency of reported adverse reactions in their specific subgroups. The API call approach provides the ability to add increasing levels of specification when comparing the relative frequency of a drug-reaction pair making a feasible tool for identifying “at-risk” demographics. In addition, this can be further stratified for specific drug dosages and built to include drug-drug interactions. In order to properly implement stratification, the Reddit text needs to be mined for more than just keyword associations and be analyzed with sentiment analysis as well. Future projects around this work could start to build word tagging features into the pipeline based on the expected available stratifications.

The most common problem identified comes from the lack of standardization among the different adverse reactions. This issue is best demonstrated in table 7.3 where several adverse reactions for Lisinopril have a low level of association to the drug, but through a simple search it is easy to identify potential other terminology matches that all had EB05 measures above the given cutoff. As such a solution for increasing the ability to use this tool could be two-fold. First by generating a dictionary of medical terminology word stems that help to identify different tenses or use of a similar complication. Second by creating an overarching text connection between similar side effects that can all be grouped together.

Disclaimer: The information and results reported in this paper are not intended or implied to be a substitute for professional medical advice, diagnosis, or treatment. Do not delay seeking medical treatment or advice because of something you have read in this paper.

BIBLIOGRAPHY

Alatawi, Y. M., & Hansen, R. A. (2017). Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). *Expert Opinion on Drug Safety*, *16*(7), 761–767. https://doi.org/10.1080/14740338.2017.1323867

Bayer. (2013). Aspirin [package insert]. Retrieved from https://www.accessdata.fda.gov/drugsatfda{_}docs/label/2013/203697Orig1s000lbl.pdf

Dumouchel, W. (1999). Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *The American Statistician*, *53*(3), 177–190. https://doi.org/10.1080/00031305.1999.10474456

Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C., & Song, M. (2019). Mining of textual health information from Reddit: Analysis of chronic diseases with extracted entities and their relations. *Journal of Medical Internet Research*, *21*(6), e12876. https://doi.org/10.2196/12876

Fram, D. M., Almenoff, J. S., & DuMouchel, W. (2003). Empirical Bayesian Data Mining for Discovering Patterns in Post-Marketing Drug Safety. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 359–368). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/956750.956792

Freifeld, C. C., Brownstein, J. S., Menone, C. M., Bao, W., Filice, R., Kass-Hout, T., & Dasgupta, N. (2014). Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter. *Drug Safety*, *37*(5), 343–350. https://doi.org/10.1007/s40264-014-0155-x

Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., & Shah, N. H. (2013). Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clinical Pharmacology & Therapeutics*, *93*(6), 539–546.

Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics*, *91*(6), 1010–1021. https://doi.org/10.1038/clpt.2012.50

Health, U. S. D. of, & of Inspector General, H. S. O. (2012). *Hospital incident reporting systems do not capture most patient harm*. Department of Health; Human Services, Office of Inspector General.

Hesha J. Duggirala (CVM), Joseph M. Tonning (CDER), Ella Smith (CFSAN), Roselie A.Bright (OITI), John D. Baker (CVM), Robert Ball (CBER), Carlos Bell (CDER). (2018). Data Mining at FDA. Retrieved from https://www.fda.gov/science-research/data-mining/data-mining-fda-white-paper

Huang, E. S., Strate, L. L., Ho, W. W., Lee, S. S., & Chan, A. T. (2011). Long-term use of aspirin and the risk of gastrointestinal bleeding. *The American Journal of Medicine*, *124*(5), 426–433. https://doi.org/10.1016/j.amjmed.2010.12.022

Janamanchi, B., Katsamakas, E., Raghupathi, W., & Gao, W. (2009). The State and Profile of Open Source Software Projects in health and medical informatics. *International Journal of Medical Informatics*, *78*(7), 457–472. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2009.02.006

Kadoyama, K., Sakaeda, T., Tamon, A., & Okuno, Y. (2012). Adverse Event Profile of Tigecycline: Data Mining of the Public Version of the U.S. Food and Drug Administration Adverse Event Reporting System. *Biological & Pharmaceutical Bulletin*, *35*, 967–970. https://doi.org/10.1248/bpb.35.967

Kass-Hout, T. A., Xu, Z., Mohebbi, M., Nelsen, H., Baker, A., Levine, J., … Bright, R. A. (2016). OpenFDA: an innovative platform providing access to a wealth of FDAs publicly available data. *Journal of the American Medical Informatics Association : JAMIA*, *23*, 596–600.

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2015). The SIDER database of drugs and side effects. *Nucleic Acids Research*, *44*(D1), D1075–D1079. https://doi.org/10.1093/nar/gkv1075

Leape, L. L. (2002). Reporting of adverse events. *The New England Journal of Medicine*, *347*(20), 1633.

Lisinopril. (2020). Retrieved from https://reference.medscape.com/drug/prinivil-zestril-lisinopril-342321

Lisinopril Prices, Coupons & Savings Tips. (n.d.). Retrieved from https://www.goodrx.com/lisinopril

MAUDE - Manufacturer and User Facility Device Experience. (n.d.). Retrieved from https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

Merck. (2016). Prinivil (lisinopril) [package insert]. Retrieved from https://www.accessdata.fda.gov/drugsatfda{_}docs/label/2014/019777s064lbl.pdf

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs RxNorm at 6 years. *Journal of the American Medical Informatics Association*, *18*(4), 441–448. https://doi.org/10.1136/amiajnl-2011-000116

Puijenbroek, E. P. van, Bate, A., Leufkens, H. G. M., Lindquist, M., Orre, R., & Egberts, A. C. G. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety*, *11*(1), 3–10. https://doi.org/10.1002/pds.668

Ryan Hood Vikram Anand, & Rocamora, D. (2017). Analyze OpenFDA Data in R with Amazon S3 and Amazon Athena. *Amazon Athena, Amazon Simple Storage Services (S3), AWS Big*. Retrieved from https://aws.amazon.com/blogs/big-data/analyze-openfda-data-in-r-with-amazon-s3-and-amazon-athena/

Sean P. Kane, P. (n.d.-a). Aspirin. Retrieved from https://clincalc.com/DrugStats/Drugs/Aspirin

Sean P. Kane, P. (n.d.-b). Lisinopril. Retrieved from https://clincalc.com/DrugStats/Drugs/Lisinopril

Shin, J. (2014). Investigating the accuracy of the openFDA API using the FDA Adverse Event Reporting System (FAERS). In *2014 ieee international conference on big data (big data)* (pp. 48–53). https://doi.org/10.1109/BigData.2014.7004412

Takeuchi, T., Tatsuki, Y., Nogami, Y., Ishiguro, N., Tanaka, Y., Yamanaka, H., … Others. (2008). Postmarketing surveillance of the safety profile of infliximab in 5000 Japanese patients with rheumatoid arthritis. *Annals of the Rheumatic Diseases*, *67*(2), 189–194.

Tigecycline. (2020). Retrieved from https://reference.medscape.com/drug/tygacil-tigecycline-342527

Wright, A. E., Davis, E. D., Khan, M., & Chaaban, M. R. (2020). Exploring Balloon Sinuplasty Adverse Events With the Innovative OpenFDA Database. *American Journal of Rhinology & Allergy*, *34*(5), 626–631. https://doi.org/10.1177/1945892420920505

Wysowski, D. K., & Swartz, L. (2005). Adverse Drug Event Surveillance and Drug Withdrawals in the United States, 1969-2002: The Importance of Reporting Suspected Reactions. *Archives of Internal Medicine*, *165*(12), 1363–1369. https://doi.org/10.1001/archinte.165.12.1363

Yang, M., Kiang, M., & Shang, W. (2015). Filtering big data from social media – Building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, *54*, 230–240. https://doi.org/https://doi.org/10.1016/j.jbi.2015.01.011

Zeng, D., Chen, H., Lusch, R., & Li, S. (2010). Social Media Analytics and Intelligence. *IEEE Intelligent Systems*, *25*(6), 13–16. https://doi.org/10.1109/MIS.2010.151

APPENDIX

The list of subreddits used, the complete list of all calculated drug-adverse event profiles, and the full code used to develop the pipeline are available on Github. AdverseEventPipeline Repository