Adverse Drug events and Pharmaceutical Company profits[*]

Adam Tammariello
University of Colorado Boulder
Boulder, Colorado

ACM Reference Format:
Adam Tammariello. 2018. Adverse Drug events and Pharmaceutical Company profits. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

The Federal Food and Drug Association provides data on the outcomes of specific adverse events pertaining to patients who report such incidents. Since the health care industry in the United States is a widely debated topic, it would be interesting to find specific data about Pharmaceutical companies providing drugs that have severe outcomes. With the combination of stock reports and profit margins, compared to the data provided by the FDA, it would be interesting to draw specific correlations between profit margins and years that the FDA reported high mortality rates related to a drug, or other levels of severity. The hope would be to provide specific reports that would contribute to the discourse of privatized health care.

With more progress made, it would be interesting as well to see if we find the frequency in which adverse events are reported. If there is any pressure on pharmaceutical companies to properly investigate whether a Drug has high rates of Adverse events, and whether or not that specific drug was fixed and or recalled from clinical trials. If we can see the rate of which action is taken on these drugs, it can give us a more rounded look on the motivations when pharmaceutical companies release certain products into the public market.

In general, I'm hoping to find a proper evidence of the lack of attention that big pharma puts on their products. There seems to be a high profit for the industry, yet certain drugs that are often over prescribed by these companies show some sort of a profit agenda rather then providing a medical solution to a specific patient's problems. This would really reveal more doubt in the health care system, especially the affect Big Pharma has on doctors and patients.

2 LITERATURE SURVEY

Prior Work done within this topic have shown to correspond to prescription drugs and drug abuse in and out side the United States. Theses studies are done with FDA Data and without.
In a Study done by Brooke Belong of the Odyssey, Health Care corporations circulate highly addictive drugs as prescription drugs, and draws a correlation between these prescriptions and increasing drug abuse rates within the country. These statistics are drawn from the FDA.
https://www.theodysseyonline.com/legal-drugs-are-more-dangerous-than-illegal-
In a Study done by Donald W. Light, he draws a correlation of adverse outcomes from FDA reported new prescription drugs to pharmaceutical companies creating high risk prescription drugs to push large profit margins. This data incorporates FDA adverse Event reporting.
Harvard Study

3 PROPOSED WORK

3.1 Data Collection

To begin this investigation, we will need to collect specific data from the FDA Adverse Event Reporting portal. This can be accessed either directly through the FDA website, or through the data collection website Engima.org. From here we can pick specific years, in this case we will choose a period of 3-5 years. This will depend on what access we have to specific stock and profit information of specific Pharmaceutical companies that we have access to. With a time frame decided we can then find the corresponding data for Drug Information within Adverse Event Reporting, as well as Patient outcomes with the corresponding Id's. Then once we have collected these two data sets, we find the stock information or profits of the specific corporation that has the highest rates of severity amongst their products or specific product.

Other data that would need to be collected is the manufacturing records of drugs such as quantity, and if possible amount distributed over a 3 year span. This information is useful in outlining the amount of drugs created of high severity or low severity in which the specific company spends money on.

If possible too, I will search if there is specific data of specific products and their profits and show how much of a companies income has been from these severe drugs.

3.2 Data Cleaning

To properly use the specific dataset, we will need to make sure we can process it easily to better provide visualizations. Within the FDA Adverse Event Reporting - Drug Information, and FDA Adverse Event Reporting - Patient Outcome, There are varying aspects to the data sets that can be cleaned. Since we are centralizing on levels of severity, data inputs that do not have any data corresponding to an outcome in Patient Outcome, will not be useful to us. Other non-specific descriptors of drugs as well as variations of missing descriptors. This allows us to easily gather the specific data that we need so as to create a system of evaluation. We have found

[*] Produces the permission block, and copyright information

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

that we need to also fill in specific missing information that has been incomplete from both the FDA Drug Information and Patient outcome.

Since applying specific numbers to the Outcomes, some information has been left incomplete when filing a report about a specific drug where the outcome hasn't been properly recorded. We must make a decision whether to exclude this from the data set, or take other means of filling in the incomplete data. We also must correspond specific dates of the patient outcomes and create system to fit that over an easily compatible time frame with our stock information and our profit information as well as the possible manufacturing numbers we can find.

3.3 Data Preprocessing

To process the data after pruning unimportant data, we must change specific aspects of the data to be able to properly visualize it. Within the Patient Outcome dataset, the attribute "Outcome" comes with a small initial, as well as a description. For our use, we will assign each specific outcome to a specific number so as to understand the frequency of outcomes of specific drugs. This will make it easier to compare frequencies. We then combine the Patient Outcome Data set with it's corresponding ID number within the Drug information data set.

To apply this knowledge and compare it to our Profit information, we need to create a specific system in which we can assign a specific drug with "severity score". This score will be compared to the Rise and Fall in profits of the corresponding pharmaceutical company that manufactures the specific Drug, There by showing a correlation between the Profits over the three year span, and the severity score of the individual drugs reported those years.

We can use this score as well so as to set up a k-means cluster of both the severity of specific drugs, and the frequencies of specific drugs reported.

3.4 Data Evaluation

Finding the most frequently reported drug by using priori algorithm From the most frequently reported drugs, we find the most frequent outcome (death, disability, minor adverse effect etc.) and looking at the most severe outcomes. Showing stock behavior of manufacturing company with the most severe outputs and compare to the frequency of Adverse reports made those years.

We may even show which specific drugs within the list of Adverse Events have been manufactured more frequently then others. This will show the attention that pharmaceutical companies put onto these FDA reports.

We then can create visualizations of severe drugs versus the profits of a specific quarter and 3 year span, and the stock information over a specific quarter and 3 year span. This information shows wether severe drugs have a positive or negative correlation to investors, as well as profit.

We can create a few k-means graphs that show us the similarties of specific types of drugs. This can also be applied to the frequencies of these specific types of drugs. This will provide specific information of each drug.

This information will then be applied to the manufacutring numbers. If specific sever drugs have any correlation of quantity compared to other mass produced drugs, we can show some importance these companies have on producing severe drugs as part of their profits.

3.5 Differences

In comparison to the prior work done, our work will be more generalized on how the public domain views pharmaceutical companies during periods of high mortality or injury, as well as the profits they gain from high risk drugs. This will connect or our main goal, which is contributing to the discourse of privatized health care, as well corroborate previous studies done in the same field. Our study will only be for The United states as well, where as the Harvard study was more global. Since the United States is one of few countries with privatized health care, we can see how domestically based pharmaceuticals gain profit during the death of domestic drug related deaths.

Another correlation can be drawn as well on the importance of Adverse Events reporting information to these specific pharmaceutical companies. If a drug is reported to be very severe, or by our standards have a high severity score, yet over the same time frame shows an increase or decrease in quantity produced, then we can see the companies do not spend time incorporating their business interest with the products they are manufacturing.

4 DATA SET

4.1 U.S FDA - Adverse Event Reporting - Drug Information

This Data set contains over 1,020,344 data points pertaining to specific attributes. The overall purpose of the data set is to show specific case ID's of a specific adverse event, and the corresponding drug information used in that event. There are 14 Attributes in the set which have the following meaning:

- IDr - specific case ID
- Drug Sequence Identifier No.- Another identifier for specificity
- Drug Role - Whether or not the Drug is suspected to be the cause of the even
- Drug Name -Name of medicinal product
- Validated/Verbatim - Whether or not a trade name is used or a Verbatim name
- Route - method in which drug was taken
- Dose - how much of a drug was taken
- Other identifiers ·

The link to the Data can be found here: Drug Information

4.2 US FDA - Adverse Event Reporting - Patient Outcome

this data set contains specific outcomes for a corresponding identifiers from other tables within the Adverse Event reporting data collection. From here we can less data points, however there are multiple points within the Drug information Data set which have the same identifiers. These correspond with those specific drugs and their outcomes. This Data set has 4 attributes:

- Idr - Specific Identification number
- outcome - and abbreviation for an Outcome (Death, Disability, Hospitalization etc.)
- Outcome Definition - corresponding defintions to outcome abbreviations
- Quarter - which quarter of the year the event happened.

The link to the data can be viewed here: Patient Outcome

4.3 Historical Quotes - NASDAQ

This data set will be specified once the initial work has been put in to decipher specifically which drug has the most adverse affect, as well as the main distributor of said drug. From there we can pick the stock information and hopefully profits of that specific company during the time frame specified and look at the information provided to see a correlation. this data set has 4 attributes:

- Time - the time in which the quote was recorded
- Open - The price of stock at Open
- High - The highest sold stock
- Low - The lowest Sold Stock
- Close- Price at closing time
- Volume- the amount of transactions that happened at this time

The data can be viewed here: Pfizer Historical Quotes

5 EVALUATION METHODS

Correlating our data will be fairly easy, the time consuming part will be processing our data to determine which specific drug has the most frequent severity reports. To do this we will use an apriori algorithm to determine the types of drugs that have the most sever outcomes. Then we will use this to determine which specific drug has either the wides ranging severity, or the most sever outcome by probability. From this list we will widdle down a small selection of drugs and their corresponding companies and find stock reports for each company during a specific time frame. We can find a correlation to each specific companies profits or stocks during a year with a high severity drug reporting. This will show us whether or not companies gain or lose profit if a drug has been shown to hospitalize patients or even kill them.

We will try to visualize which specific type of drugs have a higher frequency to better understand the importance companies put on specific drugs, and compare this with the amount manufactured that year, to see if there is a correlation between severe drugs and manufacturing. This could show us that companies recall or diminish distribution of these drugs, or that they tend to pay less attention to these reports

6 TOOLS

6.1 Excel

Using Excel we can visualize cvs files from the ouputs created by our data processing algorithms. This will help us check our work and even create simple visualizations of data that we have so far. this will be especially helpful in collaboration so that we can share results with eachother without having to run written programs in Jupyter and keep us organized when we finalize our results.

6.2 Python

In class we have had a lot of practice with the python programming language. This has proven to be especially efficient and easy to use to process data in very versatile ways. Using libraries such as numpy, pandas, cvs etc.. we can read in our data sets and clean our data sets of arbitrary information and even produce detailed visualizations of our frequency vs drug tables, as well as the correlations between profit rates and severe drugs. This will be our workbench for properly parsing data and sorting it.

6.3 Jupyter

Jupyter Lab and Jupyter Notebook will be very essential for the group to use python. This will create a basis in which regardless of the system, each member will be able to use and run python algorithms that we create. We can also use this to easily save files and push to our version control platform.

6.4 Github

We will be using github as our main source of version control. This will allow us to develop in tandem and contribute to a main product which will be the collection of data and visualizations as well as the programs we create to parse specific information. It also gives us the ability to look at participation in the group as well as a timeline of our progress.

6.5 Slack/Trello

We will strive to use Slack as a means of communication to properly discuss topics and share specific files and resources. It will also allow us to know when a member of the group pushes anything to the github so that we can approve any pull requests. We may also use trello as a resource to keep track of what work is still needed and what work has been completed to keep on track with our milestones.

6.6 Dataiku

Dataiku has been recently integrated as an alternative to the limitations of excel. This tool allows us to look at the data files in a more readable format, which includes better tools to properly visualize our data sets. It also has integrated IDE's for python, SQL, as well as short cuts that can greatly increase the process of cleaning and pre-processing. It also allows us to store our data on their cloud servers.

7 MILESTONE

- Data cleaning done by March 23rd
- Data Preprocessing done by April 20th
- Visualization of Frequent Drugs Reported
- Visualization of Specific Drug Severity

- Visualization of Drug Quantity
- NASDAQ Pharma Data correlation by April 27th
- Drug Severity versus Profits
- Drug Severity versus Stocks
- Drug severity and Quantity
- Write up and Slides created
- Presentation ready by April 6th

Work has already begin on data cleaning. However with spring break as well as different classwork we will try to prioritize the aspects of the work that will take longer than others. Most do these dates will be subject to change to adjust for different roadblocks along the way.

8 COMPLETED WORK

Since most of my team has dropped the class this has become a singular effort to mine this data. So far there are only a few things I have completed, but alot of work is left to be done so far.

8.1 Data Cleaning

The data cleaning has proven to be the most time consuming part of the process. Working with over 3 million data points, I've had to isolate specific aspects of each Data set, specifically within the Drug information data sets. I've stripped certain information such us Experiation data as well as certain attributes like "Rechallenge" and "Dechallenge". These specific attributes are irrelevant to my reports.

I've also removed certain information and incomplete data that lack any descriptor of a "drug name" that does not contain a specific corresponding outcome. This is partly due to the inability to fill in specific information. This data set has specific avents that are isolated and since I've had to apply my own numerical system for these specific outcomes, It would be innacurate to create an outcome for an event. This means however that I have to scale down my severity scale system, which I haven't properly created.

I've also managed to combine the two different data sets for FDA - Drug Information, and Patient Outcome. This however took some more cleaning since certain ID's numbers would not correlate completely with specific drug information that was missing from either data set. I'm currently figuring out a way to solve these issues.

I've also sorted out Which names are recorded as "1" verbatim "2" trade name. This is important since a non-specific drug name may make it harder to distinguish which company manufactures which drug. This requires more research within the FDA public data to see if I can properly find a way to correlate a drug name with it's trade name, which can lead me to a manufacturer.

I've also corresponded the dates in which Adverse reports were made into ranges of months. Since the dates are recorded in Quarters, I've had to take corresponding stock data and separate their quotes into quarters, however their profits of specific years are separated by specific quarters so I need to choose a better method in which to draw my correlation.

9 WORK TO BE DONE

There is alot of work to be done before I am ready to properly evaluate this information. So far I've only managed to clean most of the data, but the real task will be in processing this data since I need to create a system in which I can rate each specific drug with a severity score.

9.1 Data Pre-processing

The severity score will probably be a make up of averages for each specific drug created. The main method I think would be to take a specific drug name and count the amount of times the most severe outcomes have been associated. This would be (Death, Hospitalization, Paralyzed etc.) Each outcome would correspond with a score given between 1 - 10. These scores would then be added up and taken an average which would give us a severity score. This would be repeated for each subsequent drug in the list. From there I can compare each drug and find the certain drugs that have atleast a specific severity to better centralize the most sever drugs that are analyzed.

Once the severity score has been established I'll find each specific drug manufacturer and plot a line against the profits of the corresponding manufacturer. Other visualizations I'm hoping to produces is perhaps a K-means graph of specific drugs and their severities to see if there is a correlation of specific types of drugs that have higher rates of severity then others. The taking this data produce a visualization of the companies who manufacture/endorse the most severe drugs that have been reported.

I'll also have to find the frequencies over time in which specific types and brands of drugs have been reported over these specific quarters established. If a specific drug has been reported more frequently and there is a patters such as specific brand is reported more frequently, this could provide some detail on the company that manufactures it.

9.2 Evaluation

The visualization aspect of this will be very time consuming. Picking exactly which way to visualize the data of different types as well as creating a way to prepare my data to interpret it will be very challenging. So far I've decided on some specific graphs but nothing concrete. So far the K-means graph seems to be the easiest to create since the averages of the severity scores will be easy to create.

Choosing how to compare the data with the profits and stock information that would be intuitive is somewhat challenging as well. I've considered plotting a line of specific drug severities and show the rates of profit gain and loss over the same amount of time to show how the profits have either gone up or down during this time period. However this may mean I would create individual severity scores for specific quarters over this time span, since a severity score for an entire 3 years may be too big of a generalization, or may not do an adequate job visualizing a correlation.

RESULTS SO FAR

So far the only results I've seen are notes made during the cleaning process. Certain cold medications I've noted to result in sever outcomes such as hospitalization. Since there is so much data I haven't had the time to pick out specific patterns but as the processing part of the project continues the frequencies and patterns will probably reveal themselves.

10 ABSTRACT

Pharmaceutical companies and medications are often a large topic of conversation in American Culture. Because of privatized healthcare, our own well being is often a price point for health care providers and the pharmaceutical companies. The dichotomy between personal health and price is a somewhat taboo subject, but never the less an important aspect of our economy.

There were several questions I wanted to answer by looking at specific data that is available to the public. Given that pharmaceutical companies make profits off of the medication they manufacture, I wanted to get a profile for their headliner drugs. Specifically, I wanted to look at the types of outcomes that occur with their specific products. According to the FDA Adverse Event Reporting system, each specific outcome to a drug reported in a specific year is recorded. This can tell us alot about what kind of drug has what kind of outcome for a specific year.

The first question I wanted to answer, the first being, in a specific year, what or which kind of medication is most frequently reported. The intuition I brought into this was that the most severe drugs, or drugs with the highest mortality rate would be the most frequently reported. However, according to the research, the least severe drugs, specifically "over the counter" drugs such as ibuprofen, a common painkiller, and Folic Acid, Vitamin often used during pregnancy, were among the most frequently reported drugs.

The second question I wanted to answer was which medications have the highest severity in a given year. Among the results, the most sever drugs ranged between Anaspren, a blood cell drug, and Fentanyl, which is prescribed as a painkiller in low doses. These scores ranged from 2.9 -3.5 in severity out of 6.

The third question I wanted to answer was of the companies that sold severe drugs, how much profit was gained or lossed over a 3 year span, and did this information affect their stock patterns. The results showed that over a 3 year span, most companies experienced a 2 to 20 percent increase in profit, however there did not appear to be any trend in their stock prices.

11 INTRODUCTION

These question, as stated in the abstract give a general idea of what information is tied to specific medicinal products. That fact that this knowledge is public is a powerful tool. The FDA has specific and extensive data on outcomes related to pharmaceutical companies that can be used to hole them accountable.

In general, understanding the which kinds of drugs are frequently reported provides useful information on medications that are sold in high volumes. Ibuprofen is such a common "over the counter" drug that most would assume that it would have a minimal risk. Though this might be true, the amount of complications implies that more care should be taken when consuming something that we have such liberal access to.

Medications that have high risk is something that is usually discussed between a doctor and patient before being perscribed. However, most people don't think to actually look at the outcomes, or that most people do not think that these outcomes are recorded. Even though something being a highly severe drug could be obvious for those being prescribed the medication, the data allowed me to find certain products that contain a highly addictive and completely destructive drug Fentanyl that is sold on the market. This is important especially in the opioid epidemic we are currently living in.

Though correlation does not imply causality, it is interesting to see how companies are affected by Adverse events and their products. To some, this assumption could already be made, or at the very least not be surprising, however seeing the trends between increasing rates of reported adverse events and profit puts exactly how lucrative the pharmaceutical world, and gives a general idea how much people gain when people go through complications.

12 RELATED WORK

12.1 Data Set

The data set to answer these specific questions were largely provided by the FDA Adverse Event Reporting System. The FDA records specific drug information and patient information tied to reported Adverse events of that year. These seperate data sets are linked together by case id's, and have years available for each year as far back as 2004. The stock information was provided by NASDAQ historical quotes. Profit information had to be collected manually from annual reports of each company.

12.1.1 FDA Adverse Event Reporting - Drug Information. This dataset has 22 Attributes which specifically report the conditions in which the event was reported.

- Primary ID: is used as unique identifier to link specific data sets together.
- Case ID: used as a specific identifier to link data sets together
- Drug Sequence NO. : used as nominal value to identify a specific drug.
- Drug Role: Nominal value which shows how involved the drug was in the event. this ranges from "Primary Suspect" "Secondary Suspect" "Concomitant" and "Interacting"
- Drug Name: Nominal value for a drug name, either the chemical makeup if not a valid trade name, otherwise contains the actual product name"
  Prod Ai : Specific chemical within in a drug. Sometimes an actual drugname.
- Validated: Binary value whether 1 indicates Validated name, or 2 indicates verbatim name
- Route: is the route of administration
- Dose: Verbatim text for dose, frequency, and route as entered on report

- Dechallenge: Dechallenge code, indicating if reaction abated when drug therapy was stopped. "Y" is positive decallenge, "N" is negative dechallenge, "U" is unknown, and "D" does not apply.
- Rechallenge: Rechallenge code, indicating if reaction recurred when drug therapy was restarted. "Y" is positive rechallenge "N" is negative rechallenge, "N" is negative Rechallenge "U" is unknown, and "D" does not apply.
- Lot Number: Lot number of the drug
- Expiration Date: Expatriation date of the drug
- Expatriation date Unparsed: date values that are not in format.
- NDA No. : National Drug Administration number of the drug
- Dose Amount: Dose Amount in Different Data type
- Dose Unit: Unit of drug dose
- Dose Form: Form of dose reported
- Dose Frequency: Code for Frequency
- Serialid : serial number for the drug

The main attributes I based my analysis on were the Drug Names as an identifier, Drug type, and Drug Role to help with the cleaning process and consider certain cases where the outcome was not severe unless it was a Primary suspect.

12.1.2 Patient Outcome. This data set contains the general patient outcomes related to specific cases of Adverse events reported for that year. This has 5 attributes.

- Primary ID : Unique Identifier that connects corresponding data sets
- Case ID: case ID
- Outcome: Initials of patient outcome type
- Outcome Code Description.

The attributes that I used in this data set were outcome code, to Consolodate each type of drugs with there subsequent severity scores, as well as create clustering images with their outcomes.

12.1.3 Historical Quotes.

- Time - the time in which the quote was recorded
- Open - The price of stock at Open
- High - The highest sold stock
- Low - The lowest Sold Stock
- Close- Price at closing time
- Volume- the amount of transactions that happened at this time

The main attributes from this data set I used were High and Low to create averages by date.

12.2 Main Techniques Applied

The FDA data was a somewhat dirty data set. Several names had either spelling mistakes, weird characters and delimiters, as well as empty cells. The data however did not have many duplicates since each case ID was recorded uniquely and could be tracked among multiple data sets. There were several extraneous attributes that needed to be trimmed to help decrease the size of the data I was working with.

12.2.1 Data Cleaning. To clean the data, I started by clearing specific Attributes such as Dose size, Dose Unit, Serial Id.

From that point, I removed any rows that had missing information. The Data I was working with could not be filled in manually for example since to create the severity scores, It would require that I convert the Outcome Codes to numerical values, and since these were very specific instances, it would hurt the data mining outcome to infer some specific outcome with a case.

I also removed any row that did not have a "1" validated trade name. This was because I could not properly find a way to track down the trade names of every specific listed medication in the list, especially ones with low frequencies.

There were several names within the Drug name attribute that had subtle variations in names listed within the data set. Using Dataiku, I was able to sort each specific drug by matching characters and create uniform names for each drug for continuity.

The Historical Quotes Data was actually fairly clean with little no missing values since everything was very cut and dry numerical inputs.

12.2.2 Data Preprocessing. Once the data sets were cleaned I began creating severity scores for each specific drug. To do this I merged each data set with a specific Primary Id in the corresponding data sets. Once this was done, I created a scale for each specific outcome code.

- CA = 1- "Contanial Anomoly" - Pre-existing conditions that cause an adverse event
- OT = 2- "Other Serious" - Nondiscript severe event with corresponding drug.
- DS = 3 - Disability as a result from adverse event with corresponding drug
- HO = 4 - Hospitalization as a result from adverse event with corresponding drug.
- LT =5 - Life Threatening conditions as a result from adverse event with corresponding drug.
- DE = 6 - Death as a result from adverse event with corresponding drug.

These scores were then used to replace each code within a duplicate column. I then took the Aggregation Average of each column by their severity scores which outputted a corresponding Severity Score. During this process I also counted each case by Drug name within the data set which also outputted it a sum total of cases that were reported that year.

From here I further cleaned the data by using clustering to remove any thing outliers with low case counts so as to reduce the noise of the data. This helped in my visualization process because it was easier to visualize the high severity drugs in consideration with determining which ones I would track their net profit over the 3 year time span.

12.2.3 Data Evaluation. To analyze the data I had to take the severity score increases as well as the net profit growth of the specific set of drugs reported each year and calculate its correlation. This would determine whether or not there was a positive or negative correlation between severe drugs and their net profits for a fiscal year. this was also done with their case counts for each year, determining their frequencies per year and correlating that information with their specific profits that year.
this helped me produce certain visualizations to better interpret which drugs were the most severe in a given year, Which drugs were most frequently reported. This information helped to understand which drug types were the most severe, or had the most cases reported for analysis.

12.2.4 Data Warehouse. The main data storage I used was a OTLP Data warehouse server provided by Microsoft Azure in conjunction with Dataiku. This allowed me to store each of the individual data sets per year and access specific samples of each data set in which I could clean and pre-process which would be applied to the data set. I did not need to input my data into a SQL database since Dataiku automatically stores each data set as such. This also allows me to use SQL commands to merge specific data sets, and even apply aggregation methods to the data during a join to specifically tailor the data to be easily analyzed. This also came with an integrated python notebook so that I could create specific commands within the software to finely sort specific data such as Drug name and Drug Type, and create averages and breakdowns of their specific frequencies.

12.3 Clustering

I decided to use K-means clustering to split each data by their frequencies and their severity score. This allows me to look at the trends in medications such as relationships between severity and frequency as well as cluster severe drugs and drug types by their frequencies and severity scores. I used 5 specific nodes to start with and let Dataiku use an automatic training software to determine the inertia and suspension of each training session. Initially my data set was very large, with wide variance, so the intertia and suspension was incredibly high for my set.
To get around this I did more pre-processing to get a smaller sample with more unique data cells so as to get better cluster.

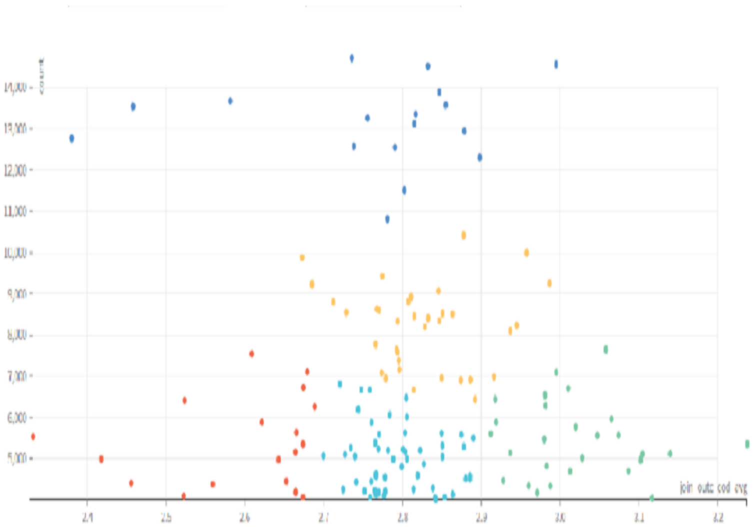

Figure 1: Clustering of Frequency and Severity

This graph shows different grouping of sepcific drugs by their frequencies and their Severities. The different color labels put each drug in to specific node categories that translated over to the final groups in which the profit analysis was done.

- Red denotes low frequency and low Severity.
- Blue denotes low Severity High Frequency.
- yellow denotes moderate severity moderate frequency
- light blue denotes moderate severity low frequency
- green denotes high severity low frequency

13 KEY RESULTS

The Results found were a breakdown of the top severe drugs tracked from one given year over a 3 year span. This was compared to the annual reports of net profits of specific companies to track the growth over that period time, and then correlated.

13.1 Frequent Drugs

The Drugs with the highest frequency shows were one with low severity. As Figure 1. showed the drugs with the lowest severities had the highest frequently reported cases. These ranged from over the counter products such as ibuprofen to Folic Acid which is a vitamin B complex often used during pregnancies.

These showed to have over 10,000 reported cases a year, however had little to no deaths within their severity scores, which ranged from 1.0 to 2.0 in severity. This countered my initial hypothesis, since even though the correlation between amount of products would relate to the amount of cases reported in year, however since they don't have a high mortality rate, it wouldn't be my assumption that it would have so many frequently reported cases.

Medications with high severities had very low frequencies, though the increased in the number of cases over time, they were marginal to those with lower frequencies, showing under 10,000 cases a year.

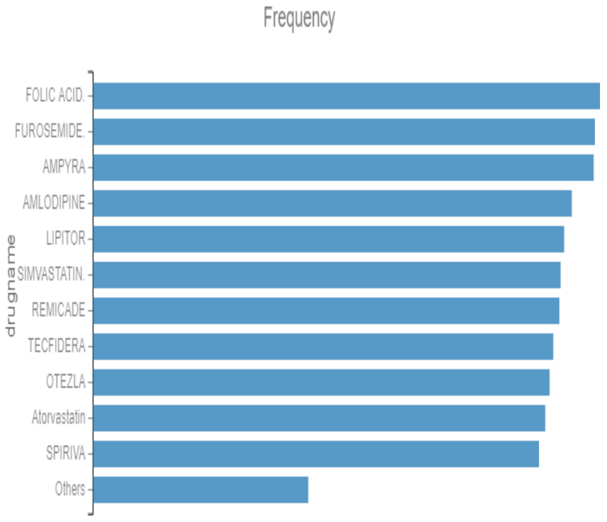

Figure 2: Frequently reported drug cases for 2016

13.2 Severe Drugs

By taking the aggregation average of each drug name when joined with the paitent outcome data set, I was able to create an average severity score for each type of medication reported between 2014, 2015, 2016. This was then visually represented to show which specific drugs were the most severe.

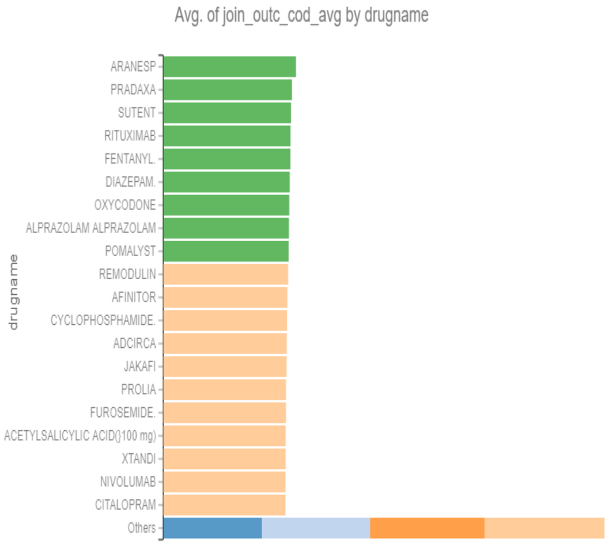

Figure 3: Severity scores of specific drugs

this contains the specific breakdowns of the top most severe drugs. The green values shows the high severity cluster, while the tan shows the moderate severity cluster.

As you can see the most severe drug was Aranesp, which is a bone marrow deficiancy medication used to promote the creating of red blood cells in the body. This ended up having a score of 3.5

The next drug was Pradaxa, blood clot medication which is prescribed to people with heart conditions. This ended up having a severity score of 3.2. the next was Sutent, which is a cancer medication used to help with Kidney cancers, as well as forms of gastrointestinal cancer. This had a severity score of 3.19 Rituximab is a used to treat auto-immune disease like Hodgkin lymphoma disease. This had a severity score of 3.1 Fentanyl is used in the form of patch used as a painkiller in low doeses this had a severity score of 3.0.

The list continues with other such popular medication like oxycodone and Alzapram, which are both painkillers and anti-anxiety medication. This was intersting to note that among the cases reported, the average weighed out to be more in the rang of disabilities and hospitalizations rather than life threatening conditions and deaths. The rates however increased consecutive almost double from the previous years. This shows a somewhat average level of risk when treating serious conditions such as heart conditions and cancer.

It is concerning however that the rate of growth increases from year to to year at almost 2x the previous year. Either this could be attributed to popularity, since subsequent profits of those specific drugs raise that same year by different margins.

13.3 Net Profits.

For this analysis I took the top drugs within those green categories and looked at their profits over the 3 year span. This information was however not available as a data set which caused me to analyze a small selection of medications listed to map trend.

- Aranesp - Amgen pharmaceuiticals experienced a 25 percent increase in sales between 2014 and 2016.
- Pradaxa - Boehringer Inglheim experienced a 8 percent increase of the sale of pradaxa between 2014 and 2016
- Sutent - Pfizer saw a 10 percent increase sales for stuent between 2014 and 2016

These were just some of the all around signs of sales growth per company based off of 2016 annual reports. These reports contain the percent change between 2014 and 2016. Some other listed drugs such as Fentanyl had to be tracked down for public usage since it is a high profile opiod that is somewhat in an epidemic. Since the drug is not widely used, however manufactured by Johnson and Johnson, there was not information to provide changes in sales.

Something that was tracked during the research process is that most of these severe drugs are high profile sales items per company. In each annual report they have specific sales for each product listed, where some medications such Fentanyl are kept out of their sales reports. Further investigation is required but I thought I thought it would be interesting to note.

13.4 Investments

After creating the severity scores and frequencies, I was able to calculate the correlation between frequency and stock rates, as well as severity score and stock rates. However this did not prove to be to frugal since the rates varied so much per year, that even when trying to use normalize the data, it did not seem to show any kind of correlation. This could be due to my methods.

However this seem to makes sense in terms of stock analysis since analyists are still trying to understand trends in the stock market, and though it would be interesting to find some correlation between the 2, the changes in prices are too rapid to properly compare with the frequency.

The same can also be said for the severity score, since the change in severities stayed relatively the same it did not seem to correspond to any correlation between stock market trends.

With more information being analyzed by other data mining researches, maybe one day this information will be useful to the public, but as it stands there is to go off of.

13.5 Conclusion

While many interesting things revealed themselves during this analysis, I found that most of what was found corresponded with my initial hypothesis. More specifically, that the frequency of reported cases and the severity of specific drugs increase as well as the net profit of those drugs increase over time. Since correlation does not prove causality, there is no clear cut evidence as to the motives behind pharmaceutical companies and their profits. However, it is intersting to see the dichotomy between human life and profit especially in today's climate.

I was disappointed to find that there was no such trend between stock trends and severity or case numbers, yet as stated earlier, as information about stock trends is revealed perhaps it will be worth revisiting the data.

14 APPLICATIONS

This information is readily available to the public. The FDA is a well respected source of information about ingestible products in the United States, however this source of information takes time to parse through. The FDA has made strides in creating a portal into which the public can access information about Adverse Events related to specific drugs, however it is not advertised as such. Here is some applications I beleive to be useful to the public if it were promoted to do research before being prescribed a drug.

14.1 Big Pharma

This information is directly related to pharmaceutical companies and their products. Since this is public information, there is nothing stopping anybody from finding this data, which could be decision between whether or not a patient wishes to take the medication. This could potentially affect sales, especially if a drug is related to high severity score as calculated. It could be in a companies best interest to use this information and promote a constant decrease in severity over time into their annual sales reports. While the have section about deaths, this information, if formatted right, could promote a competitive nature towards companies to reduce their severity scores overtime.

They can even see how their product interacts with people with pre-existing conditions overtime, as well as dosage information and drug therapy types. This is valuable knowledge for the development of pharmaceutical companies and their products. It can promote relationships with the patients prescribed to their medication and create a destigmatization of their companies.

14.2 Public Knowledge

With the revelation that certain over the counter drugs have a large frequency of reported cases, it useful knowledge to the public to understand what could potentially happen when taking certain drugs. For instance, in terms of pre-natal care, since Folic acid has a large amount of cases reported each year, it may be valuable information to OBGYN clinics to better understand certain information about recommended supplements used during pregnancy. While Folic Acid does pose a serious health risk, it is worth noting that some complications can occur using this, and therefore would inspire further research into that particular area. For a more umbrella case, the number of reported cases for certain painkillers, specifically Acetaminophen have a high number of reported cases per year. Often the combination of Alcohol and Tylenol can cause liver damage, which happens often when people treat extreme hangovers after inebriation. This is may be common knowledge, but for certain demographics they could not know. It is important to have this research easily available to public to better understand these medications in general.

14.3 Mental Health Field

This information could be especially helpful to people working in the mental health field as way to research products before prescribing them to patients.

Of mental health medications Abilify had the highest severity score and was the most frequently reported medication to have Adverse Events over the past 3 years. This is a medication used to treat schizophrenia and anxiety in extreme cases. The information in the FDA portal is easily accessible and shows how much increase and decrease in deaths and cases overtime are related to that specific drug. This is useful information for psychiatrists who may want to consider a better option to treat schizophrenia and opens up better communication with their patients about the risks involved with taking a medication.

REFERENCES