Adverse Drug events and Pharmaceutical Company profits*

Adam Tammariello
University of Colorado Boulder
Boulder, Colorado

Taicheng Song
University of Colorado Boulder
Boulder, Colorado

ACM Reference Format:
Adam Tammariello and Taicheng Song. 2018. Adverse Drug events and Pharmaceutical Company profits. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 5 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

*Produces the permission block, and copyright information

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

we need so as to create a system of evaluation. We have found 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.

REFERENCES