milestone report

Chetna Jhurani

February 27, 2018

## Intro

Hypertension is a rising problem in the United States. As such, there are many different types of hypertension medications that have many different methods of action ranging from simply decreasing blood volume (diuretics) to dilating blood vessels. The problem with having so many different medications it can be difficult to find the cheapest medication. My main goal here is to find the cheapest hypertension medication.

## Deeper Dive into the Data Set

# Important fields and information

The data set that I ended up using was the December 2012 Federal Upper Limits provided on the Medicaid website seen here <https://data.medicaid.gov/Drug-Pricing-and-Payment/Federal-Upper-Limits-2017-12/3usn-qyfh/data>. The most important information in the data set is the manufacturing cost and the cost covered by Medicaid. These columns provided the quantitative data that I would need for my data analysis. However, this dataset gives no indication of whether or not the specific medication is a hypertension medication or not. This information came from the Mayo Clinic website, which can be seen here <https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/diagnosis-treatment/drc-20373417>. Not only did it provide me with the most common hypertension medications, but it also gave me its method of action as well.

# Limitations

The dataset is limited by the medications that are covered by Medicaid. Fortunately, most of the most commonly prescribed hypertension medications were covered by Medicaid, so that was not a significant obstacle. However, the only information regarding price I am really using from the original dataset is the cost of manufacturing and the amount covered. This does not actually provide the lowest sale price, but it can be used to predict it by subtracting the two amounts provided in the dataset. Additionally, prices may vary in diffrent locations. What is provided is more of a baseline analysis.

Additionally, as Medicaid really only covers the most common medications, it is difficult to give the output much significance. What I mean by that is, while I am able to group the medication by type, each group contains at most items in it. How much significance can a statistic have when it is only analyzing sets of 1-3?

Additionally, some of the drugs required were listed as combinations, so it was difficult to find the drug by type. I am not necessarily able to answer the question of what is the cheapest medication, but I am able to find qualitative factors that may predict the price of the medication.

I was also limited by the fact that a lot of the data provided is qualitative, so my analysis was restricted in that aspect. It would have been much easier to find some kind of correlation with a more qualitative data set. I would be able to really find more trends with more quantitative data.

# Cleaning and Wrangling

As far as initial cleaning goes, there was a lot of extraneous information provided in this table such as the item code number(NDC). Often times, the same medication was listed multiple times, so removed this column from the data. Since all of the data was from December 2017, I also removed the Year and Month columns. Additional columns that were irrelevant were renamed and removed as well. Once all this was done, the distinct function was used to remove all the duplicate rows. What was left in the end was ingredient name, strength, dosage, coverage, and manufacturing price.

When it comes to data wrangling, the two most useful pieces of information here were the manufacturing prices and the amount covered by Medicaid. What I did here initially was add another column by subtracting the cost covered from the manufacturing cost. I had initially only kept this column and added more information from this point on. However, when I referred back to the initial data set, I realized the for each medication, varied with the tablet size. Thus, I added another column to find the price of the drug per milligram, making the data more meaningful. However, at this point, the same drug was still listed multiple times with different costs attached. Hence, I grouped the data by the Ingredient name and kept only the row that contained the lowest cost per milligram, as the whole purpose is to find the cheapest medication.

Now that there was only one row for each drug, I added a column for the method of action of each drug (Type) and filled it with the appropriate information. I, then, created another column for the pure substances and filled that with the appropriate informtion as well. However, the dataset still contained every single drug covered by Medicaid, when I only needed the hypertension medications. As a result, I types of hypertension medications in one variable and filtered out the rows that did not contain them.

The final data set is below.

library(readr)  
coverage\_clean <- read\_csv("coverage\_clean.csv")

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:  
## cols(  
## X1 = col\_integer(),  
## Ingredient = col\_character(),  
## Strength = col\_integer(),  
## Dosage = col\_character(),  
## AMPs = col\_double(),  
## Full = col\_double(),  
## Paid = col\_double(),  
## PPMG = col\_double(),  
## Type = col\_character(),  
## Pure = col\_character()  
## )

View(coverage\_clean)  
library(knitr)  
kable(coverage\_clean, caption = "Coverage Data")

Coverage Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X1 | Ingredient | Strength | Dosage | AMPs | Full | Paid | PPMG | Type | Pure |
| 1 | ACEBUTOLOL HYDROCHLORIDE | 200 | CAPSULE | 0.134564 | 0.235487 | 0.100923 | 0.0005046 | beta blocker | N |
| 2 | ATENOLOL | 100 | TABLET | 0.079799 | 0.139648 | 0.059849 | 0.0005985 | beta blocker | Y |
| 3 | CAPTOPRIL | 100 | TABLET | 0.580250 | 1.777050 | 1.196800 | 0.0119680 | ACE inhibitor | Y |
| 4 | HYDROCHLOROTHIAZIDE | 25 | TABLET | 0.003227 | 0.009900 | 0.006673 | 0.0002669 | diuretic | Y |
| 5 | LISINOPRIL | 40 | TABLET | 0.031287 | 0.054752 | 0.023465 | 0.0005866 | ACE inhibitor | Y |
| 6 | DILTIAZEM HYDROCHLORIDE | 360 | CAPSULE, EXTENDED RELEASE | 0.333229 | 0.616180 | 0.282951 | 0.0007860 | CCB | N |
| 7 | BENAZEPRIL HYDROCHLORIDE | 40 | TABLET | 0.029580 | 0.059590 | 0.030010 | 0.0007503 | ACE inhibitor | N |
| 8 | AMLODIPINE BESYLATE | 10 | TABLET | 0.019519 | 0.034158 | 0.014639 | 0.0014639 | CCB | N |
| 9 | LOSARTAN POTASSIUM | 100 | TABLET | 0.038030 | 0.066553 | 0.028523 | 0.0002852 | ARB | N |
| 10 | CANDESARTAN CILEXETIL | 32 | TABLET | 1.190418 | 2.083232 | 0.892814 | 0.0279004 | ARB | N |

## Preliminary Exploration

I explored the correlation between the method of action of the medication and the price per milligram of each medication (PPMG). I did this by finding the mean PPMG for each medication type and graphing it in a scatterplot. Visually, ARBs were by far the most expensive type of hypertension medication. However, as stated in the limitations, it is unclear how much statistical significance this has. There were only two datapoints that contributed to the mean, so the ARB average price may have been skewed by an extremely high price. Meaning, it may not be a good indicator of other ARB prices.

I also explored the correlation between the purity of the substance and the price per milligram. I also found the mean of pure substances and compared it with the unpure substances. This was also graphed in a scatterplot. As I had predicted, the cost was much more for the one that was not the pure hypertension drug. This has more stastical significance as there were more data points contributing to this statistic.

## Approach

My initial approach to this problem was to simply find the cheapest hypertension medication, which I was able to do with the given data set. At least, I was able to find the net cost based on the manufacturing price and the Medicaid coverage.

When I referred back to the Mayo Clinic website, I found that thiazide diuretics were the most prescribed hypertension medication, and it was coincidentally, the cheapest medication after Medicaid coverage. As such, I think that it may be of interest to take this further and compare the cost with the frequency that it is prescribed, the symptoms, and percentage of the cost that is actually covered. With the frequency, I would able to do more analysis of the data, and would also be able to include graphs that are more than simple scatterplots.