Comparison of Hypertension Medication Prices

Chetna Jhurani

March 22, 2018

# Introduction

Hypertension is a rising problem in the United States. According to the CDC, almost one in every three Americans is diagnosed with hypertension, and this number shows no signs of decreasing. Hypertension can often lead to heart disease or stroke, two of the leading casues of death in the United States. Seen here: <https://www.cdc.gov/bloodpressure/faqs.htm> 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 most common types of hypertension medications are diruetics, angiotensin-converting enzyme (ACE) inhibitors, angiotensin II receptor blockers (ARBs), and calcium channel blockers (CCBs). However, the most frequently prescribed medications may not necessarily be as affordable as those that are more frequently prescribed. The main question I want to focus on is: How can we increase access/affordability of hypertension medication. Programs such as Medicaid increase access to these medications, so I will be focusing on Medicaid coverage of certain hypertension medications.

# Deeper Dive into the Dataset

## Important fields and information

The original data set that I ended up using was the December 2017 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. After wrangling this data, I still needed to know which ones were hypertension medications as well as the method of action of the specific drug. I wrangled the dataset further with the information provided by 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. Finally, I still wanted a quantitative comparison on the most common medications, so I found information on how often each type of hypertension medication was prescribed. I pulled the statistics listed by the NCBI study seen here <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4808570/> and added it to the wrangled data set.

## Limitations of the dataset

The Federal Upper Limits dataset is limited by the medications that are covered by Medicaid. Although most of the most commonly prescribed medications are covered by Medicaid, the dataset is still restrictive. For one, many of the drugs provided are cut with another drug. Medicaid simply does not cover the pure substances. If I were to only include pure substances, I would have only had five data points. As a result, I ended up attempting to include the purest version of the most commonly prescribed medications, even if they were cut with something like potassium or hydrochloride. However, even after this, as Medicaid really only covers the most common medications, the sample size is still pretty small. There is only one diuretic included, and neither of the ARBs included are pure. In an ideal dataset, there would be more medications to work with.

## Cleaning and Wrangling

As previously stated, I started out with the Federal Upper Limits (FUL) table, wrangled it and then added information from the Mayo Clinic and the NCBI study. There was a lot of extraneous information provided in the FUL table such as the item code number(NDC). Often times, the same medication was listed multiple times with a different, 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 manufacturers’ prices and the amount covered by Medicaid. What I did here initially was mutate the table by subtracting the cost covered by Medicaid (Full) from the manufacturers’ cost (AMPs), creating the column “Paid”. 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 mutated the table further to find the price of the drug per milligram (PPMG), 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 saved types of hypertension medications in one variable (“types”) and filtered out the rows that did not contain the specified types of hypertension medications.

The code can be seen here: <https://github.com/cjhurani2000/Capstone-Project/blob/master/data%20wrangling.R>

The final table is below.

## 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()  
## )

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 |

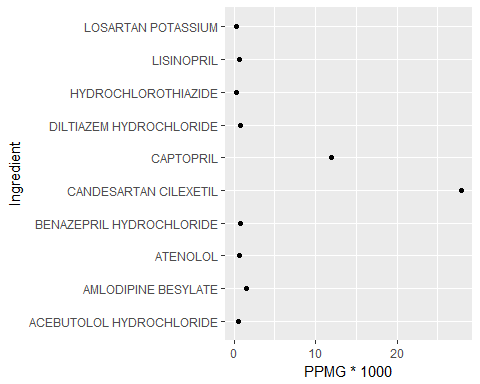
# Preliminary Analysis

I thought the best method of action following the wrangling of the data was to look at the data graphically with each drug listed individually.

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

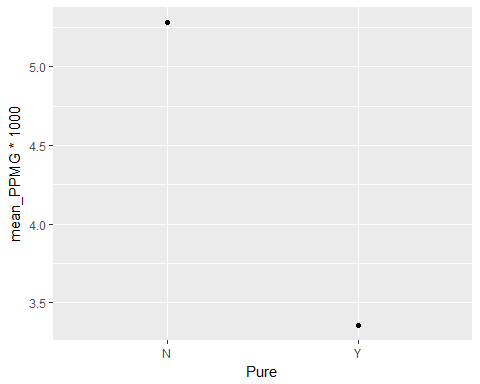
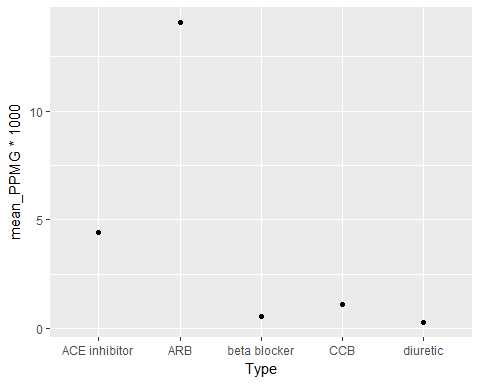
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union



Visually, captopril and candesartan cilexetil were the most expensive. In order to determine what factors contributed to this, I took the analysis a step further. I decided to focus on comparing the method of action of the drug (Type) and the purity of the substance and compared it with the price (PPMG). Before graphing these comparisons, I needed a summary statistic. Based on the limited data, I chose to examine the mean.

cov\_type <- coverage\_clean %>%   
 group\_by(Type) %>%   
 summarize(mean\_PPMG = mean(PPMG)) %>%   
 arrange(mean\_PPMG)  
  
cov\_pure <- coverage\_clean %>%  
 group\_by(Pure) %>%  
 summarize(mean\_PPMG = mean(PPMG)) %>%  
 arrange (mean\_PPMG)

The graphs using the above statistics can be seen below.

## What This Means

Visually, there is an obvious difference in the price paid for the pure substances versus the ones that are not. This makes sense, as combining two drugs will obviously cost more than one simple drug.This information by itself may not be so useful. However, when examining the data further, it may explain any apparent skew in the data. Looking at the specific types of hypertension medications, ARBs are noticably the most expensive. At this point, only scatterplots have been included as the data is not only limited, but it is also primarily qualitative. While the cost of ARBs may have been skewed by the impure Candesartan Cilexetil, I was still interested in knowing what contributed to the cost of each type of hypertension medication.

# Further Analysis

## Introducing a New Variable/Further Wrangling

So far, I have only examined qualitative data apart from the PPMG. In order to not only introduce a quantitative variable but to also further examine the types of medications, I introduced a new variable, frequency. This variable comes from the aforementioned NCBI study that contains information on how often each type of medication is prescribed. The introduction of this new variable allows the problem to be framed as a supervised regression machine learning problem: How does the frequency a medication is prescribed correlate with its price?

To do this, I wrangled the coverage\_clean table from above, even further. I grouped medications by type, then used the mean summary statistic, and saved it in a new table. I, then added a column for the frequency and filled in the values corresponding with those listed in the NCBI article.

The code can be found here: <https://github.com/cjhurani2000/Capstone-Project/blob/master/frequency.R>

The resulting table can be seen below.

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

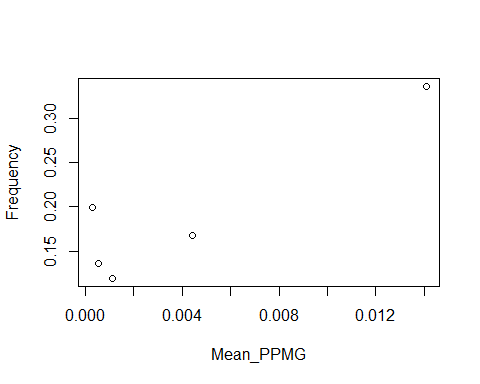
## Parsed with column specification:  
## cols(  
## X1 = col\_integer(),  
## Type = col\_character(),  
## Mean\_PPMG = col\_double(),  
## Frequency = col\_double()  
## )

Prescription Frequency

|  |  |  |  |
| --- | --- | --- | --- |
| X1 | Type | Mean\_PPMG | Frequency |
| 1 | ACE inhibitor | 0.0044350 | 0.1679 |
| 2 | ARB | 0.0140928 | 0.3357 |
| 3 | beta blocker | 0.0005516 | 0.1363 |
| 4 | CCB | 0.0011249 | 0.1191 |
| 5 | diuretic | 0.0002669 | 0.1990 |

## Machine Learning

First, let’s look at this data visually in a scatterplot.



Based on this scatterplot, the best way to proceed from here is to use a linear regression model using the code below.

#linear regression model  
freq.mod <- lm(Frequency ~ Mean\_PPMG, data=freq\_cov)  
  
#summarize results  
summary(freq.mod)

##   
## Call:  
## lm(formula = Frequency ~ Mean\_PPMG, data = freq\_cov)  
##   
## Residuals:  
## 1 2 3 4 5   
## -0.028245 0.010722 -0.008042 -0.032890 0.058455   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.13698 0.02429 5.640 0.0110 \*  
## Mean\_PPMG 13.33972 3.66222 3.643 0.0357 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04272 on 3 degrees of freedom  
## Multiple R-squared: 0.8156, Adjusted R-squared: 0.7541   
## F-statistic: 13.27 on 1 and 3 DF, p-value: 0.03568

#significance  
anova(freq.mod)

## Analysis of Variance Table  
##   
## Response: Frequency  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Mean\_PPMG 1 0.0242194 0.0242194 13.268 0.03568 \*  
## Residuals 3 0.0054762 0.0018254   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The results of the linear regression formula gives us a line around approximately y = 13.34x + 0.14 where y is the mean price paid and x is the frequency of prescription. The resulting R-squared value shows a moderately strong positive correlation between the mean price paid and the frequency it is prescribed. That means that until now, Medicaid has done a moderately good job of making sure to budget their coverage of medications based on need.

# Conclusion

## How to Use Findings

1. Patients: The customers who use hypertension medications would be able to use this information in two ways. They can use the qualitative data in order to find the cheapest type of hypertension medication. They can also use both the qualitative and quantitative data in order to determine if the current brand of medication they are using is overcharing them. In other words, they can compare the price of one brand of medication and determine if this is above or below the cost of the medication of that specific type. For example, someone on Candesartan Cilexetil may want to switch to Losartan Potassium.
2. Physicians: The main statistics here really examine the physicans. Physicans can use the data from the first half of the report to prescribe cheaper types of medications. Looking at the frequency vs cost graph, there is an obvious dip in the graph around diuretics. Had physicans noticed this, maybe they would prescribe diuretics more often for patiens who are struggling financially. They may also use this to examine why this specific type of medication is not prescribed as often and recommend that the federal budget be alloted to covering those medications that are more commonly prescribed. For example, diuretics can often lead to potassium deficiencies, so this may explain why they are not covered as often.
3. Federal Government/Medicaid: According to the NCBI article, there has been a rise in the prescription of ACE inhibitors. This change, however, has not been reflected in the Medicaid coverage of the drugs. The budget committee can use the formula provided above and can use this to determine coverage of each hypertension medication so that the coverage reflects each medications need. The Federal Upper Limits are reported every month, so it may be beneficial to examine the frequency of prescription as well.

## Broader Applications

The formula listed above can be applied to the larger, original dataset. Additional data can regarding how often the other Medicaid-covered medications are prescribed can be used to adjust the current price so that it reflects its need. Meaning, there would be increased coverage for those medications that are prescribed more often.

## Further Research

1. Future studies would involve diving deeper into certain populations. My data focused on the general United States population. However, further studies could examine those in poverty more specifically, as Medicaid coverage would be more closely tied in with that population. For example, diuretics are shown to be more effective for black and older people than ACE inhibitors and beta blockers, so this may explain why the cost does not correlate with the frequency of prescription.
2. The same general concept can also be used for Medicare coverage rather than Medicaid coverage. This, of course, would require a greater focus on older populations. For example, diuretics and ACE inhibitors are known to be much more effective in the elderly than beta blockers are. Thus, Medicare could use the formula to decide upon coverage of hypertension medications for the elderly.
3. As previously stated, the data does not show the final price the patient would have to pay for his/her medication. It only shows the baseline. A further study would go a step further and examine specific brands of hypertension medication and compare their prices with the baseline price provided in this report. The study would show which brands tend to charge more for their product. Additionally, the data from that study could be used to develop an app that customers would use in order to find the cheapest medication in their immediate vicinity.