Machine Learning

1. How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?

The main question focuses on how to find the cheapest hypertension medication. To frame this as a machine learning problem, I added the frequency each drug was prescribed as listed in this study (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4808570/>). In order to do this, I found the mean price paid for each class of hypertension medication and compared it with the frequencies listed in the article. By finding this new quantitative variable, I was able to make this a supervised problem that is a regression model. The modified table is below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Type** | **Mean\_PPMG** | **Frequency** |
| **1** | ACE inhibitor | 0.0044349583 | 0.1679 |
| **2** | ARB | 0.0140928338 | 0.3357 |
| **3** | beta blocker | 0.0005515525 | 0.1363 |
| **4** | CCB | 0.0011249375 | 0.1191 |
| **5** | diuretic | 0.0002669200 | 0.1990 |

1. What are the main features (also called independent variables or predictors) that you'll use?

The independent variable here is the mean price paid for each class of diuretic. The dependent variable will be the frequency that the drug is frequency that the drug is prescribed.

1. Which machine learning technique will you use?

As this data does fit into a regression pattern, I plan to find a formula to predict the cost of not only hypertension medications but any medication covered by Medicaid. Thus, I will be using a linear regression model comparing the mean price paid per milligram (Mean\_PPMG) and plan to compare it with the frequency that it is prescribed. This information can be used to find the price of any hypertension drug covered by the ACA, or it can be used to find the price of a newly created class of hypertension drug given the frequency it is prescribed.

1. How will you evaluate the success of your machine learning technique? What metric will you use?

I will use the ANOVA method to test the significance of the values found by the linear regression. I can use the p-value to examine the significance and create a confidence interval for the data. I can also test the success of the data by entering in a random hypertension drug to see if it cooperates with the model. For example, I can look at the frequency amlodipine (an ARB) is prescribed, and see if its price falls within the predicted confidence interval provided by the data.