World Health Organization Mortality Investigation

Bryan Arment
CSCI
University of Colorado Boulder
Boulder CO USA
brar9262@colorado.edu

Jeff Dank
CSCI
University of Colorado Boulder
Boulder CO USA
jeda7205@colorado.edu

Tony Pearo
CSCI
University of Colorado Boulder
Boulder CO USA
anthony.pearo@colorado.edu

Problem Statement/motivation

A database published by the World Health Organization (WHO) that tracks all-cause mortality with extensive classifications for cause, demographics and regions has been chosen for this analysis. The goal is to discover patterns as to how mortality presents itself in different countries across the world. Are certain diseases more prevalent in certain populations? Can having one disease predict another disease? These types of questions have real world impact. The ability to sanity check the group's findings should also be quite simple. For example, one would expect heart disease to be occurring more often in older people and not children. In addition, geographic data is available so it will be possible to see if wide cause mortality can be mapped in this manner to see if interesting, albeit, morbid patterns emerge from the data set. This dataset is very applicable to health policy as well as the distribution of resources such as foreign aid. One would expect similar patterns/analysis derived from this project could be of interest to public health professionals and governments.

Data set

The database has been compiled by the WHO and contains number of deaths by country, year, sex, age group and cause of death as far back from 1950. Data are included only for countries reporting data properly coded according to the International Classification of Diseases (ICD). Our specific data consists of two large sets that have data for the 10th revision of the ICD. The database itself can be downloaded from here: Download the raw data files of the WHO Mortality Database

The 10th revision of the ICD lists over 100 separate causes of death, breaks the age group down into one year buckets for ages five and under and then 5 year buckets for ages above five, and the data is just taken as reported to the WHO with a warning that it may not be complete for all countries.

Literature survey

An immense amount of research has been done revolving around mortality statistics and other associated health studies. While it is likely that new ground will not be broken with this project, it is an interesting and expansive data set that will allow the team to apply methods and techniques covered in the scope of this course.

Mathers and Loncar did some extensive work to try and predict global mortality out to the year 2020 (Projections of Global Mortality and Burden of Disease from 2002 to 2030).

They were able to split the data out by sex, age, and income. Also, they used simple regression equations to make their predictions. This type of analysis lines up well with our initial formulations when first stumbling upon the WHO database. Also, they were able to group some diseases together for their analysis such as “respiratory” and “digestive”. Lumping/clustering together similar causes of death like this is a great idea since the WHO mortality databases causes of death can be very specific.

Another paper focused on solely heart disease related deaths, but failed to make any predictions with generating any kind of models (Mortality From Ischemic Heart Disease). This was more of a survey of various countries and a discussion of the overall trends. One aspect that could be incorporated into the analysis is that they broke down the countries that were investigated into groups based on the countries GDP. Noting that the data may not be as accurate in still developing countries.

One last paper goes into a clustering method for US mortality data: Clustering of 27,525,663 Death Records from the United States Based on Health Conditions Associated with Death: An Example of Big Health Data Exploration. This paper is of interest because of the way that they group similar types of cause of death and then look into the health conditions of the sample in question like age, location, and demographics to better understand the people that died. This shift from a disease oriented approach to a person-centric approach gives a different lens to approach the dataset with and also goes into structuring a method that we ourselves can apply.

Proposed Work

First, we will load the dataset into pandas, take a look at the entire dataset to get an idea of if anything needs to be cleaned or tossed out. We will need to drop NAN values to allow summary statistics and modeling to be performed on our data. There may be some integration / consolidation that needs to be done, for instance generating our own key-value pairs that correspond to the various types of mortality. For example, there are various keys due to where the moralities were reported from. Additionally, we may need to group up some of the conditions in order to increase the signal in our data. One example of this is TB being spread across different codes which when we have dozens to hundreds of codes the individual code likely won’t have much effect regardless of model type. By combining related disease categories we hope to both simplify our analysis and achieve stronger modeling results.

Once we have our data cleaned and processed, we will likely start our basic data exploration and see if any interesting patterns present themselves. We will summarize the data and see mortality presents itself in various regions and age groups. From here we will identify interesting patterns we would like to focus our modeling efforts on. At this stage we will be identifying more subtle issues with the data such as factors that are heavily correlated and others that are simply outliers. We also will likely need to transform our data as we predict a lot of the data points will be heavily skewed towards advanced age groups so we will need to apply some sort of transformation to allow statistical analysis to be applicable (for example a log transformation).

As we become more familiar with the dataset and decide on what specific issues we want to focus on we will start applying various models to the data. The first and simplest will be a multi-linear regression where we take the simplified factors and see if we can extrapolate mortality information based on the data we have. This type of model will tell us what types of mortality we can expect in certain countries based on population characteristics and identify correlation/collinearity. This is useful because it may tell us what conditions are linked together which may help inform policy/aid decisions (dollars spent targeting this disease also help these other 4 conditions etc.).

Depending on the results of our regression model we may want to explore more complicated

modeling exercises such as KNN or logistic regression and other classifiers to see if we can classify individual groups/countries as high risk for various disease types or put them in other categories we decide based on the data analysis.

As part of our analysis process we will evaluate the outcome of our models and work to improve them as we go. We may need to add or subtract features as necessary and revisit the issues around correlation and transforming the data. We may need to remove items from our dataset that we missed before or may need to try transforming in a different manner than originally thought.

After we have the results of our analysis we will need to produce some data visualizations so the data will be easily digestible. We may use tools like tableau for simple charts/graphs of summary statistics but we also have the power of various python graphing libraries to generate complicated visualizations such as choropleth charts that would be very valuable for purposes of presenting global/regional findings.

Evaluation Methods

We will begin with simple analysis of all factors mortality rates across a few countries before trying to ask more difficult questions related to underlying patterns. Try to make a prediction based on factors identified in the project description. For specific mortality factors we can study relevant research on the area as our dataset is widely available and access within the medical community.

On a more micro level when looking at our individual models we will use the traditional statistical methods to assess the performance/relevance we have gained from our analysis. Using r-squared and coefficients to evaluate an overall regression as well as the impact of individual factors is one example of this.

Tools

Tableau is a useful tool for displaying our analysis visually using graphs / maps / etc. We will likely not load the entire dataset into Tableau, rather just a trimmed down CSV formatted version so that we can create a visual story. An example of Tableau output can be seen in the website tracking Covid-19 cases in Colorado: https://covid19.colorado.gov/data/case-data.

Excel can be used for quick sanity checks if needed or to create some simple plots while we are working with our dataset. It's just faster to use before making things more pretty with Tableau

Python is a good candidate to perform the brunt of our analysis. Pandas is a data analysis library available to Python that can read in large datasets, slice and dice them to how we see fit, and then perform some statistical analysis for predicting mortality. Python can also be used for its various graphing libraries such as matplotlib to create custom visualizations from our results.

Milestones

1) Have the database read into pandas and get an initial plot of a single cause of death over time for individual countries and demographics within the US
2) Gather any other supplementary data that may be useful such as pollution or traffic
3) Split dataset into training and test dataset so that we can train a model and make predictions
4) Fill in knowledge gaps – for example modeling with time series data
5) Make predictions and draw conclusions
6) Present information visually
7) Write Paper
8) Create Presentation
9) Publish Github repository