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) Clean the dataset
3) Gather any other supplementary data that may be useful such as pollution or traffic
4) Split dataset into training and test dataset so that we can train a model and make predictions
5) Fill in knowledge gaps – for example modeling with time series data
6) Make predictions and draw conclusions
7) Present information visually
8) Write Paper
9) Create Presentation

10) Publish Github repository

Updated Proposal

After performing our exploratory data analysis we have better identified the problem-set we would like to analyze. We have decided we want to perform a cluster analysis on our data in order to see if we can generate categories of countries for further analysis. We want to end up with the countries clustered by their mortality profile as we think this sort of categorization would be useful for policymakers to assess what the profile is relative to other countries to group them together to see if action can be taken to improve outcomes. We expect to start with a simpler k-means model but likely will try other hierarchical models as well to see the various outcomes. The wrinkle we are dealing with is our dataset is still pretty high dimensional with at least 103 possible mortality outcomes so we are still working through how to deal with this and what clustering model will give the best data.

Milestones Completed

We have completed two key steps in our project. The first is described above in that we have refined the type of question we are asking about the data to the clustering analysis and what we can learn from it. The second milestone is all the cleaning parsing of the data from the raw data set into the relevant dataset for our problem space. While this is only two milestones this is a significant amount of the project work as identifying the specific problem and getting the relevant data is probably at least half the battle.

The way the data is organized we have data by country with a row for each condition further divided by male and female. We also have a variety of columns that describe the number of deaths by age group for each condition. Our initial data set was huge in that it was approximately 4 million rows by 40 columns split over two csv files so we needed to parse the data down into something workable. We started by reading both files into pandas and then appending them together to have one giant dataset. We then cut down on the number of features by removing the age data from the file which eliminated around 30 of the columns. To do so we decided to only use the all age mortality column and dropped the other age columns as well as some of the more advanced age mortality and infant mortality as they weren't relevant for the question we were asking. We also removed the sub-country geographic data as we are only interested in country level data. We also confirmed that each line item has an all cause mortality without any NaN value showing up. This helped simplify our analysis as there were various age coding schemes based on a format column so we have eliminated a lot of cleaning and parsing by limiting our analysis to the total mortality data.

We now have a combined dataset with the features we think are relevant to our clustering analysis. We will likely consolidate the feature list a bit further as we iterate our analysis and see how our results turn out but we have a good solid base dataset to work off of.

Milestones Outstanding

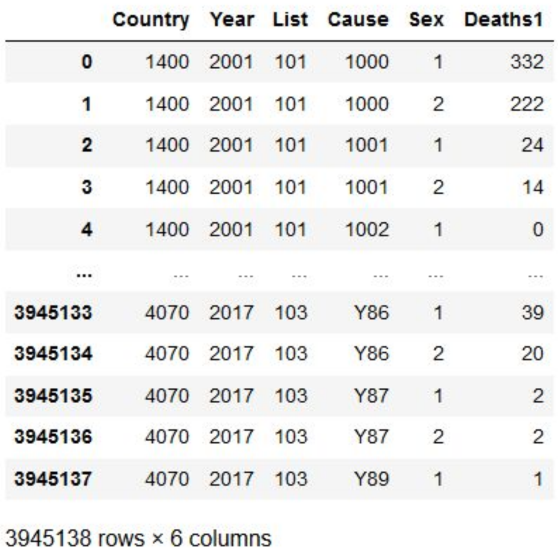
As mentioned we have a base dataset so our next steps are gathering supplemental data, performing analysis, drawing conclusions and presenting the results.

A key supplement we have found on the same site as the original data is a population add on that will allow us to normalize the values in our dataset for purposes of running the cluster models.

The next big step is deciding on a final model we want to run as there are various clustering methods available to us and we need to figure out if we will get good results with our current high dimensional base data or if we need to pare it down a bit more to run a simpler cluster model such as k-means. This decision will be made in the next couple days but once we have this working we should have presentable results and be able to report on our findings.

Results So Far

As mentioned above we have done a lot of cleaning and parsing of the dataset. This was a significant amount of work due to the volume of data we are working with (4 million rows by 40 columns to start). We have a couple different files we are working off of. First which we consider to be our “data warehouse” is the data that resulted from the discussion above in the milestone section. This is our icd10 data frame as shown below:

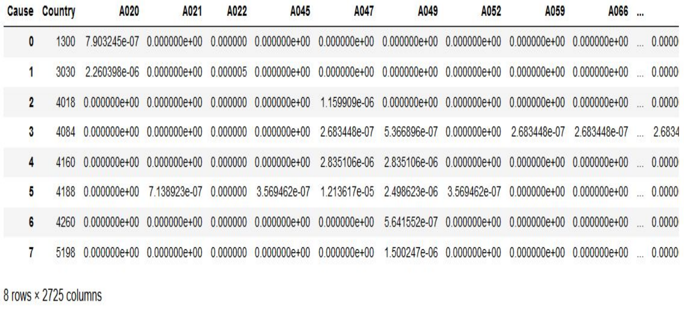


	Country	Year	List	Cause	Sex	Deaths1
0	1400	2001	101	1000	1	332
1	1400	2001	101	1000	2	222
2	1400	2001	101	1001	1	24
3	1400	2001	101	1001	2	14
4	1400	2001	101	1002	1	0
...	...	...	...	...	...	...
3945133	4070	2017	103	Y86	1	39
3945134	4070	2017	103	Y86	2	20
3945135	4070	2017	103	Y87	1	2
3945136	4070	2017	103	Y87	2	2
3945137	4070	2017	103	Y89	1	1

3945138 rows × 6 columns

As you can see we have reduced the initial dataset significantly and now have our key data points in columnar form. This is a good intermediate state for our data as we still have all the countries and all the years but we have consolidated the ages into a total and removed any Nan values up to this point. From here we can further slice and dice in order to answer specific questions.

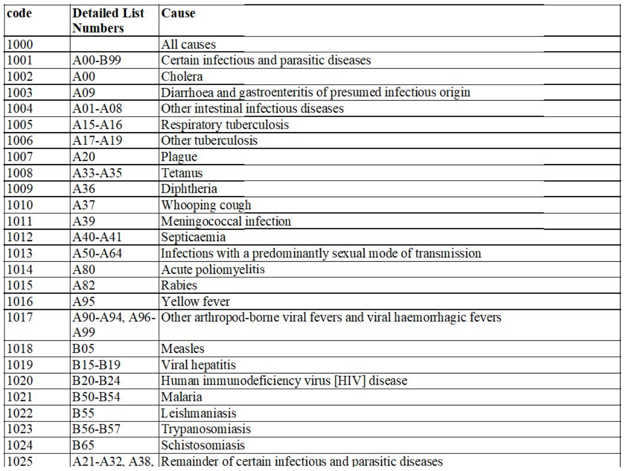
We then took our data warehouse and boiled it down further purposes of doing our initial exploratory data analysis. We filtered it into a dataframe that only kept the 2018 data and also merged in some population characteristics we will likely add into our main warehouse as they were pretty useful. The data is 2018 only with a row for each country, a column for each cause of death and then the data in the columns is a normed death number calculated by taking the deaths in each cause/ total deaths in the country for the year. We also replaced the NaNs with zero values for purposes of this initial analysis. See below:



Cause	Country	A020	A021	A022	A045	A047	A049	A052	A059	A066	...	
0	1300	7.903245e-07	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	...	0.0000
1	3030	2.260398e-06	0.000000e+00	0.000005	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	...	0.0000
2	4018	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	1.159909e-06	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	...	0.0000
3	4084	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	2.683448e-07	5.366896e-07	0.000000e+00	2.683448e-07	2.683448e-07	...	2.6834
4	4160	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	2.835106e-06	2.835106e-06	0.000000e+00	0.000000e+00	0.000000e+00	...	0.0000
5	4188	0.000000e+00	7.138923e-07	0.000000	3.569462e-07	1.213617e-05	2.498623e-06	3.569462e-07	0.000000e+00	0.000000e+00	...	0.0000
6	4260	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	5.641552e-07	0.000000e+00	0.000000e+00	0.000000e+00	...	0.0000
7	5198	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	1.500247e-06	0.000000e+00	0.000000e+00	0.000000e+00	...	0.0000

8 rows × 2725 columns

The biggest issue we have with this data is we need to come up with a way to consolidate the causes of death across the columns into a manageable amount. We currently have around 3000 columns which include every value from A00 to Z99 plus a couple other one off formats we will likely drop for consistency. Right now each code is super specific as you can see in the below snip from the data description:

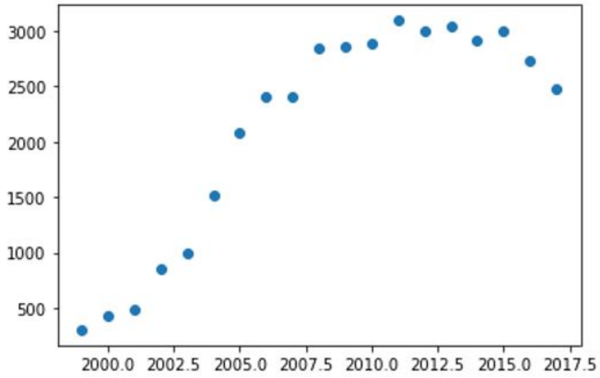


code	Detailed List Numbers	Cause
1000		All causes
1001	A00-B99	Certain infectious and parasitic diseases
1002	A00	Cholera
1003	A09	Diarrhoea and gastroenteritis of presumed infectious origin
1004	A01-A08	Other intestinal infectious diseases
1005	A15-A16	Respiratory tuberculosis
1006	A17-A19	Other tuberculosis
1007	A20	Plague
1008	A33-A35	Tetanus
1009	A36	Diphtheria
1010	A37	Whooping cough
1011	A39	Meningococcal infection
1012	A40-A41	Septicaemia
1013	A50-A64	Infections with a predominantly sexual mode of transmission
1014	A80	Acute poliomyelitis
1015	A82	Rabies
1016	A95	Yellow fever
1017	A90-A94, A96-A99	Other arthropod-borne viral fevers and viral haemorrhagic fevers
1018	B05	Measles
1019	B15-B19	Viral hepatitis
1020	B20-B24	Human immunodeficiency virus [HIV] disease
1021	B50-B54	Malaria
1022	B55	Leishmaniasis
1023	B56-B57	Trypanosomiasis
1024	B65	Schistosomiasis
1025	A21-A32, A38,	Remainder of certain infectious and parasitic diseases

For example you can see there are detailed list numbers B50-B54 which all point to various forms of malaria which could at one level be aggregated into code 1021. We think we can go even further than that though as you can see code 1001 contains all the values for infectious and parasitic diseases which are the values A00 through B99. Using the documentation we hope to come up with some sort of dictionary we can use to re code these values as described above into a manageable feature list that is hopefully in the sub 50 features range. This is still a lot but way better than 3000 and will hopefully be sufficiently reduced to allow a signal to show through.

Once we reduce the feature list we will perform additional EDA such as boxplots, correlation matrices and even just simple summary statistics by feature so we have better grasp on what our data is showing. The feature reduction will also allow us to perform outlier analysis and identify any additional transformations we need to do before running any statistical models. Right now there are still just too many features to run these efficiently and get useful data out of them.

A quick example of the types of summary plots/analysis we would like make more of is this chart showing male deaths per year in the US for the cause A047.



Once we get to a reasonable amount of features we would like to see the data in this manner on a worldwide basis and be able to drill down if necessary to investigate any interesting results we get in our main modeling as well as to sanity check the results of the more black box type algorithms.