Capstone 2 Final Report
Springboard Data Science Career Track
Maya Bhat-Gregerson
August 3, 2020

Project: Classification of poorly coded death records using text analysis

Background

Death records (aka 'death certificates') include a section where a medical professional lists one or more health conditions that resulted in the death. Ideally, this cause of death statement will consist of an etiological sequence of physiological events that led to death. Typically, medical certifiers will list multiple conditions in the cause of death section from which one is selected as the 'underlying cause of death' by the National Center for Health Statistics. The underlying cause of death is the condition or event that initiated the chain of events leading directly to death. All conditions reported on death records are assigned codes pursuant to a complex classification structure known as the International Classification of Diseases (ICD-10). Most statistics on causes of death (e.g. lung cancer, car crashes, diabetes etc.) are based on the analysis of ICD-10 codes for the underlying cause of death only. Other conditions listed on the death certificate are not usually analyzed when measuring disease burden, however, they provide valuable information on the circumstances of the death.

The problem

Often, due to inadequate training, medical certifiers do not report meaningful causes of death that identify the cause of the disease process that resulted in death. For example, it is not surprising to see 'cardiac arrest' or 'respiratory arrest' written on the death certificate as the underlying causes of death. Clearly, this is not useful in providing any actionable information that public health agencies can use to set policy because it provides no information on the disease that resulted in the cardiac or respiratory arrest.. In 2019, approximately 7% of Washington State's approximately 58,000 death records were assigned so-called 'garbage' ICD-10 codes for the underlying cause of death. The practical implication is that a significant proportion of mortality data is useless in epidemiological analysis.

Who cares and why

Public health agencies at state and federal levels want accurate and complete mortality data to be able to understand the burden of various diseases in our communities. This information allows public health entities to conduct research, implement programs, or create policies to address leading causes of death.

While public health agencies routinely try to address poorly written causes of death by contacting individual medical providers to request changes, this is obviously not an effective or

efficient way of handling the problem. One alternative that would be a more efficient way of addressing the problem is to use text analysis to reassign these problematic death records to broad but more useful cause of death groupings that can be used in routine public health epidemiological analyses.

Overview of method

I compared the relative ease and success of reclassifying death records that had garbage ICD-10 codes instead of valid underlying cause of death codes into broad cause of death categories that are useful in epidemiological analysis. I compared the effectiveness of Latent Dirichlet Allocation (LDA), an unsupervised natural language processing method in identifying topics in poorly worded death certificates. I also used two supervised methods (Multinomial Naive Bayes and Support Vector Machine) to classify death records with garbage code using records that have valid codes as the training data. I hoped to compare the classification and topic modeling results from the three algorithms to see if there's a reliable way to salvage garbage coded records.

Data set

I used Washington State mortality data in my project. I work as an epidemiologist at Washington State Department of Health and have permission to use these data to meet my capstone requirement as well as to explore novel means of making these data more useful.

I used data for deaths occurring in Washington State from January 1, 2016 through December 31, 2019. Overall, there were almost 223,000 records in the dataset.

I also used a list of ICD-10 codes that I found in published literature that are considered to be 'garbage codes' i.e. codes that are unhelpful in analysis. In the data wrangling step, I used this list of over 1,900 ICD-10 codes to identify and label death records with garbage underlying cause codes. The garbage code set is split into 9 broad groups based on the general category of disease they fell into. These 9 groups are mutually exclusive and include unspecified cancer, unspecified heart disease, unspecified cardiovascular disease, septicemia, unspecified infectious disease, ill-defined conditions, volume depletion, ill-defined injury, and undetermined intent.

Data cleaning and wrangling

I used SQL queries within Python to obtain the variables and years of interest. I performed the following data cleaning tasks:

- Checks for missing values
- Checks for out of range values for categorical variables with a fixed range of potential values

- Exclusion of deaths to residents who died outside Washington State
- Label records considered as having ‘garbage codes’ for underlying cause and also those that had ‘valid codes’
- Label records with valid underlying cause codes with the broad cause of death category to which they belonged
- Pre-processing of text fields to prepare for natural language processing algorithms
- Create functions and modules to pre-process text fields for natural language processing
- Create training and testing data sets for the supervised machine learning portion of the project.

Exploratory data analysis

In this portion of the project I examined the distribution of garbage codes by demographics (age group, gender, race, ethnicity) as well as by county of death and by the type of medical certifier who reported the cause of death (physician, medical examiner/coroner, ARNP etc.).

As expected the majority of the garbage codes occurred in older decedents as they are the most likely to die. In Washington State, the greatest proportion of garbage codes were likely to be ‘unspecified cancer’, ‘unspecified heart disease’, ‘unspecified cardiovascular disease’, or ‘septicemia’. Relatively few deaths were due to ‘unspecified infectious diseases’.

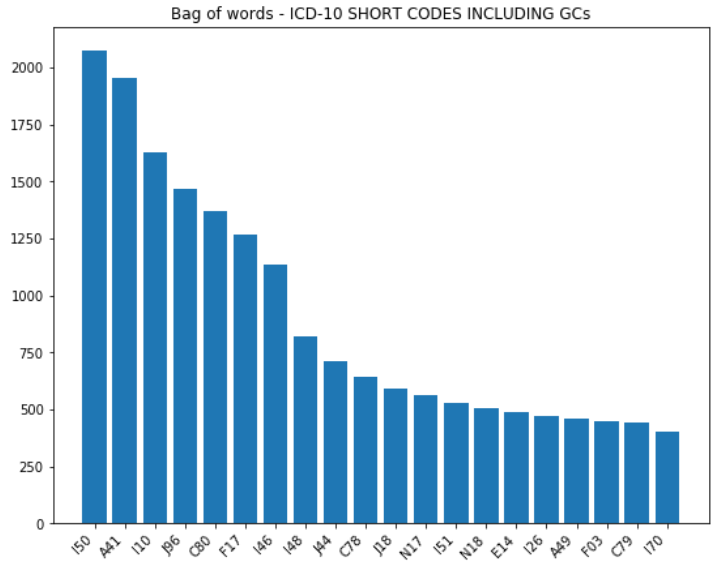
Of all certifier types, physicians certified the largest number of records with causes of death that end up receiving a garbage ICD-10 code. This is unsurprising as the majority of death records are completed by physicians. The relative distribution within each type of certifier

There were some patterns apparent in which demographic groups had greater proportions of death records with garbage underlying cause codes. Among death records for American Indian/Native Alaskan (AIAN) decedents almost 8% had garbage underlying cause codes. In comparison, the proportion of death records with the same problems among white non-Hispanic and Asian non-Hispanic decedents was 6.6% and 6.3%. The problem is particularly concerning because we know that AIAN population is undercounted in death data as they are often misclassified as another race. The problem of incomplete data due to race misclassification is now compounded by having poor data quality among those who were classified as AIAN decedents.

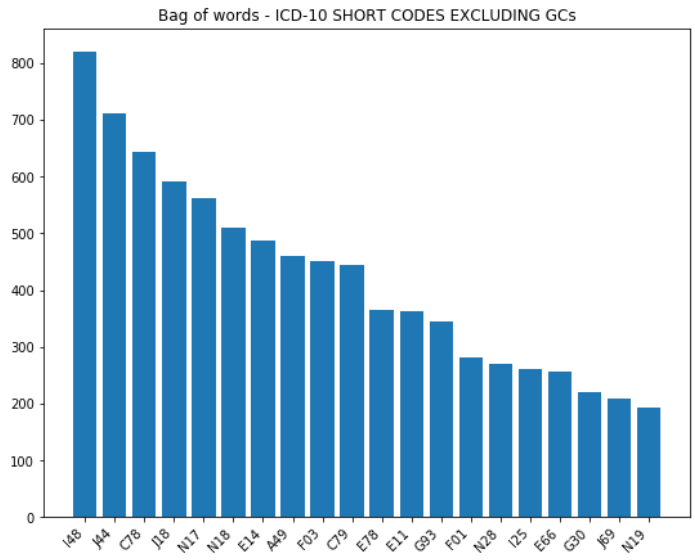
The percentage of records with problematic underlying cause codes also increases with age. Approximately 5.3% of death records for decedents who were 0-19 years old had garbage codes for underlying causes as compared with 7.0% among records for decedents who were 80 years or older. It is a well known problem that medical certifiers have greater difficulty ascertaining the true cause of death for elderly decedents and resort to phrases like “old age” or “natural causes” for this population. On the other hand, as deaths among infants and youth are less common, it is likely that they receive far greater attention when certifiers identify the cause of death.

Finally, there was a notable disparity between types of death record certifiers in the proportions of records certified with garbage underlying causes. Physician Assistants (PAs) were far more likely to report causes of death that were assigned garbage codes compared with other certifier types. Approximately 9.4% of records certified by PAs compared with 6.3% of records certified with physicians and 6.4% certified by nurse practitioners were assigned garbage underlying cause codes.

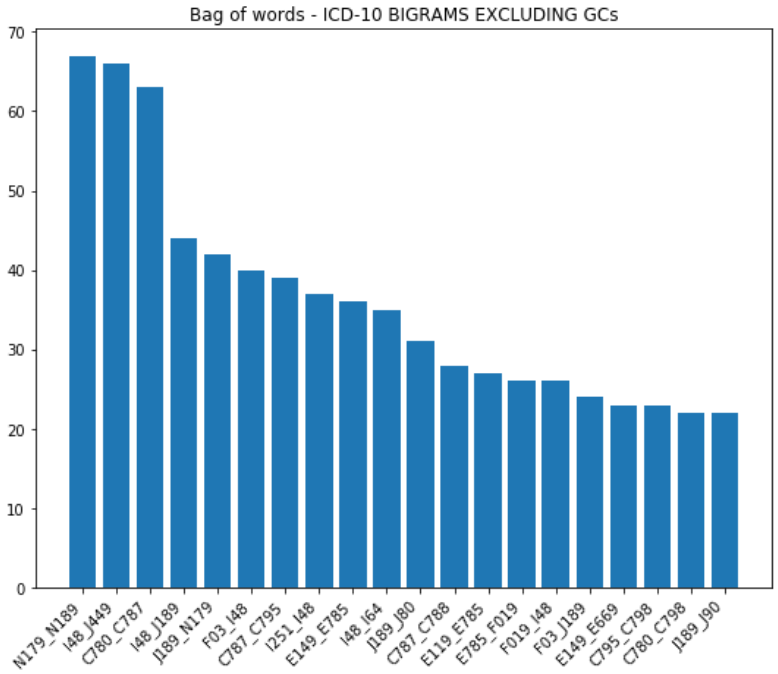
Aside from examining the distribution of garbage codes among various demographic groups and certifier types, I also used the bag of words method to examine the frequency of various ICD-10 codes in the subset of records that had garbage underlying cause codes. Bag of words showed the frequency of individual ICD-10 codes across all records with garbage underlying cause codes regardless of the position occupied by the code. In other words, these codes could have been either underlying or multiple cause codes.



Looking at all codes in the 7,124 death records with garbage underlying cause codes the ICD-10 codes for congestive heart failure, sepsis due to bacterial infection, and essential hypertension occurred most frequently. In this bag of words analysis I left the garbage codes in the corpus of ICD-10 codes.



In the next bag of words analysis I used only the multiple cause codes i.e. excluded the underlying cause code (which we know is a garbage code). In this analysis, the three multiple cause ICD-10 codes that were most common were I48 (atrial fibrillation), J44(COPD), and N17(acute renal failure).



A third bag of words plot constructed with bi-grams (two adjacent codes) and no garbage underlying cause codes showed that N17.9-N18.9 (acute renal failure and chronic kidney disease), I48-J44.9 (atrial fibrillation and COPD), and C78.0-C78.7(secondary malignant neoplasm of the lung and secondary malignant neoplasm of the liver) were the most common pairs of codes occurring across this subset of death records.

Unsupervised machine learning classifier: Latent Dirichlet Allocation

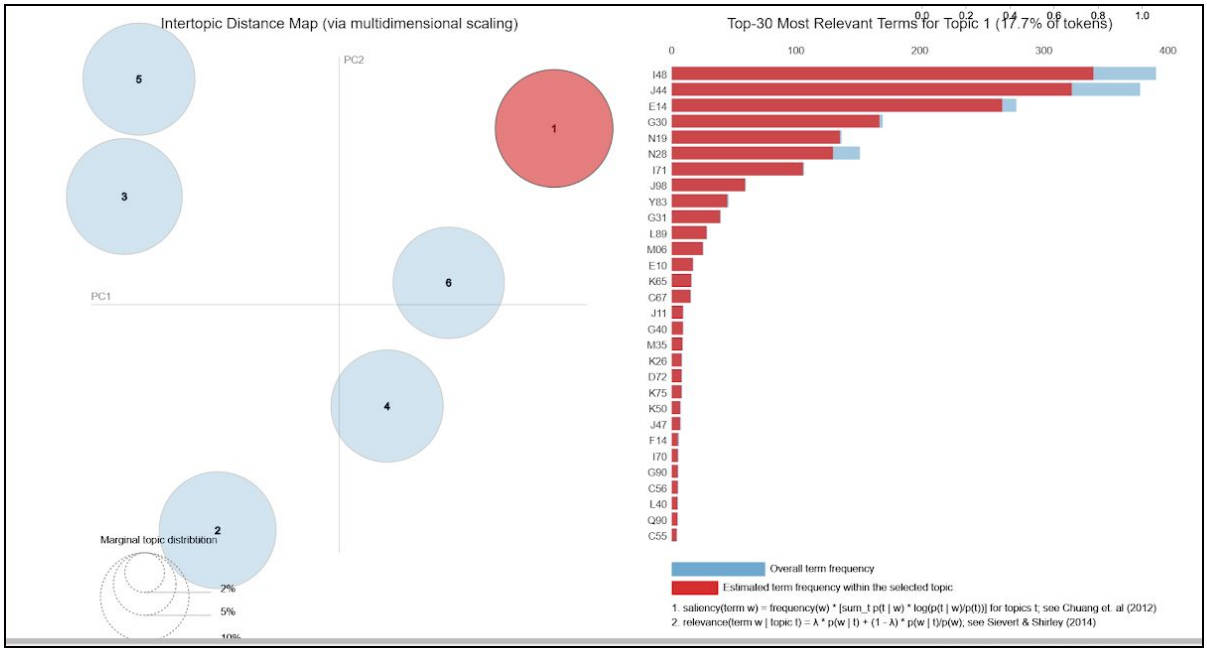
Latent Dirichlet Allocation (LDA) is an unsupervised machine learning algorithm that is used with multiple text documents to uncover the topics mentioned in the documents. To accomplish this, the algorithm compares the frequency of words appearing within topics (clusters of words) and the frequency of words appearing with documents. After the algorithm is applied we can see clusters of words belonging to a particular topic and the proportion of each document that deals with a given topic. The document can then be assigned to the 'dominant' topic appearing within it's text.

For this project, the 'documents' used with LDA were the death records (each document is one death record) and the text in the documents were the ICD-10 codes. I chose to use ICD-10 codes instead of the medical terms reported by the certifiers because ICD-10 codes are a more distilled and standardized representation of the health conditions and events described in the words and phrases entered in the death records. Some of the pre-processing steps for ICD-10 codes followed methods that would be

used with regular words. I stemmed the ICD-10 codes so that the last digit of the numeric portion of the code was excluded. This allowed for multiple, related diseases to be represented by a single root code.

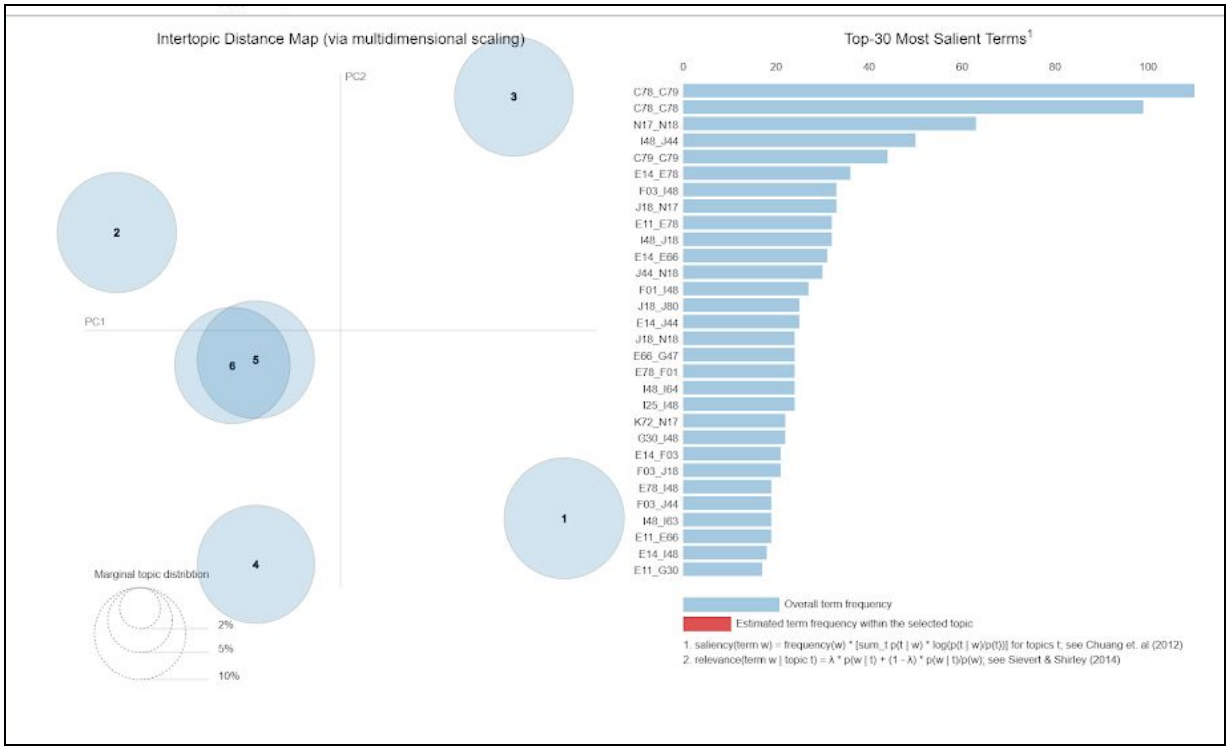
Through a grid search I determined that the optimal number of topics to specify to the LDA algorithm was 6. I also ran the algorithm with different versions of the same corpus. Early on it was obvious that including garbage codes in any position (underlying cause or any of the multiple cause positions) would yield poor results. For the remainder of the LDA analysis I used text that excluded these codes in all positions. I evaluated each model using a coherence score which is a measure of the co-occurrence of words (in this case ICD-10 scores) and therefore how well the words in a topic are related to each other. I also evaluated the models simply by examining the dominant ICD-10 codes within each one to see if the top two or three codes taken together indicate a general cause of death.

In the first LDA model, I used unigrams without any garbage codes. Visualizing the model output with pyLDAvis it is apparent that the algorithm was able to detect 6 non-overlapping topics. Each topic was most strongly associated with one or two dominant ICD-10 codes. For example, in the image below topic 1 was associated with codes I48 and J44 which were the codes for atrial fibrillation and COPD. Similarly, topic 2 in the bottom left quadrant was dominated by C78 and C79. Given the types of conditions represented, it would be interesting to investigate the role of tobacco use and obesity as risk factors for the health conditions represented by the ICD-10 codes.



I also used the model with unabridged ICD-10 codes which yielded less optimal results in terms of well-separated topics.

Finally, I also ran the models with bigrams of abridged codes (letter and two digits) and unabridged codes. These final models yielded mixed results. The models used with bigrams yielded the highest scores (0.77 out of 1.0), but the topics as visualized below were mixed e.g. two circles both had high proportion of C78 and C79 code combinations (topics 1 and 3), overlapping circles (topics 5 and 6), and less obvious diseases indicated by the code pairs in general. For a medical expert, the patterns may be obvious.



Finally, I tried to assign the dominant topic label to each death record in the data set, but this was much harder than expected. By the time I removed garbage codes from the underlying and all multiple cause fields I was left with just over 4,000 records which is not a very large set of documents to use with LDA. The length of the text in each death record was also very short consisting of 1 to 15 ICD-10 codes each.

In future attempts, I will try to use a balanced mix of records with and without garbage underlying cause codes in order to increase the number of documents available to the algorithm to create a more accurate model. I would also like to see if the patterns of codes in the multiple cause fields of well-coded (non-garbage) records used together with those of garbage coded records provide some additional information that allows

codes to cluster more cleanly into topics and perhaps even allow more specific topics to surface.

Supervised machine learning classifiers: Multinomial Naive Bayes and Support Vector Machine

In the last part of the project I used supervised machine learning algorithms to train models that can classify the garbage coded records into one of the ten leading causes of deaths in Washington State. These leading causes of death are groupings of ICD-10 codes that indicate common disease processes. I used Multinomial Naive Bayes (MNB) and separately, Support Vector Machine (SVM) to create these classifiers.

One issue I encountered at the outset of this part of the project is that I don't have access to training data consisting of garbage coded records that have been labelled with their true underlying causes. Eventually, I used records with valid codes to train the classifiers. I labeled the records as having one of the ten leading causes of death as the underlying cause (e.g. lung cancer, chronic lower respiratory disease, cerebrovascular disease, Alzheimer's disease, diabetes etc.).

Neither the MNB and SVM were particularly good at predicting classes when I tried to create models to assign the records in the test data set to one of multiple classes. MNB yielded an accuracy of 60% and linear SVM had an accuracy of almost 63%. One of the problems was that my initial set of classifications included the 10 leading causes of death and an 11th catch-all 'other' category. The 'other' category was being misclassified as heart disease or cancer and vice versa. For the rest of the analysis I removed records that were labeled as 'other' and used only the 10 causes of death.

Next, I attempted to create a series of one vs. all binary classifiers where I tried to build models that tried to classify the records in the test set as, for example, diabetes vs. non-diabetes deaths, cerebrovascular vs. non-cerebrovascular disease deaths etc. These models were vastly better in terms of their accuracy. The models ranged in accuracy from about 82% (heart disease vs. non-heart disease deaths) to almost 95% accuracy (diabetes vs. non-diabetes). Between MNB and linear SVC, the latter performed marginally better, so I performed a grid search to find the best value for the regularization parameter C (0.1). C value indicates the distance between the hyperplane and the support vectors that allows the most accurate classification.

The problem with one vs. all binary classifiers in this situation is that I have an imbalanced class problem. For example with a diabetes vs. non-diabetes classifier there will be far more records in the non-diabetes class compared with the diabetes

class. The imbalance is most noticeable in the lower recall and F1 scores for the affirmative (with disease) class of each binary classifier.

Finally, I applied the optimized linear SVM model to the garbage coded records in a series of binary classifiers. Of the garbage coded records that could be classified, the majority (over 2000) were coded as cancer deaths, followed by heart disease, and Alzheimer's disease deaths. Only one third of the roughly 15,000 records could be assigned to a meaningful cause of death.

Next Steps

There are a few major changes I would make to both the LDA and the linear SVM models.

As I mentioned above, I would like to use a balanced mix of death records with both garbage and valid underlying causes of death to increase the size of the data set and to add more information on the patterns of co-occurring ICD-10 codes that may allow for better clustering of codes.

I would also like to try the Mallet LDA algorithm which is a variation of the algorithm I used and is supposed to be better for use with shorter length texts.

With the supervised linear SVM classifier, I need to rerun the binary classifiers with balanced classes and with cross validation to improve recall and F1 scores.

I would also like to create different classes that are more narrowly defined as compared with the current 'heart disease' or 'cancer' categories. In LDA I could see certain codes for certain types of cancers (lung and liver cancers) appearing more frequently. I also saw other cause of death categories that were more narrowly defined than the leading causes of death but more meaningful than the garbage codes assigned as underlying causes to these records. I would like to recreate classes based on the information provided by LDA.

Ultimately, the goal that I would like to be able to achieve is to use different classification methods to label the garbage coded records to explore the degree to which the different algorithms agree with each other in terms of the class they would assign to each garbage coded records.

Given the new leading cause of death labels assigned by the supervised learning classifier, I will also recalculate the overall mortality rate due to each of those causes to

see if the addition of these salvaged records to these causes of death changed the rank order of the deaths in general or within specific demographic groups.